

Cloudera Release Guide



Important Notice

© 2010-2016 Cloudera, Inc. All rights reserved.

Cloudera, the Cloudera logo, Cloudera Impala, and any other product or service names or slogans contained in this document are trademarks of Cloudera and its suppliers or licensors, and may not be copied, imitated or used, in whole or in part, without the prior written permission of Cloudera or the applicable trademark holder.

Hadoop and the Hadoop elephant logo are trademarks of the Apache Software Foundation. All other trademarks, registered trademarks, product names and company names or logos mentioned in this document are the property of their respective owners. Reference to any products, services, processes or other information, by trade name, trademark, manufacturer, supplier or otherwise does not constitute or imply endorsement, sponsorship or recommendation thereof by us.

Complying with all applicable copyright laws is the responsibility of the user. Without limiting the rights under copyright, no part of this document may be reproduced, stored in or introduced into a retrieval system, or transmitted in any form or by any means (electronic, mechanical, photocopying, recording, or otherwise), or for any purpose, without the express written permission of Cloudera.

Cloudera may have patents, patent applications, trademarks, copyrights, or other intellectual property rights covering subject matter in this document. Except as expressly provided in any written license agreement from Cloudera, the furnishing of this document does not give you any license to these patents, trademarks copyrights, or other intellectual property. For information about patents covering Cloudera products, see <http://tiny.cloudera.com/patents>.

The information in this document is subject to change without notice. Cloudera shall not be liable for any damages resulting from technical errors or omissions which may be present in this document, or from use of this document.

Cloudera, Inc.
1001 Page Mill Road, Bldg 3
Palo Alto, CA 94304
info@cloudera.com
US: 1-888-789-1488
Intl: 1-650-362-0488
www.cloudera.com

Release Information

Version: Cloudera Enterprise 5.x
Date: November 21, 2016

Table of Contents

About Cloudera Enterprise 5.x Release Notes.....	12
Cloudera Enterprise 5.x Release Notes.....	13
CDH 5 Release Notes.....	14
New Features and Changes in CDH 5.....	14
<i>About Apache Hadoop MapReduce Version 1 (MRv1) and Version 2 (MRv2).....</i>	14
<i>What's New In CDH 5.9.x.....</i>	14
<i>What's New In CDH 5.8.x.....</i>	16
<i>What's New In CDH 5.7.x.....</i>	18
<i>What's New In CDH 5.6.x.....</i>	23
<i>What's New In CDH 5.5.x.....</i>	23
<i>What's New In CDH 5.4.x.....</i>	27
<i>What's New In CDH 5.3.x.....</i>	34
<i>What's New In CDH 5.2.x.....</i>	37
<i>What's New in CDH 5.1.x.....</i>	43
<i>What's New in CDH 5.0.x.....</i>	47
<i>What's New In CDH 5 Beta Releases.....</i>	50
<i>What's New in Apache Impala (incubating).....</i>	58
Incompatible Changes and Limitations.....	84
<i>Apache Avro Incompatible Changes and Limitations.....</i>	84
<i>Apache Crunch Incompatible Changes and Limitations.....</i>	85
<i>Apache DataFu Incompatible Changes and Limitations.....</i>	85
<i>Apache Flume Incompatible Changes and Limitations.....</i>	85
<i>Apache Hadoop Incompatible Changes and Limitations.....</i>	85
<i>Apache HBase Incompatible Changes and Limitations.....</i>	87
<i>Apache Hive Incompatible Changes and Limitations.....</i>	91
<i>Hue Incompatible Changes and Limitations.....</i>	93
<i>Apache Impala (incubating) Incompatible Changes and Limitations.....</i>	94
<i>Cloudera Distribution of Apache Kafka Incompatible Changes and Limitations.....</i>	105
<i>Kite Incompatible Changes and Limitations.....</i>	105
<i>Llama Incompatible Changes and Limitations.....</i>	105
<i>Apache Mahout Incompatible Changes and Limitations.....</i>	106
<i>Apache Oozie Incompatible Changes and Limitations.....</i>	107
<i>Apache Pig Incompatible Changes and Limitations.....</i>	107
<i>Cloudera Search Incompatible Changes and Limitations.....</i>	107
<i>Apache Sentry Incompatible Changes.....</i>	110

<i>Apache Spark Incompatible Changes and Limitations</i>	110
<i>Apache Sqoop Incompatible Changes and Limitations</i>	111
<i>Apache Whirr Incompatible Changes and Limitations</i>	111
<i>Apache ZooKeeper Incompatible Changes and Limitations</i>	111
Known Issues in CDH 5.....	111
<i>Operating System Known Issues</i>	111
<i>Performance Known Issues</i>	112
<i>Install and Upgrade Known Issues</i>	112
<i>Apache Flume Known Issues</i>	115
<i>Apache Hadoop Known Issues</i>	115
<i>Apache HBase Known Issues</i>	122
<i>Apache Hive Known Issues</i>	128
<i>Hue Known Issues</i>	134
<i>Apache Impala (incubating) Known Issues</i>	135
<i>Cloudera Distribution of Apache Kafka Known Issues</i>	146
<i>Apache Mahout Known Issues</i>	146
<i>Apache Oozie Known Issues</i>	146
<i>Apache Parquet Known Issues</i>	147
<i>Apache Pig Known Issues</i>	147
<i>Cloudera Search Known Issues</i>	148
<i>Apache Sentry Known Issues</i>	153
<i>Apache Spark Known Issues</i>	155
<i>Apache Sqoop Known Issues</i>	157
<i>Apache ZooKeeper Known Issues</i>	159
Issues Fixed in CDH 5.....	159
<i>Issues Fixed in CDH 5.9.x</i>	159
<i>Issues Fixed in CDH 5.8.x</i>	184
<i>Issues Fixed in CDH 5.7.x</i>	196
<i>Issues Fixed in CDH 5.6.x</i>	218
<i>Issues Fixed in CDH 5.5.x</i>	225
<i>Issues Fixed in CDH 5.4.x</i>	243
<i>Issues Fixed in CDH 5.3.x</i>	268
<i>Issues Fixed in CDH 5.2.x</i>	284
<i>Issues Fixed in CDH 5.1.x</i>	292
<i>Issues Fixed in CDH 5.0.x</i>	298
<i>Issues Fixed in CDH 5 Beta Releases</i>	305
<i>Fixed Issues in Apache Impala (incubating)</i>	306

Cloudera Manager 5 Release Notes.....	354
New Features and Changes in Cloudera Manager 5.....	354
<i>What's New in Cloudera Manager 5</i>	354
<i>Incompatible Changes in Cloudera Manager 5</i>	375
<i>Changed Features and Behaviors in Cloudera Manager 5</i>	378
Known Issues and Workarounds in Cloudera Manager 5.....	383

<i>Error when distributing parcels: No such torrent.....</i>	383
<i>Hive Replication Metadata Transfer Step fails with Temporary AWS Credential Provider.....</i>	383
<i>Hive table Views do not get restored from S3</i>	383
<i>ACLs are not replicated when restoring Hive data from S3</i>	383
<i>Snapshot diff is not working for Hive to S3 replication when data is deleted on source.....</i>	383
<i>Block agents from heartbeating to a Cloudera Manager with different UUID until agent restart.....</i>	384
<i>Cloudera Manager set catalogd default jvm memory to 4G can cause out of memory error on upgrade to Cloudera Manager 5.7 or higher.....</i>	384
<i>Cloudera Manager 5.7.4 installer does not show Key Trustee KMS.....</i>	384
<i>Class Not Found Error when upgrading to Cloudera Manager 5.7.2.....</i>	384
<i>Kerberos setup fails on Debian 8.2.....</i>	384
<i>Password in Cloudera Manager's db.properties file is not redacted.....</i>	384
<i>Cluster provisioning fails.....</i>	385
<i>Cloudera Manager can run out of memory if a remote repository URL is unreachable.....</i>	385
<i>Clients can run Hive on Spark jobs even if Hive dependency on Spark is not configured.....</i>	385
<i>Known Issues for the DSSD D5 Hadoop Plugin</i>	385
<i>The YARN NodeManager connectivity health test does not work for CDH 5.....</i>	386
<i>HDFS HA clusters see NameNode failures when KDC connectivity is bad.....</i>	386
<i>The HDFS File browser in Cloudera Manager fails when HDFS federation is enabled.....</i>	387
<i>Hive Metastore canary fails to drop database.....</i>	387
<i>Cloudera Manager upgrade fails due to incorrect Sqoop 2 path.....</i>	387
<i>NameNode incorrectly reports missing blocks during rolling upgrade.....</i>	387
<i>Using ext3 for server dirs easily hit inode limit.....</i>	387
<i>Backup and disaster recovery replication does not set MapReduce Java options</i>	388
<i>Kafka 1.2 CSD conflicts with CSD included in Cloudera Manager 5.4.....</i>	388
<i>Recommission host does not deploy client configurations.....</i>	388
<i>Hive on Spark is not supported in Cloudera Manager and CDH 5.4 and CDH 5.5.....</i>	388
<i>CDH 5 requires JDK 1.7.....</i>	388
<i>Upgrade wizard incorrectly upgrades the Sentry DB.....</i>	388
<i>Cloudera Manager does not correctly generate client configurations for services deployed using CSDs.....</i>	388
<i>Solr, Oozie and HttpFS fail when KMS and TLS/SSL are enabled using self-signed certificates.....</i>	389
<i>Cloudera Manager 5.3.1 upgrade fails if Spark standalone and Kerberos are configured.....</i>	389
<i>Adding Key Trustee KMS 5.4 to Cloudera Manager 5.5 displays warning.....</i>	389
<i>KMS and Key Trustee ACLs do not work in Cloudera Manager 5.3.....</i>	389
<i>Exporting and importing Hue database sometimes times out after 90 seconds.....</i>	389
<i>Changing the Key Trustee Server hostname requires editing keytrustee.conf.....</i>	390
<i>Hosts with Impala Llama roles must also have at least one YARN role.....</i>	390
<i>The high availability wizard does not verify that there is a running ZooKeeper service.....</i>	390
<i>Cloudera Manager Installation Path A fails on RHEL 5.7 due to PostgreSQL conflict.....</i>	390
<i>Spurious warning on Accumulo 1.6 gateway hosts.....</i>	390
<i>Accumulo 1.6 service log aggregation and search does not work.....</i>	390
<i>Cloudera Manager incorrectly sizes Accumulo Tablet Server max heap size after 1.4.4-cdh4.5.0 to 1.6.0-cdh4.6.0 upgrade.....</i>	391
<i>Accumulo installations using LZO do not indicate dependence on the GPL Extras parcel.....</i>	391
<i>Created pools are not preserved when Dynamic Resource Pools page is used to configure YARN or Impala.....</i>	391

<i>User should be prompted to add the AMON role when adding MapReduce to a CDH 5 cluster.....</i>	391
<i>Enterprise license expiration alert not displayed until Cloudera Manager Server is restarted.....</i>	391
<i>Configurations for decommissioned roles not migrated from MapReduce to YARN.....</i>	391
<i>The HDFS command Roll Edits does not work in the UI when HDFS is federated.....</i>	392
<i>Cloudera Manager reports a confusing version number if you have oozie-client, but not oozie installed on a CDH 4.4 node.....</i>	392
<i>Cloudera Manager does not work with CDH 5.0.0 Beta 1.....</i>	392
<i>On CDH 4.1 secure clusters managed by Cloudera Manager 4.8.1 and higher, the Impala Catalog server needs advanced configuration snippet update.....</i>	392
<i>Rolling Upgrade to CDH 5 is not supported.....</i>	392
<i>Error reading .zip file created with the Collect Diagnostic Data command.....</i>	392
<i>After JobTracker failover, complete jobs from the previous active JobTracker are not visible.....</i>	392
<i>After JobTracker failover, information about rerun jobs is not updated in Activity Monitor.....</i>	393
<i>Installing on AWS, you must use private EC2 hostnames.....</i>	393
<i>If HDFS uses Quorum-based Storage without HA enabled, the SecondaryNameNode cannot checkpoint.....</i>	393
<i>Changing the rack configuration may temporarily cause mis-replicated blocks to be reported.....</i>	393
<i>Cannot use '/' as a mount point with a Federated HDFS Nameservice.....</i>	393
<i>Historical disk usage reports do not work with federated HDFS.....</i>	394
<i>(CDH 4 only) Activity monitoring does not work on YARN activities.....</i>	394
<i>HDFS monitoring configuration applies to all Nameservices.....</i>	394
<i>Supported and Unsupported Replication Scenarios and Limitations.....</i>	394
<i>Restoring snapshot of a file to an empty directory does not overwrite the directory.....</i>	394
<i>HDFS Snapshot appears to fail if policy specifies duplicate directories.....</i>	394
<i>Hive replication fails if "Force Overwrite" is not set.....</i>	394
<i>Cloudera Manager set cataloged default JVM memory to 4G can cause an out of memory error during upgrade to Cloudera Manager 5.7 and higher.....</i>	395
<i>Issues Fixed in Cloudera Manager 5.....</i>	395
<i>Issues Fixed in Cloudera Manager 5.9.....</i>	395
<i>Issues Fixed in Cloudera Manager 5.8.3.....</i>	399
<i>Issues Fixed in Cloudera Manager 5.8.2.....</i>	400
<i>Issues Fixed in Cloudera Manager 5.8.1.....</i>	401
<i>Issues Fixed in Cloudera Manager 5.8.0.....</i>	401
<i>Issues Fixed in Cloudera Manager 5.7.5.....</i>	408
<i>Issues Fixed in Cloudera Manager 5.7.4.....</i>	408
<i>Issues Fixed in Cloudera Manager 5.7.2.....</i>	409
<i>Issues Fixed in Cloudera Manager 5.7.1.....</i>	410
<i>Issues Fixed in Cloudera Manager 5.7.0.....</i>	412
<i>Issues Fixed in Cloudera Manager 5.6.1.....</i>	415
<i>Issues Fixed in Cloudera Manager 5.6.0.....</i>	415
<i>Issues Fixed in Cloudera Manager 5.5.5.....</i>	416
<i>Issues Fixed in Cloudera Manager 5.5.4.....</i>	416
<i>Issues Fixed in Cloudera Manager 5.5.3.....</i>	417
<i>Issues Fixed in Cloudera Manager 5.5.2.....</i>	417
<i>Issues Fixed in Cloudera Manager 5.5.1.....</i>	420
<i>Issues Fixed in Cloudera Manager 5.5.0.....</i>	421

<i>Issues Fixed in Cloudera Manager 5.4.11</i>	424
<i>Issues Fixed in Cloudera Manager 5.4.10</i>	424
<i>Issues Fixed in Cloudera Manager 5.4.9</i>	425
<i>Issues Fixed in Cloudera Manager 5.4.8</i>	426
<i>Issues Fixed in Cloudera Manager 5.4.7</i>	426
<i>Issues Fixed in Cloudera Manager 5.4.6</i>	427
<i>Issues Fixed in Cloudera Manager 5.4.5</i>	427
<i>Issues Fixed in Cloudera Manager 5.4.3</i>	430
<i>Issues Fixed in Cloudera Manager 5.4.1</i>	434
<i>Issues Fixed in Cloudera Manager 5.4.0</i>	436
<i>Issues Fixed in Cloudera Manager 5.3.10</i>	437
<i>Issues Fixed in Cloudera Manager 5.3.9</i>	437
<i>Issues Fixed in Cloudera Manager 5.3.8</i>	438
<i>Issues Fixed in Cloudera Manager 5.3.7</i>	438
<i>Issues Fixed in Cloudera Manager 5.3.6</i>	438
<i>Issues Fixed in Cloudera Manager 5.3.4</i>	439
<i>Issues Fixed in Cloudera Manager 5.3.3</i>	439
<i>Issues Fixed in Cloudera Manager 5.3.2</i>	440
<i>Issues Fixed in Cloudera Manager 5.3.1</i>	441
<i>Issues Fixed in Cloudera Manager 5.3.0</i>	443
<i>Issues Fixed in Cloudera Manager 5.2.7</i>	444
<i>Issues Fixed in Cloudera Manager 5.2.6</i>	444
<i>Issues Fixed in Cloudera Manager 5.2.5</i>	445
<i>Issues Fixed in Cloudera Manager 5.2.2</i>	445
<i>Issues Fixed in Cloudera Manager 5.2.1</i>	446
<i>Issues Fixed in Cloudera Manager 5.2.0</i>	446
<i>Issues Fixed in Cloudera Manager 5.1.6</i>	448
<i>Fixed Issues in Cloudera Manager 5.1.5</i>	448
<i>Fixed Issues in Cloudera Manager 5.1.4</i>	448
<i>Issues Fixed in Cloudera Manager 5.1.3</i>	449
<i>Issues Fixed in Cloudera Manager 5.1.2</i>	449
<i>Issues Fixed in Cloudera Manager 5.1.1</i>	450
<i>Issues Fixed in Cloudera Manager 5.1.0</i>	451
<i>Fixed Issues in Cloudera Manager 5.0.7</i>	453
<i>Fixed Issues in Cloudera Manager 5.0.6</i>	453
<i>Fixed Issues in Cloudera Manager 5.0.5</i>	453
<i>Issues Fixed in Cloudera Manager 5.0.2</i>	454
<i>Issues Fixed in Cloudera Manager 5.0.1</i>	454
<i>Issues Fixed in Cloudera Manager 5.0.0</i>	456
<i>Issues Fixed in Cloudera Manager 5.0.0 Beta 2</i>	459
<i>Issues Fixed in Cloudera Manager 5.0.0 Beta 1</i>	460

Cloudera Navigator 2 Data Management Release Notes.....	461
New Features and Changes in Cloudera Navigator 2 Data Management.....	461

<i>New Features in Cloudera Navigator 2</i>	461
<i>Changed Features in Cloudera Navigator 2</i>	466
Known Issues and Workarounds in Cloudera Navigator 2 Data Management.....	467
Issues Fixed in Cloudera Navigator 2 Data Management.....	469
<i>Issues Fixed in Cloudera Navigator 2.8.0</i>	469
<i>Issues Fixed in Cloudera Navigator 2.7.3</i>	470
<i>Issues Fixed in Cloudera Navigator 2.7.2</i>	471
<i>Issues Fixed in Cloudera Navigator 2.7.1</i>	471
<i>Issues Fixed in Cloudera Navigator 2.7.0</i>	471
<i>Issues Fixed in Cloudera Navigator 2.6.5</i>	472
<i>Issues Fixed in Cloudera Navigator 2.6.2</i>	473
<i>Issues Fixed in Cloudera Navigator 2.6.1</i>	473
<i>Issues Fixed in Cloudera Navigator 2.6.0</i>	474
<i>Issues Fixed in Cloudera Navigator 2.5.0</i>	475
<i>Issues Fixed in Cloudera Navigator 2.4.4</i>	475
<i>Issues Fixed in Cloudera Navigator 2.4.2</i>	475
<i>Issues Fixed in Cloudera Navigator 2.4.1</i>	477
<i>Issues Fixed in Cloudera Navigator 2.4.0</i>	477
<i>Issues Fixed in Cloudera Navigator 2.3.9</i>	478
<i>Issues Fixed in Cloudera Navigator 2.3.8</i>	479
<i>Issues Fixed in Cloudera Navigator 2.3.3</i>	479
<i>Issues Fixed in Cloudera Navigator 2.3.1</i>	480
<i>Issues Fixed in Cloudera Navigator 2.3.0</i>	480
<i>Issues Fixed in Cloudera Navigator 2.2.9</i>	480
<i>Issues Fixed in Cloudera Navigator 2.2.4</i>	481
<i>Issues Fixed in Cloudera Navigator 2.2.3</i>	481
<i>Issues Fixed in Cloudera Navigator 2.2.2</i>	481
<i>Issues Fixed in Cloudera Navigator 2.2.1</i>	481
<i>Issues Fixed in Cloudera Navigator 2.2.0</i>	482
<i>Issues Fixed in Cloudera Navigator 2.1.6</i>	482
<i>Issues Fixed in Cloudera Navigator 2.1.5</i>	482
<i>Issues Fixed in Cloudera Navigator 2.1.4</i>	482
<i>Issues Fixed in Cloudera Navigator 2.1.2</i>	482
<i>Issues Fixed in Cloudera Navigator 2.1.1</i>	483
<i>Issues Fixed in Cloudera Navigator 2.0.5</i>	483
<i>Issues Fixed in Cloudera Navigator 2.0.3</i>	483
<i>Issues Fixed in Cloudera Navigator 2.0.2</i>	483
<i>Issues Fixed in Cloudera Navigator 2.0.1</i>	483
<i>Issues Fixed in Cloudera Navigator 2.0.0</i>	484
<i>Issues Fixed in Cloudera Navigator 1.2.0</i>	484

Cloudera Navigator Key Trustee Server Release Notes.....485

New Features and Changes in Cloudera Navigator Key Trustee Server.....	485
<i>What's New in Cloudera Navigator Key Trustee Server</i>	485

<i>Changed Features and Behaviors in Cloudera Navigator Key Trustee Server</i>	486
Known Issues and Workarounds in Cloudera Navigator Key Trustee Server.....	486
Issues Fixed in Cloudera Navigator Key Trustee Server.....	487
<i>Issues Fixed in Cloudera Navigator Key Trustee Server 5.9.0</i>	488
<i>Issues Fixed in Cloudera Navigator Key Trustee Server 5.8.0</i>	488
<i>Issues Fixed in Cloudera Navigator Key Trustee Server 5.7.0</i>	488
<i>Issues Fixed in Cloudera Navigator Key Trustee Server 5.5.2</i>	488
<i>Issues Fixed in Cloudera Navigator Key Trustee Server 5.5.0</i>	489
<i>Issues Fixed in Cloudera Navigator Key Trustee Server 5.4.9</i>	491
<i>Issues Fixed in Cloudera Navigator Key Trustee Server 5.4.3</i>	491
<i>Issues Fixed in Cloudera Navigator Key Trustee Server 5.4.0</i>	492

Cloudera Navigator Key HSM Release Notes.....493

New Features and Changes in Cloudera Navigator Key HSM.....	493
<i>What's New in Cloudera Navigator Key HSM</i>	493
Known Issues and Workarounds in Cloudera Navigator Key HSM.....	493
Issues Fixed in Cloudera Navigator Key HSM.....	494
<i>Issues Fixed in Cloudera Navigator Key HSM 1.8.0</i>	494
<i>Issues Fixed in Cloudera Navigator Key HSM 1.7.0</i>	494
<i>Issues Fixed in Cloudera Navigator Key HSM 1.6.0</i>	494
<i>Issues Fixed in Cloudera Navigator Key HSM 1.5.1</i>	495
<i>Issues Fixed in Cloudera Navigator Key HSM 1.5.0</i>	495
<i>Issues Fixed in Cloudera Navigator Key HSM 1.4.0</i>	495

Key Trustee KMS Release Notes.....496

New Features in Key Trustee KMS.....	496
<i>What's New in Key Trustee KMS</i>	496
Known Issues and Workarounds in Key Trustee KMS.....	497
Issues Fixed in Key Trustee KMS.....	498
<i>Issues Fixed in Key Trustee KMS 5.9.0</i>	498
<i>Issues Fixed in Key Trustee KMS 5.8.2</i>	498
<i>Issues Fixed in Key Trustee KMS 5.8.0</i>	498
<i>Issues Fixed in Key Trustee KMS 5.7.4</i>	498
<i>Issues Fixed in Key Trustee KMS 5.7.1</i>	498
<i>Issues Fixed in Key Trustee KMS 5.7.0</i>	498
<i>Issues Fixed in Key Trustee KMS 5.5.4</i>	499
<i>Issues Fixed in Key Trustee KMS 5.5.0</i>	499
<i>Issues Fixed in Key Trustee KMS 5.4.3</i>	499

Cloudera Navigator Encrypt Release Notes.....500

New Features and Changes in Cloudera Navigator Encrypt.....	500
<i>What's New in Cloudera Navigator Encrypt</i>	500

<i>What's Changed in Cloudera Navigator Encrypt</i>	500
Known Issues and Workarounds in Cloudera Navigator Encrypt.....	501
Issues Fixed in Cloudera Navigator Encrypt.....	501
<i>Issues Fixed in Cloudera Navigator Encrypt 3.10.0</i>	501
<i>Issues Fixed in Cloudera Navigator Encrypt 3.9.0</i>	501
<i>Issues Fixed in Cloudera Navigator Encrypt 3.8.0</i>	502
<i>Issues Fixed in Cloudera Navigator Encrypt 3.7.1</i>	502
<i>Issues Fixed in Cloudera Navigator Encrypt 3.7.0</i>	503

Version and Download Information.....504

CDH 5 and Cloudera Manager 5 Requirements and Supported Versions.....505

CDH Requirements for Cloudera Manager.....	505
CDH and Cloudera Manager Supported Operating Systems.....	505
<i>CDH and Cloudera Manager 5.9.x Supported Operating Systems</i>	506
<i>CDH and Cloudera Manager 5.8.x Supported Operating Systems</i>	506
<i>CDH and Cloudera Manager 5.7.x Supported Operating Systems</i>	507
<i>CDH and Cloudera Manager 5.6.x Supported Operating Systems</i>	507
<i>CDH and Cloudera Manager 5.5.x Supported Operating Systems</i>	508
<i>CDH and Cloudera Manager 5.4.x Supported Operating Systems</i>	508
<i>CDH and Cloudera Manager 5.3.x Supported Operating Systems</i>	509
<i>CDH and Cloudera Manager 5.2.x Supported Operating Systems</i>	509
<i>CDH and Cloudera Manager 5.1.x Supported Operating Systems</i>	510
<i>CDH and Cloudera Manager 5.0.x Supported Operating Systems</i>	510
Filesystem Requirements.....	511
<i>Supported Filesystems</i>	511
<i>File Access Time</i>	511
CDH and Cloudera Manager Supported Databases.....	511
CDH and Cloudera Manager Supported JDK Versions.....	513
Cloudera Manager Supported Browsers.....	513
Supported Network Protocols.....	513
Multihoming Support.....	514
CDH and Cloudera Manager Supported Transport Layer Security Versions.....	514
Cloudera Manager Resource Requirements.....	515
CDH and Cloudera Manager Networking and Security Requirements.....	515
Product Compatibility Matrix for Apache Accumulo.....	519
Product Compatibility Matrix for Impala.....	519
Product Compatibility Matrix for Cloudera Distribution of Apache Kafka.....	521
Product Compatibility Matrix for Cloudera Navigator.....	521
<i>Cloudera Navigator Supported Databases</i>	528
<i>Cloudera Navigator Supported Browsers</i>	528
<i>Cloudera Navigator Supported CDH and Managed Service Versions</i>	528

Product Compatibility Matrix for Cloudera Navigator Encryption.....	530
<i>Cloudera Navigator Key Trustee Server</i>	531
<i>Key Trustee KMS</i>	532
<i>Cloudera Navigator Key HSM</i>	533
<i>Cloudera Navigator Encrypt</i>	534
Product Compatibility Matrix for Apache Sentry.....	535
Product Compatibility Matrix for Apache Spark.....	535
Product Compatibility Matrix for EMC DSSD D5.....	536
Product Compatibility for EMC Isilon.....	536
Product Compatibility Matrix for Backup and Disaster Recovery.....	537
<i>Supported Replication Scenarios for Clusters using Isilon Storage</i>	540
Supported Configurations with Virtualization and Cloud Platforms.....	540
<i>Amazon Web Services</i>	540
<i>Google Cloud Platform</i>	540
<i>Microsoft Azure</i>	540
<i>VMware</i>	541
Deprecated Items.....	542

About Cloudera Enterprise 5.x Release Notes

This guide contains release and download information for installers and administrators. It includes release notes as well as information about versions and downloads. The guide also provides a release matrix that shows which major and minor release version of a product is supported with which release version of Cloudera Manager, CDH and, if applicable, Cloudera Impala.

Cloudera Enterprise 5.x Release Notes

Continue reading:

- [CDH 5 Release Notes](#) on page 14
- [Cloudera Manager 5 Release Notes](#) on page 354
- [Cloudera Navigator 2 Data Management Release Notes](#) on page 461
- [Cloudera Navigator Key Trustee Server Release Notes](#) on page 485
- [Cloudera Navigator Key HSM Release Notes](#) on page 493
- [Key Trustee KMS Release Notes](#) on page 496
- [Cloudera Navigator Encrypt Release Notes](#) on page 500
- [Version and Download Information](#) on page 504
- [CDH 5 and Cloudera Manager 5 Requirements and Supported Versions](#) on page 505
- [Deprecated Items](#) on page 542

CDH 5 Release Notes

For links to the detailed change lists that describe the bug fixes and improvements to all of the CDH 5 projects, see the packaging section of [CDH Version and Packaging Information](#).

For more information about installing and configuring CDH 5, see [Cloudera Installation](#).

New Features and Changes in CDH 5



Note: There is no CDH 5.0.7, 5.1.1, 5.1.6, 5.2.2, 5.2.7, 5.3.7, or 5.4.6 release.

About Apache Hadoop MapReduce Version 1 (MRv1) and Version 2 (MRv2)



Important: Cloudera recommends that you use YARN (now production-ready) with CDH 5.

- **MapReduce 2.0 (MRv2):** CDH 5 includes MapReduce 2.0 (MRv2) running on YARN. The fundamental idea of the YARN architecture is to split up the two primary responsibilities of the JobTracker — resource management and job scheduling/monitoring — into separate daemons: a global ResourceManager (RM) and per-application ApplicationMasters (AM). With MRv2, the ResourceManager (RM) and per-node NodeManagers (NM), form the data-computation framework. The ResourceManager service effectively replaces the functions of the JobTracker, and NodeManagers run on worker nodes instead of TaskTracker daemons. The per-application ApplicationMaster is, in effect, a framework-specific library and is tasked with negotiating resources from the ResourceManager and working with the NodeManager(s) to execute and monitor the tasks. For details of the new architecture, see [Apache Hadoop NextGen MapReduce \(YARN\)](#).
- **MapReduce Version 1 (MRv1):** For backward compatibility, CDH 5 continues to support the original MapReduce JobTracker and TaskTrackers, but you should begin migrating to MRv2.



Note:

Cloudera does not support running MRv1 and YARN daemons on the same nodes at the same time.

- **Deprecated properties:**

In Hadoop 2.0.0 and later (MRv2), a number of Hadoop and HDFS properties have been deprecated. (The change dates from Hadoop 0.23.1, on which the Beta releases of CDH 4 were based). A list of deprecated properties and their replacements can be found at [Hadoop Deprecated Properties](#).



Note: All of these deprecated properties continue to work in MRv1. Conversely the newmapreduce*properties listed do not work in MRv1.

What's New In CDH 5.9.x

Apache Hadoop

- CDH 5.9 allows you to use temporary credentials to log in to Amazon S3. You can obtain temporary credentials from Amazon's Security Token Service (STS).

Apache HBase

- A tool has been added--org.apache.hadoop.hbase.replication.regionserver.DumpReplicationQueues--to dump existing replication peers, configurations, and queues when using HBase replication. The tool includes two flags:
 - --distributed - Polls each replication server for information about the replication queues being processed on this replication server. By default, this is not enabled, and the information about the replication queues and configuration is obtained from ZooKeeper.
 - --hdfs When --distributed is used, this flag attempts to calculate the total size of the WAL files used by the replication queues. Because multiple peers can be configured, this value can be overestimated.

For more information, see [Class DumpReplicationQueues](#).

- Metrics have been added that expose the amount of replayed work occurring in the HBase replication system. For more information on these metrics, see [Replication Metrics](#) in the *Apache HBase Reference Guide*.

Apache Hive

- [HIVE-14270](#) : Added parameters to optimize write performance for Hive tables and partitions that are stored on Amazon S3. See [Optimizing Hive Write Performance on Amazon S3](#).

Hue

- [HUE-2915](#): Integrates Hue with [Amazon S3](#). You can now access both S3 and HDFS in the File Browser, create tables from files in S3, and save query results in S3. See how to [Enable S3 Cloud Storage](#).
- [HUE-4039](#): Improves SQL Autocompleter. The new Autocompleter deeply understands Hive and Impala SQL dialects and provides smart suggestions based on your statement structure and cursor position. See how to manually [Enable and Disable Autocompleter](#).
- [HUE-3877](#): Adds support for Amazon RDS. You can now deploy Hue against an Amazon RDS database instance with MySQL, PostgreSQL, and Oracle engines.
- Rebase of Hue on upstream [Hue 3.11](#).

Apache Impala (incubating)

- Performance improvements:
 - [\[IMPALA-3206\]](#) Speedup for queries against DECIMAL columns in Avro tables. The code that parses DECIMAL values from Avro now uses native code generation.
 - [\[IMPALA-3674\]](#) Improved efficiency in LLVM code generation can reduce codegen time, especially for short queries.
 - [\[IMPALA-2979\]](#) Improvements to scheduling on worker nodes, enabled by the REPLICA_PREFERENCE query option. See [REPLICA_PREFERENCE Query Option \(or higher only\)](#) for details.
- [\[IMPALA-1683\]](#) The REFRESH statement can be applied to a single partition, rather than the entire table. See [REFRESH Statement](#) and [Refreshing a Single Partition](#) for details.
- Improvements to the Impala web user interface:
 - [\[IMPALA-2767\]](#) You can now force a session to expire by clicking a link in the web UI, on the **/sessions** tab.
 - [\[IMPALA-3715\]](#) The **/memz** tab includes more information about Impala memory usage.
 - [\[IMPALA-3716\]](#) The **Details** page for a query now includes a **Memory** tab.
- [\[IMPALA-3499\]](#) Scalability improvements to the catalog server. Impala handles internal communication more efficiently for tables with large numbers of columns and partitions, where the size of the metadata exceeds 2 GiB.

CDH 5 Release Notes

- [IMPALA-3677] You can send a `SIGUSR1` signal to any Impala-related daemon to write a Breakpad minidump. For advanced troubleshooting, you can now produce a minidump without triggering a crash. See [Breakpad Minidumps for Impala \(or higher only\)](#) for details about the Breakpad minidump feature.
- [IMPALA-3687] The schema reconciliation rules for Avro tables have changed slightly for `CHAR` and `VARCHAR` columns. Now, if the definition of such a column is changed in the Avro schema file, the column retains its `CHAR` or `VARCHAR` type as specified in the SQL definition, but the column name and comment from the Avro schema file take precedence. See [Creating Avro Tables](#) for details about column definitions in Avro tables.
- [IMPALA-3575] Some network operations now have additional timeout and retry settings. The extra configuration helps avoid failed queries for transient network problems, to avoid hangs when a sender or receiver fails in the middle of a network transmission, and to make cancellation requests more reliable despite network issues.

Apache Sentry

- Sentry adds support for securing data on Amazon RDS. As a result, Sentry will now be able to secure URIs with an RDS schema.
- [SENTRY-1233](#) - Logging improvements for `SentryConfigToolSolr`.
- [SENTRY-1119](#) - Allow data engines to obtain the `ActionFactory` directly from the configuration, instead of having hardcoded component-specific classes. This will allow external data engines to integrate with Sentry easily.
- [SENTRY-1229](#) - Added a basic configurable cache to `SentryGenericProviderBackend`.

Apache Spark

- You can now set up AWS credentials for Spark with the Hadoop credential provider, to avoid exposing the AWS secret key in configuration files.

Apache Sqoop

- The mainframe import module extension has been added to support data sets on tape.

Cloudera Search

- The Solr watchdog is now configured to use the fully qualified domain name (FQDN) of the host on which the Solr process is running (instead of `127.0.0.1`). You can override this configuration by setting `SOLR_HOSTNAME` environment variable to appropriate value (before starting the Solr server).
- Cloudera Search adds support for index snapshots. For more information on how to back up, migrate, or restore your indexed data, see [Backing Up and Restoring Cloudera Search](#).

What's New In CDH 5.8.x

What's New in CDH 5.8.3

This is a maintenance release that fixes some important issues. For details, see [Issues Fixed in CDH 5.8.x](#) on page 184.

What's New in CDH 5.8.2

This is a maintenance release that fixes some important issues. For details, see [Issues Fixed in CDH 5.8.x](#) on page 184.

What's New in CDH 5.8.0

The following sections describe new features introduced in CDH 5.8.0.

Operating System Support

- **Operating Systems** - Support for Debian 8.2.

Apache HBase

- Additional metrics have been added to monitor garbage collection pauses and other external pauses that might cause a server process to momentarily block any request processing.
- New throughput and Bloom filter metrics have been added to the existing HBase microbenchmarks.

- Downstream users of CDH who build on top of Apache HBase can now pull in fewer transitive dependencies by relying on shaded client artifacts. To do so, use the maven artifact `hbase-shaded-client` where you would normally use the `hbase-client` artifact. All HBase APIs remain the same.



Note: In a previous release of CDH, this maven artifact exists but does not contain any of the needed classes to interact with an HBase cluster.

- The HMaster Web UI now shows the aggregate of all space being used by snapshots.

Hue

General Features:

- Rebase of Hue on upstream [Hue 3.10](#).
- Refactor of Hue Infrastructure:
 - Performance is further optimized for large numbers of databases and tables.
 - Exporting and importing documents is improved.

SQL Editor and Browser:

- Revamp of the Hue SQL Application:
 - Editor is redesigned to create a single-page experience.
 - Code editor is redesigned with enhanced auto-complete, keyboard shorts, search and replace, and more.
 - A live status of the query history displays as an icon

.)

Admin and Security:

- [HUE-3386](#): Users are auto-logged out when TTL Expires. See [Securing Sessions](#).
- [Hue-3808](#): Users can do live DEBUG log toggling. See [Enable DEBUG](#).
- A new step in the Cloudera Manager Add Service wizard helps configure and test an external Hue database.

Search:

- [SENTRY-1217](#): Users can grant Sentry Solr privileges.

Oozie:

- [Hue-3464](#): The dashboard and editor are decoupled for granular access.
- Saved Hive queries can be dragged and dropped into a workflow.

Apache Impala (incubating)

See [New Features in Impala 2.6.x / CDH 5.8.x](#) on page 60.

Apache Oozie

- [OOZIE-2330](#) : The Spark Action now allows <file> and <archive> elements. It also omits <job-tracker> and <name-node> elements to allow for the use of the global or default values.
- The launcher job no longer uses YARN's uber mode by default.

Cloudera Search

- Cloudera Search adds support for storing permissions in the Sentry service. You can enable storing permissions in the Sentry service by [Enabling the Sentry Service for Solr](#). If you have already configured Sentry's policy file-based approach, you can migrate existing authorization settings as described in [Migrating from Sentry Policy Files to the Sentry Service](#). `solrctl` has been extended to support:
 - Migrating existing policy files to the Sentry service
 - Managing managing permissions in the Sentry service

Apache Sentry

- Sentry adds support for securing data on Amazon S3. As a result, Sentry will now be able to secure URIs with an S3 schema.
- Cloudera Search adds support for storing permissions in the Sentry service. You can enable storing permissions in the Sentry service by [Enabling the Sentry Service for Solr](#). If you have already configured Sentry's policy file-based approach, you can migrate existing authorization settings as described in [Migrating from Sentry Policy Files to the Sentry Service](#). `solrctl` has been extended to support:
 - Migrating existing policy files to the Sentry service
 - Managing managing permissions in the Sentry service
- [SENTRY-1175](#): Improved usability for Sentry URIs and URI privileges. If URIs in Hive DDL statements or URI privileges lack scheme and authority components, Sentry automatically completes such URLs by applying the default scheme and authority based on the HDFS configuration provided to Sentry.
- **Performance Improvements**
 - [SENTRY-1293](#): `ResourceAuthorizationProvider.doHasAccess` no longer performs expensive operations to convert string permissions to Privilege objects.
 - [SENTRY-1292](#): Reordered the `DBModelAction` `EnumSet` to improve authorization performance.

Key Trustee Server

Auto backup on install of Key Trustee Server

Key Trustee server backs up automatically on first run when using parcels and Cloudera Manager, and sets up a cron job (hourly) for continuous backup.

Auto backup on install of KMS

Key Trustee KMS backs up automatically on first run when using parcels and Cloudera Manager.

What's New In CDH 5.7.x

What's New in CDH 5.7.0

The following sections describe new features introduced in CDH 5.7.0.

Operating System Support

- **Operating Systems** - Support for:
 - RHEL/CentOS 6.6, 6.7, 7.1, and 7.2
 - Oracle Enterprise Linux (OEL) 7.1 and 7.2
 - SUSE Linux Enterprise Server (SLES) 11 with Service Packs 2, 3, 4
 - Debian: Wheezy 7.0, 7.1, and 7.8



Important: Cloudera supports RHEL 7 with the following limitations:

- Only RHEL 7.2 and 7.1 are supported. RHEL 7.0 is not supported.
- RHEL 7.1 is only supported with CDH 5.5 and higher.
- RHEL 7.2 is only supported with CDH 5.7 and higher.
- Only new installations of RHEL 7.2 and 7.1 are supported by Cloudera. For upgrades to RHEL 7.1 or 7.2, contact your OS vendor and see [Does Red Hat support upgrades between major versions of Red Hat Enterprise Linux?](#).

Apache Hadoop

- Improve support for heterogeneous storage. [Heterogeneous Storage Management](#) (HSM) is now supported natively in Cloudera Manager.

- [HADOOP-10651](#) - Service access can be restricted by IP and hostname. You can now define a whitelist and blacklist per service. The default whitelist is * and the default blacklist is empty.
- [HADOOP-12764](#) - You can configure KMS maxHttpHeaderSize. The default value of KMS maxHttpHeaderSize increased from 4096 to 65536 and is now configurable in `service.xml`.
- [HDFS-7279](#) - In CDH 5.5.0 and higher, DataNode WebHDFS implementation uses Netty as an HTTP server instead of Jetty. With improved buffer and connection management, Netty lowers the risk for DataNode latency and OutOfMemoryError (OOM).
- [HDFS-8873](#) - You can rate-limit the directoryScanner. A new configuration property, `dfs.datanode.directoryscan.throttle.limit.ms.per.sec`, allows you to reduce the impact on disk performance of directory scanning. The default value is 1000.
- [HDFS-9260](#) - Garbage collection of full block reports is improved. Data structures for `BlockInfo` and replicas were changed to keep them sorted. This allows for faster and easier garbage collection of full block reports.

Apache HBase

See also [Apache HBase Known Issues](#) on page 122.

- CDH 5.7.0 adds support for a snapshot owner for a table. To configure a table snapshot owner, set the `OWNER` attribute on the table to a valid HBase user. By default, the table owner is the user who created the table. The table owner and the table creator can restore the table from a snapshot.
- You can set an HDFS storage policy to store write-ahead logs (WALs) on solid-state drives (SSDs) or a mix of SSDs and spinning disks.
- Configuration for snapshot timeouts has been simplified. A single configuration option, `hbase.snapshot.master.timeout.millis`, controls how long the master waits for a response from the RegionServer before timing out. The default timeout value has been increased from 60,000 milliseconds to 300,000 to accommodate larger tables.
- You can optionally balance a table's regions by size by setting the `hbase.normalizer.enabled` property to `true`. The default value is `false`. To configure the time interval for the HMaster to check for region balance, set the `hbase.normalizer.period` property to the interval, in milliseconds. The default value is `1800000`, or 30 minutes.
- You can configure parallel request cancellation for multi-get operations using the `hbase.client.replica.interrupt.multiget` configuration property, using an advanced configuration snippet in Cloudera Manager, or in `hbase-site.xml` if you do not use Cloudera Manager.
- The `hbase-spark` module has been added, which provides support for using HBase data in Spark jobs using `HBaseContext` and `JavaHBaseContext` contexts. See the [HBase and Spark chapter](#) of the Apache HBase Reference Guide for details about building a Spark application with HBase support.
- The REST API now supports creating, reading, updating, and deleting namespaces.
- If you use the G1 garbage collector, you can disable the `BoundedByteBufferPool`.
- The HBase web UI includes graphical tools for managing MemStore and StoreFile details for a region.
- The HBase web UI displays the number of regions a table has.
- Two new methods of the Scan API, `setColumnFamilyTimeRange` and `getColumnFamilyTimeRange`, allow you to limit Scan results to versions of columns within a specified timestamp range.
- A new API, `MultiRowRangeFilter`, allows you to scan multiple row ranges in a single scan.
- A new client API, `SecurityCapability`, enables you to check whether the HBase server supports cell-level security.
- When using the `scan` command in HBase Shell, you can use the `ROWPREFIXFILTER` option to include only rows matching a given prefix, in addition to any other filters you specify. For example:

```
hbase> scan 't1', {ROWPREFIXFILTER => 'row2',
                   FILTER => (QualifierFilter (>=, 'binary:xyz')) AND
                   (TimestampsFilter ( 123, 456))}
```

- The new `get_splits` HBase Shell command returns the split points for a table. For example:

```
hbase> get_splits 't2'
Total number of splits = 5
```

```
=> [ "", "10", "20", "30", "40" ]
```

- Three new commands relating to the region normalizer have been added to the HBase Shell:
 - `normalizer_enabled` checks whether the region normalizer is enabled.
 - `normalizer_switch` toggles the region normalizer on or off.
 - `normalize` runs the region normalizer if it is enabled.
- When configuring a ZooKeeper quorum for HBase, you can now specify the port for each ZooKeeper host separately, instead of using the same port for each ZooKeeper host. The configuration property `hbase.zookeeper.clientPort` is no longer required. For example:

```
<property>
  <name>hbase.zookeeper.quorum</name>
  <value>zk1.example.com:2181,zk2.example.com:20000,zk3.example.com:31111</value>
</property>
```

- A new configuration option, `hbase.regionserver.hostname`, was added, but Cloudera recommends against its use. See [Compatibility Notes for CDH 5.7](#) on page 89.
- Two new configuration options allow you to disable loading of all coprocessors or only table-level coprocessors: `hbase.coprocessor.enabled` (which defaults to `true`) and `hbase.coprocessor.user.enabled` (which also defaults to `true`). Cloudera does not recommend disabling HBase-wide coprocessors, because security functionality is implemented using coprocessors. However, disabling table-level coprocessors may be appropriate in some cases.
- A new configuration option, `hbase.loadincremental.validate.hfile`, allows you to skip validation of HFiles during a bulk load operation when set to `false`. The default setting is `true`.
- The default `PermSize` for HBase processes is now set to 128 MB. This setting is effective only for JDK 7, because JDK 8 ignores the `PermSize` setting. To change this setting, edit the **HBase Client Environment Advanced Configuration Snippet (Safety Valve) for `hbase-env.sh`** if you use Cloudera Manager or `conf/hbase-env.sh` otherwise.
- In CDH 5.6 and lower, the `HBaseFsck#checkRegionConsistency()` method would throw an `IOException` if a region repair operation timed out after `hbase.hbck.assign.timeout` (which defaults to 120 seconds). This exception would cause the entire `hbck` operation to fail. In CDH 5.7.0, if the region being repaired is not `hbase:meta` or another system table, the region is skipped, an error is logged, and the `hbck` operation continues. This new behavior is disabled by default; to enable it, set the `hbase.hbck.skipped.regions.limit` option to an integer greater than 0. If more than this number of regions is skipped, the `hbck` operation fails.
- Two new options, `-exclusive` and `-disableBalancer`, have been added to the `hbck` utility. The `hbck` utility now runs without locks unless in fixer mode, and the balancer is only disabled in fixer mode, by default. You can disable these options to retain the old behavior, but Cloudera recommends using the new default behavior.
- Two new MapReduce jobs, `SyncTable` and `HashTable`, allow you to synchronize two different HBase tables that are each receiving live writes. To print usage instructions, run the job with no arguments. These examples show how to run these jobs:

```
$ bin/hbase org.apache.hadoop.hbase.mapreduce.SyncTable \
          --dryrun=true \
--sourcezkcluster=zk1.example.com,zk2.example.com,zk3.example.com:2181:/hbase \
          hdfs://nn:9000/ hashes/tableA \
          tableA tableA

$ bin/hbase org.apache.hadoop.hbase.mapreduce.HashTable \
          --batchsize=32000 \
          --numhashfiles=50 \
          --starttime=1265875194289 \
          --endtime=1265878794289 \
          --families=cf2,cf3 \
          TestTable /hashes/testTable
```

- The CopyTable command now allows you to override the `org.apache.hadoop.hbase.mapreduce.TableOutputFormat` property by prefixing the property keys with the `hbase.mapred.output.` prefix. For example, `hbase.mapred.output.hbase.security.authentication` is passed to `CopyTable` as `hbase.security.authentication`. This is useful when directing output to a peer cluster with different configuration settings.
- Several improvements have been made to the HBase canary, including:
 - Sniffing of regions and RegionServers is now parallel to improve performance in large clusters with more than 1000 regions and more than 500 RegionServers.
 - The canary sets `cachecheck` to `false` when performing Gets and Scans to avoid influencing the BlockCache.
 - `FirstKeyOnlyFilter` is used during Gets and Scans to improve performance in a flat wide table.
 - A region is selected at random when sniffing a RegionServer.
 - The sink class used by the canary is now configurable.
 - A new flag, `-allRegions`, sniffs all regions on a RegionServer if running in RegionServer mode.
- Distributed log replay has been disabled in CDH 5.7 and higher, due to reliability issues. Distributed log replay is unsupported.
- A new configuration option, `hbase.hfile.drop.behind.compaction`, causes the OS-level filesystem cache to be dropped behind compactions. This provides significant performance improvements on large compactions. To disable this behavior, set the option to `false`. It defaults to `true`.
- Two new configuration options, `hbase.hstore.compaction.max.size` and `hbase.hstore.compaction.max.size.offpeak`, allow you to specify a maximum compaction size during normal hours and during off-peak hours, if the off-peak feature is used. If unspecified, `hbase.hstore.compaction.max.size` defaults to 9,223,372,036,854,775,807 (`Long.MAX_VALUE`), which is essentially unbounded.
- The HDFS replication factor can now be specified per column family by setting the `DFS_REPLICATION` attribute of the column family to the desired number of replicas, either at table creation or by altering an existing table schema. If set to 0, the default replication factor is used. If fewer than the desired number of replicas exist, the HDFS FSCK utility reports it.
- If you use the ChaosMonkey tool, you can low-load a custom ChaosMonkey implementation by passing the class to the `-m` or `--monkey` option of the ChaosMonkey tool, in the same way that you would normally pass `SLOW` or `CALM`.
- A new configuration option, `hbase.use.dynamic.jars`, allows you to disable the dynamic classloader if set to `false`.

Apache Hive

- [HIVE-9298](#) - Supports reading alternate timestamp formats. The SerDe property, `timestamp.formats`, is added to allow you to pass in a comma-delimited list of alternate timestamp formats. For example, the following ALTER TABLE statement (in this case, with Joda date-time parsing) adds: `yyyy-MM-dd'T'HH:mm:ss` and milliseconds since [Unix epoch](#), represented by the special case pattern `millis`.

```
ALTER TABLE timestamp_formats SET SERDEPROPERTIES
("timestamp.formats"="yyyy-MM-dd'T'HH:mm:ss,millis");
```

- [HIVE-7292](#) - Hive on Spark (HoS) General Availability. This release supports running Hive jobs using Spark as an additional execution backend. This improves performance for Hive queries, especially those involving multiple reduce stages.
- [HIVE-12338](#) - Add Web UI to HiveServer2. Exposes a Web UI on the HiveServer2 service to surface process metrics (JVM stats, open session, and similar) and per-query runtime information (query plan, performance logs). The Web UI is available on port 10002 by default.
- [HIVE-10115](#) - Handle LDAP and Kerberos authentication on the same HiveServer2 interface. In a Kerberized cluster when alternate authentication is enabled on HiveServer2, it accepts Kerberos authentication. Before this enhancement, when LDAP authentication to HiveServer2 was enabled, the service blocked acceptance of other authentication mechanisms.
- Hive Metastore / HiveServer2 metadata scalability improvements. Fixes to improve scalability and performance include support for DirectSQL by default for metastore operations and incremental memory usage improvements.

CDH 5 Release Notes

- [HIVE-12271](#) - Add additional catalog and query execution metrics for Hive Metastore / HiveServer2 and integrate with Cloudera Manager. Additional Hive metrics include catalog information (number of databases, tables, and partitions in Hive Metastore) and query execution stats (number of worker threads used, job submission times, and statistics for planning and compilation times). All metrics are available in Cloudera Manager.
- The DATE datatype is supported for Parquet and Avro tables.

Hue

- Hive metastore service improvements make browsing Hive data faster and easier.
- SQL editor improvements:
 - SQL assist scales to thousands of tables and databases.
 - Improved row and column headers.
 - Button to delete the query history.
 - Button to format SQL.
- Security improvements:
 - Timeout property for idle sessions.
 - Customizable splash screen on the login page.
 - Support for password-protected SAML certificates.
 - Custom xmlsec1 binary for multiple SAML protocols.
 - Ability to synchronize groups on login for any authentication backend.
- Oozie application improvements:
 - Display graph of external workflow.
 - Option to dry run jobs on submission.
 - User timezone recognition.
 - Automatic email on failure.
 - Ability to execute individual actions.

Apache Impala (incubating)

See [New Features in Impala 2.5.x / CDH 5.7.x](#) on page 62.

MapReduce

- The `mapred job -history` command provides an efficient way to fetch history for MapReduce jobs.

Apache Oozie

- [OOZIE-2411](#) - The Oozie email action now supports BCC:, as well as To: and CC:.

Cloudera Search

- Improvements to `loadSolr`. For additional information about these improvements, see the [Morphlines Reference Guide](#). Improvements include:
 - `loadSolr` retries SolrJ requests if exceptions, such as connection resets, occur. This is useful in cases such as temporary Solr server overloads or transient connectivity failures. This behavior is enabled by default, but can be disabled by setting the Java system property `org.kitesdk.morphline.solr.LoadSolrBuilder.disableRetryPolicyByDefault=true`.
 - `loadSolr` includes an optional rate-limiting parameter that is used to set the maximum number of morphline records to ingest per second. The default value is no limit.

Apache Spark

- Spark is rebased on Apache Spark 1.6.0.
- [SPARK-10000](#) - Spark 1.6.0 includes a new unified memory manager. The new memory manager is turned off by default (unlike Apache Spark 1.6.0), to make it easier for users to migrate existing workloads, but it is supported.

- [SPARK-9999](#) - Spark 1.6.0 introduces a new Dataset API. However this API is experimental, likely to undergo some changes, and unsupported.
- [SPARK-6028](#) and [SPARK-6230](#) - Encryption support for Spark RPC
- [SPARK-2750](#) - HTTPS support for History Server and web UI
- Added support for Spark SQL (DataFrames) in PySpark.
- Added support for the following MLlib features:
 - spark.ml
 - ML pipeline APIs
- The `hbase-spark` module has been added, which provides support for using HBase data in Spark jobs using `HBaseContext` and `JavaHBaseContext` contexts. See the [HBase and Spark chapter](#) of the Apache HBase Reference Guide for details about building a Spark application with HBase support.
- The `DATE` datatype is supported for Parquet and Avro tables.

Apache Sentry

[SENTRY-510](#) - Added support for collection of metrics for the Sentry/HDFS plugin

YARN

- Preemption guarantees that important tasks are not starved for resources, while allowing the CDH cluster to be used for experimental and research tasks. In FairScheduler, you can now disable preemption for a specific queue to ensure that its resources are not taken for other tasks.

What's New in CDH 5.7.1

This is a maintenance release that fixes some important issues. For details, see [Issues Fixed in CDH 5.7.1](#) on page 205.

What's New in CDH 5.7.2

This is a maintenance release that fixes some important issues. For details, see [Issues Fixed in CDH 5.7.2](#) on page 202.

What's New in CDH 5.7.3

This is a maintenance release that fixes some important issues. For details, see [Issues Fixed in CDH 5.7.3](#) on page 200.

What's New in CDH 5.7.4

This is a maintenance release that fixes some important issues. For details, see [Issues Fixed in CDH 5.7.4](#) on page 198.

What's New in CDH 5.7.5

This is a maintenance release that fixes some important issues. For details, see [Issues Fixed in CDH 5.7.5](#) on page 196.

What's New In CDH 5.6.x

What's New in CDH 5.6.0

CDH 5.6.0 is a maintenance release that introduces [EMC DSSD D5 Storage Appliance Integration for Hadoop DataNodes](#).

What's New in CDH 5.6.1

This is a maintenance release that fixes some important issues. For details, see [Issues Fixed in CDH 5.6.1](#) on page 218.

What's New In CDH 5.5.x

What's New in CDH 5.5.0

The following sections describe new features introduced in CDH 5.5.0.

Operating System and Database Support

- **Operating Systems** - Support for RHEL/CentOS 6.6 (in SE Linux mode), 6.7, and 7.1, and Oracle Enterprise Linux 7.1.



Important: Cloudera supports RHEL 7 with the following limitations:

- Only RHEL 7.1 is supported. RHEL 7.0 is not supported.
- Only new installations of RHEL 7.1 are supported by Cloudera. For upgrades to RHEL 7.1, contact your OS vendor and see [Does Red Hat support upgrades between major versions of Red Hat Enterprise Linux?](#)

- **Databases** - Supports MariaDB 5.5, Oracle 12c, and PostgreSQL 9.4.

Apache Flume

- CDH 5.5 release is rebased on Flume 1.6.
- [FLUME-2498](#) Taildir source.
- [FLUME-2215](#) ResettableFileInputStream support for ucs-4 character.
- [FLUME-2729](#) PollableSource backoff times made configurable.
- [FLUME-2628](#) Netcat source support for different source encodings.
- [FLUME-2753](#) Support for empty replace string in Search and Replace interceptor.
- [FLUME-2763](#) Flume_env script support to handle JVM parameters.
- [FLUME-2095](#) JMS source support for username and password.

Apache Hadoop

- [HADOOP-1540](#) - DistCp supports file exclusions with a new filter option, `-exclusions <argument>`, to prevent files from being copied. The argument is a file that contains a list of Java regex patterns (one per line). If an exclusion pattern is matched, the file is not copied. To use, pass `-filters <pathToFileterFile>` to the `distcp` command.
- [HADOOP-8989](#) - The Hadoop shell now has a `find` utility, like that in UNIX, that allows users to search for files by name. Run `hadoop fs -help find` for more info.
- [HADOOP-11219](#), [HADOOP-7280](#) - WebImageViewer was upgraded to Netty 4. This does not affect the external classpath of Hadoop.
- [HADOOP-11827](#) - DistCp `buildListing()` now uses a threadpool to improve performance. To use, pass `--numListstatusThreads <numThreads>` to the `distcp` command. The default value is 1.
- [HDFS-6133](#) - HDFS balancer supports the exclusion of subtrees because running the HDFS balancer can destroy local data that is important for applications such as the HBase RegionServer.
- [HDFS-8828](#) - DistCp leverages HDFS snapshot diff to more easily build file and directory lists. The snapshot diff report provides diff information between two snapshots or between a snapshot and a non-HDFS directory.
- Improvements to HDFS scalability and performance:
 - [HDFS-7279](#) - In CDH 5.5.0 and higher, DataNode WebHDFS implementation uses Netty as an HTTP server instead of Jetty. With improved buffer and connection management, Netty lowers the risk for DataNode latency and OutOfMemoryError (OOM).
 - [HDFS-7435](#) and [HDFS-8867](#) add more efficient over-the-wire encoding.
 - [HDFS-7923](#) adds rate-limiting for block reports so that the NameNode is not swamped by DataNodes sending too many block reports at once.
 - [HDFS-7923](#) and [HDFS-7999](#) eliminate some cases on the DataNode side where I/O errors lead to scans being repeated on the local disks.
 - [HDFS-8581](#) fixes some cases where a lock is held for too long.
 - [HDFS-8792](#) and [HDFS-7609](#) optimize data structures on the NameNode side.
 - [HDFS-9107](#) fixes a bug that could limit scalability on larger clusters by causing the NameNode to falsely consider DataNodes to be dead.
 - Other bugs included: [HADOOP-11785](#), [HADOOP-12172](#), [HADOOP-11659](#), [HDFS-8845](#)

Apache HBase

- CDH now includes a *scanner heartbeat check*, which enforces a time limit on the execution of scan RPC requests. When the server receives a scan RPC request, a time limit is calculated to be half of the smaller of the two values `hbase.client.scanner.timeout.period` and `hbase.rpc.timeout`. When the time limit is reached, the server will return the results it has accumulated up to that point. For more information, see [Configuring the HBase Scanner Heartbeat](#).

Cloudera Search

- Cloudera Search adds support for Kerberos authentication for hosts running Solr behind a proxy server.
- Cloudera Search adds support for using LDAP and Active Directory for authentication.
- `solrctl` supports the Config API.

`solrctl` includes a `config` command that uses the Config API to directly manage configurations represented in Config objects. Config objects represent collection configuration information as specified by the `solrctl collection --create -c configName` command. `instancedirs` and Config objects handle the same information, meeting the same need from the Solr server perspective, but there are a number of differences between these two implementations.

Table 1: Config and instancedir Comparison

Attribute	Config	instancedir
Security	<p>Security support provided.</p> <ul style="list-style-type: none"> In a Kerberos-enabled cluster, the ZooKeeper hosts associated with configurations created using the Config API automatically has proper ZooKeeper ACLs. Because <code>instancedir</code> updates ZooKeeper directly, it is the client's responsibility to add the proper ACLs, which is cumbersome. Sentry can be used to control access to the Config API, providing access control. For more information, see Configuring Sentry Authorization for Solr. 	No ZooKeeper security support. Any user can create, delete, or modify <code>instancedirs</code> directly in ZooKeeper.
Creation method	Generated from existing configs or <code>instancedirs</code> in ZooKeeper using the ConfigSet API.	Manually edited locally and re-uploaded directly to ZooKeeper using <code>solrctl instancedir</code> .
Template support	<p>Several predefined templates are available. These can be used as the basis for creating additional configs. Additional templates can be created by creating configs that are immutable.</p> <p>Mutable templates that use a Managed Schema can be modified using the Schema API as opposed to being manually edited. As a result, configs are less flexible, but they are</p>	One standard template.

Attribute	Config	instancedir
	also less error-prone than instancedirs.	
Sentry support	Configs include a number of templates, each with Sentry-enabled and non-Sentry-enabled versions. To enable Sentry, choose a Sentry-enabled template.	instancedirs include a single template that supports enabling Sentry. To enable Sentry with instancedirs, overwrite the original <code>solrconfig.xml</code> file with <code>solrconfig.xml.secure</code> as described in Enabling Solr as a Client for the Sentry Service Using the Command Line .

- Solr includes a set of built-in immutable configurations.

These templates are instantiated when Solr is initialized. This means these templates are not automatically available after an upgrade. To enable these templates on upgraded installations, use `solrctl init` or initialize Solr using Cloudera Manager. The newly included templates and the functionality each template supports are as follows:

Table 2: Available Config Templates and Attributes

Template Name	Supports Schema API	Uses Schemaless Solr	Supports Sentry
predefinedTemplate			
managedTemplate	■		
schemalessTemplate	■	■	
predefinedTemplateSecure			■
managedTemplateSecure	■		■
schemalessTemplateSecure	■	■	■

Apache Sentry (incubating)

- Sentry is rebased on Apache Sentry 1.5.1.
- Sentry introduces column-level access control for tables in Hive and Impala. Previously, Sentry supported privilege granularity only at the table level. To restrict access to a column of sensitive data, you needed to first create a view for a subset of columns, and then grant privileges on that view. Instead, Sentry now allows you to assign the `SELECT` privilege on a subset of columns in a table.
- Support for enabling Kerberos authentication for the Sentry web server.

Apache Spark

- Spark is rebased on Apache Spark 1.5.0.
- Dynamic allocation is enabled by default. You can explicitly disable dynamic allocation by using the option: `spark.dynamicAllocation.enabled = false`. Dynamic allocation is implicitly disabled if `--num-executors` is specified in the job.
- The following Spark libraries are now supported:
 - Spark SQL (including DataFrames). The following Spark SQL features are not supported:
 - Thrift JDBC/ODBC server
 - Spark SQL CLI
 - MLlib. The following MLlib features are not supported:
 - `spark.ml`

- ML pipeline APIs

–

Apache Sqoop

- Sqoop is rebased on Apache Sqoop 1.4.6.

What's New in CDH 5.5.1

This is a maintenance release that fixes important issues in Apache Commons and Apache HBase; for details, see [Issues Fixed in CDH 5.5.1](#) on page 238.

What's New in CDH 5.5.2

This is a maintenance release that fixes some important issues. For details, see [Issues Fixed in CDH 5.5.2](#) on page 233.

What's New in CDH 5.5.4



Note:

There is no CDH 5.5.3 release. This skip in the CDH 5.x sequence allows the CDH and Cloudera Manager components of Cloudera Enterprise 5.5.4 to have consistent numbering.

This is a maintenance release that fixes some important issues. For details, see [Issues Fixed in CDH 5.5.4](#) on page 228.

What's New in CDH 5.5.5

This is a maintenance release that fixes some important issues. For details, see [Issues Fixed in CDH 5.5.5](#) on page 225

.

What's New In CDH 5.4.x

What's New in CDH 5.4.0



Important:

Upgrading to CDH 5.4.0 and later from any earlier release requires an HDFS metadata upgrade. Be careful to follow all of the upgrade steps as instructed.

For the latest Impala features, see [New Features in Impala 2.2.x / CDH 5.4.x](#) on page 67.

Operating System Support

CDH 5.4.0 adds support for RHEL and CentOS 6.6.

Security

The following summarizes new security capabilities in CDH 5.4.0:

- Secure Hue impersonation support for the Hue HBase application.
- Redaction of sensitive data from logs, centrally managed by Cloudera Manager, which prevents the WHERE clause in queries from leaking sensitive data into logs and management UIs.
- Cloudera Manager support for custom Kerberos principals.
- Kerberos support for Sqoop 2.
- Kerberos and TLS/SSL support for Flume Thrift source and sink.
- Navigator SAML support (requires Cloudera Manager).
- Navigator Key Trustee can now be installed and monitored by Cloudera Manager.
- Search can be configured to use SSL.

CDH 5 Release Notes

- Search supports protecting Solr and Lily HBase Indexer metadata using ZooKeeper ACLs in a Kerberos-enabled environment.

Apache Crunch

New HBase-related features:

- `HBaseTypes.cells()` was added to support serializing HBase Cell objects.
- All of the `HFileUtils` methods now support `PCollection` extends `Cell`, which includes both `PCollectionKeyValue` and `PCollectionCell`, on their method signatures.
- `HFileTarget`, `HBaseTarget`, and `HBaseSourceTarget` all support any subclass of `Cell` as an output type. `HFileSource` and `HBaseSourceTarget` still return `KeyValue` as the input type for backward compatibility with existing Crunch pipelines.

Developers can use Cell-based APIs in the same way as `KeyValue`-based APIs if they are not ready to update their code, but will probably have to change code inside `DoFns` because HBase 0.99 and later APIs deprecated or removed a number of methods from the HBase 0.96 API.

Apache Flume

CDH 5.4.0 adds SSL and Kerberos support for the Thrift source and sink, and implements DatasetSink 2.0.

Apache Hadoop

HDFS

- CDH 5.4.0 implements HDFS 2.6.0.
- CDH 5.4.0 HDFS provides hot-swap capability for DataNode disk drives. You can add or replace HDFS data volumes without shutting down the DataNode host ([HDFS-1362](#)).
- CDH 5.4.0 introduces cluster-wide redaction of sensitive data in logs and SQL queries. See [Sensitive Data Redaction](#).
- CDH 5.4.0 adds support for [Heterogenous Storage Policies](#).
- HDFS 2.6.0+ supports the option to configure AES encryption for block data transfer, using the property `dfs.encrypt.data.transfer.algorithm`. AES offers improved cryptographic strength and performance over the prior options of 3DES and RC4.

MapReduce

CDH 5.4.0 implements [MAPREDUCE-5785](#), which simplifies MapReduce job configuration. Instead of having to set both the heap size (`mapreduce.map.java.opts` or `mapreduce.reduce.java.opts`) and the container size (`mapreduce.map.memory.mb` or `mapreduce.reduce.memory.mb`), you can now choose to set only one of them; the other is inferred from `mapreduce.job.heap.memory-mb.ratio`. If you do not specify either of them, the container size defaults to 1 GB and the heap size is inferred.

For jobs that do not set the heap size, the JVM size increases from 200 MB to a default 820 MB. This is adequate for most jobs, but streaming tasks might need more memory because the Java process causes total usage to exceed the container size. This typically occurs only for those tasks relying on aggressive garbage collection to keep the heap under 200 MB.

YARN

- [YARN-2990](#) improves application launch time by 6 seconds when using FairScheduler (with the default Cloudera Manager settings shown in [YARN \(MR2 Included\) Properties in CDH 5.4.0](#)).

Apache HBase

CDH 5.4.0 implements HBase 1.0.

MultiWAL Support for HBase

CDH 5.4.0 introduces MultiWAL support for HBase region servers, allowing you to increase throughput when a region writes the write-ahead log (WAL).

doAs Impersonation for HBase

CDH 5.4.0 introduces `doAs` impersonation for the HBase Thrift server. `doAs` impersonation allows a client to authenticate to HBase as any user, and re-authenticate at any time, instead of as a static user only.

Read Replicas for HBase

CDH 5.4.0 introduces read replicas, along with a new timeline consistency model. This feature allows you to balance consistency and availability on a per-read basis, and provides a measure of high availability for reads if a RegionServer becomes unavailable.

Storing Medium Objects (MOBs) in HBase

CDH 5.4.0 HBase MOB allows you to store objects up to 10 MB (medium objects, or MOBs) directly in HBase while maintaining read and write performance.

Apache Hive

CDH 5.4.0 implements Hive 1.1.0. New capabilities include:

- A test-only version of [Hive on Spark](#) with the following limitations:
 - Parquet does not currently support vectorization; it simply ignores the setting of `hive.vectorized.execution.enabled`.
 - Hive on Spark does not yet support dynamic partition pruning.
 - Hive on Spark does not yet support HBase. If you want to interact with HBase, Cloudera recommends that you use Hive on MapReduce.

To deploy and test Hive on Spark in a test environment, use Cloudera Manager.



Important: Hive on Spark is included in CDH 5.4 and higher but is not currently supported nor recommended for production use. To try this feature, use it in a test environment until Cloudera resolves currently existing issues and limitations to make it ready for production use.

- Support for JAR files changes without scheduled maintenance.

To implement this capability, proceed as follows:

1. Set `hive.reloadable.aux.jars.path` in `/etc/hive/conf/hive-site.xml` to the directory that contains the JAR files.
2. Execute the `reload;` statement on HiveServer2 clients such as Beeline and the Hive JDBC.

- Beeline support for retrieving and printing query logs.

Some features in the upstream release are not yet supported for production use in CDH; these include:

- [HIVE-7935](#) - Support dynamic service discovery for HiveServer2
- [HIVE-6455](#) - Scalable dynamic partitioning and bucketing optimization
- [HIVE-5317](#) - Implement insert, update, and delete in Hive with full ACID support
- [HIVE-7068](#) - Integrate AccumuloStorageHandler
- [HIVE-7090](#) - Support session-level temporary tables in Hive
- [HIVE-7341](#) - Support for Table replication across HCatalog instances
- [HIVE-4752](#) - Add support for HiveServer2 to use Thrift over HTTP

Hue

CDH 5.4.0 adds the following:

- New Oozie editor
- Performance improvements
- New Search facets
- HBase impersonation

Kite

Kite in CDH has been rebased on the 1.0 release upstream. This breaks backward compatibility with existing APIs. The APIs are documented at <http://kitesdk.org/docs/1.0.0/apidocs/index.html>.

Notable changes are:

- Dataset writers that implement flush and sync now extend interfaces (Flushable and Syncable). Writers that no longer have misleading flush and sync methods.
- `DatasetReaderException`, `DatasetWriterException`, and `DatasetRepositoryException` have been removed and replaced with more specific exceptions, such as `IncompatibleSchemaException`. Exception classes now indicate what went wrong instead of what threw the exception.
- The `partition` API is no longer exposed; use the `view` API instead.
- `kite-data-hcatalog` is now `kite-data-hive`.



Note:

From 1.0 on, Kite will be strict about breaking compatibility and will use [semantic versioning](#) to signal which compatibility guarantees you can expect from a release (for example, incompatible changes require increasing the major version number). For more information, see the [Hello, Kite SDK 1.0](#) blog post.

Apache Oozie

- Added Spark action which lets you run Spark applications from Oozie workflows. See the [Oozie documentation](#) for more details.
- The Hive2 action now collects and reports Hadoop Job IDs for MapReduce jobs launched by Hive Server 2.
- The launcher job now uses YARN uber mode for all but the Shell action; this reduces the overhead (time and resources) of running these Oozie actions.

Apache Parquet

- Parquet memory manager now changes the row group size if the current size is expected to cause out-of-memory (OOM) errors because too many files are open. This causes a `WARN` message to be printed in the logs. A new setting, `parquet.memory.pool.ratio`, controls the percentage of the JVM's heap memory Parquet attempts to use.
- To improve job startup time, footers are no longer read by default for MapReduce jobs ([PARQUET-139](#)).



Note:

To revert to the old behavior (`ParquetFileReader` reads in all the files to obtain the footers), set `parquet.task.side.metadata` to `false` in the job configuration.

- The Parquet Avro object model can now read lists and maps written by Hive, Avro, and Thrift (similar capabilities were added to Hive in CDH 5.3). This compatibility fix does not change behavior. The extra record layer wrapping the list elements when Avro reads lists written by Hive can now be removed; to do this, set the expected Avro schema or set `parquet.avro.add-list-element-records` to `false`.
- Avro's map representation now writes null values correctly.
- The Parquet Thrift object model can now read data written by other object models (such as Hive, Impala, or Parquet-Avro), given a Thrift class for the data; compile a Thrift definition into an object, and supply it when creating the job.

Cloudera Search

- Solr metadata stored in ZooKeeper can now be protected by Zookeeper ACLs. In a Kerberos-enabled environment, Solr metadata stored in ZooKeeper is owned by the `solr` user and cannot be modified by other users.



Note:

- The Solr principal name can be configured in Cloudera Manager. The default name is `solr`, although other names can be specified.
- Collection configuration information stored under the `/solr/configs` znode is not affected by this change. As a result, collection configuration behavior is unchanged.

Administrators who modify Solr ZooKeeper metadata through operations like `solrctl init` or `solrctl cluster --put-solrxml` must now supply `solrctl` with a JAAS configuration using the `--jaas` configuration parameter. The JAAS configuration must specify the principal, typically `solr`, that the solr process uses.

End users, who typically do not need to modify Solr metadata, are unaffected by this change.

- Lily HBase Indexer metadata stored in ZooKeeper can now be protected by Zookeeper ACLs. In a Kerberos-enabled environment, Lily HBase Indexer metadata stored in ZooKeeper is owned by the Solr user and cannot be modified by other users.

End users, who typically do not manage the Lily HBase Indexer, are unaffected by this change.

- The Lily HBase Indexer supports restricting access using Sentry.
- Services included with Search for CDH 5.4.0, including Solr, Key-Value Store Indexer, and Flume, now support SSL.
- The Spark Indexer and the Lily HBase Batch Indexer support delegation tokens for mapper-only jobs.
- Search for CDH 5.4.0 implements [SOLR-5746](#), which improves `solr.xml` file parsing. Error checking for duplicated options or unknown option names was added. These checks can help identify mistakes made during manual edits of the `solr.xml` file. User-modified `solr.xml` files may cause errors on startup due to these parsing improvements.
- By default, CloudSolrServer now uses multiple threads to add documents.



Note: Note: Due to multithreading, if document addition is interrupted by an exception, some documents, in addition to the one being added when the failure occurred, may be added.

To get the old, single-threaded behavior, set parallel updates to false on the CloudSolrServer instance.

Related JIRA: [SOLR-4816](#).

- Updates are routed directly to the correct shard leader, eliminating document routing at the server. This allows for near linear indexing throughput scalability. Document routing requires that the `solrj` client must know each document's unique identifier. The unique identifiers allow the client to route the update directly to the correct shard. For additional information, see [Shards and Indexing Data in SolrCloud](#).

Related JIRA: [SOLR-4816](#).

- The `loadSolr` morphline command supports nested documents. For more information, see [Morphlines Reference Guide](#).
- `Navigator` can be used to audit Cloudera Search activity.
- Search for CDH 5.4 supports logging queries before they are executed. This allows you to identify queries that could increase resource consumption. This also enables improving schemas or filters to meet your performance requirements. To enable this feature, set the `SolrCore` and `SolrCore.Request` log level to DEBUG.

Related JIRA: [SOLR-6919](#)

- `UniqFieldsUpdateProcessorFactory`, which Solr Server implements, has been improved to support all of the `FieldMutatingUpdateProcessorFactory` selector options. The `<lst named="fields">` init param option is deprecated. Replace this option with `<arr name="fieldName">`.

If the `<lst named="fields">` init param option is used, Solr logs a warning.

Related JIRA: [SOLR-4249](#).

- Configuration information was previously available using `FieldMutatingUpdateProcessorFactory` (`oneOrMany` or `getBooleanArg`). Those methods are now deprecated. The methods have been moved to `NamedList` and renamed to `removeConfigArgs` and `removeBooleanArg`, respectively.

If the `oneOrMany` or `getBooleanArg` methods of `FieldMutatingUpdateProcessorFactory` are used, Solr logs a warning.

Related JIRA: [SOLR-5264](#).

Apache Spark

CDH 5.4.0 Spark is rebased on Apache Spark 1.3.0 and provides the following new capabilities:

CDH 5 Release Notes

- Spark Streaming WAL (write-ahead log) on HDFS, preventing any data loss on driver failure
- Kafka connector for Spark Streaming to avoid the need for the HDFS WAL
- Spark Streaming recovery is supported for production use
- Spark external shuffle service
- Improvements in automatically setting CDH classpaths for Avro, Parquet, Flume, and Hive
- Improvements in the collection of task metrics

The following is not supported in a production environment because of its immaturity:

- Spark SQL (which now includes dataframes)

See also [Apache Spark Known Issues](#) on page 155 and [Apache Spark Incompatible Changes and Limitations](#) on page 110.

Apache Sqoop

- **Sqoop 2:**
 - Implements Sqoop 2 version 1.99.5.
 - Sqoop 2 supports Kerberos.
 - Sqoop 2 supports PostgreSQL as the repository database.

What's New in CDH 5.4.1

This is a maintenance release that fixes the following issue; for details of other important fixes, see [Issues Fixed in CDH 5.4.1](#) on page 262.

Upgrades to CDH 5.4.1 from Releases Earlier than 5.4.0 May Fail

Problem: Because of a change in the implementation of the NameNode metadata upgrade mechanism, upgrading to CDH 5.4.1 from a version lower than 5.4.0 can take an inordinately long time. In a cluster with NameNode high availability (HA) configured and a large number of edit logs, the upgrade can fail, with errors indicating a timeout in the pre-upgrade step on JournalNodes.

What to do:

To avoid the problem: Do not upgrade to CDH 5.4.1; upgrade to CDH 5.4.2 instead.

If you experience the problem: If you have already started an upgrade and seen it fail, contact Cloudera Support. This problem involves no risk of data loss, and manual recovery is possible.

If you have already completed an upgrade to CDH 5.4.1, or are installing a new cluster: In this case you are not affected and can continue to run CDH 5.4.1.

Cloudera Search

- Beginning with CDH 5.4.1, Search for CDH supports configurable transaction log replication levels for replication logs stored in HDFS.

Apache Spark

- Spark supports submitting python applications in cluster mode.

What's New in CDH 5.4.2

This is a maintenance release that fixes the following issue; for details of other important fixes, see [Issues Fixed in CDH 5.4.2](#) on page 262.

Upgrades to CDH 5.4.1 from Releases Earlier than 5.4.0 May Fail

Problem: Because of a change in the implementation of the NameNode metadata upgrade mechanism, upgrading to CDH 5.4.1 from a version lower than 5.4.0 can take an inordinately long time. In a cluster with NameNode high availability (HA) configured and a large number of edit logs, the upgrade can fail, with errors indicating a timeout in the pre-upgrade step on JournalNodes.

What to do:

To avoid the problem: Do not upgrade to CDH 5.4.1; upgrade to CDH 5.4.2 instead.

If you experience the problem: If you have already started an upgrade and seen it fail, contact Cloudera Support. This problem involves no risk of data loss, and manual recovery is possible.

If you have already completed an upgrade to CDH 5.4.1, or are installing a new cluster: In this case you are not affected and can continue to run CDH 5.4.1.

What's New in CDH 5.4.3

This is a maintenance release that fixes the following issue; for details of other important fixes, see [Issues Fixed in CDH 5.4.3](#) on page 260.

NameNode Incorrectly Reports Missing Blocks During Rolling Upgrade

Problem: During a rolling upgrade to any of the releases listed below, the NameNode may report missing blocks after rolling back multiple DataNodes. This is caused by a race condition with block reporting between the DataNode and the NameNode. No permanent data loss occurs, but data can be unavailable for up to six hours before the problem corrects itself.

Releases affected: CDH 5.0.6, 5.1.5, 5.2.5, 5.3.3, 5.4.1, 5.4.2

What to do:

To avoid the problem: Cloudera advises skipping the affected releases and installing a release containing the fix. For example, do not upgrade to CDH 5.4.2; upgrade to CDH 5.4.3 instead.

The releases containing the fix are: CDH 5.3.4, 5.4.3

If you have already completed an upgrade to an affected release, or are installing a new cluster: You can continue to run the release, or upgrade to a release that is not affected.

What's New in CDH 5.4.4

This is a maintenance release that fixes some important issues. For details, see [Issues Fixed in CDH 5.4.4](#) on page 259.

What's New in CDH 5.4.5

This is a maintenance release that fixes some important issues. For details, see [Issues Fixed in CDH 5.4.5](#) on page 256.

What's New in CDH 5.4.7



Note:

There is no CDH 5.4.6 release. This skip in the CDH 5.x sequence allows the CDH and Cloudera Manager components of Cloudera Enterprise 5.4.7 to have consistent numbering.

This is a maintenance release that fixes some important issues. For details, see [Issues Fixed in CDH 5.4.7](#) on page 254

What's New in CDH 5.4.8

This is a maintenance release that fixes some important issues; for details, see [Issues Fixed in CDH 5.4.8](#) on page 251

What's New in CDH 5.4.9

This is a maintenance release that fixes some important issues. For details, see [Issues Fixed in CDH 5.4.9](#) on page 249.

What's New in CDH 5.4.10

This is a maintenance release that fixes some important issues. For details, see [Issues Fixed in CDH 5.4.10](#) on page 246.

What's New in CDH 5.4.11

This is a maintenance release that fixes some important issues. For details, see [Issues Fixed in CDH 5.4.11](#) on page 243

.

What's New In CDH 5.3.x

What's New in CDH 5.3.0

Oracle JDK 8 Support

CDH 5.3 supports Oracle JDK 1.8.

Apache Hadoop

HDFS

CDH 5.3 provides the following new capabilities:

- **HDFS Data At Rest Encryption** - This feature is now ready for use in production environments.



Important:

Client hosts may need a more recent version of `libcrypto.so`. See [Apache Hadoop Known Issues](#) on page 115 for more information.



Important:

Cloudera provides two solutions:

- **Navigator Encrypt** is production ready and available to Cloudera customers licensed for Cloudera Navigator. Navigator Encrypt operates at the Linux volume level, so it can encrypt cluster data inside and outside HDFS. Consult your Cloudera account team for more information.
- **HDFS Encryption** is production ready and operates at the HDFS directory level, enabling encryption to be applied only to HDFS folders where needed.

- **S3A** - S3A is an HDFS implementation of the Simple Storage Service (S3) from Amazon Web Services. It is similar to S3N, which is the other implementation of this functionality. The key difference is that S3A relies on the officially-supported AWS Java SDK for communicating with S3, while S3N uses a best-effort-supported `jets3t` library to do the same. For a listing of the parameters, see [HADOOP-10400](#).

YARN

YARN now provides a way for long-running applications to get new delegation tokens.

Apache Flume

CDH 5.3 provides a Kafka Channel ([FLUME-2500](#)).

Apache HBase

CDH 5.3 provides `checkAndMutate(RowMutations)`, in addition to existing support for atomic `checkAndPut` as well as `checkAndDelete` operations on individual rows ([HBASE-11796](#)).

Apache Hive

- Hive can use multiple HDFS encryption zones.
- Hive-HBase integration contains many fixes and new features such as reading HBase snapshots.
- Many Hive Parquet fixes.
- Hive Server 2 can handle multiple LDAP domains for authentication.

Hue

New Features:

- Hue is re-based on Hue 3.7
- [SAML authentication](#) has been revamped.
- CDH 5.3 simplifies the task of configuring Hue to store data in an Oracle database by bundling the Oracle Install Client.

Apache Oozie

- You can now update the definition and properties of an already running Coordinator. See the [documentation](#) for more information.
- A new `poll` command in the Oozie client polls a Workflow Job, Coordinator Job, Coordinator Action, or Bundle Job until it finishes. See the [documentation](#) for more information.

Apache Parquet

- [PARQUET-132](#): Add type parameter to AvroParquetInputFormat for Spark
- [PARQUET-107](#): Add option to disable summary metadata files
- [PARQUET-64](#): Add support for new type annotations (date, time, timestamp, etc.)

Cloudera Search

New Features:

- Cloudera Search includes a version of Kite 0.15.0, which includes all morphlines-related backports of all fixes and features in Kite 0.17.1. Morphlines now includes functionality that enables partially updating document as well as deleting documents. Partial updating or deleting can be completed by unique IDs or by documents that match a query. For additional information on Kite, see:
 - [Kite repository](#)
 - [Kite Release Notes](#)
 - [Kite documentation](#)
 - [Kite examples](#)
- CrunchIndexerTool now sends a commit to Solr on job success.
- Added support for deleting documents stored in Solr [by unique id](#) as well as [by query](#).

Apache Sentry (incubating)

- Sentry HDFS Plugin - Allows you to configure synchronization of Sentry privileges to HDFS ACLs for specific HDFS directories. This simplifies the process of sharing table data between Hive or Impala and other clients (such as MapReduce, Pig, Spark), by automatically updating the ACLs when a GRANT or REVOKE statement is executed. It also allows all roles and privileges to be managed in a central location (by Sentry).
- Metrics - CDH 5.3 supports metrics for the Sentry service. These metrics can be reported either through JMX or the console; configure this by setting the property `sentry.service.reporter` to `jmx` or `console`. A Sentry web server listening by default on port 51000 can expose the metrics in json format. Web reporting is disabled by default; enable it by setting `sentry.service.web.enable` to `true`. You can configure the port on which Sentry web server listens by means of the `sentry.service.web.port` property .

Apache Spark

- CDH Spark has been rebased on Apache Spark 1.2.0.
- Spark Streaming can now save incoming data to a WAL (write-ahead log) on HDFS, preventing any data loss on driver failure.



Important:

This feature is currently in Beta; Cloudera includes it in CDH Spark but does not support it.

- The YARN back end now supports dynamic allocation of executors. See [Job Scheduling](#) for more information.
- Native library paths (set via Spark configuration options) are correctly propagated to executors in YARN mode ([SPARK-1719](#)).
- The Snappy codec should now work out-of-the-box on Linux distributions with older `glibc` versions such as CentOS 5.
- Spark SQL now includes the Spark Thrift Server in CDH.



Important:

Spark SQL remains an experimental and unsupported feature in CDH.

See [Apache Spark Incompatible Changes and Limitations](#) on page 110 and [Apache Spark Known Issues](#) on page 155 for additional important information.

Apache Sqoop

- **Sqoop 1:**

- The MySQL connector now fetches on a row-by row-basis.
- The SQL server now has *upsert* (insert or update) support ([SQOOP-1403](#)).
- The Oracle direct connector now works with index-organized tables ([SQOOP-1632](#)). To use this capability, you must set the chunk method to PARTITION:

```
-Doraoop.chunk.method=PARTITION
```

- **Sqoop 2:**

- FROM/TO re-factoring is now supported ([SQOOP-1367](#)).

What's New in CDH 5.3.1

This is a maintenance release that fixes some important issues; for details, see [Issues Fixed in CDH 5.3.1](#) on page 283.

What's New in CDH 5.3.2

This is a maintenance release that fixes some important issues; for details, see [Issues Fixed in CDH 5.3.2](#) on page 280.

What's New in CDH 5.3.3

This is a maintenance release that fixes some important issues; for details, see [Issues Fixed in CDH 5.3.3](#) on page 279.

What's New in CDH 5.3.4

This is a maintenance release that fixes some important issues; for details, see [Issues Fixed in CDH 5.3.4](#) on page 277.

What's New in CDH 5.3.5

This is a maintenance release that fixes the following issue. For details of other fixes, see [Issues Fixed in CDH 5.3.5](#) on page 276:

Potential job failures during YARN rolling upgrades to CDH 5.3.4

Problem: A MapReduce security fix introduced a compatibility issue that results in job failures during YARN rolling upgrades from CDH 5.3.3 to CDH 5.3.4.

Release affected: CDH 5.3.4

Release containing the fix: CDH 5.3.5

Workarounds: You can use any one of the following workarounds for this issue:

- Upgrade to CDH 5.3.5.
- Restart any jobs that might have failed during the upgrade.
- Explicitly set the version of MapReduce to be used so it is picked on a per-job basis.

1. Update the YARN property, **MR Application Classpath** (`mapreduce.application.classpath`), either in Cloudera Manager or in the `mapred-site.xml` file. Remove all existing values and add a new entry:
`<parcel-path>/lib/hadoop-mapreduce/*`, where `<parcel-path>` is the absolute path to the parcel

installation. For example, the default installation path for the CDH 5.3.3 parcel would be:
`/opt/cloudera/parcels/CDH-5.3.3-1.cdh5.3.3.p0.5/lib/hadoop-mapreduce/*.`

2. Wait until jobs submitted with the above client configuration change have run to completion.
3. Upgrade to CDH 5.3.4.
4. Update the **MR Application Classpath** (`mapreduce.application.classpath`) property to point to the new CDH 5.3.4 parcel.

Do not delete the old parcel until after all jobs submitted prior to the upgrade have finished running.

What's New in CDH 5.3.6

This is a maintenance release that fixes several issues. For details, see [Issues Fixed in CDH 5.3.6](#) on page 275.

What's New in CDH 5.3.8

This is a maintenance release that fixes some important issues; for details, see [Issues Fixed in CDH 5.3.8](#) on page 272.

What's New in CDH 5.3.9

This is a maintenance release that fixes some important issues; for details, see [Issues Fixed in CDH 5.3.9](#) on page 271.

What's New in CDH 5.3.10

This is a maintenance release that fixes some important issues; for details, see [Issues Fixed in CDH 5.3.10](#) on page 268.

What's New In CDH 5.2.x

What's New in CDH 5.2.0



Important:

Upgrading to CDH 5.2.0 and later from any earlier release requires an HDFS metadata upgrade and other steps not usually required for a minor-release upgrade.

CDH 5.2.0 is a minor release which includes new features and bug fixes.

Go to [Issues Fixed In CDH 5.2.0](#) or keep reading for [New Features and Changes](#) on page 37.

See also [Known Issues in CDH 5](#) on page 111.

New Features and Changes

New features and changes are grouped by area.

Operating System Support

CDH 5.2.0 adds support for Ubuntu Trusty (version 14.04).



Important:

Installing CDH by adding a repository entails an additional step on Ubuntu Trusty, to ensure that you get the CDH version of ZooKeeper, rather than the version that is bundled with Trusty.

Apache Avro

CDH 5.2 implements Avro version 1.7.6, with backports from 1.7.7. Important changes include:

- [AVRO-1398](#): Increase default sync interval from 16k to 64k. There is a very small chance this could cause an incompatibility in some cases, but you can control the interval by setting `avro.mapred.sync.interval` in the MapReduce job configuration. For example, set it to 16000 to get the old behavior.

CDH 5 Release Notes

- [AVRO-1355](#): Record schema should reject duplicate field names. This change rejects schemas with duplicate field names. This could affect some applications, but if schemas have duplicate field names then they are unlikely to work properly in any case. The workaround is to make sure a record's field names are unique within the record.

Apache Hadoop HDFS

CDH 5.2 provides the following new capabilities:

- **HDFS Data at Rest Encryption**



Note: Cloudera provides the following two solutions for data at rest encryption:

- **Navigator Encrypt** - Is production ready and available for Cloudera customers licensed for Cloudera Navigator. Navigator Encrypt operates at the Linux volume level, so it can encrypt cluster data inside and outside HDFS. Talk to your Cloudera account team for more information about this capability.
- **HDFS Encryption** - Included in CDH 5.2.0 and operates at the HDFS folder level, enabling encryption to be applied only to HDFS folders where needed. This feature has several known limitations. Therefore, Cloudera does not currently support this feature in CDH 5.2 and it is *not* recommended for production use. To try the feature, upgrade to the latest version of CDH 5.

HDFS now implements transparent, end-to-end encryption of data read from and written to HDFS by creating encryption zones. An encryption zone is a directory in HDFS with all of its contents, that is, every file and subdirectory in it, encrypted. You can use either the **KMS** or the **Key Trustee** service to store, manage, and access encryption zone keys.

HDFS now implements transparent, end-to-end encryption of data read from and written to HDFS by creating encryption zones. An encryption zone is a directory in HDFS with all of its contents, that is, every file and subdirectory in it, encrypted.

- Extended attributes: HDFS XAttrs allow extended attributes to be stored per file (<https://issues.apache.org/jira/browse/HDFS-2006>).
- Authentication improvements when using an HTTP proxy server.
- A new Hadoop Metrics sink that allows writing directly to Graphite.
- Specification for Hadoop Compatible Filesystem effort.
- `OfflineImageViewer` to browse an `fsimage` via the WebHDFS API.
- Supportability improvements and bug fixes to the NFS gateway.
- Modernized web UIs (HTML5 and JavaScript) for HDFS daemons.

MapReduce

CDH 5.2 provides an optimized implementation of the mapper side of the MapReduce shuffle. The optimized implementation may require tuning different from the original implementation, and so it is considered experimental and is not enabled by default.

You can select this new implementation on a per-job basis by setting the job configuration value `mapreduce.job.map.output.collector.class` to `org.apache.hadoop.mapred.nativetask.NativeMapOutputCollectorDelegator`, or use enable Cloudera Manager to enable it.

Some jobs which use custom writable types or comparators may not be able to take advantage of the optimized implementation.

the following new capabilities and improvements:

YARN

CDH 5.2 provides the following new capabilities and improvements:

- New features and improvements in the Fair Scheduler:
 - New features:
 - Fair Scheduler now allows setting the `fairsharePreemptionThreshold` per queue (leaf and non-leaf). This threshold is a decimal value between 0 and 1; if a queue's usage is under (`preemption-threshold * fairshare`) for a configured duration, resources from other queues are preempted to satisfy this queue's request. Set this value in `fair-scheduler.xml`. The default value is 0.5.
 - Fair Scheduler now allows setting the `fairsharePreemptionTimeout` per queue (leaf and non-leaf). For a starved queue, this timeout determines when to trigger preemption from other queues. Set this value in `fair-scheduler.xml`.
 - Fair Scheduler now shows the **Steady Fair Share** in the Web UI. The Steady Fair Share is the share of the cluster resources a particular queue or pool would get if all existing queues had running applications.
 - Improvements:
 - Fair Scheduler uses **Instantaneous Fair Share** (`fairshare`) that considers only active queues) for scheduling decisions to improve the time to achieve steady state (`fairshare`).
 - The default for `maxAMShare` is now 0.5, meaning that only half the cluster's resources can be taken up by Application Masters. You can change this value in `fair-scheduler.xml`.
- YARN's REST APIs support submitting and killing applications.
- YARN's timeline store is integrated with Kerberos.

Apache Crunch

CDH 5.2 provides the following new capabilities:

- Improvements in [Scrunch](#), including:
 - New join API that matches the one in Crunch
 - New aggregation API, including support for [Algebird](#)-based aggregations
 - Built-in serialization support for all tuple types as well as case classes.
- A new module, `crunch-hive`, for reading and writing **Optimized Row Columnar** (ORC) Files with Crunch.

Apache Flume

CDH 5.2 provides the following new capabilities:

- [Kafka](#) Integration: Flume can now accept data from Kafka via the `KafkaSource` ([FLUME-2250](#)) and push to Kafka using the `KafkaSink` ([FLUME-2251](#)).
- Kite Sink can now write to Hive and HBase datasets ([FLUME-2463](#)).
- Flume agents can now be configured via Zookeeper (experimental, [FLUME-1491](#))
- Embedded Agents now support Interceptors ([FLUME-2426](#))
- `syslog` Sources now support configuring which fields should be kept ([FLUME-2438](#))
- File Channel replay is now much faster ([FLUME-2450](#))
- New regular-expression search-and-replace interceptor ([FLUME-2431](#))
- Backup checkpoints can be optionally compressed ([FLUME-2401](#))

Hue

CDH 5.2 provides the following new capabilities:

- New application for editing Sentry roles and Privileges on databases and tables
- Search App
- Heatmap, Tree, Leaflet widgets
- Micro-analysis of fields
- Exclusion facets
- Oozie Dashboard: bulk actions, faster display
- File Browser: drag-and-drop upload, history, ACLs edition

CDH 5 Release Notes

- Hive and Impala: LDAP pass-through, query expiration, TLS/SSL (Hive), new graphs
- Job Browser: YARN kill application button

Apache HBase

CDH 5.2 implements HBase 0.98.6, which represents a minor upgrade to HBase. This upgrade introduces new features and moves some features which were previously marked as experimental to fully supported status. For detailed information and instructions on how to use the new capabilities, see [New Features and Changes for HBase in CDH 5](#).

Apache Hive

CDH 5.2 introduces the following important changes in Hive.

- CDH 5.2 implements Hive 0.13, providing the following new capabilities:
 - Sub-queries in the `WHERE` clause
 - Common table expressions (CTE)
 - Parquet supports `timestamp`
 - HiveServer2 can be configured with a `hiverc` file that is automatically run when users connect
 - Permanent UDFs
 - HiveServer2 session and operation timeouts
 - Beeline accepts a `-i` option to initialize with a SQL file
 - New join syntax (implicit joins)
- As of CDH 5.2.0, you can create Avro-backed tables simply by using `STORED AS AVRO` in a DDL statement. The AvroSerDe takes care of creating the appropriate Avro schema from the Hive table schema, making it much easier to use Avro with Hive.
- Hive supports additional datatypes, as follows:
 - Hive can read `char` and `varchar` datatypes written by Hive, and `char` and `varchar` datatypes written by Impala.
 - Impala can read `char` and `varchar` datatypes written by Hive and Impala.

These new types have been enabled by expanding the supported DDL, so they are backward compatible. You can add `varchar(n)` columns by creating new tables with that type, or changing a `string` column in existing tables to `varchar`.



Note:

`char(n)` columns are not stored in a fixed-length representation, and do not improve performance (as they do in some other databases). Cloudera recommends that in most cases you use `text` or `varchar` instead.

- `DESCRIBE DATABASE` returns additional fields: `owner_name` and `owner_type`. The command will continue to behave as expected if you identify the field you're interested in by its (string) name, but could produce unexpected results if you use a numeric index to identify the field(s).

Impala

Impala in CDH 5.2.0 includes major new features such as spill-to-disk for memory-intensive queries, subquery enhancements, analytic functions, and new `CHAR` and `VARCHAR` data types. For the full feature list and more details, see [What's New in Apache Impala \(incubating\)](#) on page 58.

Kite

Kite is an open source set of libraries, references, tutorials, and code samples for building data-oriented systems and applications. For more information about Kite, see the [Kite SDK Development Guide](#).

Kite has been rebased to version 0.15.0 in CDH 5.2.0, from the base version 0.10.0 in CDH 5.1. `kite-morphlines` modules are backward-compatible, but this change breaks backward-compatibility for the `kite-data` API.

Kite Data

The Kite data API has had substantial updates since the version included in CDH 5.1.

Changes from 0.15.0

The Kite version in CDH 5.2 is based on 0.15.0, but includes some newer changes. Specifically, it includes support for dataset namespaces, which can be used to set the database in the Hive Metastore.

The introduction of namespaces changed the file system repository layout; now there is an additional namespace directory for datasets stored in HDFS (`repository/namespace/dataset/`). There are no compatibility problems when you use `Dataset URIs`, but all datasets created with the `DatasetRepository API` will be located in a namespace directory. This new directory level is not expected in Kite 0.15.0 or 0.16.0 and will prevent the dataset from being loaded. The work-around is to switch to using `Dataset URIs` (see below) that include the namespace component. Existing datasets will work without modification.

Except as noted above, Kite 0.15.0 in CDH 5.2 is fully backward-compatible. It can load datasets written with any previous Kite version.

Dataset URLs

Datasets are identified with a single URI, rather than a repository URI and dataset name. The dataset URI contains all the information Kite needs to determine which implementation (Hive, HBase, or HDFS) to use for the dataset, and includes both the dataset's name and a namespace.

The Kite API has been updated so that developers call methods in the `Datasets` utility class as they would use `DatasetRepository` methods. The `Datasets` methods are recommended, and the `DatasetRepository API` is deprecated.

Views

The Kite data API now allows you to select a view of the dataset by setting constraints. These constraints are used by Kite to automatically prune unnecessary partitions and filter records.

MapReduce input and output formats

The `kite-data-mapreduce` module has been added. It provides both `DatasetKeyInputFormat` and `DatasetKeyOutputFormat` that allow you to run MapReduce jobs over datasets or views. Spark is also supported by the input and output formats.

Dataset CLI tool

Kite now includes a command-line utility that can run common maintenance tasks, like creating a dataset, migrating a dataset's schema, copying from one dataset to another, and importing CSV data. It also has helpers that can create Avro schemas from data files and other Kite-related configuration.

Flume DatasetSink

The Flume `DatasetSink` has been updated for the `kite-data` API changes. It supports all previous configurations without modification.

In addition, the `DatasetSink` now supports dataset URIs with the configuration option `kite.dataset.uri`.

Apache Mahout

Mahout jobs launched from the `bin/mahout` script will now use cluster's default parameters, rather than hard-coded parameters from the library. This may change the algorithms' run-time behavior, possibly for the better. ([MAHOUT-1565](#).)

Apache Oozie

CDH 5.2 introduces the following important changes:

CDH 5 Release Notes

- A new [Hive 2 Action](#) allows Oozie to run HiveServer2 scripts. Using the Hive Action with HiveServer2 is now deprecated; you should switch to the new Hive 2 Action as soon as possible.
- The MapReduce action can now also be configured by Java code

This gives users the flexibility of using their own driver Java code for configuring the MR job, while also getting the advantages of the MapReduce action (instead of using the Java action). See the [documentation](#) for more info.

- The PurgeService is now able to remove completed child jobs from long running coordinator jobs
- ALL can now be set for `oozie.service.LiteWorkflowStoreService.user.retry.error.code.ext` to make Oozie retry actions automatically for every type of error
- All Oozie servers in an Oozie HA group now synchronize on the same randomly generated rolling secret for signing auth tokens
- You can now upgrade from CDH 4.x to CDH 5.2 and later with jobs in RUNNING and SUSPENDED states. (An upgrade from CDH 4.x to a CDH 5.x release *earlier* than CDH 5.2.0 would still require that no jobs be in either of those states).

[Apache Parquet \(incubating\)](#)

CDH 5.2 Parquet is rebased on Parquet 1.5 and Parquet-format 2.1.0.

[Cloudera Search](#)

New Features:

- Cloudera Search adds support for Spark indexing using the CrunchIndexerTool.
- Cloudera Search adds fault tolerance for single-shard deployments. This fault tolerance is enabled with a new `-a` option in `solrctl`, which configures shards to automatically be re-added on an existing, healthy node if the node hosting the shard become unavailable.
- Components of Cloudera Search include Kite 0.15.0. This includes all morphlines-related backports of all fixes and features in Kite 0.17.0. For additional information on Kite, see:
 - [Kite repository](#)
 - [Kite Release Notes](#)
 - [Kite documentation](#)
 - [Kite examples](#)
- Search adds support for multi-threaded faceting on fields. This enables parallelizing operations, allowing them to run more quickly on highly concurrent hardware. This is especially helpful in cases where faceting operations apply to large datasets over many fields.
- Search adds support for distributed pivot faceting, enabling faceting on multi-shard collections.

[Apache Sentry \(incubating\)](#)

CDH 5.2 introduces the following changes to Sentry.

Sentry Service:

- If you are using the database-backed Sentry service, upgrading from CDH 5.1 to CDH 5.2 will require a schema upgrade.
- **Hive SQL Syntax:**
 - GRANT and REVOKE statements have been expanded to include `WITH GRANT OPTION`, thus allowing you to delegate granting and revoking privileges.
 - The `SHOW GRANT ROLE` command has been updated to allow non-admin users to list grants for roles that are currently assigned to them.
 - The `SHOW ROLE GRANT GROUP <groupName>` command has been updated to allow non-admin users that are part of the group specified by `<groupName>` to list all roles assigned to this group.

Apache Spark

CDH 5.2 Spark is rebased on Apache Spark/Streaming 1.1 and provides the following new capabilities:

- Stability and performance improvements.
- New sort-based shuffle implementation (disabled by default).
- Better performance monitoring through the Spark UI.
- Support for arbitrary Hadoop `InputFormats` in PySpark.
- Improved Yarn support with several bug fixes.

Apache Sqoop

CDH 5.2 Sqoop 1 is rebased on Sqoop 1.4.5 and includes the following changes:

- Mainframe connector added.
- Parquet support added.

There are no changes for Sqoop 2.

What's New in CDH 5.2.1

CDH 5.2.1 maintenance release that fixes the “POODLE” and Apache Hadoop Distributed Cache vulnerabilities described below. All CDH 5.2.0 users should upgrade to 5.2.1 as soon as possible.

Go to [Issues Fixed In CDH 5.2.1](#).

“POODLE” Vulnerability on TLS/SSL enabled ports

The POODLE (Padding Oracle On Downgraded Legacy Encryption) attack forces the use of the obsolete SSLv3 protocol and then exploits a cryptographic flaw in SSLv3. The only solution is to disable SSLv3 entirely. This requires changes across a wide variety of components of CDH and Cloudera Manager in 5.2.0 and all earlier versions. CDH 5.2.1 provides these changes for CDH 5.2.0 deployments. For more information, see the [Cloudera Security Bulletin](#).

Apache Hadoop Distributed Cache Vulnerability

The Distributed Cache Vulnerability allows a malicious cluster user to expose private files owned by the user running the YARN NodeManager process. For more information, see the [Cloudera Security Bulletin](#).

What's New in CDH 5.2.3

This is a maintenance release that fixes some important issues; for details, see [Issues Fixed in CDH 5.2.3](#).

What's New in CDH 5.2.4

This is a maintenance release that fixes some important issues; for details, see [Issues Fixed in CDH 5.2.4](#).

What's New in CDH 5.2.5

This is a maintenance release that fixes some important issues; for details, see [Issues Fixed in CDH 5.2.5](#).

What's New in CDH 5.2.6

This is a maintenance release that fixes some important issues; for details, see [Issues Fixed in CDH 5.2.6](#).

What's New in CDH 5.1.x

The following sections describe what's new in CDH 5.1.x releases. You can also review [Known Issues in CDH 5](#) on page 111 and [Incompatible Changes and Limitations](#) on page 84.

What's New in CDH 5.1.0

This is a minor release, which includes new features, changes, and fixed issues. See also [Issues Fixed in CDH 5.1.0](#) on page 297.

CDH 5 Release Notes

New Features and Changes in CDH 5.1.0

This is a minor release which introduces the following new features and changes, organized by component. See also [What's New in Apache Impala \(incubating\)](#) on page 58.

Operating System Support

CDH 5.1 adds support for version 6.5 of RHEL and related platforms.

Apache Crunch

- CDH 5.1.0 implements Crunch 0.10.0.

Apache Flume

- CDH 5.1.0 implements Flume 1.5.0.

Apache Hadoop

HDFS

POSIX Access Control Lists: As of CDH 5.1, HDFS supports POSIX Access Control Lists (ACLs), an addition to the traditional POSIX permissions model already supported. ACLs provide fine-grained control of permissions for HDFS files by providing a way to set different permissions for specific named users or named groups.

NFS Gateway Improvements: CDH 5.1 makes the following improvements to the HDFS NFS gateway capability:

- Subdirectory mounts :
 - Previously, clients could mount only the HDFS root directory.
 - As of CDH 5.1, a single mount point, configured via the `nfs.export.point` property in `hdfs-site.xml` on the NFS gateway node, is available to clients.
- Improved support for Kerberized clusters ([HDFS-5898](#)):
 - Previously the NFS Gateway could connect to a secure cluster, but didn't support logging in from a keytab.
 - As of CDH 5.1, set the `nfs.kerberos.principal` and `nfs.keytab.file` properties in `hdfs-site.xml` to allow users to log in from a keytab.
- Support for port monitoring ([HDFS-6406](#)):
 - Previously, the NFS Gateway would always accept connections from any client.
 - As of CDH 5.1, set `nfs.port.monitoring.disabled` to `false` in `hdfs-site.xml` to allow connections only from privileged ports (those with root access).
- Static uid/gid mapping for NFS clients that are not in synch with the NFS Gateway ([HDFS-6435](#)):
 - NFS sends UIDs and GIDs over the network from client to server, meaning that the UIDs and GIDs must be in synch between clients and server machines in order for users and groups to be set appropriately for file access and file creation; this is usually but not always the case.
 - As of CDH 5.1, you can configure a static UID/GID mapping file, by default `/etc/nfs.map`.
 - You can change the default (to use a different file path) by means of the `nfs.static.mapping.file` property in `hdfs-site.xml`.
 - The following sample entries illustrate the format of the file:

```
uid 10 100 # Map the remote UID 10 to the local UID 100
gid 11 101 # Map the remote GID 11 to the local GID 101
```

- Hadoop portmap, or insecure system portmap, no longer required:
 - Many supported OS have portmap bugs detailed [here](#).
 - CDH 5.1 allows you to circumvent the problems by starting the NFS gateway as root, whether you install CDH from packages or parcels.

**Note:**

After initially registering with the system portmap as root, the NFS Gateway drops privileges and runs as a regular user.

- Cloudera Manager starts the gateway as root by default.
- Support for AIX NFS clients ([HDFS-6549](#)):
 - To deploy AIX NFS clients, set `nfs.aix.compatibility.mode.enabled` to `true` in `hdfs-site.xml`.
 - This enables code that handles bugs in the AIX implementation of NFS.

MapReduce and YARN

YARN with Impala supports Dynamic Prioritization.

Apache HBase

- CDH 5.1.0 implements HBase 0.98.
- As of CDH 5.1.0, HBase fully supports BucketCache, which was introduced as an experimental feature in CDH 5 Beta 1.
- HBase now supports access control for `EXEC` permissions.
- CDH 5.1.0 HBase introduces a reverse scan API; allowing you to scan a table in reverse.
- You can now run a MapReduce job over a snapshot from HBase, rather than being limited to live data.
- A new stateless streaming scanner is available over the REST API.
- The `delete*` methods of the Delete class of the HBase Client API now use the timestamp from the constructor, the same behavior as the Put class. (In HBase versions before CDH 5.1, the `delete*` methods ignored the constructor's timestamp, and used the value of `HConstants.LATEST_TIMESTAMP`. This behavior was different from the behavior of the `add()` methods of the Put class.)
- The `SnapshotInfo` tool has been enhanced in the following ways:
 - A new option, `-list-snapshots`, has been added to the `SnapshotInfo` command. This option allows you to list snapshots on either a local or remote server.
 - You can now pass the `-size-in-bytes` flag to print the size of snapshot files in bytes rather than the default human-readable format.
 - The size of each snapshot file in bytes is checked against the size reported in the manifest, and if the two sizes differ, the tool reports the file as corrupt.
- A new `-target` option for `ExportSnapshot` allows you to specify a different name for the target cluster from the snapshot name on the source cluster.

In addition, Cloudera has fixed some binary incompatibilities between HBase 0.96 and 0.98. As a result, the incompatibilities introduced by [HBASE-10452](#) and [HBASE-10339](#) do not affect CDH 5.1 HBase, as explained below:

- HBASE-10452 introduced a new exception and error message in `setTimeStamp()`, for an extremely unlikely event when where getting a `TimeRange` could fail because of an integer overflow. CDH 5.1 suppresses the new exception to retain compatibility with HBase 0.96, but logs the error.
- HBASE-10339 contained code which inadvertently changed the signatures of the `getFamilyMap` method. CDH 5.1 restores these signatures to those used in HBase 0.96, to retain compatibility.

Apache Hive

- Permission inheritance fixes
- Support for decimal computation, and for reading and writing decimal-format data from and to Parquet and Avro

Hue

CDH 5.1.0 implements Hue 3.6.

New Features:

CDH 5 Release Notes

- Search App v2:
 - 100% Dynamic dashboard
 - Drag-and-Drop dashboard builder
 - Text, Timeline, Pie, Line, Bar, Map, Filters, Grid and HTML widgets
 - Solr Index creation wizard (from a file)
- Ability to view compressed Snappy, Avro and Parquet files
- Impala HA
- Close Impala and Hive sessions queries and commands

Apache Mahout

- CDH 5.1.0 implements Mahout 0.9.

See also [Apache Mahout Incompatible Changes and Limitations](#) on page 106.

Apache Oozie

- You can now submit Sqoop jobs from the Oozie command line.
- LAST_ONLY execution mode now works correctly ([OOZIE-1319](#)).

Cloudera Search

New Features:

- A Quick Start script that automates using Search to query data from the Enron Email dataset. The script downloads the data, expands it, moves it to HDFS, indexes, and pushes the results live. The documentation now also includes a companion quick start guide, which describes the tasks the script completes, as well as customization options.
- `solrctl` now has built-in support for schema-less Solr.
- Sentry-based document-level security for role-based access control of a collection. Document-level access control associates authorization tokens with each document in the collection, enabling granting Sentry roles access to sets of documents in a collection.
- Cloudera Search includes a version of Kite 0.10.0, which includes all morphlines-related backports of all fixes and features in Kite 0.15.0. For additional information on Kite, see:
 - [Kite repository](#)
 - [Kite Release Notes](#)
 - [Kite documentation](#)
 - [Kite examples](#)
- Support for the Parquet file format is included with this version of Kite 0.10.0.
- Inclusion of hbase-indexer-1.5.1, a new version of the Lily HBase Indexer. This new version of the indexer includes the 0.10.0 version of Kite mentioned above. This 0.10.0 version of Kite includes the backports and fixes included in Kite 0.15.0.

Apache Sentry (incubating)

- CDH 5.1.0 implements Sentry 1.2. This includes a database-backed Sentry service which uses the more traditional GRANT/REVOKE statements instead of the previous policy file approach making it easier to maintain and modify privileges.
- Revised authorization privilege model for Hive and Impala.

Apache Spark

- CDH 5.1.0 implements Spark 1.0.
- The `spark-submit` command abstracts across the variety of deployment modes that Spark supports and takes care of assembling the classpath for you.
- Application History Server (SparkHistoryServer) improves monitoring capabilities.
- You can launch PySpark applications against YARN clusters. PySpark currently only works in YARN Client mode.

Other improvements include:

- Streaming integration with Kerberos
- Addition of more algorithms to MLLib (Sparse Vector Support)
- Improvements to Avro integration
- Spark SQL alpha release (new SQL engine). Spark SQL allows you to run SQL statements inside a Spark application that manipulate and produce RDDs.

**Note:**

Because of its immaturity and alpha status, Cloudera does not currently offer commercial support for Spark SQL, but bundles it with our distribution so that you can try it out.

- Authentication of all Spark communications

What's New in CDH 5.1.2

This is a maintenance release which fixes several issues. See [Issues Fixed in CDH 5.1.2](#) on page 296

**Note:**

There is no CDH 5.1.1 release. This skip in the CDH 5.x sequence allows the CDH and CM components of Cloudera Enterprise 5.1.2 to have consistent numbering.

What's New in CDH 5.1.3

This is a maintenance release that fixes several issues. See [Issues Fixed in CDH 5.1.3](#) on page 295.

What's New in CDH 5.1.4

This is a maintenance release that fixes important security issues. See [Issues Fixed in CDH 5.1.4](#) on page 294,

What's New in CDH 5.1.5

This is a maintenance release that fixes several issues. See [Issues Fixed in CDH 5.1.5](#) on page 293.

What's New in CDH 5.0.x

Use these links to go to a specific release.

What's New in CDH 5.0.0

This is a major release which includes new features, changes, and fixed issues. See also [Issues Fixed in CDH 5.0.x](#) on page 298 and [Known Issues in CDH 5](#) on page 111.

For information about CDH 5 Beta releases, see [What's New In CDH 5 Beta Releases](#) on page 50.

New Features and Changes in CDH 5.0.0***Apache Hadoop*****HDFS****New Features:**

- [HDFS-5776](#)- Hedged reads in HDFS for improved HBase MTTR.
- [HDFS-4685](#)- Implementation of extended file access control lists in HDFS.

Notable Bug Fixes:

- [HDFS-5339](#) - WebHDFS URI does not accept logical nameservices when security is enabled.
- [HDFS-5898](#) - Allow NFS gateway to login/relogin from its Kerberos keytab.
- [HDFS-5921](#) - "Browse filesystem" on the Namenode UI does not work if any directory has the sticky bit set.
- HDFS and Hive replication between different Kerberos realms now works.

CDH 5 Release Notes

- [HDFS-5922](#) - DataNode heartbeat thread can get stuck in a tight loop.

MapReduce & YARN

New Feature:

- FairScheduler supports moving running applications between queries.

Notable Bug Fixes:

- Several critical fixes to stabilize ResourceManager HA - Web UI, unmanaged ApplicationMasters and secure-cluster support.
- Support for large values of `mapreduce.task.io.sort.mb`.
- JobHistory Server has information on failed MapReduce jobs.

Apache HBase

New Features:

- [HBASE-10436](#)- Restore RegionServer lists removed from HBase 0.96.0 JMX.

Many of the metrics exposed in CDH 4/0.94 were removed with the refactorization of metrics in CDH 5/0.96. This patch restores the availability of the lists of live and dead RegionServers. In 0.94 this was a large nested structure as shown below, which included the RegionServer lists and metrics from each region.

```
{  
    "name" : "hadoop:service=Master,name=Master",  
    "modelerType" : "org.apache.hadoop.hbase.master.MXBeanImpl",  
    "ZookeeperQuorum" : "localhost:2181",  
    ....  
    "RegionsInTransition" : [ ],  
    "RegionServers" : [ {  
        "key" : "localhost,48346,1390857257246",  
        "value" : {  
            "load" : 2,  
        ....  
    }  
}
```

CDH 5 Beta 1 and Beta 2 did not contain this list; they only displayed counts of the number of live and dead RegionServers. As of CDH 5.0.0, this list is now presented in a semi-colon separated field as follows:

```
{  
    "name" : "Hadoop:service=HBase,name=Master,sub=Server",  
    "modelerType" : "Master,sub=Server",  
    "tag.Context" : "master",  
    "tag.liveRegionServers" : "localhost,56196,1391992019130",  
    "tag.deadRegionServers" :  
  
    "localhost,40010,1391035309673;localhost,41408,1391990380724;localhost,38682,1390950017735",  
    ....  
}
```

- Assorted usability and compatibility improvements as well as improvements to exporting snapshots.

Apache Flume

New Feature:

- The HBase Sink now supports coalescing multiple Increment RPCs into one ([FLUME-2338](#)).

Changed Behavior:

- File Channel Write timeout has been removed and the configuration parameter is now ignored ([FLUME-2307](#)).
- Syslog UDP source can now accept larger messages ([FLUME-2130](#)).
- AsyncHBase Sink is now fully functional ([FLUME-2334](#)).
- Use standard lookup to find queue/topic in JMS Source ([FLUME-2311](#)).

Notable Bug Fixes:

- Deadlock fixed in Dataset sink ([FLUME-2320](#)).
- FileChannel Dual Checkpoint Backup Thread is now released on application stop ([FLUME-2328](#)).
- Spool Dir source now checks interrupt flag before writing to channel ([FLUME-2283](#)).
- Morphline sink increments eventDrainAttemptCount when it takes event from channel ([FLUME-2323](#)).
- Bucketwriter now permanently closed only on idle and roll timeouts ([FLUME-2325](#)).
- BucketWriter#close now cancels idleFuture ([FLUME-2305](#)).

Apache Oozie

As of CDH 5.0.0 Oozie includes a glob pattern feature ([OOZIE-1471](#)), allowing you do a move of wild cards in the FS Action. For example:

```
<fs name="archive-files">
<move source="hdfs://namenode/output/*"
target="hdfs://namenode/archive" />
<ok to="next"/>
<error to="fail"/>
</fs>
```

By default, up to 1000 files can be matched; you can change this default by means of the `oozie.action.fs.glob.max` parameter.

Cloudera Search

- Cloudera Search includes a version of Kite 0.10.0, which includes backports of all fixes and features in Kite 0.12.0. For additional information on Kite, see:
 - [Kite repository](#)
 - [Kite Release Notes](#)
 - [Kite documentation](#)
 - [Kite examples](#)

What's New in CDH 5.0.1

This is a maintenance release which fixes several issues. In addition, it introduces a change to the configuration for HTTPS communication between HDFS and YARN. See also [Issues Fixed in CDH 5.0.1](#) on page 301.

New Features and Changes in CDH 5.0.1

Enabling TLS/SSL in CDH 5: Enabling HTTPS communication in CDH 5 requires extra configuration properties to be added to YARN (`yarn-site.xml` and `mapred-site.xml`) and HDFS (`hdfs-site.xml`), in addition to the existing configuration settings described [here](#).

What's New in CDH 5.0.2

This is a maintenance release which fixes several issues. See [Issues Fixed in CDH 5.0.2](#) on page 301.

What's New in CDH 5.0.3

This is a maintenance release that fixes several issues. See [Issues Fixed in CDH 5.0.3](#) on page 300.

What's New in CDH 5.0.4

This is a maintenance release that fixes several issues. See [Issues Fixed in CDH 5.0.4](#) on page 299.

What's New in CDH 5.0.5

This is a maintenance release that fixes the “POODLE” and Apache Hadoop Distributed Cache vulnerabilities described in [“POODLE” Vulnerability on TLS/SSL enabled ports](#) on page 299, as well as other issues. All CDH 5.0.x users should upgrade to 5.0.5 as soon as possible. See [Issues Fixed in CDH 5.0.5](#) on page 299.

What's New in CDH 5.0.6

This is a maintenance release that fixes several issues. See [Issues Fixed in CDH 5.0.6](#) on page 298.

What's New In CDH 5 Beta Releases

Use these links to go to a specific release.

What's New in CDH 5 Beta 1

This is a beta release which previews new features, changes, and fixed issues.

Oracle JDK 7 Support

- CDH 5 supports Oracle JDK 1.7 and supports users running applications compiled with JDK 1.7. For CDH 5 Beta 1 the certified version is JDK 1.7.0_25. Cloudera has tested this version across all components.
- CDH 5 does not support JDK 1.6; you must install JDK 1.7.

Apache Flume

New Features:

- [FLUME-2190](#) - Includes a new Twitter Source that feeds off the Twitter firehose
- [FLUME-2109](#) - HTTP Source now supports HTTPS
- Flume now auto-detects Cloudera Search dependencies.

Apache Hadoop

HDFS

New Features:

- [HDFS-4953](#): Enable HDFS local reads via mmap.
- [HDFS-2802](#): Support for RW/RO snapshots in HDFS. See:
`hadoop-hdfs-project/hadoop-hdfs/src/site/apt/HdfsNfsGateway.apt.vm`
- [HDFS-4750](#): Support NFSv3 interface to HDFS.
- [HDFS-4817](#): Make HDFS advisory caching configurable on a per-file basis.
- [HDFS-3601](#): Add BlockPlacementPolicyWithNodeGroup to support block placement with 4-layer network topology.
- [HDFS-5122](#): Support failover and retry in WebHdfsFileSystem for NN HA.
- [HDFS-4772](#) / [HDFS-5043](#): Add number of children (of a directory) in HdfsFileStatus.
- [HDFS-4434](#): Provide a mapping from INodeId to INode. See: `/ .reserved/.inodes/<INODE_NUMBER>`
- [HDFS-2576](#): Enhances the DistributedFileSystem's Create API so that clients can specify favored DataNodes for a file's blocks.

Changed Features:

- [HDFS-4659](#): Support setting execution bit for regular files.
 - **Impact:** In CDH 5, files copied out of `copyToLocal` may now have the executable bit set if it was set when they were created or copied into HDFS.
- [HDFS-4594](#): WebHDFS open sets Content-Length header to what is specified by length parameter rather than how much data is actually returned.
 - **Impact:** In CDH 5, Content-Length header will contain the number of bytes actually returned, rather than the request length.

Changed Behavior:

- [HDFS-4645](#): Move from randomly generated block ID to sequentially generated block ID.
- [HDFS-4451](#): HDFS balancer command returns exit code 1 on success instead of 0.

MapReduce v2 (YARN)

New Features:

- **ResourceManager High Availability:** YARN now allows you to use multiple ResourceManagers so that there is no single point of failure. In-flight jobs are recovered without re-running completed tasks.
- Monitoring and enforcing memory and CPU-based resource utilization using cgroups.

- **Continuous Scheduling:** This feature decouples scheduling from the node heartbeats for improved performance in large clusters.

Changed Feature:

- **ResourceManager Restart:** Persistent implementations of the RMStateStore (filesystem-based and ZooKeeper-based) allow recovery of in-flight jobs.

Apache HBase

Summary of New Features

- Support for Hadoop 2.0
- Improved MTTR (meta first recovery, distributed log replay)
- Improved compatibility and upgradeability (Protobuf serialization format)
- Namespaces added for administrative domains
- Snapshots (ported to 0.94 / CDH4.2)
- Online region merge mechanisms added
- Major security and functional improvements made for the REST proxy server

Administrative Features

Protobuf: All of the serialization that goes across the wire between servers written to and read by HBase file formats have been converted to extensible Protobuf encodings. This breaks compatibility with previous versions but should make future extensions less likely to break compatibility in these areas. This feature is enabled by default.

- [HBASE-5305](#): Improve cross-version compatibility and upgradeability.
- [HBASE-7898](#): Serializing cells over RPC.

Namespaces: Namespaces is a new feature that groups tables into different administrative domains. An admin can be only given rights to act upon a particular namespace. This feature is enabled by default and requires file system layout changes that must be completed during upgrade.

- [HBASE-8015](#): Added support for namespaces.

MTTR Improvements: Mean time to recovery has greatly improved.

- [HBASE-7590](#): “Costless” notifications from master to rs/clients.
- [HBASE-7213](#) / [HBASE-8631](#): New .meta suffix to separate HLog file / Recover Meta before other regions in case of server crash.
- [HBASE-7006](#): Distributed log replay (Caveat).
- [HBASE-9116](#): Adds a view/edit tool for favored node mappings for regions (incomplete, likely a dot version).

Metrics: There are several new metrics and a new naming convention for metrics in HBase. This also includes metrics for each region.

- [HBASE-3614](#): Per region metrics.
- [HBASE-4050](#): Rationalize metrics; Update HBase metrics framework to metrics2.

Miscellaneous:

- [HBASE-7403](#): HBase online region merge.
- Shell improvements; tables list to be more well-rounded.
- [HBASE-5953](#): Expose the current state of the balancerSwitch.
- [HBASE-5934](#): Add the ability for Performance Evaluation to set table compression.
- [HBASE-6135](#): New Web UI.
- [HBASE-8148](#): Allow IPC to bind to a specific address (also 0.94.7)
- [HBASE-5498](#): Secure Bulk Load (also 0.94.5)

CDH 5 Release Notes

Backup and Disaster Recovery Features

Replication: Several critical bug fixes.

- [HBASE-9373](#): Replication has been hardened.
- [HBASE-9158](#): Serious bug in cyclic replication.
- [HBASE-8737](#): Changes to the replication RPC to use cell blocks.

Snapshots: HBase table snapshots were backported to 0.94.x. There are some incompatibilities between the implementation released in CDH 4 with that in CDH 5.

- [HBASE-7290](#): Online snapshots (backported to 0.94.x).
- [HBASE-8352](#): Rename snapshots folder from `.snapshot` to `.hbase-snapshots` (Incompatible change).

Copy table:

- [HBASE-8609](#): Add startRow-stopRow options to the CopyTable.

Import:

- [HBASE-7702](#): Add filtering to import jobs.

HBase Proxies

The REST server now supports Hadoop authentication and authorization mechanisms. The Avro gateway has been removed while the Thrift2 proxy has made progress but is not complete. However, it has been included as a preview feature.

REST:

- [HBASE-9347](#): Support for specifying filter in REST server requests.
- [HBASE-7803](#): Support caching on scan.
- [HBASE-7757](#): Add Web UI for Thrift and REST servers.
- [HBASE-5050](#): SPNEGO-based authentication.
- [HBASE-8661](#): Support REST over HTTPS.
- [HBASE-8662](#): Support for impersonation.
- [HBASE-7986](#): [REST] Make HTablePool size configurable.

Thrift:

- [HBASE-5879](#): Enable JMX metrics collection for the Thrift proxy.

Thrift2: Ongoing efforts to match Thrift and REST functionality. (Incomplete, only a preview feature)

Avro:

- [HBASE-5948](#): Avro gateway removed.

Stability Features

There have been several bug fixes, test fixes and configuration default changes that greatly increase our confidence in the stability of the 0.96.0 release. The main improvement comes from the use of a systematic fault-injection framework.

- [HBASE-7721](#) Atomic multi-row mutations in META
- Integration testing
- [HBASE-7977](#) - TableLocks
- [HBASE-7898 many flaky tests hardened](#)

Performance Features

Several features have been added to improve throughput and performance characteristics of HBase and its clients.

**Warning:**

Currently the 0.95.2/CDH 5 beta 1 release will suffer performance degradation when over 40 nodes are used when compared to CDH 4.

Throughput:

- [HBASE-4676](#): Prefix compression / tree encoding.
- [HBASE-8334](#): Essential column families on by default (filtering optimization).
- [HBASE-5074](#) / [HBASE-8322](#): Re-enable HBase checksums by default.
- [HBASE-6466](#): Enable multi-threaded memstore flush
- [HBASE-6783](#): Make short circuit read the default.

Predictable Performance:

- [HBASE-5959](#): Added a Stochastic LoadBalancer
- [HBASE-7842](#): Exploring compactor.
- [HBASE-7236](#): Add per-table/per-cf configuration via metadata
- [HBASE-8163](#): MemStoreChunkPool: Improvement for Java GC
- [HBASE-4391](#)/[HBASE-6567](#): Mlock / Memory locking improvements (less disk swap).
- [HBASE-4391](#): Bucket cache (untested)

Miscellaneous:

- [HBASE-6870](#): Improvement to HTable coprocessorExec scan performance.

Developer Features

These features are to aid application developers or for major changes that will enable future minor version improvements.

- [HBASE-9121](#): HTrace updates.
- [HBASE-8375](#): Durability setting per table.
 - [HBASE-7801](#): Deferred sync for WAL logs (0.94.7 and later)
- [HBASE-7897](#): Tags supported in cell interface (for future security features).
- [HBASE-5937](#): Refactor HLog into interface (allows for new HLogs in 0.96.x).
- [HBASE-4336](#): Modularization of POM / Multiple jars (many follow-ons, [HBASE-7898](#)).
- [HBASE-8224](#): Publish -hadoop1 and -hadoop2 versioned jars to Maven (CDH published jars are assumed -hadoop2).
- [HBASE-9164](#): Move towards Cell interface in client instead of KeyValue.
- [HBASE-7898](#): Serializing cells over RPC.
- [HBASE-7725](#): Add ability to create custom compaction request.

Hue**New Features:**

- With the Sqoop 2 application, data from databases can be easily exported or imported into HDFS in a scalable manner. The Job Wizard hides the complexity of creating Sqoop jobs and the dashboard offers live progress and log access.
- Zookeeper App: Navigate and browse the Znode hierarchy and content of a Zookeeper cluster. Znodes can be added, deleted and edited. Multi-clusters are supported and various statistics are available for them.
- The Hue Shell application has been removed and replaced by the Pig Editor, HBase Browser and the Sqoop 1 apps.
- Python 2.6 is required.
- Beeswax daemon has been replaced by HiveServer2.
- CDH 5 Hue will only work with HiveServer2 from CDH 5. No support for impersonation.

Hue also includes the following changed features (Updated to upstream version 3.0.0):

CDH 5 Release Notes

- [\[HUE-897\]](#) - [core] Redesign of the overall layout
- [\[HUE-1521\]](#) - [core] Improve JobTracker High Availability
- [\[HUE-1493\]](#) - [beeswax] Replace the Beeswax server with HiveServer2
- [\[HUE-1474\]](#) - [core] Upgrade Django backend version from 1.2 to 1.4
- [\[HUE-1506\]](#) - [search] Impersonation support added
- [\[HUE-1475\]](#) - [core] Switch back from the Spawning web server
- [\[HUE-917\]](#) - Support SAML based authentication to enable single sign-on (SSO)

From master:

- [\[HUE-950\]](#) - [core] Improvements to the document model
- [\[HUE-1595\]](#) - Integrate Metastore data into Hive and Impala Query UIs
- [\[HUE-1275\]](#) - [metastore] Show Metastore table details
- [\[HUE-1622\]](#) - [core] Mini tour added to Hue home page

Apache Hive and HCatalog

New Features (Updated to upstream version 0.11.0):

- [\[HIVE-446\]](#) - Implement TRUNCATE for table data
- [\[HIVE-896\]](#) - Add LEAD/LAG/FIRST/LAST analytical windowing functions to Hive
- [\[HIVE-2693\]](#) - Add DECIMAL data type
- [\[HIVE-3834\]](#) - Support ALTER VIEW AS SELECT in Hive

Performance improvements (from 0.12):

- [\[HIVE-3764\]](#) - Support metastore version consistency check
- [\[HIVE-305\]](#) - Port Hadoop streaming process's counters/status reporters to Hive Transforms
- [\[HIVE-1402\]](#) - Add parallel ORDER BY to Hive
- [\[HIVE-2206\]](#) - Add a new optimizer for query correlation discovery and optimization
- [\[HIVE-2517\]](#) - Support GROUP BY on struct type
- [\[HIVE-2655\]](#) - Ability to define functions in HQL
- [\[HIVE-4911\]](#) - Enable QOP configuration for HiveServer2 Thrift transport

Cloudera Impala

Cloudera Impala 1.2.0 is now available as part of CDH 5. For more details on Impala, refer the [Impala Documentation](#).

Llama

Llama is a system that mediates resource management between Cloudera Impala and Hadoop YARN. Llama enables Impala to reserve, use, and release resource allocations in a Hadoop cluster. Llama is only required if resource management is enabled in Impala.

Apache Mahout

New Features (Updated to Mahout 0.8):

- Numerous performance improvements to Vector and Matrix implementations, APIs and their iterators (see also [MAHOUT-1192](#), [MAHOUT-1202](#))
- Numerous performance improvements to the recommender implementations (see also [MAHOUT-1272](#), [MAHOUT-1035](#), [MAHOUT-1042](#), [MAHOUT-1151](#), [MAHOUT-1166](#), [MAHOUT-1167](#), [MAHOUT-1169](#), [MAHOUT-1205](#), [MAHOUT-1264](#))
- [MAHOUT-1088](#): Support for biased item-based recommender.
- [MAHOUT-1089](#): SGD matrix factorization for rating prediction with user and item biases.
- [MAHOUT-1106](#): Support for SVD++
- [MAHOUT-944](#): Support for converting one or more Lucene storage indexes to SequenceFiles as well as an upgrade of the supported Lucene version to Lucene 4.3.
- [MAHOUT-1154](#) and related: New streaming k-means implementation that offers online (and fast) clustering.

- [MAHOUT-833](#): Make conversion to SequenceFiles Map-Reduce. 'seqdirectory' can now be run as a MapReduce job.
- [MAHOUT-1052](#): Add an option to MinHashDriver that specifies the dimension of vector to hash (indexes or values).
- [MAHOUT-884](#): Matrix concatenate utility; presently only concatenates two matrices.

Apache Oozie

New Features:

- Updated to Oozie 4.0.0.
- **High Availability:** Multiple Oozie servers can now be utilized to provide an HA Oozie service as well as provide horizontal scalability. See upstream [documentation](#) for more details.
- **HCatalog Integration:** HCatalog table partitions can now be used as data dependencies in coordinators. See upstream [documentation](#) for more details. .
- **SLA Monitoring:** Oozie can now actively monitor SLA-sensitive jobs and send out notifications for SLA meets and misses. SLA information is also now available through a new SLA tab in the Oozie Web UI, JMS messages, and a REST API. See upstream [documentation](#).
- **JMS Notifications:** Oozie can now publish notifications to a JMS Provider about job status changes and SLA events. See upstream [documentation](#).
- The FileSystem action can now use glob patterns for file paths when doing move, delete, chmod, and chgrp.

Cloudera Search

Cloudera Search 1.0.0 is now available as part of CDH 5. For more details on Search see the [Search documentation](#).

The Cloudera Development Kit (CDK) is a set of libraries and tools that can be used with Search and other CDH components to build jobs/systems on top of the Hadoop ecosystem. See the [CDK Documentation](#) and [Release Notes](#) for more details.



Note: An existing dependency, Apache Tika, has been upgraded to version 1.4.

Apache Sentry (incubating)

CDH 5 Beta 1 includes the first upstream release of Apache Sentry, `sentry-1.2.0-incubating`.

Apache Sqoop

CDH 5 Sqoop 1 has been rebased on Apache Sqoop 1.4.4.

What's New in CDH 5 Beta 2

This is a beta release which previews new features, changes, and fixed issues. See also [Issues Fixed in CDH 5 Beta 2](#) on page 305.

New Features and Changes in CDH 5 Beta 2

Apache Crunch

The Apache Crunch™ project develops and supports Java APIs that simplify the process of creating data pipelines on top of Apache Hadoop. The Crunch APIs are modeled after FlumeJava (PDF), which is the library that Google uses for building data pipelines on top of their own implementation of MapReduce.

Apache DataFu

- Upgraded from version 0.4 to 1.1.0 (this upgrade is not backward compatible).
- New features include UDFS SHA, SimpleRandomSample, COALESCE, ReservoirSample, EmptyBagToNullFields, and many others.

Apache Flume

- [FLUME-2294](#) - Added a new sink to write Kite datasets.
- [FLUME-2056](#) - Spooling Directory Source can now only pass the name of the file in the event headers.

CDH 5 Release Notes

- [FLUME-2155](#) - File Channel is indexed during replay to improve replay performance for faster startup.
- [FLUME-2217](#) - Syslog Sources can optionally preserve all syslog headers in the message body.
- [FLUME-2052](#) - Spooling Directory Source can now replace or ignore malformed characters in input files.

Apache Hadoop HDFS

New Features/Improvements:

- As of CDH 5 Beta 2, you can upgrade HDFS with high availability (HA) enabled, if you are using Quorum-based storage. (Quorum-based storage is the only method available in CDH 5; NFS shared storage is not supported.)
- [HDFS-4949](#) - CDH 5 Beta 2 supports HDFS caching.
- As of CDH 5 Beta 2, you can configure an NFSv3 gateway that allows any NFSv3-compatible client to mount HDFS as a file system on the client's local file system.
- [HDFS-5709](#) - Improve upgrade with existing files and directories named .snapshot.

Major Bug Fixes:

- [HDFS-5449](#) - Fix WebHDFS compatibility break.
- [HDFS-5671](#) - Fix socket leak in `DFSInputStream#getBlockReader`.
- [HDFS-5353](#) - Short circuit reads fail when `dfs.encrypt.data.transfer` is enabled.
- [HDFS-5438](#) - Flaws in block report processing can cause data loss.

Changed Behavior:

- As of CDH 5 Beta 2, in order for the NameNode to start up on a secure cluster, you should have the `dfs.web.authentication.kerberos.principal` property defined in `hdfs-site.xml`. For clusters managed by Cloudera Manager, you do not need to explicitly define this property.
- [HDFS-5037](#) - Active NameNode should trigger its own edit log rolls. Clients will now retry for a configurable period when encountering a NameNode in Safe Mode.
- The default behavior of the `mkdir` command has changed. As of CDH 5 Beta 2, if the parent folder does not exist, the `-p` switch must be explicitly mentioned otherwise the command fails.

MapReduce (MRv1 and YARN)

- Fair Scheduler (in YARN and MRv1) now supports advance configuration to automatically place applications in queues.
- MapReduce now supports running multiple reducers in `über` mode and in local job runner.

Apache HBase

- **Online Schema Change** is now a supported feature.
- **Online Region Merge** is now a supported feature.
- **Namespaces:** CDH 5 Beta 2 includes the namespaces feature which enables different sets of tables to be administered by different administrative users. All upgraded tables will live in the default "hbase" namespace. Administrators may create new namespaces and create tables users with rights to the namespace may administer permissions on the tables within the namespace.
- There have been several improvements to HBase's **mean time to recovery** (mttr) in the face of Master or RegionServer failures.
 - Distributed log splitting has matured, and is always activated. The option to use the old slower splitting mechanism no longer exists.
 - Failure detection time has been improved. New notifications are now sent when RegionServers or Masters fail which triggers corrective action quickly.
 - The Meta table has a dedicated write ahead log which enables faster recovery region recovery if the RegionServer serving meta goes down.
- The **Region Balancer** has been significantly updated to take more load attributes into account.

- Added **TableSnapshotInputFormat** and **TableSnapshotScanner** to perform scans over HBase table snapshots from the client side, bypassing the HBase servers. The former configures a MapReduce job, while the latter does a single client-side scan over snapshot files. Can also be used with offline HBase with in-place or exported snapshot files.
- The **KeyValue** API has been deprecated for applications in favor of the **Cell** interface. Users upgrading to HBase 0.96 may still use **KeyValue** by future upgrades may remove the class or parts of its functionality. Users are encouraged to update their applications to use the new Cell interface.
- Currently Experimental features:
 - **Distributed log replay:** This mechanism allows for faster recovery from RegionServer failures but has one special case where it will violate ACID guarantees. Cloudera does not currently recommend activating this feature.
 - **Bucket cache:** This is an offheap caching mechanism that use extra RAM and block devices (such as flash drives) to greatly increase the read caching capabilities provided by the BlockCache. Cloudera does not currently recommend activating this feature.
 - **Favored nodes:** This feature enables HBase to better control where its data is written to in HDFS in order to better preserve performance after a failure. This is disabled currently because it doesn't interact well with the HBase Balancer or HDFS Balancer. Cloudera does not currently recommend activating this feature.

See this [blog post](#) for more details.

Apache Hive

New Features:

- Improved JDBC specification coverage:
 - Improvements to `getDatabaseMajorVersion()`, `getDatabaseMinorVersion()` APIs ([HIVE-3181](#))
 - Added JDBC support for new datatypes: Char ([HIVE-5683](#)), Decimal ([HIVE-5355](#)) and Varchar ([HIVE-5209](#))
 - You can now specify the database for a session in the HiveServer2 connection URL ([HIVE-4256](#))
- Encrypted communication between the Hive Server and Clients. This includes TLS/SSL encryption for non-Kerberos connections to HiveServer2 ([HIVE-5351](#)).
- A native Parquet SerDe is now available as part of the CDH 5 Beta 2 package. Users can directly create a Parquet format table without any external package dependency.

Changed Behavior:

- [HIVE-4256](#) - With Sentry enabled, the use `<database>` command is now executed as part of the connection to HiveServer2. Hence, a user with no privileges to access a database will not be allowed to connect to HiveServer2.

Hue

- Hue has been upgraded to version 3.5.0.
- Impala and Hive Editor are now one-page apps. The Editor, Progress, Table list and Results are all on the same page
- Result graphing for the Hive and Impala Editors.
- Editor and Dashboard for Oozie SLA, crontab and credentials.
- The Sqoop2 app supports autocomplete of database and table names/fields.
- [DBQuery App](#): MySQL and PostgreSQL Query Editors.
- New Search feature: [Graphical facets](#)
- Integrate external Web applications in any language. See this [blog post](#) for more details.
- Create Hive tables and load quoted CSV data. Tutorial available [here](#).
- Submit any Oozie jobs directly from HDFS. Tutorial available [here](#)
- New [SAML backend](#) enables single sign-on (SSO) with Hue.

Apache Oozie

- Oozie now supports cron-style scheduling capability.
- Oozie now supports High Availability with security.

Apache Pig

- AvroStorage rewritten for better performance, and moved from piggybank to core Pig
- ASSERT, IN, and CASE operators added
- ParquetStorage added for integration with Parquet

Cloudera Search

- The Cloudera CDK has been renamed and updated to Kite version 0.11.0. For additional information on Kite, see:
 - [Kite repository](#)
 - [Kite Release Notes](#)
 - [Kite documentation](#)
 - [Kite examples](#)

Apache Spark (incubating)

Spark is a fast, general engine for large-scale data processing. For installation and configuration instructions, see [Spark Installation](#).

Apache Sqoop

Sqoop 2 has been upgraded from version 1.99.2 to 1.99.3.

What's New in Apache Impala (incubating)

This release of Impala contains the following changes and enhancements from previous releases.

New Features in Impala 2.8.x / CDH 5.10.x

- Performance improvements:
 - The COMPUTE STATS statement can take advantage of multithreading.
- Integration with Apache Kudu:
 - The experimental Impala support for the Kudu storage layer has been folded into the main Impala development branch. Impala can now directly access Kudu tables, opening up new capabilities such as enhanced DML operations and continuous ingestion.
 - The DELETE statement is a flexible way to remove data from a Kudu table. Previously, removing data from an Impala table involved removing or rewriting the underlying data files, dropping entire partitions, or rewriting the entire table. This Impala statement only works for Kudu tables.
 - The UPDATE statement is a flexible way to modify data within a Kudu table. Previously, updating data in an Impala table involved replacing the underlying data files, dropping entire partitions, or rewriting the entire table. This Impala statement only works for Kudu tables.
 - The UPSERT statement is a flexible way to ingest, modify, or both data within a Kudu table. Previously, ingesting data that might contain duplicates involved an inefficient multi-stage operation, and there was no built-in protection against duplicate data. The UPSERT statement, in combination with the primary key designation for Kudu tables, lets you add or replace rows in a single operation, and automatically avoids creating any duplicate data.
 - The CREATE TABLE statement gains some new clauses that are specific to Kudu tables. DISTRIBUTED BY PRIMARY KEY, NOT NULL, and STORED AS KUDU. These clauses replace the explicit TBLPROPERTIES settings that were required in the early experimental phases of integration between Impala and Kudu.
 - Not all Impala data types are supported in Kudu tables. In particular, currently the Impala TIMESTAMP type is not allowed in a Kudu table. Impala does not recognize the UNIXTIME_MICROS Kudu type when it is present in a Kudu table. (These two representations of date/time data use different units and are not directly compatible.) You cannot create columns of type TIMESTAMP, DECIMAL, VARCHAR, or CHAR within a Kudu table. Within a query, you can cast values in a result set to these types.

- Currently, Kudu tables are not interchangeable between Impala and Hive the way other kinds of Impala tables are. Although the metadata for Kudu tables is stored in the metastore database, currently Hive cannot access Kudu tables.
- The `INSERT` statement can deal efficiently with the `INSERT ... VALUES` syntax, or `INSERT ... SELECT` involving a small number of rows. Kudu tables are not subject to the kinds of performance and scalability issues that affect tables containing many small HDFS data files.
-
- **Security:**
 - Impala can take advantage of the S3 encrypted credential store, to avoid exposing the secret key when accessing data stored on S3.
- [IMPALA-1654] Several kinds of DDL operations can now work on a range of partitions. The partitions can be specified using operators such as `<`, `>=`, and `!=` rather than just an equality predicate applying to a single partition. This new feature extends the syntax of several clauses of the `ALTER TABLE` statement (`DROP PARTITION`, `SET [UN]CACHED`, `SET FILEFORMAT | SERDEPROPERTIES | TBLPROPERTIES`), the `SHOW FILES` statement, and the `COMPUTE INCREMENTAL STATS` statement. It does not apply to statements that are defined to only apply to a single partition, such as `LOAD DATA`, `ALTER TABLE ... ADD PARTITION`, `SET LOCATION`, and `INSERT` with a static partitioning clause.

New Features in Impala 2.7.x / CDH 5.9.x

- Performance improvements:
 - [IMPALA-3206] Speedup for queries against `DECIMAL` columns in Avro tables. The code that parses `DECIMAL` values from Avro now uses native code generation.
 - [IMPALA-3674] Improved efficiency in LLVM code generation can reduce codegen time, especially for short queries.
 - [IMPALA-2979] Improvements to scheduling on worker nodes, enabled by the `REPLICA_PREFERENCE` query option. See [REPLICA_PREFERENCE Query Option \(or higher only\)](#) for details.
- [IMPALA-1683] The `REFRESH` statement can be applied to a single partition, rather than the entire table. See [REFRESH Statement](#) and [Refreshing a Single Partition](#) for details.
- Improvements to the Impala web user interface:
 - [IMPALA-2767] You can now force a session to expire by clicking a link in the web UI, on the [/sessions](#) tab.
 - [IMPALA-3715] The [/memz](#) tab includes more information about Impala memory usage.
 - [IMPALA-3716] The [Details](#) page for a query now includes a **Memory** tab.
- [IMPALA-3499] Scalability improvements to the catalog server. Impala handles internal communication more efficiently for tables with large numbers of columns and partitions, where the size of the metadata exceeds 2 GiB.
- [IMPALA-3677] You can send a `SIGUSR1` signal to any Impala-related daemon to write a Breakpad minidump. For advanced troubleshooting, you can now produce a minidump without triggering a crash. See [Breakpad Minidumps for Impala \(or higher only\)](#) for details about the Breakpad minidump feature.
- [IMPALA-3687] The schema reconciliation rules for Avro tables have changed slightly for `CHAR` and `VARCHAR` columns. Now, if the definition of such a column is changed in the Avro schema file, the column retains its `CHAR` or `VARCHAR` type as specified in the SQL definition, but the column name and comment from the Avro schema file take precedence. See [Creating Avro Tables](#) for details about column definitions in Avro tables.

- [IMPALA-3575] Some network operations now have additional timeout and retry settings. The extra configuration helps avoid failed queries for transient network problems, to avoid hangs when a sender or receiver fails in the middle of a network transmission, and to make cancellation requests more reliable despite network issues.

New Features in Impala 2.6.x / CDH 5.8.x

- Improvements to Impala support for the Amazon S3 filesystem:
 - Impala can now write to S3 tables through the `INSERT` or `LOAD DATA` statements. See [Using Impala with the Amazon S3 Filesystem](#) for general information about using Impala with S3.
 - A new query option, `S3_SKIP_INSERT_STAGING`, lets you trade off between fast `INSERT` performance and slower `INSERTS` that are more consistent if a problem occurs during the statement. The new behavior is enabled by default. See [S3_SKIP_INSERT_STAGING Query Option \(or higher only\)](#) for details about this option.
- Performance improvements for the runtime filtering feature:
 - The default for the `RUNTIME_FILTER_MODE` query option is changed to `GLOBAL` (the highest setting). See [RUNTIME_FILTER_MODE Query Option \(or higher only\)](#) for details about this option.
 - The `RUNTIME_BLOOM_FILTER_SIZE` setting is now only used as a fallback if statistics are not available; otherwise, Impala uses the statistics to estimate the appropriate size to use for each filter. See [RUNTIME_BLOOM_FILTER_SIZE Query Option \(or higher only\)](#) for details about this option.
 - New query options `RUNTIME_FILTER_MIN_SIZE` and `RUNTIME_FILTER_MAX_SIZE` let you fine-tune the sizes of the Bloom filter structures used for runtime filtering. If the filter size derived from Impala internal estimates or from the `RUNTIME_FILTER_BLOOM_SIZE` falls outside the size range specified by these options, any too-small filter size is adjusted to the minimum, and any too-large filter size is adjusted to the maximum. See [RUNTIME_FILTER_MIN_SIZE Query Option \(or higher only\)](#) and [RUNTIME_FILTER_MAX_SIZE Query Option \(or higher only\)](#) for details about these options.
 - Runtime filter propagation now applies to all the operands of `UNION` and `UNION ALL` operators.
 - Runtime filters can now be produced during join queries even when the join processing activates the spill-to-disk mechanism.

See [Runtime Filtering for Impala Queries \(or higher only\)](#) for general information about the runtime filtering feature.

- Admission control and dynamic resource pools are enabled by default. See [Admission Control and Query Queuing](#) for details about admission control.
- Impala can now manually set column statistics, using the `ALTER TABLE` statement with a `SET COLUMN STATS` clause. See [Setting Column Stats Manually through ALTER TABLE](#) for details.
- Impala can now write lightweight “minidump” files, rather than large core files, to save diagnostic information when any of the Impala-related daemons crash. This feature uses the open source `breakpad` framework. See [Breakpad Minidumps for Impala \(or higher only\)](#) for details.
- New query options improve interoperability with Parquet files:
 - The `PARQUET_FALLBACK_SCHEMA_RESOLUTION` query option lets Impala locate columns within Parquet files based on column name rather than ordinal position. This enhancement improves interoperability with applications that write Parquet files with a different order or subset of columns than are used in the Impala table. See [PARQUET_FALLBACK_SCHEMA_RESOLUTION Query Option \(or higher only\)](#) for details.
 - The `PARQUET_ANNOTATE_STRINGS_UTF8` query option makes Impala include the `UTF-8` annotation metadata for `STRING`, `CHAR`, and `VARCHAR` columns in Parquet files created by `INSERT` or `CREATE TABLE AS SELECT` statements. See [PARQUET_ANNOTATE_STRINGS_UTF8 Query Option \(or higher only\)](#) for details.

See [Using the Parquet File Format with Impala Tables](#) for general information about working with Parquet files.

- Improvements to security and reduction in overhead for secure clusters:
 - Overall performance improvements for secure clusters. (TPC-H queries on a secure cluster were benchmarked at roughly 3x as fast as the previous release.)
 - Impala now recognizes the `auth_to_local` setting, specified through the HDFS configuration setting `hadoop.security.auth_to_local`. This feature is disabled by default; to enable it, specify `--load_auth_to_local_rules=true` in the `impalad` configuration settings. See [Mapping Kerberos Principals to Short Names for Impala](#) for details.
 - Timing improvements in the mechanism for the `impalad` daemon to acquire Kerberos tickets. This feature spreads out the overhead on the KDC during Impala startup, especially for large clusters.
 - For Kerberized clusters, the Catalog service now uses the Kerberos principal instead of the operating system user that runs the `catalogd` daemon. This eliminates the requirement to configure a `hadoop.user.group.static.mapping.overrides` setting to put the OS user into the Sentry administrative group, on clusters where the principal and the OS user name for this user are different.
- Overall performance improvements for join queries, by using a prefetching mechanism while building the in-memory hash table to evaluate join predicates. See [PREFETCH_MODE Query Option \(or higher only\)](#) for the query option to control this optimization.
- The `impala-shell` interpreter has a new command, `SOURCE`, that lets you run a set of SQL statements or other `impala-shell` commands stored in a file. You can run additional `SOURCE` commands from inside a file, to set up flexible sequences of statements for use cases such as schema setup, ETL, or reporting. See [impala-shell Command Reference](#) for details and [Running Commands and SQL Statements in impala-shell](#) for examples.
- The `millisecond()` built-in function lets you extract the fractional seconds part of a `TIMESTAMP` value. See [Impala Date and Time Functions](#) for details.
- If an Avro table is created without column definitions in the `CREATE TABLE` statement, and columns are later added through `ALTER TABLE`, the resulting table is now queryable. Missing values from the newly added columns now default to `NULL`. See [Using the Avro File Format with Impala Tables](#) for general details about working with Avro files.
- The mechanism for interpreting `DECIMAL` literals is improved, no longer going through an intermediate conversion step to `DOUBLE`:
 - Casting a `DECIMAL` value to `TIMESTAMP DOUBLE` produces a more precise value for the `TIMESTAMP` than formerly.
 - Certain function calls involving `DECIMAL` literals now succeed, when formerly they failed due to lack of a function signature with a `DOUBLE` argument.
 - Faster runtime performance for `DECIMAL` constant values, through improved native code generation for all combinations of precision and scale.

See [DECIMAL Data Type \(or higher only\)](#) for details about the `DECIMAL` type.

- Improved type accuracy for `CASE` return values. If all `WHEN` clauses of the `CASE` expression are of `CHAR` type, the final result is also `CHAR` instead of being converted to `STRING`. See [Impala Conditional Functions](#) for details about the `CASE` function.
- Uncorrelated queries using the `NOT EXISTS` operator are now supported. Formerly, the `NOT EXISTS` operator was only available for correlated subqueries.
- Improved performance for reading Parquet files.
- Improved performance for `top-N` queries, that is, those including both `ORDER BY` and `LIMIT` clauses.

- Impala optionally skips an arbitrary number of header lines from text input files on HDFS based on the `skip.header.line.count` value in the `TBLPROPERTIES` field of the table metadata. See [Data Files for Text Tables](#) for details.
- Trailing comments are now allowed in queries processed by the `impala-shell` options `-q` and `-f`.
- Impala can run `COUNT` queries for RCFFile tables that include complex type columns. See [Complex Types \(or higher only\)](#) for general information about working with complex types, and [ARRAY Complex Type \(or higher only\)](#), [MAP Complex Type \(or higher only\)](#), and [STRUCT Complex Type \(or higher only\)](#) for syntax details of each type.

New Features in Impala 2.5.x / CDH 5.7.x



Note: Impala 2.5.x is available as part of CDH 5.7.x and is not available for CDH 4. Cloudera does not intend to release future versions of Impala for CDH 4 outside patch and maintenance releases if required. Given the end-of-maintenance status for CDH 4, Cloudera recommends all customers to migrate to a recent CDH 5 release.

- Dynamic partition pruning. When a query refers to a partition key column in a `WHERE` clause, and the exact set of column values are not known until the query is executed, Impala evaluates the predicate and skips the I/O for entire partitions that are not needed. For example, if a table was partitioned by year, Impala would apply this technique to a query such as `SELECT c1 FROM partitioned_table WHERE year = (SELECT MAX(year) FROM other_table)`.

The dynamic partition pruning optimization technique lets Impala avoid reading data files from partitions that are not part of the result set, even when that determination cannot be made in advance. This technique is especially valuable when performing join queries involving partitioned tables. For example, if a join query includes an `ON` clause and a `WHERE` clause that refer to the same columns, the query can find the set of column values that match the `WHERE` clause, and only scan the associated partitions when evaluating the `ON` clause.

Dynamic partition pruning is controlled by the same settings as the runtime filtering feature. By default, this feature is enabled at a medium level, because the maximum setting can use slightly more memory for queries than in previous releases. To fully enable this feature, set the query option `RUNTIME_FILTER_MODE=GLOBAL`.

- Runtime filtering. This is a wide-ranging set of optimizations that are especially valuable for join queries. Using the same technique as with dynamic partition pruning, Impala uses the predicates from `WHERE` and `ON` clauses to determine the subset of column values from one of the joined tables could possibly be part of the result set. Impala sends a compact representation of the filter condition to the hosts in the cluster, instead of the full set of values or the entire table.

By default, this feature is enabled at a medium level, because the maximum setting can use slightly more memory for queries than in previous releases. To fully enable this feature, set the query option `RUNTIME_FILTER_MODE=GLOBAL`.

This feature involves some new query options: `RUNTIME_FILTER_MODE`, `MAX_NUM_RUNTIME_FILTERS`, `RUNTIME_BLOOM_FILTER_SIZE`, `RUNTIME_FILTER_WAIT_TIME_MS`, and `DISABLE_ROW_RUNTIME_FILTERING`.

- More efficient use of the HDFS caching feature, to avoid hotspots and bottlenecks that could occur if heavily used cached data blocks were always processed by the same host. By default, Impala now randomizes which host processes each cached HDFS data block, when cached replicas are available on multiple hosts. (Remember to use the `WITH REPLICATION` clause with the `CREATE TABLE` or `ALTER TABLE` statement when enabling HDFS caching for a table or partition, to cache the same data blocks across multiple hosts.) The new query option `SCHEDULE_RANDOM_REPLICA` lets you fine-tune the interaction with HDFS caching even more.
- The `TRUNCATE TABLE` statement now accepts an `IF EXISTS` clause, making `TRUNCATE TABLE` easier to use in setup or ETL scripts where the table might or might not exist.
- Improved performance and reliability for the `DECIMAL` data type:
 - Using `DECIMAL` values in a `GROUP BY` clause now triggers the native code generation optimization, speeding up queries that group by values such as prices.

- Checking for overflow in `DECIMAL` multiplication is now substantially faster, making `DECIMAL` a more practical data type in some use cases where formerly `DECIMAL` was much slower than `FLOAT` or `DOUBLE`.
- Multiplying a mixture of `DECIMAL` and `FLOAT` or `DOUBLE` values now returns the `DOUBLE` rather than `DECIMAL`. This change avoids some cases where an intermediate value would underflow or overflow and become `NULL` unexpectedly.
- For UDFs written in Java, or Hive UDFs reused for Impala, Impala now allows parameters and return values to be primitive types. Formerly, these things were required to be one of the “Writable” object types.
- Performance improvements for HDFS I/O. Impala now caches HDFS file handles to avoid the overhead of repeatedly opening the same file.
- Performance improvements for queries involving nested complex types. Certain basic query types, such as counting the elements of a complex column, now use an optimized code path.
- Improvements to the memory reservation mechanism for the Impala admission control feature. You can specify more settings, such as the timeout period and maximum aggregate memory used, for each resource pool instead of globally for the Impala instance. The default limit for concurrent queries (the `max requests` setting) is now unlimited instead of 200. The Cloudera Manager user interface for admission control has been reworked, with the settings available under the **Dynamic Resource Pools** window.
- Performance improvements related to code generation. Even in queries where code generation is not performed for some phases of execution (such as reading data from Parquet tables), Impala can still use code generation in other parts of the query, such as evaluating functions in the `WHERE` clause.
- Performance improvements for queries using aggregation functions on high-cardinality columns. Formerly, Impala could do unnecessary extra work to produce intermediate results for operations such as `DISTINCT` or `GROUP BY` on columns that were unique or had few duplicate values. Now, Impala decides at run time whether it is more efficient to do an initial aggregation phase and pass along a smaller set of intermediate data, or to pass raw intermediate data back to next phase of query processing to be aggregated there. This feature is known as **streaming pre-aggregation**. In case of performance regression, this feature can be turned off using the `DISABLE_STREAMING_PREAGGREGATIONS` query option.
- Spill-to-disk feature now always recommended. In earlier releases, the spill-to-disk feature could be turned off using a pair of configuration settings, `enable_partitioned_aggregation=false` and `enable_partitioned_hash_join=false`. The latest improvements in the spill-to-disk mechanism, and related features that interact with it, make this feature robust enough that disabling it is now no longer needed or supported. In particular, some new features in and higher do not work when the spill-to-disk feature is disabled.
- Improvements to scripting capability for the `impala-shell` command, through user-specified substitution variables that can appear in statements processed by `impala-shell`:
 - The `--var` command-line option lets you pass key-value pairs to `impala-shell`. The shell can substitute the values into queries before executing them, where the query text contains the notation `${var:varname}`. For example, you might prepare a SQL file containing a set of DDL statements and queries containing variables for database and table names, and then pass the applicable names as part of the `impala-shell -f filename` command.
 - The `SET` and `UNSET` commands within the `impala-shell` interpreter now work with user-specified substitution variables, as well as the built-in query options. The two kinds of variables are divided in the `SET` output. As with variables defined by the `--var` command-line option, you refer to the user-specified substitution variables in queries by using the notation `${var:varname}` in the query text. Because the substitution variables are processed by `impala-shell` instead of the `impalad` backend, you cannot define your own substitution variables through the `SET` statement in a JDBC or ODBC application.
- Performance improvements for query startup. Impala better parallelizes certain work when coordinating plan distribution between `impalad` instances, which improves startup time for queries involving tables with many partitions on large clusters, or complicated queries with many plan fragments.

- Performance and scalability improvements for tables with many partitions. The memory requirements on the coordinator node are reduced, making it substantially faster and less resource-intensive to do joins involving several tables with thousands of partitions each.
- Whitelisting for access to internal APIs. For applications that need direct access to Impala APIs, without going through the HiveServer2 or Beeswax interfaces, you can specify a list of Kerberos users who are allowed to call those APIs. By default, the `impala` and `hdfs` users are the only ones authorized for this kind of access. Any users not explicitly authorized through the `internal_principals_whitelist` configuration setting are blocked from accessing the APIs. This setting applies to all the Impala-related daemons, although currently it is primarily used for HDFS to control the behavior of the catalog server.
- Improvements to Impala integration and usability for Hue. (The code changes are actually on the Hue side.)
 - The list of tables now refreshes dynamically.
- Usability improvements for case-insensitive queries. You can now use the operators `ILIKE` and `IREGEXP` to perform case-insensitive wildcard matches or regular expression matches, rather than explicitly converting column values with `UPPER` or `LOWER`.
- Performance and reliability improvements for DDL and insert operations on partitioned tables with a large number of partitions. Impala only re-evaluates metadata for partitions that are affected by a DDL operation, not all partitions in the table. While a DDL or insert statement is in progress, other Impala statements that attempt to modify metadata for the same table wait until the first one finishes.
- Reliability improvements for the `LOAD DATA` statement. Previously, this statement would fail if the source HDFS directory contained any subdirectories at all. Now, the statement ignores any hidden subdirectories, for example `_impala_insert_staging`.
- A new operator, `IS [NOT] DISTINCT FROM`, lets you compare values and always get a `true` or `false` result, even if one or both of the values are `NULL`. The `IS NOT DISTINCT FROM` operator, or its equivalent `<=>` notation, improves the efficiency of join queries that treat key values that are `NULL` in both tables as equal.
- Security enhancements for the `impala-shell` command. A new option, `--ldap_password_cmd`, lets you specify a command to retrieve the LDAP password. The resulting password is then used to authenticate the `impala-shell` command with the LDAP server.
- The `CREATE TABLE AS SELECT` statement now accepts a `PARTITIONED BY` clause, which lets you create a partitioned table and insert data into it with a single statement.
- User-defined functions (UDFs and UDAFs) written in C++ now persist automatically when the `catalogd` daemon is restarted. You no longer have to run the `CREATE FUNCTION` statements again after a restart.
- User-defined functions (UDFs) written in Java can now persist when the `catalogd` daemon is restarted, and can be shared transparently between Impala and Hive. You must do a one-time operation to recreate these UDFs using new `CREATE FUNCTION` syntax, without a signature for arguments or the return value. Afterwards, you no longer have to run the `CREATE FUNCTION` statements again after a restart. Although Impala does not have visibility into the UDFs that implement the Hive built-in functions, user-created Hive UDFs are now automatically available for calling through Impala.
- Reliability enhancements for memory management. Some aggregation and join queries that formerly might have failed with an out-of-memory error due to memory contention, now can succeed using the spill-to-disk mechanism.
- The `SHOW DATABASES` statement now returns two columns rather than one. The second column includes the associated comment string, if any, for each database. Adjust any application code that examines the list of databases and assumes the result set contains only a single column.
- A new optimization speeds up aggregation operations that involve only the partition key columns of partitioned tables. For example, a query such as `SELECT COUNT(DISTINCT k), MIN(k), MAX(k) FROM t1` can avoid reading any data files if `t1` is a partitioned table and `K` is one of the partition key columns. Because this technique can produce different results in cases where HDFS files in a partition are manually deleted or are empty, you must enable the optimization by setting the query option `OPTIMIZE_PARTITION_KEY_SCANS`.

- The `DESCRIBE` statement can now display metadata about a database, using the syntax `DESCRIBE DATABASE db_name`.
- The `uuid()` built-in function generates an alphanumeric value that you can use as a guaranteed unique identifier. The uniqueness applies even across tables, for cases where an ascending numeric sequence is not suitable.

New Features in Impala 2.4.x / CDH 5.6.x



Note: Impala 2.4.0 is available as part of CDH 5.6.0 and is not available for CDH 4. Cloudera does not intend to release future versions of Impala for CDH 4 outside patch and maintenance releases if required. Given the end-of-maintenance status for CDH 4, Cloudera recommends all customers to migrate to a recent CDH 5 release.

- Impala can be used on the DSSD D5 Storage Appliance. From a user perspective, the Impala features are the same as in .

New Features in Impala 2.3.x / CDH 5.5.x



Note: Impala 2.3.x is available as part of CDH 5.5.x and is not available for CDH 4. Cloudera does not intend to release future versions of Impala for CDH 4 outside patch and maintenance releases if required. Given the end-of-maintenance status for CDH 4, Cloudera recommends all customers to migrate to a recent CDH 5 release.

The following are the major new features in Impala 2.3.x. This major release, available as part of CDH 5.5.x, contains improvements to SQL syntax (particularly new support for complex types), performance, manageability, security.

- Complex data types: `STRUCT`, `ARRAY`, and `MAP`. These types can encode multiple named fields, positional items, or key-value pairs within a single column. You can combine these types to produce nested types with arbitrarily deep nesting, such as an `ARRAY` of `STRUCT` values, a `MAP` where each key-value pair is an `ARRAY` of other `MAP` values, and so on. Currently, complex data types are only supported for the Parquet file format.
- Column-level authorization lets you define access to particular columns within a table, rather than the entire table. This feature lets you reduce the reliance on creating views to set up authorization schemes for subsets of information. See [Column-level Authorization](#) for background details, and [GRANT Statement \(or higher only\)](#) and [REVOKE Statement \(or higher only\)](#) for Impala-specific syntax.
- The `TRUNCATE TABLE` statement removes all the data from a table without removing the table itself.
- Nested loop join queries. Some join queries that formerly required equality comparisons can now use operators such as `<` or `>=`. This same join mechanism is used internally to optimize queries that retrieve values from complex type columns.
- Reduced memory usage and improved performance and robustness for spill-to-disk feature.
- Performance improvements for querying Parquet data files containing multiple row groups and multiple data blocks:
 - For files written by Hive, SparkSQL, and other Parquet MR writers and spanning multiple HDFS blocks, Impala now scans the extra data blocks locally when possible, rather than using remote reads.
 - Impala queries benefit from the improved alignment of row groups with HDFS blocks for Parquet files written by Hive, MapReduce, and other components in CDH 5.5 and higher. (Impala itself never writes multiblock Parquet files, so the alignment change does not apply to Parquet files produced by Impala.) These Parquet writers now add padding to Parquet files that they write to align row groups with HDFS blocks. The `parquet.writer.max-padding` setting specifies the maximum number of bytes, by default 8 megabytes, that can be added to the file between row groups to fill the gap at the end of one block so that the next row group starts at the beginning of the next block. If the gap is larger than this size, the writer attempts to fit another entire row group in the remaining space. Include this setting in the `hive-site` configuration file to

CDH 5 Release Notes

influence Parquet files written by Hive, or the `hdfs-site` configuration file to influence Parquet files written by all non-Impala components.

- Many new built-in scalar functions, for convenience and enhanced portability of SQL that uses common industry extensions.

Math functions:

- ATAN2
- COSH
- COT
- DCEIL
- DEXP
- DFLOOR
- DLOG10
- DPOW
- DROUND
- DSQRT
- DTRUNC
- FACTORIAL, and corresponding ! operator
- FPOW
- RADIANS
- RANDOM
- SINH
- TANH

String functions:

- BTRIM
- CHR
- REGEXP_LIKE
- SPLIT_PART

Date and time functions:

- INT_MONTHS_BETWEEN
- MONTHS_BETWEEN
- TIMEOFDAY
- TIMESTAMP_CMP

Bit manipulation functions:

- BITAND
- BITNOT
- BITOR
- BITXOR
- COUNTSET
- GETBIT
- ROTATELEFT
- ROTATERIGHT
- SETBIT
- SHIFTLEFT
- SHIFTRIGHT

Type conversion functions:

- TYPEOF

The `effective_user()` function.

- New built-in analytic functions: `PERCENT_RANK`, `NTILE`, `CUME_DIST`.
 - The `DROP DATABASE` statement now works for a non-empty database. When you specify the optional `CASCADE` clause, any tables in the database are dropped before the database itself is removed.
 - The `DROP TABLE` and `ALTER TABLE DROP PARTITION` statements have a new optional keyword, `PURGE`. This keyword causes Impala to immediately remove the relevant HDFS data files rather than sending them to the HDFS trashcan. This feature can help to avoid out-of-space errors on storage devices, and to avoid files being left behind in case of a problem with the HDFS trashcan, such as the trashcan not being configured or being in a different HDFS encryption zone than the data files.
 - The `impala-shell` command has a new feature for live progress reporting. This feature is enabled through the `--live_progress` and `--live_summary` command-line options, or during a session through the `LIVE_SUMMARY` and `LIVE_PROGRESS` query options.
 - The `impala-shell` command also now displays a random “tip of the day” when it starts.
 - The `impala-shell` option `-f` now recognizes a special filename – to accept input from stdin.
 - Format strings for the `unix_timestamp()` function can now include numeric timezone offsets.
 - Impala can now run a specified command to obtain the password to decrypt a private-key PEM file, rather than having the private-key file be unencrypted on disk.
 - Impala components now can use SSL for more of their internal communication. SSL is used for communication between all three Impala-related daemons when the configuration option `ssl_server_certificate` is enabled. SSL is used for communication with client applications when the configuration option `ssl_client_ca_certificate` is enabled.
- Currently, you can only use one of server-to-server TLS/SSL encryption or Kerberos authentication. This limitation is tracked by the issue [IMPALA-2598](#).
- Improved flexibility for intermediate data types in user-defined aggregate functions (UDAFs).

In CDH 5.5.2 / Impala 2.3.2, the bug fix for [IMPALA-2598](#) removes the restriction on using both Kerberos and SSL for internal communication between Impala components.

New Features in Impala 2.2.x for CDH 5.4.3 and 5.4.4

No new features added to the Impala code. The certification of Impala with EMC Isilon under CDH 5.4.4 means that now you can query data stored on Isilon storage devices through Impala. See [Using CDH with Isilon Storage](#) for details. The same level of Impala is included with both CDH 5.4.3 and 5.4.4.



Note: The Impala 2.2.x maintenance releases now use the CDH 5.4.x numbering system rather than increasing the Impala version numbers. Impala 2.2 and higher are not available under CDH 4.

New Features in Impala 2.2.x / CDH 5.4.x



Note: Impala 2.2.0 is available as part of CDH 5.4.0 and is not available for CDH 4. Cloudera does not intend to release future versions of Impala for CDH 4 outside patch and maintenance releases if required. Given the end-of-maintenance status for CDH 4, Cloudera recommends all customers to migrate to a recent CDH 5 release.

The following are the major new features in Impala 2.2.x. This release, available as part of CDH 5.4.x, contains improvements to performance, manageability, security, and SQL syntax.

- Several improvements to date and time features enable higher interoperability with Hive and other database systems, provide more flexibility for handling time zones, and future-proof the handling of `TIMESTAMP` values:

- The `WITH REPLICATION` clause for the `CREATE TABLE` and `ALTER TABLE` statements lets you control the replication factor for HDFS caching for a specific table or partition. By default, each cached block is only present on a single host, which can lead to CPU contention if the same host processes each cached block. Increasing the replication factor lets Impala choose different hosts to process different cached blocks, to better distribute the CPU load.
- Startup flags for the `impalad` daemon enable a higher level of compatibility with `TIMESTAMP` values written by Hive, and more flexibility for working with date and time data using the local time zone instead of UTC. To enable these features, set the `impalad` startup flags
`-use_local_tz_for_unix_timestamp_conversions=true` and
`-convert_legacy_hive_parquet_utc_timestamps=true`.

The `-use_local_tz_for_unix_timestamp_conversions` setting controls how the `unix_timestamp()`, `from_unixtime()`, and `now()` functions handle time zones. By default (when this setting is turned off), Impala considers all `TIMESTAMP` values to be in the UTC time zone when converting to or from Unix time values. When this setting is enabled, Impala treats `TIMESTAMP` values passed to or returned from these functions to be in the local time zone. When this setting is enabled, take particular care that all hosts in the cluster have the same timezone settings, to avoid inconsistent results depending on which host reads or writes `TIMESTAMP` data.

The `-convert_legacy_hive_parquet_utc_timestamps` setting causes Impala to convert `TIMESTAMP` values to the local time zone when it reads them from Parquet files written by Hive. This setting only applies to data using the Parquet file format, where Impala can use metadata in the files to reliably determine that the files were written by Hive. If in the future Hive changes the way it writes `TIMESTAMP` data in Parquet, Impala will automatically handle that new `TIMESTAMP` encoding.

See [TIMESTAMP Data Type](#) for details about time zone handling and the configuration options for Impala / Hive compatibility with Parquet format.

- In Impala 2.2.0 and higher, built-in functions that accept or return integers representing `TIMESTAMP` values use the `BIGINT` type for parameters and return values, rather than `INT`. This change lets the date and time functions avoid an overflow error that would otherwise occur on January 19th, 2038 (known as the [“Year 2038 problem” or “Y2K38 problem”](#)). This change affects the `from_unixtime()` and `unix_timestamp()` functions. You might need to change application code that interacts with these functions, change the types of columns that store the return values, or add `CAST()` calls to SQL statements that call these functions.

See [Impala Date and Time Functions](#) for the current function signatures.

- The `SHOW FILES` statement lets you view the names and sizes of the files that make up an entire table or a specific partition. See [SHOW FILES Statement](#) for details.
- Impala can now run queries against Parquet data containing columns with complex or nested types, as long as the query only refers to columns with scalar types.
- Performance improvements for queries that include `IN()` operators and involve partitioned tables.
- The new `-max_log_files` configuration option specifies how many log files to keep at each severity level. The default value is 10, meaning that Impala preserves the latest 10 log files for each severity level (`INFO`, `WARNING`, and `ERROR`) for each Impala-related daemon (`impalad`, `statestored`, and `catalogd`). Impala checks to see if any old logs need to be removed based on the interval specified in the `logbufsecs` setting, every 5 seconds by default. See [Rotating Impala Logs](#) for details.
- Redaction of sensitive data from Impala log files. This feature protects details such as credit card numbers or tax IDs from administrators who see the text of SQL statements in the course of monitoring and troubleshooting a Hadoop cluster. See [Redacting Sensitive Information from Impala Log Files](#) for background information for Impala users, and [Sensitive Data Redaction](#) for usage details.
- Lineage information is available for data created or queried by Impala. This feature lets you track who has accessed data through Impala SQL statements, down to the level of specific columns, and how data has been propagated between tables. See [Viewing Lineage Information for Impala Data](#) for background information for Impala users,

[Managing Hive and Impala Lineage Properties](#) for usage details, and [Cloudera Navigator Lineage Diagrams](#) for how to interpret the lineage information.

- Impala tables and partitions can now be located on the Amazon Simple Storage Service (S3) filesystem, for convenience in cases where data is already located in S3 and you prefer to query it in-place. Queries might have lower performance than when the data files reside on HDFS, because Impala uses some HDFS-specific optimizations. Impala can query data in S3, but cannot write to S3. Therefore, statements such as `INSERT` and `LOAD DATA` are not available when the destination table or partition is in S3. See [Using Impala with the Amazon S3 Filesystem](#) for details.



Important:

Impala query support for Amazon S3 is included in CDH 5.4.0, but is not currently supported or recommended for production use. To try this feature, use it in a test environment until Cloudera resolves currently existing issues and limitations to make it ready for production use.

- Improved support for HDFS encryption. The `LOAD DATA` statement now works when the source directory and destination table are in different encryption zones.
- Additional arithmetic function `mod()`. See [Impala Mathematical Functions](#) for details.
- Flexibility to interpret `TIMESTAMP` values using the UTC time zone (the traditional Impala behavior) or using the local time zone (for compatibility with `TIMESTAMP` values produced by Hive).
- Enhanced support for ETL using tools such as Flume. Impala ignores temporary files typically produced by these tools (filenames with suffixes `.copying` and `.tmp`).
- The CPU requirement for Impala, which had become more restrictive in Impala 2.0.x and 2.1.x, has now been relaxed.

The prerequisite for CPU architecture has been relaxed in Impala 2.2.0 and higher. From this release onward, Impala works on CPUs that have the SSSE3 instruction set. The SSE4 instruction set is no longer required. This relaxed requirement simplifies the upgrade planning from Impala 1.x releases, which also worked on SSSE3-enabled processors.

- Enhanced support for `CHAR` and `VARCHAR` types in the `COMPUTE STATS` statement.
- The amount of memory required during setup for “spill to disk” operations is greatly reduced. This enhancement reduces the chance of a memory-intensive join or aggregation query failing with an out-of-memory error.
- Several new conditional functions provide enhanced compatibility when porting code that uses industry extensions. The new functions are: `isfalse()`, `isnotfalse()`, `isnottrue()`, `istrue()`, `nonnullvalue()`, and `nullvalue()`. See [Impala Conditional Functions](#) for details.
- The Impala debug web UI now can display a visual representation of the query plan. On the `/queries` tab, select **Details** for a particular query. The **Details** page includes a **Plan** tab with a plan diagram that you can zoom in or out (using scroll gestures through mouse wheel or trackpad).

New Features in Impala 2.1.8 / CDH 5.3.10

This point release is exclusively a bug fix release.



Note: This Impala maintenance release is only available as part of CDH 5, not under CDH 4.

New Features in Impala 2.1.7 / CDH 5.3.9

This point release is exclusively a bug fix release.



Note: This Impala maintenance release is only available as part of CDH 5, not under CDH 4.

New Features in Impala 2.1.6 / CDH 5.3.8

This point release is exclusively a bug fix release.



Note: This Impala maintenance release is only available as part of CDH 5, not under CDH 4.

New Features in Impala 2.1.5 / CDH 5.3.6

This point release is exclusively a bug fix release.



Note: This Impala maintenance release is only available as part of CDH 5, not under CDH 4.

New Features in Impala 2.1.4 / CDH 5.3.4

No new features. This point release is exclusively a bug fix release. Because CDH 5.3.5 does not include any code changes for Impala, Impala 2.1.4 is included with both CDH 5.3.4 and 5.3.5.



Note: This Impala maintenance release is only available as part of CDH 5, not under CDH 4.

New Features in Impala 2.1.3 / CDH 5.3.3

No new features. This point release is exclusively a bug fix release.



Note: Impala 2.1.3 is available as part of CDH 5.3.3, not under CDH 4.

New Features in Impala 2.1.2 / CDH 5.3.2

No new features. This point release is exclusively a bug fix release.



Note: Impala 2.1.2 is available as part of CDH 5.3.2, not under CDH 4.

New Features in Impala 2.1.1 / CDH 5.3.1

No new features. This point release is exclusively a bug fix release.

New Features in Impala 2.1.0 / CDH 5.3.0

This release contains the following enhancements to query performance and system scalability:

- Impala can now collect statistics for individual partitions in a partitioned table, rather than processing the entire table for each `COMPUTE STATS` statement. This feature is known as incremental statistics, and is controlled by the `COMPUTE INCREMENTAL STATS` syntax. (You can still use the original `COMPUTE STATS` statement for nonpartitioned tables or partitioned tables that are unchanging or whose contents are entirely replaced all at once.) See [COMPUTE STATS Statement](#) and [Table and Column Statistics](#) for details.

- Optimization for small queries lets Impala process queries that process very few rows without the unnecessary overhead of parallelizing and generating native code. Reducing this overhead lets Impala clear small queries quickly, keeping YARN resources and admission control slots available for data-intensive queries. The number of rows considered to be a “small” query is controlled by the `EXEC_SINGLE_NODE_ROWS_THRESHOLD` query option. See [EXEC_SINGLE_NODE_ROWS_THRESHOLD Query Option \(or higher only\)](#) for details.
- An enhancement to the statestore component lets it transmit heartbeat information independently of broadcasting metadata updates. This optimization improves reliability of health checking on large clusters with many tables and partitions.
- The memory requirement for querying gzip-compressed text is reduced. Now Impala decompresses the data as it is read, rather than reading the entire gzipped file and decompressing it in memory.

New Features in Impala 2.0.5 / CDH 5.2.6

No new features. This point release is exclusively a bug fix release.



Note: Impala 2.0.5 is available as part of CDH 5.2.6, not under CDH 4.

New Features in Impala 2.0.4 / CDH 5.2.5

No new features. This point release is exclusively a bug fix release.



Note: Impala 2.0.4 is available as part of CDH 5.2.5, not under CDH 4.

New Features in Impala 2.0.3 / CDH 5.2.4

No new features. This point release is exclusively a bug fix release.



Note: Impala 2.0.3 is available as part of CDH 5.2.4, not under CDH 4.

New Features in Impala 2.0.2 / CDH 5.2.3

No new features. This point release is exclusively a bug fix release.



Note: Impala 2.0.2 is available as part of CDH 5.2.3, not under CDH 4.

New Features in Impala 2.0.1 / CDH 5.2.1

No new features. This point release is exclusively a bug fix release.

New Features in Impala 2.0.0 / CDH 5.2.0

The following are the major new features in Impala 2.0. This major release, available both with CDH 5.2 and for CDH 4, contains improvements to performance, scalability, security, and SQL syntax.

- Queries with joins or aggregation functions involving high volumes of data can now use temporary work areas on disk, reducing the chance of failure due to out-of-memory errors. When the required memory for the intermediate result set exceeds the amount available on a particular node, the query automatically uses a temporary work area on disk. This “spill to disk” mechanism is similar to the `ORDER BY` improvement from Impala 1.4. For details, see [SQL Operations that Spill to Disk](#).
- Subquery enhancements:

- Subqueries are now allowed in the WHERE clause, for example with the IN operator.
- The EXISTS and NOT EXISTS operators are available. They are always used in conjunction with subqueries.
- The IN and NOT IN queries can now operate on the result set from a subquery, not just a hardcoded list of values.
- Uncorrelated subqueries let you compare against one or more values for equality, IN, and EXISTS comparisons. For example, you might use WHERE clauses such as WHERE column = (SELECT MAX(some_other_column FROM table) or WHERE column IN (SELECT some_other_column FROM table WHERE conditions).
- Correlated subqueries let you cross-reference values from the outer query block and the subquery.
- Scalar subqueries let you substitute the result of single-value aggregate functions such as MAX(), MIN(), COUNT(), or AVG(), where you would normally use a numeric value in a WHERE clause.

For details about subqueries, see [Subqueries in Impala SELECT Statements](#). For information about new and improved operators, see [EXISTS Operator](#) and [IN Operator](#).

- Analytic functions such as RANK(), LAG(), LEAD(), and FIRST_VALUE() let you analyze sequences of rows with flexible ordering and grouping. Existing aggregate functions such as MAX(), SUM(), and COUNT() can also be used in an analytic context. See [Impala Analytic Functions](#) for details. See [Impala Aggregate Functions](#) for enhancements to existing aggregate functions.
- New data types provide greater compatibility with source code from traditional database systems:
 - VARCHAR is like the STRING data type, but with a maximum length. See [VARCHAR Data Type \(or higher only\)](#) for details.
 - CHAR is like the STRING data type, but with a precise length. Short values are padded with spaces on the right. See [CHAR Data Type \(or higher only\)](#) for details.
- Security enhancements:
 - Formerly, Impala was restricted to using either Kerberos or LDAP / Active Directory authentication within a cluster. Now, Impala can freely accept either kind of authentication request, allowing you to set up some hosts with Kerberos authentication and others with LDAP or Active Directory. See [Using Multiple Authentication Methods with Impala](#) for details.
 - GRANT statement. See [GRANT Statement \(or higher only\)](#) for details.
 - REVOKE statement. See [REVOKE Statement \(or higher only\)](#) for details.
 - CREATE ROLE statement. See [CREATE ROLE Statement \(or higher only\)](#) for details.
 - DROP ROLE statement. See [DROP ROLE Statement \(or higher only\)](#) for details.
 - SHOW ROLES and SHOW ROLE GRANT statements. See [SHOW Statement](#) for details.
 - To complement the HDFS encryption feature, a new Impala configuration option, --disk_spill_encryption secures sensitive data from being observed or tampered with when temporarily stored on disk.

The new security-related SQL statements work along with the Sentry authorization framework. See [Enabling Sentry Authorization for Impala](#) for details.

- Impala can now read compressed text files compressed by gzip, bzip, or Snappy. These files do not require any special table settings to work in an Impala text table. Impala recognizes the compression type automatically based on file extensions of .gz, .bz2, and .snappy respectively. These types of compressed text files are intended for convenience with existing ETL pipelines. Their non-splittable nature means they are not optimal for high-performance parallel queries. See [Using gzip, bzip2, or Snappy-Compressed Text Files](#) for details.
- Query hints can now use comment notation, /* +hint_name */ or -- +hint_name, at the same places in the query where the hints enclosed by [] are recognized. This enhancement makes it easier to reuse Impala queries on other database systems. See [Query Hints in Impala SELECT Statements](#) for details.
- A new query option, QUERY_TIMEOUT_S, lets you specify a timeout period in seconds for individual queries.

The working of the --idle_query_timeout configuration option is extended. If no QUERY_OPTION_S query option is in effect, --idle_query_timeout works the same as before, setting the timeout interval. When the QUERY_OPTION_S query option is specified, its maximum value is capped by the value of the --idle_query_timeout option.

That is, the system administrator sets the default and maximum timeout through the `--idle_query_timeout` startup option, and then individual users or applications can set a lower timeout value if desired through the `QUERY_TIMEOUT_S` query option. See [Setting Timeout Periods for Daemons, Queries, and Sessions](#) and [QUERY_TIMEOUT_S Query Option \(or higher only\)](#) for details.

- New functions `VAR_SAMP()` and `VAR_POP()` are aliases for the existing `VARIANCE_SAMP()` and `VARIANCE_POP()` functions.
- A new date and time function, `DATE_PART()`, provides similar functionality to `EXTRACT()`. You can also call the `EXTRACT()` function using the SQL-99 syntax, `EXTRACT(unit FROM timestamp)`. These enhancements simplify the porting process for date-related code from other systems. See [Impala Date and Time Functions](#) for details.
- New approximation features provide a fast way to get results when absolute precision is not required:
 - The `APPX_COUNT_DISTINCT` query option lets Impala rewrite `COUNT(DISTINCT)` calls to use `NDV()` instead, which speeds up the operation and allows multiple `COUNT(DISTINCT)` operations in a single query. See [APPX_COUNT_DISTINCT Query Option \(or higher only\)](#) for details.

The `APPX_MEDIAN()` aggregate function produces an estimate for the median value of a column by using sampling. See [APPX_MEDIAN Function](#) for details.

- Impala now supports a `DECODE()` function. This function works as a shorthand for a `CASE()` expression, and improves compatibility with SQL code containing vendor extensions. See [Impala Conditional Functions](#) for details.
- The `STDDEV()`, `STDDEV_POP()`, `STDDEV_SAMP()`, `VARIANCE()`, `VARIANCE_POP()`, `VARIANCE_SAMP()`, and `NDV()` aggregate functions now all return `DOUBLE` results rather than `STRING`. Formerly, you were required to `CAST()` the result to a numeric type before using it in arithmetic operations.
- The default settings for Parquet block size, and the associated `PARQUET_FILE_SIZE` query option, are changed. Now, Impala writes Parquet files with a size of 256 MB and an HDFS block size of 256 MB. Previously, Impala attempted to write Parquet files with a size of 1 GB and an HDFS block size of 1 GB. In practice, Impala used a conservative estimate of the disk space needed for each Parquet block, leading to files that were typically 512 MB anyway. Thus, this change will make the file size more accurate if you specify a value for the `PARQUET_FILE_SIZE` query option. It also reduces the amount of memory reserved during `INSERT` into Parquet tables, potentially avoiding out-of-memory errors and improving scalability when inserting data into Parquet tables.
- Anti-joins are now supported, expressed using the `LEFT ANTI JOIN` and `RIGHT ANTI JOIN` clauses. These clauses returns results from one table that have no match in the other table. You might use this type of join in the same sorts of use cases as the `NOT EXISTS` and `NOT IN` operators. See [Joins in Impala SELECT Statements](#) for details.
- The `SET` command in `impala-shell` has been promoted to a real SQL statement. You can now set query options such as `PARQUET_FILE_SIZE`, `MEM_LIMIT`, and `SYNC_DDL` within JDBC, ODBC, or any other kind of application that submits SQL without going through the `impala-shell` interpreter. See [SET Statement](#) for details.
- The `impala-shell` interpreter now reads settings from an optional configuration file, named `$HOME/.impalarc` by default. See [impala-shell Configuration File](#) for details.
- The library used for regular expression parsing has changed from Boost to Google RE2. This implementation change adds support for non-greedy matches using the `*?` notation. This and other changes in the way regular expressions are interpreted means you might need to re-test queries that use functions such as `regexp_extract()` or `regexp_replace()`, or operators such as `REGEXP` or `RLIKE`. See [Apache Impala \(incubating\) Incompatible Changes and Limitations](#) on page 94 for those details.

New Features in Impala 1.4.4 / CDH 5.1.5

No new features. This point release is exclusively a bug fix release.



Note: Impala 1.4.4 is available as part of CDH 5.1.5, not under CDH 4.

New Features in Impala 1.4.3 / CDH 5.1.4

No new features. This point release is exclusively a bug fix release for an SSL security issue.



Note: Impala 1.4.3 is available as part of CDH 5.1.4, and under CDH 4.

New Features in Impala 1.4.2 / CDH 5.1.3

Impala 1.4.2 is purely a bug-fix release. It does not include any new features.



Note: Impala 1.4.2 is only available as part of CDH 5.1.3, not under CDH 4.

New Features in Impala 1.4.1 / CDH 5.1.2

Impala 1.4.1 is purely a bug-fix release. It does not include any new features.

New Features in Impala 1.4.0 / CDH 5.1.0

The following are the major new features in Impala 1.4.

- The `DECIMAL` data type lets you store fixed-precision values, for working with currency or other fractional values where it is important to represent values exactly and avoid rounding errors. This feature includes enhancements to built-in functions, numeric literals, and arithmetic expressions.
- On CDH 5, Impala can take advantage of the HDFS caching feature to “pin” entire tables or individual partitions in memory, to speed up queries on frequently accessed data and reduce the CPU overhead of memory-to-memory copying. When HDFS files are cached in memory, Impala can read the cached data without any disk reads, and without making an additional copy of the data in memory. Other Hadoop components that read the same data files also experience a performance benefit.
- Impala can now use Sentry-based authorization based either on the original policy file, or on rules defined by `GRANT` and `REVOKE` statements issued through Hive.
- For interoperability with Parquet files created through other Hadoop components, such as Pig or MapReduce jobs, you can create an Impala table that automatically sets up the column definitions based on the layout of an existing Parquet data file.
- `ORDER BY` queries no longer require a `LIMIT` clause. If the size of the result set to be sorted exceeds the memory available to Impala, Impala uses a temporary work space on disk to perform the sort operation.
- LDAP connections can be secured through either SSL or TLS.
- The following new built-in scalar and aggregate functions are available:
 - A new built-in function, `EXTRACT()`, returns one date or time field from a `TIMESTAMP` value.
 - A new built-in function, `TRUNC()`, truncates date/time values to a particular granularity, such as year, month, day, hour, and so on.
 - `ADD_MONTHS()` built-in function, an alias for the existing `MONTHS_ADD()` function.
 - A new built-in function, `ROUND()`, rounds `DECIMAL` values to a specified number of fractional digits.
 - Several built-in aggregate functions for computing properties for statistical distributions: `STDDEV()`, `STDDEV_SAMP()`, `STDDEV_POP()`, `VARIANCE()`, `VARIANCE_SAMP()`, and `VARIANCE_POP()`.
 - Several new built-in functions, such as `MAX_INT()`, `MIN_SMALLINT()`, and so on, let you conveniently check whether data values are in an expected range. You might be able to switch a column to a smaller type, saving memory during processing.

- New built-in functions, `IS_INF()` and `IS_NAN()`, check for the special values infinity and “not a number”. These values could be specified as `inf` or `nan` in text data files, or be produced by certain arithmetic expressions.
- The `SHOW PARTITIONS` statement displays information about the structure of a partitioned table.
- New configuration options for the `impalad` daemon let you specify initial memory usage for all queries. The initial resource requests handled by Llama and YARN can be expanded later if needed, avoiding unnecessary over-allocation and reducing the chance of out-of-memory conditions.
- The Impala `CREATE TABLE` statement now has a `STORED AS AVRO` clause, allowing you to create Avro tables through Impala.
- New `impalad` configuration options let you fine-tune the calculations Impala makes to estimate resource requirements for each query. These options can help avoid problems due to overconsumption due to too-low estimates, or underutilization due to too-high estimates.
- A new `SUMMARY` command in the `impala-shell` interpreter provides a high-level summary of the work performed at each stage of the explain plan. The summary is also included in output from the `PROFILE` command.
- Performance improvements for the `COMPUTE STATS` statement:
 - The `NDV` function is speeded up through native code generation.
 - Because the `NULL` count is not currently used by the Impala query planner, in Impala 1.4.0 and higher, `COMPUTE STATS` does not count the `NULL` values for each column. (The `#Nulls` field of the stats table is left as `-1`, signifying that the value is unknown.)
- Performance improvements for partition pruning. This feature reduces the time spent in query planning, for partitioned tables with thousands of partitions. Previously, Impala typically queried tables with up to approximately 3000 partitions. With the performance improvement in partition pruning, now Impala can comfortably handle tables with tens of thousands of partitions.
- The documentation provides additional guidance for planning tasks.
- The `impala-shell` interpreter now supports UTF-8 characters for input and output. You can control whether `impala-shell` ignores invalid Unicode code points through the `--strict_unicode` option. (Although this option is removed in Impala 2.0.)

New Features in Impala 1.3.3 / CDH 5.0.5

No new features. This point release is exclusively a bug fix release for an SSL security issue.



Note: Impala 1.3.3 is only available as part of CDH 5.0.5, not under CDH 4.

New Features in Impala 1.3.2 / CDH 5.0.4

No new features. This point release is exclusively a bug fix release for the IMPALA-1019 issue related to HDFS caching.



Note: Impala 1.3.2 is only available as part of CDH 5.0.4, not under CDH 4.

New Features in Impala 1.3.1 / CDH 5.0.3

This point release is primarily a vehicle to deliver bug fixes. Any new features are minor changes resulting from fixes for performance, reliability, or usability issues.

Because 1.3.1 is the first 1.3.x release for CDH 4, if you are on CDH 4, also consult [New Features in Impala 1.3.0 / CDH 5.0.0](#) on page 76 for more features that are new to you.

**Note:**

- The Impala 1.3.1 release is available for both CDH 4 and CDH 5. This is the first release in the 1.3.x series for CDH 4.

- A new `impalad` startup option, `--insert_inherit_permissions`, causes Impala `INSERT` statements to create each new partition with the same HDFS permissions as its parent directory. By default, `INSERT` statements create directories for new partitions using default HDFS permissions. See [INSERT Statement](#) for examples of `INSERT` statements for partitioned tables.
- The `SHOW FUNCTIONS` statement now displays the return type of each function, in addition to the types of its arguments. See [SHOW Statement](#) for examples.
- You can now specify the clause `FIELDS TERMINATED BY '\0'` with a `CREATE TABLE` statement to use text data files that use ASCII 0 (`\nul`) characters as a delimiter. See [Using Text Data Files with Impala Tables](#) for details.
- In Impala 1.3.1 and higher, the `REGEEXP` and `RLIKE` operators now match a regular expression string that occurs anywhere inside the target string, the same as if the regular expression was enclosed on each side by `.*`. See [REGEXP Operator](#) for examples. Previously, these operators only succeeded when the regular expression matched the entire target string. This change improves compatibility with the regular expression support for popular database systems. There is no change to the behavior of the `regexp_extract()` and `regexp_replace()` built-in functions.

New Features in Impala 1.3.0 / CDH 5.0.0

**Note:**

- The Impala 1.3.1 release is available for both CDH 4 and CDH 5. This is the first release in the 1.3.x series for CDH 4.

- The admission control feature lets you control and prioritize the volume and resource consumption of concurrent queries. This mechanism reduces spikes in resource usage, helping Impala to run alongside other kinds of workloads on a busy cluster. It also provides more user-friendly conflict resolution when multiple memory-intensive queries are submitted concurrently, avoiding resource contention that formerly resulted in out-of-memory errors. See [Admission Control and Query Queuing](#) for details.
- Enhanced `EXPLAIN` plans provide more detail in an easier-to-read format. Now there are four levels of verbosity: the `EXPLAIN_LEVEL` option can be set from 0 (most concise) to 3 (most verbose). See [EXPLAIN Statement](#) for syntax and [Understanding Impala Query Performance - EXPLAIN Plans and Query Profiles](#) for usage information.
- The `TIMESTAMP` data type accepts more kinds of input string formats through the `UNIX_TIMESTAMP` function, and produces more varieties of string formats through the `FROM_UNIXTIME` function. The documentation now also lists more functions for date arithmetic, used for adding and subtracting `INTERVAL` expressions from `TIMESTAMP` values. See [Impala Date and Time Functions](#) for details.
- New conditional functions, `NULIF()`, `NULIFZERO()`, and `ZEROIFNULL()`, simplify porting SQL containing vendor extensions to Impala. See [Impala Conditional Functions](#) for details.
- New utility function, `CURRENT_DATABASE()`. See [Impala Miscellaneous Functions](#) for details.
- Integration with the YARN resource management framework. Only available in combination with CDH 5. This feature makes use of the underlying YARN service, plus an additional service (Llama) that coordinates requests to YARN for Impala resources, so that the Impala query only proceeds when all requested resources are available. See [Resource Management for Impala](#) for full details.

On the Impala side, this feature involves some new startup options for the `impalad` daemon:

- `-enable_rm`

- `--llama_host`
- `--llama_port`
- `--llama_callback_port`
- `--cgroup_hierarchy_path`

For details of these startup options, see [Modifying Impala Startup Options](#).

This feature also involves several new or changed query options that you can set through the `impala-shell` interpreter and apply within a specific session:

- `MEM_LIMIT`: the function of this existing option changes when Impala resource management is enabled.
- `REQUEST_POOL`: a new option. (Renamed to `RESOURCE_POOL` in Impala 1.3.0.)
- `V_CPU_CORES`: a new option.
- `RESERVATION_REQUEST_TIMEOUT`: a new option.

For details of these query options, see [impala-shell Query Options for Resource Management](#).

New Features in Impala 1.2.4



Note: Impala 1.2.4 works with CDH 4. It is primarily a bug fix release for Impala 1.2.3, plus some performance enhancements for the catalog server to minimize startup and DDL wait times for Impala deployments with large numbers of databases, tables, and partitions.

- On Impala startup, the metadata loading and synchronization mechanism has been improved and optimized, to give more responsiveness when starting Impala on a system with a large number of databases, tables, or partitions. The initial metadata loading happens in the background, allowing queries to be run before the entire process is finished. When a query refers to a table whose metadata is not yet loaded, the query waits until the metadata for that table is loaded, and the load operation for that table is prioritized to happen first.
- Formerly, if you created a new table in Hive, you had to issue the `INVALIDATE_METADATA` statement (with no table name) which was an expensive operation that reloaded metadata for all tables. Impala did not recognize the name of the Hive-created table, so you could not do `INVALIDATE_METADATA new_table` to get the metadata for just that one table. Now, when you issue `INVALIDATE_METADATA table_name`, Impala checks to see if that name represents a table created in Hive, and if so recognizes the new table and loads the metadata for it. Additionally, if the new table is in a database that was newly created in Hive, Impala also recognizes the new database.
- If you issue `INVALIDATE_METADATA table_name` and the table has been dropped through Hive, Impala will recognize that the table no longer exists.
- New startup options let you control the parallelism of the metadata loading during startup for the `catalogd` daemon:
 - `--load_catalog_in_background` makes Impala load and cache metadata using background threads after startup. It is true by default. Previously, a system with a large number of databases, tables, or partitions could be unresponsive or even time out during startup.
 - `--num_metadata_loading_threads` determines how much parallelism Impala devotes to loading metadata in the background. The default is 16. You might increase this value for systems with huge numbers of databases, tables, or partitions. You might lower this value for busy systems that are CPU-constrained due to jobs from components other than Impala.

New Features in Impala 1.2.3



Note: Impala 1.2.3 works with CDH 4 and with CDH 5 beta 2. The resource management feature requires CDH 5 beta.

Impala 1.2.3 contains exactly the same feature set as Impala 1.2.2. Its only difference is one additional fix for compatibility with Parquet files generated outside of Impala by components such as Hive, Pig, or MapReduce. If you are upgrading from Impala 1.2.1 or earlier, see [New Features in Impala 1.2.2](#) on page 78 for the latest added features.

New Features in Impala 1.2.2



Note: Impala 1.2.2 works with CDH 4. Its feature set is a superset of features in the Impala 1.2.0 beta, with the exception of resource management, which relies on CDH 5.

Impala 1.2.2 includes new features for performance, security, and flexibility. The major enhancements over 1.2.1 are performance related, primarily for join queries.

New user-visible features include:

- Join order optimizations. This highly valuable feature automatically distributes and parallelizes the work for a join query to minimize disk I/O and network traffic. The automatic optimization reduces the need to use query hints or to rewrite join queries with the tables in a specific order based on size or cardinality. The new `COMPUTE STATS` statement gathers statistical information about each table that is crucial for enabling the join optimizations. See [Performance Considerations for Join Queries](#) for details.
- `COMPUTE STATS` statement to collect both table statistics and column statistics with a single statement. Intended to be more comprehensive, efficient, and reliable than the corresponding Hive `ANALYZE TABLE` statement, which collects statistics in multiple phases through MapReduce jobs. These statistics are important for query planning for join queries, queries on partitioned tables, and other types of data-intensive operations. For optimal planning of join queries, you need to collect statistics for each table involved in the join. See [COMPUTE STATS Statement](#) for details.
- Reordering of tables in a join query can be overridden by the `STRAIGHT_JOIN` operator, allowing you to fine-tune the planning of the join query if necessary, by using the original technique of ordering the joined tables in descending order of size. See [Overriding Join Reordering with STRAIGHT_JOIN](#) for details.
- The `CROSS JOIN` clause in the `SELECT` statement to allow Cartesian products in queries, that is, joins without an equality comparison between columns in both tables. Because such queries must be carefully checked to avoid accidental overconsumption of memory, you must use the `CROSS JOIN` operator to explicitly select this kind of join. See [Cross Joins and Cartesian Products with the CROSS JOIN Operator](#) for examples.
- The `ALTER TABLE` statement has new clauses that let you fine-tune table statistics. You can use this technique as a less-expensive way to update specific statistics, in case the statistics become stale, or to experiment with the effects of different data distributions on query planning.
- LDAP username/password authentication in JDBC/ODBC. See [Enabling LDAP Authentication for Impala](#) for details.
- `GROUP_CONCAT()` aggregate function to concatenate column values across all rows of a result set.
- The `INSERT` statement now accepts hints, `[SHUFFLE]` and `[NOSHUFFLE]`, to influence the way work is redistributed during `INSERT...SELECT` operations. The hints are primarily useful for inserting into partitioned Parquet tables, where using the `[SHUFFLE]` hint can avoid problems due to memory consumption and simultaneous open files in HDFS, by collecting all the new data for each partition on a specific node.
- Several built-in functions and operators are now overloaded for more numeric data types, to reduce the requirement to use `CAST()` for type coercion in `INSERT` statements. For example, the expression `2+2` in an `INSERT` statement formerly produced a `BIGINT` result, requiring a `CAST()` to be stored in an `INT` variable. Now, addition, subtraction, and multiplication only produce a result that is one step “bigger” than their arguments, and numeric and conditional functions can return `SMALLINT`, `FLOAT`, and other smaller types rather than always `BIGINT` or `DOUBLE`.
- New `fnv_hash()` built-in function for constructing hashed values. See [Impala Mathematical Functions](#) for details.
- The clause `STORED AS PARQUET` is accepted as an equivalent for `STORED AS PARQUETFILE`. This more concise form is recommended for new code.

Because Impala 1.2.2 builds on a number of features introduced in 1.2.1, if you are upgrading from an older 1.1.x release straight to 1.2.2, also review [New Features in Impala 1.2.1](#) on page 79 to see features such as the SHOW TABLE STATS and SHOW COLUMN STATS statements, and user-defined functions (UDFs).

New Features in Impala 1.2.1



Note: Impala 1.2.1 works with CDH 4. Its feature set is a superset of features in the Impala 1.2.0 beta, with the exception of resource management, which relies on CDH 5.

Impala 1.2.1 includes new features for security, performance, and flexibility.

New user-visible features include:

- SHOW TABLE STATS *table_name* and SHOW COLUMN STATS *table_name* statements, to verify that statistics are available and to see the values used during query planning.
- CREATE TABLE AS SELECT syntax, to create a new table and transfer data into it in a single operation.
- OFFSET clause, for use with the ORDER BY and LIMIT clauses to produce “paged” result sets such as items 1-10, then 11-20, and so on.
- NULLS FIRST and NULLS LAST clauses to ensure consistent placement of NULL values in ORDER BY queries.
- New [built-in functions](#): least(), greatest(), initcap().
- New aggregate function: ndv(), a fast alternative to COUNT(DISTINCT *col*) returning an approximate result.
- The LIMIT clause can now accept a numeric expression as an argument, rather than only a literal constant.
- The SHOW CREATE TABLE statement displays the end result of all the CREATE TABLE and ALTER TABLE statements for a particular table. You can use the output to produce a simplified setup script for a schema.
- The --idle_query_timeout and --idle_session_timeout options for impalad control the time intervals after which idle queries are cancelled, and idle sessions expire. See [Setting Timeout Periods for Daemons, Queries, and Sessions](#) for details.
- User-defined functions (UDFs). This feature lets you transform data in very flexible ways, which is important when using Impala as part of an ETL or ELT pipeline. Prior to Impala 1.2, using UDFs required switching into Hive. Impala 1.2 can run scalar UDFs and user-defined aggregate functions (UDAs). Impala can run high-performance functions written in C++, or you can reuse existing Hive functions written in Java.

You create UDFs through the CREATE FUNCTION statement and drop them through the DROP FUNCTION statement. See [Impala User-Defined Functions \(UDFs\)](#) for instructions about coding, building, and deploying UDFs, and [CREATE FUNCTION Statement](#) and [DROP FUNCTION Statement](#) for related SQL syntax.

- A new service automatically propagates changes to table data and metadata made by one Impala node, sending the new or updated metadata to all the other Impala nodes. The automatic synchronization mechanism eliminates the need to use the INVALIDATE METADATA and REFRESH statements after issuing Impala statements such as CREATE TABLE, ALTER TABLE, DROP TABLE, INSERT, and LOAD DATA.

For even more precise synchronization, you can enable the [SYNC_DDL](#) query option before issuing a DDL, INSERT, or LOAD DATA statement. This option causes the statement to wait, returning only after the catalog service has broadcast the applicable changes to all Impala nodes in the cluster.

**Note:**

Because the catalog service only monitors operations performed through Impala, `INVALIDATE METADATA` and `REFRESH` are still needed on the Impala side after creating new tables or loading data through the Hive shell or by manipulating data files directly in HDFS. Because the catalog service broadcasts the result of the `REFRESH` and `INVALIDATE METADATA` statements to all Impala nodes, when you do need to use those statements, you can do so a single time rather than on every Impala node.

This service is implemented by the `catalogd` daemon. See [The Impala Catalog Service](#) for details.

- `CREATE TABLE ... AS SELECT` syntax, to create a table and copy data into it in a single operation. See [CREATE TABLE Statement](#) for details.
- The `CREATE TABLE` and `ALTER TABLE` statements have new clauses `TBLPROPERTIES` and `WITH SERDEPROPERTIES`. The `TBLPROPERTIES` clause lets you associate arbitrary items of metadata with a particular table as key-value pairs. The `WITH SERDEPROPERTIES` clause lets you specify the serializer/deserializer (SerDes) classes that read and write data for a table; although Impala does not make use of these properties, sometimes particular values are needed for Hive compatibility. See [CREATE TABLE Statement](#) and [ALTER TABLE Statement](#) for details.
- Delegation support lets you authorize certain OS users associated with applications (for example, `hue`), to submit requests using the credentials of other users. Only available in combination with CDH 5. See [Configuring Impala Delegation for Hue and BI Tools](#) for details.
- Enhancements to `EXPLAIN` output. In particular, when you enable the new `EXPLAIN_LEVEL` query option, the `EXPLAIN` and `PROFILE` statements produce more verbose output showing estimated resource requirements and whether table and column statistics are available for the applicable tables and columns. See [EXPLAIN Statement](#) for details.
- `SHOW CREATE TABLE` summarizes the effects of the original `CREATE TABLE` statement and any subsequent `ALTER TABLE` statements, giving you a `CREATE TABLE` statement that will re-create the current structure and layout for a table.
- The `LIMIT` clause for queries now accepts an arithmetic expression, in addition to numeric literals.

New Features in Impala 1.2.0 (Beta)



Note: The Impala 1.2.0 beta release only works in combination with the beta version of CDH 5. The Impala 1.2.0 software is bundled together with the CDH 5 beta 1 download.

The Impala 1.2.0 beta includes new features for security, performance, and flexibility.

New user-visible features include:

- User-defined functions (UDFs). This feature lets you transform data in very flexible ways, which is important when using Impala as part of an ETL or ELT pipeline. Prior to Impala 1.2, using UDFs required switching into Hive. Impala 1.2 can run scalar UDFs and user-defined aggregate functions (UDAs). Impala can run high-performance functions written in C++, or you can reuse existing Hive functions written in Java.

You create UDFs through the `CREATE FUNCTION` statement and drop them through the `DROP FUNCTION` statement. See [Impala User-Defined Functions \(UDFs\)](#) for instructions about coding, building, and deploying UDFs, and [CREATE FUNCTION Statement](#) and [DROP FUNCTION Statement](#) for related SQL syntax.

- A new service automatically propagates changes to table data and metadata made by one Impala node, sending the new or updated metadata to all the other Impala nodes. The automatic synchronization mechanism eliminates the need to use the `INVALIDATE METADATA` and `REFRESH` statements after issuing Impala statements such as `CREATE TABLE`, `ALTER TABLE`, `DROP TABLE`, `INSERT`, and `LOAD DATA`.

**Note:**

Because this service only monitors operations performed through Impala, `INVALIDATE METADATA` and `REFRESH` are still needed on the Impala side after creating new tables or loading data through the Hive shell or by manipulating data files directly in HDFS. Because the catalog service broadcasts the result of the `REFRESH` and `INVALIDATE METADATA` statements to all Impala nodes, when you do need to use those statements, you can do so a single time rather than on every Impala node.

This service is implemented by the `catalogd` daemon. See [The Impala Catalog Service](#) for details.

- Integration with the YARN resource management framework. Only available in combination with CDH 5. This feature makes use of the underlying YARN service, plus an additional service (Llama) that coordinates requests to YARN for Impala resources, so that the Impala query only proceeds when all requested resources are available. See [Resource Management for Impala](#) for full details.

On the Impala side, this feature involves some new startup options for the `impalad` daemon:

- `-enable_rm`
- `-llama_host`
- `-llama_port`
- `-llama_callback_port`
- `-cgroup_hierarchy_path`

For details of these startup options, see [Modifying Impala Startup Options](#).

This feature also involves several new or changed query options that you can set through the `impala-shell` interpreter and apply within a specific session:

- `MEM_LIMIT`: the function of this existing option changes when Impala resource management is enabled.
- `YARN_POOL`: a new option. (Renamed to `RESOURCE_POOL` in Impala 1.3.0.)
- `V_CPU_CORES`: a new option.
- `RESERVATION_REQUEST_TIMEOUT`: a new option.

For details of these query options, see [impala-shell Query Options for Resource Management](#).

- `CREATE TABLE ... AS SELECT` syntax, to create a table and copy data into it in a single operation. See [CREATE TABLE Statement](#) for details.
- The `CREATE TABLE` and `ALTER TABLE` statements have a new `TBLPROPERTIES` clause that lets you associate arbitrary items of metadata with a particular table as key-value pairs. See [CREATE TABLE Statement](#) and [ALTER TABLE Statement](#) for details.
- Delegation support lets you authorize certain OS users associated with applications (for example, `hue`), to submit requests using the credentials of other users. Only available in combination with CDH 5. See [Configuring Impala Delegation for Hue and BI Tools](#) for details.
- Enhancements to `EXPLAIN` output. In particular, when you enable the new `EXPLAIN_LEVEL` query option, the `EXPLAIN` and `PROFILE` statements produce more verbose output showing estimated resource requirements and whether table and column statistics are available for the applicable tables and columns. See [EXPLAIN Statement](#) for details.

New Features in Impala 1.1.1

Impala 1.1.1 includes new features for security and stability.

New user-visible features include:

- Additional security feature: auditing. New startup options for `impalad` let you capture information about Impala queries that succeed or are blocked due to insufficient privileges. To take full advantage of this feature with Cloudera Manager, upgrade to Cloudera Manager 4.7 or higher. For details, see [Overview of Impala Security](#).
- Parquet data files generated by Impala 1.1.1 are now compatible with the Parquet support in Hive. See [Apache Impala \(incubating\) Incompatible Changes and Limitations](#) on page 94 for the procedure to update older Impala-created Parquet files to be compatible with the Hive Parquet support.
- Additional improvements to stability and resource utilization for Impala queries.
- Additional enhancements for compatibility with existing file formats.

New Features in Impala 1.1

Impala 1.1 includes new features for security, performance, and usability.

New user-visible features include:

- Extensive new security features, built on top of the Sentry open source project. Impala now supports fine-grained authorization based on roles. A policy file determines which privileges on which schema objects (servers, databases, tables, and HDFS paths) are available to users based on their membership in groups. By assigning privileges for views, you can control access to table data at the column level. For details, see [Overview of Impala Security](#).
- Impala 1.1 works with Cloudera Manager 4.6 or higher. To use Cloudera Manager to manage authorization for the Impala web UI (the web pages served from port 25000 by default), use Cloudera Manager 4.6.2 or higher.
- Impala can now create, alter, drop, and query views. Views provide a flexible way to set up simple aliases for complex queries; hide query details from applications and users; and simplify maintenance as you rename or reorganize databases, tables, and columns. See the overview section [Overview of Impala Views](#) and the statements [CREATE VIEW Statement](#), [ALTER VIEW Statement](#), and [DROP VIEW Statement](#).
- Performance is improved through a number of automatic optimizations. Resource consumption is also reduced for Impala queries. These improvements apply broadly across all kinds of workloads and file formats. The major areas of performance enhancement include:
 - Improved disk and thread scheduling, which applies to all queries.
 - Improved hash join and aggregation performance, which applies to queries with large build tables or a large number of groups.
 - Dictionary encoding with Parquet, which applies to Parquet tables with short string columns.
 - Improved performance on systems with SSDs, which applies to all queries and file formats.
- Some new built-in functions are implemented: [translate\(\)](#) to substitute characters within strings, [user\(\)](#) to check the login ID of the connected user.
- The new `WITH` clause for `SELECT` statements lets you simplify complicated queries in a way similar to creating a view. The effects of the `WITH` clause only last for the duration of one query, unlike views, which are persistent schema objects that can be used by multiple sessions or applications. See [WITH Clause](#).
- An enhancement to `DESCRIBE` statement, `DESCRIBE FORMATTED table_name`, displays more detailed information about the table. This information includes the file format, location, delimiter, ownership, external or internal, creation and access times, and partitions. The information is returned as a result set that can be interpreted and used by a management or monitoring application. See [DESCRIBE Statement](#).
- You can now insert a subset of columns for a table, with other columns being left as all `NULL` values. Or you can specify the columns in any order in the destination table, rather than having to match the order of the corresponding columns in the source. `VALUES` clause. This feature is known as “column permutation”. See [INSERT Statement](#).
- The new `LOAD DATA` statement lets you load data into a table directly from an HDFS data file. This technique lets you minimize the number of steps in your ETL process, and provides more flexibility. For example, you can bring data into an Impala table in one step. Formerly, you might have created an external table where the data files are not entirely under your control, or copied the data files to Impala data directories manually, or loaded the original data into one table and then used the `INSERT` statement to copy it to a new table with a different file format, partitioning scheme, and so on. See [LOAD DATA Statement](#).
- Improvements to Impala-HBase integration:
 - New query options for HBase performance: [HBASE_CACHE_BLOCKS](#) and [HBASE_CACHING](#).
 - Support for binary data types in HBase tables. See [Supported Data Types for HBase Columns](#) for details.

- You can issue `REFRESH` as a SQL statement through any of the programming interfaces that Impala supports. `REFRESH` formerly had to be issued as a command through the `impala-shell` interpreter, and was not available through a JDBC or ODBC API call. As part of this change, the functionality of the `REFRESH` statement is divided between two statements. In Impala 1.1, `REFRESH` requires a table name argument and immediately reloads the metadata; the new `INVALIDATE METADATA` statement works the same as the Impala 1.0 `REFRESH` did: the table name argument is optional, and the metadata for one or all tables is marked as stale, but not actually reloaded until the table is queried. When you create a new table in the Hive shell or through a different Impala node, you must enter `INVALIDATE METADATA` with no table parameter before you can see the new table in `impala-shell`. See [REFRESH Statement](#) and [INVALIDATE METADATA Statement](#).

New Features in Impala 1.0.1

The primary enhancements in Impala 1.0.1 are internal, for compatibility with the new Cloudera Manager 4.6 release. Try out the new **Impala Query Monitoring** feature in Cloudera Manager 4.6, which requires Impala 1.0.1.

New user-visible features include:

- The `VALUES` clause lets you `INSERT` one or more rows using literals, function return values, or other expressions. For performance and scalability, you should still use `INSERT ... SELECT` for bringing large quantities of data into an Impala table. The `VALUES` clause is a convenient way to set up small tables, particularly for initial testing of SQL features that do not require large amounts of data. See [VALUES Clause](#) for details.
- The `-B` and `-o` options of the `impala-shell` command can turn query results into delimited text files and store them in an output file. The plain text results are useful for using with other Hadoop components or Unix tools. In benchmark tests, it is also faster to produce plain rather than pretty-printed results, and write to a file rather than to the screen, giving a more accurate picture of the actual query time.
- Several bug fixes. See [Issues Fixed in the 1.0.1 Release](#) on page 346 for details.

New Features in Impala 1.0

This version has multiple performance improvements and adds the following functionality:

- Several bug fixes. See [Issues Fixed in the 1.0 GA Release](#) on page 347.
- [ALTER TABLE](#) statement.
- [Hints](#) to allow specifying a particular join strategy.
- [REFRESH](#) for a single table.
- Dynamic resource management, allowing high concurrency for Impala queries.

New Features in Version 0.7 of the Impala Beta Release

This version has multiple performance improvements and adds the following functionality:

- Several bug fixes. See [Issues Fixed in Version 0.7 of the Beta Release](#) on page 349.
- Support for the Parquet file format. For more information on file formats, see [How Impala Works with Hadoop File Formats](#).
- Added support for Avro.
- Support for the memory limits. For more information, see the example on modifying memory limits in [Modifying Impala Startup Options](#).
- Bigger and faster joins through the addition of partitioned joins to the already supported broadcast joins.
- Fully distributed aggregations.
- Fully distributed top-n computation.
- Support for creating and altering tables.
- Support for GROUP BY with floats and doubles.

In this version, both CDH 4.1 and 4.2 are supported, but due to performance improvements added, we highly recommend you use CDH 4.2 or higher to see the full benefit. If you are using Cloudera Manager, version 4.5 is required.

New Features in Version 0.6 of the Impala Beta Release

- Several bug fixes. See [Issues Fixed in Version 0.6 of the Beta Release](#) on page 350.

- Added support for Impala on SUSE and Debian/Ubuntu. Impala is now supported on:
 - RHEL5.7/6.2 and Centos5.7/6.2
 - SUSE 11 with Service Pack 1 or higher
 - Ubuntu 10.04/12.04 and Debian 6.03
- Cloudera Manager 4.5 and CDH 4.2 support Impala 0.6.
- Support for the RCFile file format. For more information on file formats, see [Understanding File Formats](#).

New Features in Version 0.5 of the Impala Beta Release

- Several bug fixes. See [Issues Fixed in Version 0.5 of the Beta Release](#) on page 351.
- Added support for a JDBC driver that allows you to access Impala from a Java client. To use this feature, follow the instructions in [Configuring Impala to Work with JDBC](#) to install the JDBC driver JARs on the client machine and modify the CLASSPATH on the client to include the JARs.

New Features in Version 0.4 of the Impala Beta Release

- Several bug fixes. See [Issues Fixed in Version 0.4 of the Beta Release](#) on page 352.
- Added support for Impala on RHEL5.7/Centos5.7. Impala is now supported on RHEL5.7/6.2 and Centos5.7/6.2.
- Cloudera Manager 4.1.3 supports Impala 0.4.
- The Impala debug webserver now has the ability to serve static files from \${IMPALA_HOME}/www. This can be disabled by setting --enable_webserver_doc_root=false on the command line. As a result, Impala now uses the Twitter Bootstrap library to style its debug webpages, and the /queries page now tracks the last 25 queries run by each Impala daemon.
- Additional metrics available on the Impala Debug Webpage.

New Features in Version 0.3 of the Impala Beta Release

- Several bug fixes. See [Issues Fixed in Version 0.3 of the Beta Release](#) on page 352.
- The state-store-service binary has been renamed statestored.
- The location of the Impala configuration files has changed from the /usr/lib/impala/conf directory to the /etc/impala/conf directory.

New Features in Version 0.2 of the Impala Beta Release

- Several bug fixes. See [Issues Fixed in Version 0.2 of the Beta Release](#) on page 352.
- **Added Default Query Options** Default query options override all default QueryOption values when starting impalad. The format is:

```
-default_query_options='key=value;key=value'
```

Incompatible Changes and Limitations



Important:

For changes in operating system support and other major requirements, see [CDH and Cloudera Manager Supported Operating Systems](#) on page 505.

Apache Avro Incompatible Changes and Limitations

See [Apache Avro](#) on page 37 section of [What's New In CDH 5.2.x](#) on page 37 for two changes that could possibly affect you when you upgrade to CDH 5.2.0.

Apache Crunch Incompatible Changes and Limitations

The following changes introduced in CDH 5.2 are not backward compatible:

- The `MemPipeline` now checks to ensure that any `DoFns` that are passed to it are serializable. This is designed to catch non-serializable `DoFns` during testing.
- Scala's `Iterable` has been replaced by `TraversableOnce` inside `Scrunch flatMap` functions in order to support functions that return Iterators.

CDH 5.4.0 introduces new HBase APIs, which will probably require some changes to Crunch code developed against HBase 0.96 APIs. For more information, see the section on [Apache Crunch](#) on page 28 under "What's New in CDH 5.4.0".

Apache DataFu Incompatible Changes and Limitations

- Upgraded from version 0.4 to 1.1.0 (this upgrade is not backwards compatible).
- Removed `ApplyQuantiles`, `AliasBagFields`.
- Renamed package `datafu.pig.numbers` to `datafu.pig.random`.
- Renamed package `datafu.pig.bag.sets` to `datafu.pig.sets`.
- Renamed `TimeCount` to `SessionCount`, moved to `datafu.pig.sessions`.

Apache Flume Incompatible Changes and Limitations

There are no incompatible changes at this point.

Apache Hadoop Incompatible Changes and Limitations

HDFS

The following incompatible changes have been introduced in CDH 5:

- [HDFS-6434](#): Default permission for creating file should be 644 for WebHdfs/HttpFS.
- [HDFS-9085](#): Show renewer information in `DelegationTokenIdentifier#toString`.
- The `getSnapshottableDirListing()` method returns `null` when there are no snapshottable directories. This is a change from CDH 5 Beta 2 where the method returns an empty array instead.
- Files named `.snapshot` or `.reserved` must not exist within HDFS.
- [HADOOP-10020](#): Disable symlinks temporarily.
- [HDFS-2832](#) - The HDFS internal layout version has changed between CDH 5 Beta 1 and CDH 5 Beta 2, so a file system upgrade is required to move an existing Beta 1 cluster to Beta 2.
- [HDFS-4451](#): HDFS balancer command returns exit code 0 on success instead of 1.
- [HDFS-4594](#): WebHDFS open sets Content-Length header to what is specified by length parameter rather than how much data is actually returned.
 - **Impact:** In CDH 5, Content-Length header will contain the number of bytes actually returned, rather than the request length.
- [HDFS-4659](#): Support setting execution bit for regular files.
 - **Impact:** In CDH 5, files copied out of `copyToLocal` may now have the executable bit set if it was set when they were created or copied into HDFS.
- [HDFS-4997](#) - libhdfs functions now return correct error codes in `errno` in case of an error, instead of always returning 255.
- [HDFS-5138](#) - The `-finalize` NameNode startup option has been removed. To finalize an in-progress upgrade, you should instead use the `hdfs dfsadmin -finalizeUpgrade` command while your NameNode is running, or while both NameNodes are running in a High Availability setup.
- [HDFS-7279](#) - In CDH 5.5.0 and higher, DataNode WebHDFS implementation uses Netty as an HTTP server instead of Jetty.

Change in High-Availability Support

In CDH 5, the only high-availability (HA) implementation is Quorum-based storage; shared storage using NFS is no longer supported.

MapReduce



Important: There is no separate tarball for MRv1. Instead, the MRv1 binaries, examples, and other contents are delivered in the Hadoop tarball itself. The scripts for running MRv1 are in the `bin-mapreduce1` directory in the tarball, and the MRv1 examples are in the `examples-mapreduce1` directory. You need to do some additional configuration; follow the directions below.

To use MRv1 from a tarball installation, complete the following steps:

1. Extract the files from the tarball.



Note: In the steps that follow, `install_dir` is the name of the directory into which you extracted the files.

2. Create a symbolic link as follows:

```
ln -s install_dir/bin-mapreduce1 install_dir/share/hadoop/mapreduce1/bin
```

3. Create a second symbolic link as follows:

```
ln -s install_dir/etc/hadoop-mapreduce1 install_dir/share/hadoop/mapreduce1/conf
```

4. Set the `HADOOP_HOME` and `HADOOP_CONF_DIR` environment variables in your execution environment as follows:

```
$ export HADOOP_HOME=install_dir/share/hadoop/mapreduce1  
$ export HADOOP_CONF_DIR=$HADOOP_HOME/conf
```

5. Copy your existing `start-dfs.sh` and `stop-dfs.sh` scripts to `install_dir/bin-mapreduce1`

6. For convenience, add `install_dir/bin` to the `PATH` variable in your execution environment.

Apache MapReduce 2.0 (YARN) Incompatible Changes

The following incompatible changes occurred for Apache MapReduce 2.0 (YARN) between CDH 4.x and CDH 5 Beta 2:

- The `CATALINA_BASE` variable no longer determines whether a component is configured for YARN or MRv1. Use the `alternatives` command instead, and make sure `CATALINA_BASE` is not set.
- [YARN-1288](#) - YARN Fair Scheduler ACL change. Root queue defaults to everybody, and other queues default to nobody.
- YARN High Availability configurations have changed. Configuration keys have been renamed among other changes.
- The `YARN_HOME` property has been changed to `HADOOP_YARN_HOME`.
- Note the following changes to configuration properties in `yarn-site.xml`:
 - The value of `yarn.nodemanager.aux-services` should be changed from `mapreduce.shuffle` to `mapreduce_shuffle`.
 - `yarn.nodemanager.aux-services.mapreduce.shuffle.class` has been renamed to `yarn.nodemanager.aux-services.mapreduce_shuffle.class`
 - `yarn.resourcemanager.resourcemanager.connect.max.wait.secs` has been renamed to `yarn.resourcemanager.connect.max-wait.secs`
 - `yarn.resourcemanager.resourcemanager.connect.retry_interval.secs` has been renamed to `yarn.resourcemanager.connect.retry-interval.secs`
 - `yarn.resourcemanager.am.max-retries` is renamed to `yarn.resourcemanager.am.max-attempts`

- The `YARN_HOME` environment variable used in the `yarn.application.classpath` has been renamed to `HADOOP_YARN_HOME`. Make sure you include `$HADOOP_YARN_HOME/*,$HADOOP_YARN_HOME/lib/*` in the classpath.
- A CDH 4 client cannot be used against a CDH 5 cluster and vice-versa. Note that YARN in CDH 4 is experimental, and suffers from the following major incompatibilities.
 - Almost all of the proto files have been renamed.
 - Several user-facing APIs have been modified as part of an API stabilization effort.

Apache MapReduce 2.0 (YARN) Limitations

DockerContainerExecutor not supported in YARN

Cloudera does not support DockerContainerExecutor in YARN.

Apache HBase Incompatible Changes and Limitations

Compatibility Notes for CDH 5

This section contains information that is relevant for all releases within the CDH 5 family. See the sections below for information which pertains to specific releases within CDH 5. If you are upgrading through more than one version (for instance, from CDH 5.0 to CDH 5.2), read the sections for each version, as most of the information listed applies to the given version and newer releases.

General Notes

- Rolling upgrades from CDH 4 to CDH 5 are not possible because existing CDH 4 HBase clients cannot make requests to CDH 5 servers and CDH 5 HBase clients cannot make requests to CDH 4 servers. Replication between CDH 4 and CDH 5 is not currently supported. Exposed JMX metrics in CDH 4 have been refactored and some have been removed.
- The upgrade from CDH 4 HBase to CDH 5 HBase is irreversible and requires HBase to be shutdown completely.
- As of CDH4.2, the default Split Policy changed from `ConstantSizeRegionSplitPolicy` to `IncreasingToUpperBoundRegionSplitPolicy` (`ITUBRSP`). This affects upgrades from CDH 4.1 or earlier to CDH 5.
- `FilterBase` no longer implements `Writable`. This means that you do not need to implement `readFields()` and `write()` methods when writing your own custom fields. Instead, put this logic into the `toByteArray` and `parseFrom` methods. See [this page](#) for an example.
- The default number of retained cell versions is reduced from 3 to 1. To increase the number of versions, you can specify the `VERSIONS` option at table creation or by altering existing tables. Starting with CDH 5.2, you can specify a global default number of versions, which will be applied to all newly created tables where the number of versions is not otherwise specified, by setting `hbase.column.max.version` to the desired number of versions in `hbase-site.xml`.
- In CDH 5 prior to 5.1.3, a Put submitted with a `KeyValue`, `KeyValue.Type.Delete` does not delete the cell. This is different from the behavior in CDH 4. In CDH 5.1.3, this behavior is changed, so that a Put submitted with a `KeyValue`, `KeyValue.Type.Delete` does delete the cell. This fix is provided in [HBASE-11788](#).

Developer API Changes

- The set of exposed APIs has been solidified. If you are using APIs outside of the [user API](#), we cannot guarantee compatibility with future minor versions.
- CDH 5 introduces a new layout for HBase build artifacts and requires POM changes if you use Maven, or JAR changes otherwise.

Previously, in CDH 4 you only needed to add a dependency for the HBase JAR:

```
<dependency>
<groupId> org.apache.hbase </groupId>
<artifactId> hbase </artifactId>
```

```
<optional> true </optional>
</dependency>
```

Now, when building against CDH 5 you will need to add a dependency for the `hbase-client` JAR. The `hbase` module continues to exist as a convenient top-level wrapper for existing clients, and it pulls in all the sub-modules automatically. But it is only a simple wrapper, so its repository directory will carry no actual jars.

```
<dependency>
  <groupId>org.apache.hbase</groupId>
  <artifactId>hbase-client</artifactId>
  <version>${hbase.version}</version>
</dependency>
```

If your code uses the HBase minicluster, you can pull in the `hbase-testing-util` dependency:

```
<dependency>
  <groupId>org.apache.hbase</groupId>
  <artifactId>hbase-testing-util</artifactId>
  <version>${cdh.hbase.version}</version>
</dependency>
```

If you need to obtain all HBase JARs required to build a project, copy them from the CDH installation directory (typically `/usr/lib/hbase` for an RPM install, or `/opt/cloudera/parcels/CDH/lib/hbase` if you install using Parcels), or from the [CDH 5 HBase tarballs](#). However, for building client applications, Cloudera recommends using build tools such as Maven, rather than manually referencing JARs.

- CDH 5 introduces support for addressing cells with an empty column qualifier (a string of 0 bytes in length), but not all edge services handle that scenario correctly. In some cases, attempting to address a cell at [`rowkey`, `fam`] results in interaction with the entire column family, rather than the empty column qualifier.

Users of the HBase Shell, MapReduce, REST, and Thrift must use `family` instead of `family:` (notice the omitted ":"), to interact with an entire column family, rather than an empty column qualifier. Including the ":" will be interpreted as an interaction with the empty qualifier in the `family` column family.

- **API Removals**

- [HBASE-7315/HBASE-7263](#) - Row lock user API has been removed.
- [HBASE-6706](#) - Removed total order partitioner.

Operator API Changes

- Many of the default configurations from CDH 4 in `hbase-default.xml` have been changed to new values in CDH 5. See [HBASE-8450](#) for a complete list of changes.
- [HBASE-6553](#) - Removed Avro Gateway. This feature was less robust and not used as much as the Thrift gateways. It has been removed upstream.
- HBase provides a metrics framework based on JMX beans. Between HBase 0.94 and 0.96, the metrics framework underwent many changes. Some beans were added and removed, some metrics were moved from one bean to another, and some metrics were renamed or removed. Click [here](#) to download the CSV spreadsheet which provides a mapping.

User API Changes

- The HBase User API (Get, Put, Result, Scanner etc; see [Apache HBase API documentation](#)) has evolved and attempts have been made to make sure the HBase Clients are source code compatible and thus should recompile without needing any source code modifications. This cannot be guaranteed however, since with the conversion to ProtoBufs, some relatively obscure APIs have been removed. Rudimentary efforts have also been made to preserve recompile compatibility with advanced APIs such as Filters and Coprocessors. These advanced APIs are still evolving and our guarantees for API compatibility are weaker here.
- As of 0.96, the User API has been marked and all attempts at compatibility in future versions will be made. A version of the javadoc that only contains the User API can be found [here](#).

- Other changes to CDH 5 HBase that require the upgrade include:
 - [HBASE-8015](#): The HBase Namespaces feature has changed HBase HDFS file layout.
 - [HBASE-4451](#): Renamed ZooKeeper nodes.
 - [HBASE-3171](#): The `META` table in CDH 4 has been renamed to be `hbase:meta`. Similarly the `ACL` table has been renamed to `hbase:acl`. The `.ROOT` table has been removed.
 - [HBASE-8352](#): HBase snapshots are now saved to the `/<hbase>/ .hbase-snapshot` dir instead of the `/ .snapshot` dir. This should be handled before upgrading HDFS.
 - [HBASE-7660](#): Removed support for HFile V1. All internal HBase files in the HFile v1 format must be converted to the HFile v2 format.
 - [HBASE-6170/HBASE-8909](#) - The `hbase.regionserver.lease.period` configuration parameter has been deprecated. Use `hbase.client.scanner.timeout.period` instead.
- The behavior of the filter `MUST_PASS_ALL` changed between CDH 4 and CDH 5. In CDH 4, a `FilterList` with the default `MUST_PASS_ALL` operator return all rows (not filtering the results). In CDH 5, no results are returned when the `FilterList` is empty with the `MUST_PASS_ALL` operator. To continue using the CDH 4 behavior, modify your code to use the `scan.setLoadColumnFamiliesOnDemand(false)` method.

Compatibility Notes for CDH 5.9

- The default RPC scheduler has been changed from 'deadline' to 'fifo'. To reenable 'deadline', set `hbase.ipc.server.callqueue.type` to `deadline` in the `hbase-site.xml` file.
- Apache HBase no longer includes XSS defense or encoding for filters. Due to licensing issues, HBase no longer includes a prior XSS defense nor an encoding for filters. Additionally, several dependencies have been removed. Downstream users relying on transitive inclusion of the following will need to directly rely on the appropriate dependency themselves: jsr305 (from the FindBugs project), Apache Commons Fileupload, nekohtml, beanshell core, Apache xml graphics, OWASP antisamy, OWASP esapi, Xalan, Apache Xerces, and Xom.

Compatibility Notes for CDH 5.7

- Cloudera recommends not using the new advanced configuration option `hbase.regionserver.hostname`, added in HBase 1.2 (CDH 5.7.0), which allows you to specify a separate external-facing hostname for a RegionServer.

Compatibility Notes for CDH 5.4

- The ports used by Apache HBase 1.0 changed from the 600XX range to the 160XX range. HBase in CDH reverted the change, and continues to use the 600XX port range, to maintain compatibility.
- If you used visibility labels prior to CDH 5.4 and assigned superuser privileges to HBase users by adding the `system` label to their set of labels, these users will no longer be superusers in CDH 5.4. To be sure that cached credentials are cleared, use the HBase Shell command `clear_auths <username>`, for each affected user. To grant users superuser privileges, add them to the **HBase Superusers** group in Cloudera Manager, or add them to the `hbase.superuser` property in `hbase-site.xml`, and restart the HMaster.
- HTrace is experimental in CDH 5.4.0. Artifacts and package names cannot be relied upon.
- Jersey was updated from 1.8 to 1.9. This has the following implications.
 - The Jersey version is now consistent with Apache HBase and other CDH components.
 - If your project relies upon `jersey-server`, you may need to make modifications.
- Curator in Hadoop was updated from 2.6.0 to 2.7.1. This has the following implications for HBase.
 - `PathUtils.validatePath(String)` changed return types, which will cause runtime errors for code compiled against the older version.
 - The `SharedCountReader` and `SharedValueReader` interfaces each added a method, which will cause compilation errors for code made to use the old version.
- `commons-codec` was upgraded from 1.7 to 1.9. This has the following implications for HBase.

CDH 5 Release Notes

- The class `org.apache.commons.codec.net.QuotedPrintableCodec` has a constructor that throws additional exceptions. See the [API reference](#) for details.
- `commons-logging` was updated from version 1.1.1. to 1.2. This has the following implications for HBase.
 - `org.apache.commons.logging.LogSource.setLogImplementation(String)` no longer throws `ExceptionInInitializerError`, which may change behavior of code that expects it.
- API changes: CDH reverted API changes in HBase 1.0 which broke compatibility with HBase in CDH 5.0, 5.1, 5.2, and 5.3. If you have written applications using Apache HBase 1.0 APIs, you may need to modify these applications to run in CDH 5.4.

Differences between CDH 5.4 HBase 1.0 and Apache HBase 1.0:

- CDH 5.4.0 keeps `commons-math` at version 2.1 to maintain compatibility with earlier CDH releases, whereas Apache HBase 1.0 uses `commons-math` 2.2.
- CDH 5.4.0 keeps Netty at version 3 to maintain compatibility with earlier CDH releases, whereas Apache HBase 1.0 uses Netty 4.

Compatibility Notes for CDH 5.3

- The `Put` class no longer implements `Writable`. Instead, you can change the definition to `comparable` if you have only Puts, or `comparable` if you have a mix of Puts, Gets, and Deletes.

Compatibility Notes for CDH 5.2

- In HBase in CDH 5.1, the default value for `hbase.security.access.early_out` was set to `false`. In CDH 5.2, the default value has been changed to `true`, to maintain consistency with the behavior in CDH 4. When set to `true`, if a user is not granted access to a column family qualifier, the AccessController immediately throws an `AccessDeniedException`. This change to the default behavior will affect users who enabled HFile version 3 and the AccessController coprocessor in CDH 5.1, and then upgrade to CDH 5.2. In this case, if you prefer `hbase.security.access.early_out` to be disabled, explicitly set it to `false` in `hbase-site.xml`.
- Starting with CDH 5.2, you can specify a global default number of versions, which will be applied to all newly created tables where the number of versions is not otherwise specified, by setting `hbase.column.max.version` to the desired number of versions in `hbase-site.xml`.
- HBase in CDH 5.2 differs from Apache HBase 0.98.6 in that CDH does not include [HBASE-11546](#), which provides ZooKeeper-less region assignment. CDH omits this feature because it is an incompatible change that prevents an upgraded cluster from being rolled back to a previous version.

Developer Interface Changes

- HBase 0.98.5 removed `ClientSmallScanner` from the public API. HBase in CDH 5.2 restores the constructor to maintain backward compatibility, but in future releases of HBase, this class will no longer be public. You should change your code to use the `Scan.setSmall(true)` method instead.

Compatibility Notes for CDH 5.1

General Notes

- [HBASE-8218](#) changes `AggregationClient` by replacing the `byte[] tablename` parameters with `HTable table`. This means that coprocessors compiled against CDH 5.0.x won't run or compile in CDH 5.1 and later.
- In CDH 5.1 and later, `delete*` methods of the `Delete` class of the HBase Client API use the timestamp from the constructor, the same behavior as the `Put` class. (In previous versions, the `delete*` methods ignored the constructor's timestamp, and used the value of `HConstants.LATEST_TIMESTAMP`. This behavior was different from the behavior of the `add()` methods of the `Put` class.) See [HBASE-10964](#).

- In CDH 5 prior to 5.1.3, a Put submitted with a `KeyValue`, `KeyValue.Type.Delete` does not delete the cell. This is different from the behavior in CDH 4. In CDH 5.1.3, this behavior is changed, so that a Put submitted with a `KeyValue`, `KeyValue.Type.Delete` does delete the cell. This fix is provided in [HBASE-11788](#).
- In CDH 5.1 and newer, HBase introduces a new snapshot format ([HBASE-7987](#)). A snapshot created in HBase 0.98 cannot be read by HBase 0.96. HBase 0.98 can read snapshots produced in previous versions of HBase, and no conversion is necessary.
- In CDH 5.1, the default value for `hbase.security.access.early_out` was changed from `true` to `false`. A setting of `true` means that if a user is not granted access to a column family qualifier, the AccessController immediately throws an `AccessDeniedException`. *This behavior change was reverted for CDH 5.2.*

Developer Interface Changes

- `HTablePool` is no longer supported in CDH 5.1 and later. The `HConnection` object is the replacement. You create the connection once and pass it around, as with the old table pool.

```
HConnection connection = HConnectionManager.createConnection(config);
HTableInterface table = connection.getTable(tableName);
table.put(put);
table.close();
connection.close();
```

You can set the `hbase.hconnection.threads.max` property in `hbase-site.xml` to control the pool size or you can pass an `ExecutorService` to `HConnectionManager.createConnection()`.

```
ExecutorService pool = ...;
HConnection connection = HConnectionManager.createConnection(conf, pool);
```

Compatibility Notes for CDH 5 Beta Releases



Warning:

CDH 5 Beta 1 and Beta 2 are not intended for production use, and have been superseded by official releases in the CDH 5 family.

The HBase client from CDH 5 Beta 1 is not wire compatible with CDH 5 Beta 2 because of changes introduced in [HBASE-9612](#). As a consequence, CDH 5 Beta 1 users will not be able to execute a rolling upgrade to CDH 5 Beta 2 (or later). This patch unifies the way the HBase clients make requests and simplifies the internals, but breaks wire compatibility. Developers may need to recompile applications built upon the CDH 5 Beta 1 API.

As of CDH 5 Beta 1 (HBase 0.95), the value of `hbase.regionserver.checksum.verify` defaults to `true`; in earlier releases the default is `false`.

API Removals

- See [API Differences between CDH 4.5 and CDH 5 Beta 2](#).

Compatibility between CDH Beta and Apache HBase Releases

- Apache HBase 0.95.2 is not wire compatible with CDH 5 Beta 1 HBase 0.95.2.
- Apache HBase 0.96.x should be wire compatible with CDH 5 Beta 2 HBase 0.96.1.1.

Apache Hive Incompatible Changes and Limitations



Note: As of CDH 5, HCatalog is part of Apache Hive; incompatible changes in HCatalog are included below.

Metastore schema upgrade: CDH 5.2.0 includes Hive version 0.13.1. Upgrading from an earlier Hive version to Hive 0.13.1 or later requires a metastore schema upgrade.



Warning:

You must upgrade the metastore schema before starting the new version of Hive. Failure to do so may result in metastore corruption.

CDH 5 includes a new offline tool called `schematool`; Cloudera recommends you use this tool to upgrade your metastore schema.

Hive upgrade: Upgrading Hive from CDH 4 to CDH 5, or from an earlier CDH 5.x release to CDH 5.2 or later, requires several manual steps. Follow the upgrade guide closely.

Incompatible changes between CDH 4 and CDH 5:

- The CDH 4 JDBC client is not compatible with CDH 5 HiveServer2. JDBC applications connecting to the CDH 5 HiveServer2 will require the CDH 5 JDBC client driver.
- JDBC applications will require the newer CDH 5 JDBC packages in order to connect to HiveServer2. You do not need to recompile applications for this change.
- Because of security and concurrency issues, the original Hive server (HiveServer1) and the Hive command-line interface (CLI) are deprecated in current versions of CDH 5 and will be removed in a future release. Cloudera strongly encourages you to migrate to HiveServer2 and Beeline.
- CDH 5 Hue will not work with HiveServer2 from CDH 4.
- The `npath` function has been removed.
- Cloudera recommends that custom ObjectInspectors created for use with custom SerDes have a no-argument constructor in addition to their normal constructors, for serialization purposes. See [HIVE-5380](#) for more details.
- The SerDe interface has been changed which requires the custom SerDe modules to be reworked.
- The decimal data type format has changed as of CDH 5 Beta 2 and is not compatible with CDH 4.
- From CDH 5 Beta 2 onwards, the Parquet SerDe is part of the Hive package. The SerDe class name has changed as a result. However, there is a wrapper class for backward compatibility, so any existing Hive tables created with the Parquet SerDe will continue to work with CDH 5 Beta 2 and later Hive versions.

Incompatible changes between any earlier CDH version and CDH 5.4.x:

- CDH 5.2.0 and later clients cannot communicate with CDH 5.1.x and earlier servers. This means that you must upgrade the server before the clients.
- As of CDH 5.2.0, `DESCRIBE DATABASE` returns additional fields: `owner_name` and `owner_type`. The command will continue to behave as expected if you identify the field you're interested in by its (string) name, but could produce unexpected results if you use a numeric index to identify the field(s).
- CDH 5.2.0 implements [HIVE-6248](#), which includes some backward-incompatible changes to the HCatalog API.
- The CDH 5.2 Hive JDBC driver is not wire-compatible with the CDH 5.1 version of HiveServer2. Make sure you upgrade Hive clients and all other Hive hosts in tandem: the server first, and then the clients.
- HiveServer 1 is deprecated as of CDH 5.3, and will be removed in a future release of CDH. Users of HiveServer 1 should upgrade to HiveServer 2 as soon as possible.
- `org.apache.hcatalog` is deprecated as of CDH 5.3. All client-facing classes were moved from `org.apache.hcatalog` to `org.apache.hive.hcatalog` as of CDH 5.0 and the deprecated classes in `org.apache.hcatalog` will be removed altogether in a future release. If you are still using `org.apache.hcatalog`, you should move to `org.apache.hive.hcatalog` immediately.
- **Date partition columns:** as of Hive version 13, implemented in CDH 5.2, Hive validates the format of dates in partition columns, if they are stored as dates. A partition column with a date in invalid form can neither be used nor dropped once you upgrade to CDH 5.2 or higher. To avoid this problem, do one of the following:
 - Fix any invalid dates before you upgrade. Hive expects dates in partition columns to be in the form YYYY-MM-DD.
 - Store dates in partition columns as strings or integers.

You can use the following SQL query to find any partition-column values stored as dates:

```
SELECT "DBS"."NAME", "TBLS"."TBL_NAME", "PARTITION_KEY_VALS"."PART_KEY_VAL"
FROM "PARTITION_KEY_VALS"
INNER JOIN "PARTITIONS" ON "PARTITION_KEY_VALS"."PART_ID" = "PARTITIONS"."PART_ID"
INNER JOIN "PARTITION_KEYS" ON "PARTITION_KEYS"."TBL_ID" = "PARTITIONS"."TBL_ID"
INNER JOIN "TBLS" ON "TBLS"."TBL_ID" = "PARTITIONS"."TBL_ID"
INNER JOIN "DBS" ON "DBS"."DB_ID" = "TBLS"."DB_ID"
AND "PARTITION_KEYS"."INTEGER_IDX" = "PARTITION_KEY_VALS"."INTEGER_IDX"
AND "PARTITION_KEYS"."PKEY_TYPE" = 'date';
```

- **Decimal precision and scale:** As of CDH 5.4, Hive support for decimal precision and scale changes as follows:

1. When `decimal` is used as a type, it means `decimal(10, 0)` rather than a precision of 38 with a variable scale.
2. When Hive is unable to determine the precision and scale of a decimal type (for example in the case of non-generic User-Defined Function (UDF) that has an `evaluate()` method that returns `decimal`), a precision and scale of (38, 18) is assumed. In previous versions, a precision of 38 and a variable scale were assumed. Cloudera recommends you develop generic UDFs instead, and specify exact precision and scale.
3. When a decimal value is assigned or cast to a different decimal type, rounding is used to handle cases in which the precision of the value is greater than that of the target decimal type, as long as the integer portion of the value can be preserved. In previous versions, if the value's precision was greater than 38 (the only allowed precision for the `decimal` type), the value was set to null, regardless of whether the integer portion could be preserved.

- **Deprecation of HivePassThrough serde formats:** As of CDH 5.4, [HIVE-8910](#) changes how the storage handler uses the `HivePassThroughOutputFormat` class. It removes the empty default constructor, which breaks `org.apache.hadoop.util.ReflectionUtils.newInstance` and throws a `NoSuchMethodException`. The workaround is to re-create the Hive tables without HivePassThrough serde formats.

Hue Incompatible Changes and Limitations

- You will need to upgrade any custom applications after you upgrade to CDH 5.4.0.
- In [HUE-1859](#), the LDAP synchronization backend was moved to a generic middleware. If your code uses `DesktopSynchronizationBackendBase`, you will need to create your own middleware, and extend the new `LdapSynchronizationMiddleware`. Put that new custom middleware class in the `middleware=` line of the [desktop] section of `hue.ini`. The following example uses a middleware called `desktop.auth.backend.my_middleware`.

```
[desktop]
...
# Comma-separated list of Django middleware classes to use.
# See https://docs.djangoproject.com/en/1.4/ref/middleware/ for more details on
middlewares in Django.
middleware=desktop.auth.backend.LdapSynchronizationBackend,desktop.auth.backend.my_middleware
...
```

- [HUE-1658 \[oozie\]](#) Hue depends on [OOZIE-1306](#) which is in CDH 5 Beta 2 but has not been included in any other release yet. Set the following backward compatibility flag to false to use the old frequency number/unit representation instead of the new crontab.

```
enable_cron_scheduling = false
```

- Hue 3.0.0 was a major revision of Hue. The user interface changed significantly.
- CDH 5 Hue works only with the default system Python version of the operating system it is being installed on. For example, on RHEL/CentOS 6, you need Python 2.6 to start Hue.



Note: RHEL 5 and CentOS 5 users will have to download Python 2.6 from the EPEL repository.

- The Beeswax daemon has been replaced by HiveServer2. Hue should therefore point to a running HiveServer2. This change involves removing the Beeswaxd code entirely and the following major updates to the [beeswax] section of the Hue configuration file, hue.ini.

```
[beeswax]
# Host where Hive server Thrift daemon is running.
# If Kerberos security is enabled, use fully-qualified domain name (FQDN).
## hive_server_host=<FQDN of Hive Server>

# Port where HiveServer2 Thrift server runs on.
## hive_server_port=10000
```

- Search bind authentication is now used by default instead of direct bind. To revert to the previous settings, use the new search_bind_authentication configuration property.

```
[desktop]
[[ldap]]
search_bind_authentication=false
```

- The Hue Shell app has been removed completely. This includes removing both the Shell app code and the [shell] section from hue.ini.
- YARN should be used by default.

Apache Impala (incubating) Incompatible Changes and Limitations

The Impala version covered by this documentation library contains the following incompatible changes. These are things such as file format changes, removed features, or changes to implementation, default configuration, dependencies, or prerequisites that could cause issues during or after an Impala upgrade.

Even added SQL statements or clauses can produce incompatibilities, if you have databases, tables, or columns whose names conflict with the new keywords.

Incompatible Changes Introduced in Impala for CDH 5.9.x / Impala 2.7.x

- Bug fixes related to parsing of floating-point values (IMPALA-1731 and IMPALA-3868) can change the results of casting strings that represent invalid floating-point values. For example, formerly a string value beginning or ending with inf, such as 1.23inf or infinite, now are converted to NULL when interpreted as a floating-point value. Formerly, they were interpreted as the special “infinity” value when converting from string to floating-point. Similarly, now only the string NaN (case-sensitive) is interpreted as the special “not a number” value. String values containing multiple dots, such as 3..141 or 3.1.4.1, are now interpreted as NULL rather than being converted to valid floating-point values.
- The column types shown in the DESCRIBE FORMATTED output are in uppercase, where formerly they were in lowercase. This is not an intended change and could be reverted in the future, when IMPALA-4372 is resolved.

Incompatible Changes Introduced in Impala for CDH 5.8.x / Impala 2.6.x

- The default for the RUNTIME_FILTER_MODE query option is changed to GLOBAL (the highest setting).
- The RUNTIME_BLOOM_FILTER_SIZE setting is now only used as a fallback if statistics are not available; otherwise, Impala uses the statistics to estimate the appropriate size to use for each filter.
- Admission control and dynamic resource pools are enabled by default. When upgrading from an earlier release, you must turn on these settings yourself if they are not already enabled. See [Admission Control and Query Queuing](#) for details about admission control.
- Impala reserves some new keywords, in preparation for support for Kudu syntax: buckets, delete, distribute, hash, ignore, split, and update.
- For Kerberized clusters, the Catalog service now uses the Kerberos principal instead of the operating system user that runs the catalogd daemon. This eliminates the requirement to configure a

`hadoop.user.group.static.mapping.overrides` setting to put the OS user into the Sentry administrative group, on clusters where the principal and the OS user name for this user are different.

- The mechanism for interpreting `DECIMAL` literals is improved, no longer going through an intermediate conversion step to `DOUBLE`:
 - Casting a `DECIMAL` value to `TIMESTAMP DOUBLE` produces a more precise value for the `TIMESTAMP` than formerly.
 - Certain function calls involving `DECIMAL` literals now succeed, when formerly they failed due to lack of a function signature with a `DOUBLE` argument.
- Improved type accuracy for `CASE` return values. If all `WHEN` clauses of the `CASE` expression are of `CHAR` type, the final result is also `CHAR` instead of being converted to `STRING`.
- The initial release of CDH 5.7 / Impala 2.5 sometimes has a higher peak memory usage than in previous releases while reading Parquet files. The following query options might help to reduce memory consumption in the Parquet scanner:
 - Reduce the number of scanner threads, for example: `set num_scanner_threads=30`
 - Reduce the batch size, for example: `set batch_size=512`
 - Increase the memory limit, for example: `set mem_limit=64g`

You can track the status of the fix for this issue at [IMPALA-3662](#).

- The `S3_SKIP_INSERT_STAGING` query option, which is enabled by default, increases the speed of `INSERT` operations for S3 tables. The speedup applies to regular `INSERT`, but not `INSERT OVERWRITE`. The tradeoff is the possibility of inconsistent output files left behind if a node fails during `INSERT` execution. See [S3_SKIP_INSERT_STAGING Query Option \(or higher only\)](#) for details.

Certain features are turned off by default, to avoid regressions or unexpected behavior following an upgrade. Consider turning on these features after suitable testing:

- Impala now recognizes the `auth_to_local` setting, specified through the HDFS configuration setting `hadoop.security.auth_to_local`. This feature is disabled by default; to enable it, specify `--load_auth_to_local_rules=true` in the `impalad` configuration settings.
- A new query option, `PARQUET_ANNOTATE_STRINGS_UTF8`, makes Impala include the UTF-8 annotation metadata for `STRING`, `CHAR`, and `VARCHAR` columns in Parquet files created by `INSERT` or `CREATE TABLE AS SELECT` statements.
- A new query option, `PARQUET_FALLBACK_SCHEMA_RESOLUTION`, lets Impala locate columns within Parquet files based on column name rather than ordinal position. This enhancement improves interoperability with applications that write Parquet files with a different order or subset of columns than are used in the Impala table.

Incompatible Changes Introduced in Impala for CDH 5.7.x / Impala 2.5.x

- The admission control default limit for concurrent queries (the `max requests` setting) is now unlimited instead of 200.
- Multiplying a mixture of `DECIMAL` and `FLOAT` or `DOUBLE` values now returns `DOUBLE` rather than `DECIMAL`. This change avoids some cases where an intermediate value would underflow or overflow and become `NULL` unexpectedly. The results of multiplying `DECIMAL` and `FLOAT` or `DOUBLE` might now be slightly less precise than before. Previously, the intermediate types and thus the final result depended on the exact order of the values of different types being multiplied, which made the final result values difficult to reason about.
- Previously, the `_` and `%` wildcard characters for the `LIKE` operator would not match characters on the second or subsequent lines of multi-line string values. The fix for issue [IMPALA-2204](#) causes the wildcard matching to apply to the entire string for values containing embedded `\n` characters. This could cause different results than in previous Impala releases for identical queries on identical data.

- Formerly, all Impala UDFs and UDAFs required running the `CREATE FUNCTION` statements to re-create them after each `catalogd` restart. In 5.6 and higher, functions written in C++ are persisted across restarts, and the requirement to re-create functions only applies to functions written in Java. Adapt any function-reloading logic that you have added to your Impala environment.
- `CREATE TABLE LIKE` no longer inherits HDFS caching settings from the source table.
- The `SHOW DATABASES` statement now returns two columns rather than one. The second column includes the associated comment string, if any, for each database. Adjust any application code that examines the list of databases and assumes the result set contains only a single column.
- The output of the `SHOW FUNCTIONS` statement includes two new columns, showing the kind of the function (for example, `BUILTIN`) and whether or not the function persists across catalog server restarts. For example, the `SHOW FUNCTIONS` output for the `_impala_builtins` database starts with:

return type	signature	binary type	is persistent
<code>BIGINT</code>	<code>abs(BIGINT)</code>	<code>BUILTIN</code>	<code>true</code>
<code>DECIMAL(*,*)</code>	<code>abs(DECIMAL(*, *))</code>	<code>BUILTIN</code>	<code>true</code>
<code>DOUBLE</code>	<code>abs(DOUBLE)</code>	<code>BUILTIN</code>	<code>true</code>
...			

Incompatible Changes Introduced in Impala for CDH 5.6.x / Impala 2.4.x

Other than support for DSSD storage, the Impala feature set for CDH 5.6 is the same as for CDH 5.5. Therefore, there are no incompatible changes for Impala introduced in CDH 5.6.

Incompatible Changes Introduced in Impala for CDH 5.5.x / Impala 2.3.x



Note:

The use of the Llama component for integrated resource management within YARN is no longer supported with 5.5 and higher.

For clusters running Impala alongside other data management components, you define static service pools to define the resources available to Impala and other components. Then within the area allocated for Impala, you can create dynamic service pools, each with its own settings for the Impala admission control feature.

- If Impala encounters a Parquet file that is invalid because of an incorrect magic number, the query skips the file. This change is caused by the fix for issue [IMPALA-2130](#). Previously, Impala would attempt to read the file despite the possibility that the file was corrupted.
- Previously, calls to overloaded built-in functions could treat parameters as `DOUBLE` or `FLOAT` when no overload had a signature that matched the exact argument types. Now Impala prefers the function signature with `DECIMAL` parameters in this case. This change avoids a possible loss of precision in function calls such as `greatest(0, 99999.8888)`; now both parameters are treated as `DECIMAL` rather than `DOUBLE`, avoiding any loss of precision in the fractional value. This could cause slightly different results than in previous Impala releases for certain function calls.
- Formerly, adding or subtracting a large interval value to a `TIMESTAMP` could produce a nonsensical result. Now when the result goes outside the range of `TIMESTAMP` values, Impala returns `NULL`.

- Formerly, it was possible to accidentally create a table with identical row and column delimiters. This could happen unintentionally, when specifying one of the delimiters and using the default value for the other. Now an attempt to use identical delimiters still succeeds, but displays a warning message.
- Formerly, Impala could include snippets of table data in log files by default, for example when reporting conversion errors for data values. Now any such log messages are only produced at higher logging levels that you would enable only during debugging.

Incompatible Changes Introduced in Impala for CDH 5.4.x



Note: Impala 2.2.0 is available as part of CDH 5.4.0 and is not available for CDH 4. Cloudera does not intend to release future versions of Impala for CDH 4 outside patch and maintenance releases if required. Given the end-of-maintenance status for CDH 4, Cloudera recommends all customers to migrate to a recent CDH 5 release.

Changes to File Handling

Impala queries ignore files with extensions commonly used for temporary work files by Hadoop tools. Any files with extensions .tmp or .copying are not considered part of the Impala table. The suffix matching is case-insensitive, so for example Impala ignores both .copying and .COPYING suffixes.

The log rotation feature in Impala 2.2.0 and higher means that older log files are now removed by default. The default is to preserve the latest 10 log files for each severity level, for each Impala-related daemon. If you have set up your own log rotation processes that expect older files to be present, either adjust your procedures or change the Impala -max_log_files setting.

Changes to Prerequisites

The prerequisite for CPU architecture has been relaxed in Impala 2.2.0 and higher. From this release onward, Impala works on CPUs that have the SSSE3 instruction set. The SSE4 instruction set is no longer required. This relaxed requirement simplifies the upgrade planning from Impala 1.x releases, which also worked on SSSE3-enabled processors.

Incompatible Changes Introduced in Impala for CDH 5.3.x

Changes to Prerequisites

Currently, Impala 2.1.x does not function on CPUs without the SSE4.1 instruction set. This minimum CPU requirement is higher than in previous versions, which relied on the older SSSE3 instruction set. Check the CPU level of the hosts in your cluster before upgrading to Impala 2.1.x or CDH 5.3.x.

Changes to Output Format

The “small query” optimization feature introduces some new information in the EXPLAIN plan, which you might need to account for if you parse the text of the plan output.

New Reserved Words

New SQL syntax introduces additional reserved words: FOR, GRANT, REVOKE, ROLE, ROLES, INCREMENTAL.

Incompatible Changes Introduced in Impala 2.0.5 / CDH 5.2.6

No incompatible changes.



Note: Impala 2.0.5 is available as part of CDH 5.2.6, not under CDH 4.

CDH 5 Release Notes

Incompatible Changes Introduced in Impala 2.0.4 / CDH 5.2.5

No incompatible changes.



Note: Impala 2.0.4 is available as part of CDH 5.2.5, not under CDH 4.

Incompatible Changes Introduced in Impala 2.0.3 / CDH 5.2.4



Note: Impala 2.0.3 is available as part of CDH 5.2.4, not under CDH 4.

Incompatible Changes Introduced in Impala 2.0.2 / CDH 5.2.3

No incompatible changes.



Note: Impala 2.0.2 is available as part of CDH 5.2.3, not under CDH 4.

Incompatible Changes Introduced in Impala 2.0.1 / CDH 5.2.1

- The `INSERT` statement has always left behind a hidden work directory inside the data directory of the table. Formerly, this hidden work directory was named `.impala_insert_staging`. In Impala 2.0.1 and later, this directory name is changed to `_impala_insert_staging`. (While HDFS tools are expected to treat names beginning either with underscore and dot as hidden, in practice names beginning with an underscore are more widely supported.) If you have any scripts, cleanup jobs, and so on that rely on the name of this work directory, adjust them to use the new name.
- The `abs()` function now takes a broader range of numeric types as arguments, and the return type is the same as the argument type.
- Shorthand notation for character classes in regular expressions, such as `\d` for digit, are now available again in regular expression operators and functions such as `regexp_extract()` and `regexp_replace()`. Some other differences in regular expression behavior remain between Impala 1.x and Impala 2.x releases. See [Incompatible Changes Introduced in Impala 2.0.0 / CDH 5.2.0](#) on page 98 for details.

Incompatible Changes Introduced in Impala 2.0.0 / CDH 5.2.0

Changes to Prerequisites

Currently, Impala 2.0.x does not function on CPUs without the SSE4.1 instruction set. This minimum CPU requirement is higher than in previous versions, which relied on the older SSSE3 instruction set. Check the CPU level of the hosts in your cluster before upgrading to Impala 2.0.x or CDH 5.2.x.

Changes to Query Syntax

The new syntax where query hints are allowed in comments causes some changes in the way comments are parsed in the `impala-shell` interpreter. Previously, you could end a `--` comment line with a semicolon and `impala-shell` would treat that as a no-op statement. Now, a comment line ending with a semicolon is passed as an empty statement to the Impala daemon, where it is flagged as an error.

Impala 2.0 and later uses a different support library for regular expression parsing than in earlier Impala versions. Now, Impala uses the [Google RE2 library](#) rather than Boost for evaluating regular expressions. This implementation change causes some differences in the allowed regular expression syntax, and in the way certain regex operators are interpreted. The following are some of the major differences (not necessarily a complete list):

- `.*?` notation for non-greedy matches is now supported, where it was not in earlier Impala releases.

- By default, ^ and \$ now match only begin/end of buffer, not begin/end of each line. This behavior can be overridden in the regex itself using the `m` flag.
- By default, . does not match newline. This behavior can be overridden in the regex itself using the `s` flag.
- \Z is not supported.
- < and > for start of word and end of word are not supported.
- Lookahead and lookbehind are not supported.
- Shorthand notation for character classes, such as \d for digit, is not recognized. (This restriction is lifted in Impala 2.0.1, which restores the shorthand notation.)

Changes to Output Format

In Impala 2.0 and later, `user()` returns the full Kerberos principal string, such as `user@example.com`, in a Kerberized environment.

The changed format for the user name in secure environments is also reflected where the user name is displayed in the output of the `PROFILE` command.

In the output from `SHOW FUNCTIONS`, `SHOW AGGREGATE FUNCTIONS`, and `SHOW ANALYTIC FUNCTIONS`, arguments and return types of arbitrary `DECIMAL` scale and precision are represented as `DECIMAL(*,*)`. Formerly, these items were displayed as `DECIMAL(-1,-1)`.

Changes to Query Options

The `PARQUET_COMPRESSION_CODEC` query option has been replaced by the `COMPRESSION_CODEC` query option.

Changes to Configuration Options

The meaning of the `--idle_query_timeout` configuration option is changed, to accommodate the new `QUERY_TIMEOUT_S` query option. Rather than setting an absolute timeout period that applies to all queries, it now sets a maximum timeout period, which can be adjusted downward for individual queries by specifying a value for the `QUERY_TIMEOUT_S` query option. In sessions where no `QUERY_TIMEOUT_S` query option is specified, the `--idle_query_timeout` timeout period applies the same as in earlier versions.

The `--strict_unicode` option of `impala-shell` was removed. To avoid problems with Unicode values in `impala-shell`, define the following locale setting before running `impala-shell`:

```
export LC_CTYPE=en_US.UTF-8
```

New Reserved Words

Some new SQL syntax requires the addition of new reserved words: `ANTI`, `ANALYTIC`, `OVER`, `PRECEDING`, `UNBOUNDED`, `FOLLOWING`, `CURRENT`, `ROWS`, `RANGE`, `CHAR`, `VARCHAR`.

Changes to Data Files

The default Parquet block size for Impala is changed from 1 GB to 256 MB. This change could have implications for the sizes of Parquet files produced by `INSERT` and `CREATE TABLE AS SELECT` statements.

Although older Impala releases typically produced files that were smaller than the old default size of 1 GB, now the file size matches more closely whatever value is specified for the `PARQUET_FILE_SIZE` query option. Thus, if you use a non-default value for this setting, the output files could be larger than before. They still might be somewhat smaller than the specified value, because Impala makes conservative estimates about the space needed to represent each column as it encodes the data.

CDH 5 Release Notes

When you do not specify an explicit value for the `PARQUET_FILE_SIZE` query option, Impala tries to keep the file size within the 256 MB default size, but Impala might adjust the file size to be somewhat larger if needed to accommodate the layout for **wide** tables, that is, tables with hundreds or thousands of columns.

This change is unlikely to affect memory usage while writing Parquet files, because Impala does not pre-allocate the memory needed to hold the entire Parquet block.

Incompatible Changes Introduced in Impala 1.4.4 / CDH 5.1.5

No incompatible changes.



Note: Impala 1.4.4 is available as part of CDH 5.1.5, not under CDH 4.

Incompatible Changes Introduced in Impala 1.4.3 / CDH 5.1.4

No incompatible changes. The TLS/SSL security fix does not require any change in the way you interact with Impala.



Note: Impala 1.4.3 is available as part of CDH 5.1.4, and under CDH 4.

Incompatible Changes Introduced in Impala 1.4.2 / CDH 5.1.3

None. Impala 1.4.2 is purely a bug-fix release. It does not include any incompatible changes.



Note: Impala 1.4.2 is only available as part of CDH 5.1.3, not under CDH 4.

Incompatible Changes Introduced in Impala 1.4.1 / CDH 5.1.2

None. Impala 1.4.1 is purely a bug-fix release. It does not include any incompatible changes.

Incompatible Changes Introduced in Impala 1.4.0 / CDH 5.1.0

- There is a slight change to required security privileges in the Sentry framework. To create a new object, now you need the `ALL` privilege on the parent object. For example, to create a new table, view, or function requires having the `ALL` privilege on the database containing the new object.
- With the ability of `ORDER BY` queries to process unlimited amounts of data with no `LIMIT` clause, the query options `DEFAULT_ORDER_BY_LIMIT` and `ABORT_ON_DEFAULT_LIMIT_EXCEEDED` are now deprecated and have no effect.
- There are some changes to the list of reserved words. The following keywords are new:

- `API_VERSION`
- `BINARY`
- `CACHED`
- `CLASS`
- `PARTITIONS`
- `PRODUCED`
- `UNCACHED`

The following were formerly reserved keywords, but are no longer reserved:

- `COUNT`
- `GROUP_CONCAT`
- `NDV`
- `SUM`

- The fix for issue [IMPALA-973](#) changes the behavior of the `INVALIDATE METADATA` statement regarding nonexistent tables. In Impala 1.4.0 and higher, the statement returns an error if the specified table is not in the metastore database at all. It completes successfully if the specified table is in the metastore database but not yet recognized by Impala, for example if the table was created through Hive. Formerly, you could issue this statement for a completely nonexistent table, with no error.

Incompatible Changes Introduced in Impala 1.3.3 / CDH 5.0.5

No incompatible changes. The TLS/SSL security fix does not require any change in the way you interact with Impala.



Note: Impala 1.3.3 is only available as part of CDH 5.0.5, not under CDH 4.

Incompatible Changes Introduced in Impala 1.3.2 / CDH 5.0.4

With the fix for IMPALA-1019, you can use HDFS caching for files that are accessed by Impala.



Note: Impala 1.3.2 is only available as part of CDH 5.0.4, not under CDH 4.

Incompatible Changes Introduced in Impala 1.3.1 / CDH 5.0.3

- In Impala 1.3.1 and higher, the `REGEXP` and `RLIKE` operators now match a regular expression string that occurs anywhere inside the target string, the same as if the regular expression was enclosed on each side by `.*`. See [REGEXP Operator](#) for examples. Previously, these operators only succeeded when the regular expression matched the entire target string. This change improves compatibility with the regular expression support for popular database systems. There is no change to the behavior of the `regexp_extract()` and `regexp_replace()` built-in functions.
- The result set for the `SHOW FUNCTIONS` statement includes a new first column, with the data type of the return value.

Incompatible Changes Introduced in Impala 1.3.0 / CDH 5.0.0

- The `EXPLAIN_LEVEL` query option now accepts numeric options from 0 (most concise) to 3 (most verbose), rather than only 0 or 1. If you formerly used `SET EXPLAIN_LEVEL=1` to get detailed explain plans, switch to `SET EXPLAIN_LEVEL=3`. If you used the mnemonic keyword (`SET EXPLAIN_LEVEL=verbose`), you do not need to change your code because now level 3 corresponds to verbose.
- The keyword `DECIMAL` is now a reserved word. If you have any databases, tables, columns, or other objects already named `DECIMAL`, quote any references to them using backticks (`` ``) to avoid name conflicts with the keyword.



Note: Although the `DECIMAL` keyword is a reserved word, currently Impala does not support `DECIMAL` as a data type for columns.

- The query option named `YARN_POOL` during the CDH 5 beta period is now named `REQUEST_POOL` to reflect its broader use with the Impala admission control feature.
- There are some changes to the list of reserved words.
 - The names of aggregate functions are no longer reserved words, so you can have databases, tables, columns, or other objects named `AVG`, `MIN`, and so on without any name conflicts.
 - The internal function names `DISTINCTPC` and `DISTINCTPCSA` are no longer reserved words, although `DISTINCT` is still a reserved word.
 - The keywords `CLOSE_FN` and `PREPARE_FN` are now reserved words.

- The HDFS property `dfs.client.file-block-storage-locations.timeout` was renamed to `dfs.client.file-block-storage-locations.timeout.millis`, to emphasize that the unit of measure is milliseconds, not seconds. Impala requires a timeout of at least 10 seconds, making the minimum value for this setting 10000. On systems not managed by Cloudera Manager, you might need to edit the `hdfs-site.xml` file in the Impala configuration directory for the new name and minimum value.

Incompatible Changes Introduced in Impala 1.2.4

There are no incompatible changes introduced in Impala 1.2.4.

Previously, after creating a table in Hive, you had to issue the `INVALIDATE METADATA` statement with no table name, a potentially expensive operation on clusters with many databases, tables, and partitions. Starting in Impala 1.2.4, you can issue the statement `INVALIDATE METADATA table_name` for a table newly created through Hive. Loading the metadata for only this one table is faster and involves less network overhead. Therefore, you might revisit your setup DDL scripts to add the table name to `INVALIDATE METADATA` statements, in cases where you create and populate the tables through Hive before querying them through Impala.

Incompatible Changes Introduced in Impala 1.2.3

Because the feature set of Impala 1.2.3 is identical to Impala 1.2.2, there are no new incompatible changes. See [Incompatible Changes Introduced in Impala 1.2.2](#) on page 102 if you are upgrading from Impala 1.2.1 or 1.1.x.

Incompatible Changes Introduced in Impala 1.2.2

The following changes to SQL syntax and semantics in Impala 1.2.2 could require updates to your SQL code, or schema objects such as tables or views:

- With the addition of the `CROSS JOIN` keyword, you might need to rewrite any queries that refer to a table named `CROSS` or use the name `CROSS` as a table alias:

```
-- Formerly, 'cross' in this query was an alias for t1  
-- and it was a normal join query.  
-- In 1.2.2 and higher, CROSS JOIN is a keyword, so 'cross'  
-- is not interpreted as a table alias, and the query  
-- uses the special CROSS JOIN processing rather than a  
-- regular join.  
select * from t1 cross join t2...  
  
-- Now if CROSS is used in other context such as a table or column name,  
-- use backticks to escape it.  
create table `cross` (x int);  
select * from `cross`;
```

- Formerly, a `DROP DATABASE` statement in Impala would not remove the top-level HDFS directory for that database. The `DROP DATABASE` has been enhanced to remove that directory. (You still need to drop all the tables inside the database first; this change only applies to the top-level directory for the entire database.)
- The keyword `PARQUET` is introduced as a synonym for `PARQUETFILE` in the `CREATE TABLE` and `ALTER TABLE` statements, because that is the common name for the file format. (As opposed to `SequenceFile` and `RCFile` where the “File” suffix is part of the name.) Documentation examples have been changed to prefer the new shorter keyword. The `PARQUETFILE` keyword is still available for backward compatibility with older Impala versions.
- New overloads are available for several operators and built-in functions, allowing you to insert their result values into smaller numeric columns such as `INT`, `SIMALLINT`, `TINYINT`, and `FLOAT` without using a `CAST()` call. If you remove the `CAST()` calls from `INSERT` statements, those statements might not work with earlier versions of Impala.

Because many users are likely to upgrade straight from Impala 1.x to Impala 1.2.2, also read [Incompatible Changes Introduced in Impala 1.2.1](#) on page 103 for things to note about upgrading to Impala 1.2.x in general.

In a Cloudera Manager environment, the catalog service is not recognized or managed by Cloudera Manager versions prior to 4.8. Cloudera Manager 4.8 and higher require the catalog service to be present for Impala. Therefore, if you upgrade to Cloudera Manager 4.8 or higher, you must also upgrade Impala to 1.2.1 or higher. Likewise, if you upgrade Impala to 1.2.1 or higher, you must also upgrade Cloudera Manager to 4.8 or higher.

Incompatible Changes Introduced in Impala 1.2.1

The following changes to SQL syntax and semantics in Impala 1.2.1 could require updates to your SQL code, or schema objects such as tables or views:

- In Impala 1.2.1 and higher, all `NULL` values come at the end of the result set for `ORDER BY ... ASC` queries, and at the beginning of the result set for `ORDER BY ... DESC` queries. In effect, `NULL` is considered greater than all other values for sorting purposes. The original Impala behavior always put `NULL` values at the end, even for `ORDER BY ... DESC` queries. The new behavior in Impala 1.2.1 makes Impala more compatible with other popular database systems. In Impala 1.2.1 and higher, you can override or specify the sorting behavior for `NULL` by adding the clause `NULLS FIRST` or `NULLS LAST` at the end of the `ORDER BY` clause.

Impala 1.2.1 goes along with CDH 4.5 and Cloudera Manager 4.8. If you used the beta version Impala 1.2.0 that came with the beta of CDH 5, Impala 1.2.1 includes all the features of Impala 1.2.0 except for resource management, which relies on the YARN framework from CDH 5.

The new `catalogd` service might require changes to any user-written scripts that stop, start, or restart Impala services, install or upgrade Impala packages, or issue `REFRESH` or `INVALIDATE_METADATA` statements:

- See [Impala Installation](#), [Upgrading Impala](#) and [Starting Impala](#), for usage information for the `catalogd` daemon.
- The `REFRESH` and `INVALIDATE_METADATA` statements are no longer needed when the `CREATE TABLE`, `INSERT`, or other table-changing or data-changing operation is performed through Impala. These statements are still needed if such operations are done through Hive or by manipulating data files directly in HDFS, but in those cases the statements only need to be issued on one Impala node rather than on all nodes. See [REFRESH Statement](#) and [INVALIDATE_METADATA Statement](#) for the latest usage information for those statements.
- See [The Impala Catalog Service](#) for background information on the `catalogd` service.

In a Cloudera Manager environment, the catalog service is not recognized or managed by Cloudera Manager versions prior to 4.8. Cloudera Manager 4.8 and higher require the catalog service to be present for Impala. Therefore, if you upgrade to Cloudera Manager 4.8 or higher, you must also upgrade Impala to 1.2.1 or higher. Likewise, if you upgrade Impala to 1.2.1 or higher, you must also upgrade Cloudera Manager to 4.8 or higher.

Incompatible Changes Introduced in Impala 1.2.0 (Beta)

There are no incompatible changes to SQL syntax in Impala 1.2.0 (beta).

Because Impala 1.2.0 is bundled with the CDH 5 beta download and depends on specific levels of Apache Hadoop components supplied with CDH 5, you can only install it in combination with the CDH 5 beta.

The new `catalogd` service might require changes to any user-written scripts that stop, start, or restart Impala services, install or upgrade Impala packages, or issue `REFRESH` or `INVALIDATE_METADATA` statements:

- See [Impala Installation](#), [Upgrading Impala](#) and [Starting Impala](#), for usage information for the `catalogd` daemon.
- The `REFRESH` and `INVALIDATE_METADATA` statements are no longer needed when the `CREATE TABLE`, `INSERT`, or other table-changing or data-changing operation is performed through Impala. These statements are still needed if such operations are done through Hive or by manipulating data files directly in HDFS, but in those cases the statements only need to be issued on one Impala node rather than on all nodes. See [REFRESH Statement](#) and [INVALIDATE_METADATA Statement](#) for the latest usage information for those statements.
- See [The Impala Catalog Service](#) for background information on the `catalogd` service.

The new resource management feature interacts with both YARN and Llama services, which are available in CDH 5. These services are set up for you automatically in a Cloudera Manager (CM) environment. For information about setting up the YARN and Llama services, see the instructions for [YARN](#) and [Llama](#) in the [CDH 5 Documentation](#).

Incompatible Changes Introduced in Impala 1.1.1

There are no incompatible changes in Impala 1.1.1.

CDH 5 Release Notes

Previously, it was not possible to create Parquet data through Impala and reuse that table within Hive. Now that Parquet support is available for Hive 10, reusing existing Impala Parquet data files in Hive requires updating the table metadata. Use the following command if you are already running Impala 1.1.1:

```
ALTER TABLE table_name SET FILEFORMAT PARQUETFILE;
```

If you are running a level of Impala that is older than 1.1.1, do the metadata update through Hive:

```
ALTER TABLE table_name SET SERDE 'parquet.hive.serde.ParquetHiveSerDe';
ALTER TABLE table_name SET FILEFORMAT
  INPUTFORMAT "parquet.hive.DeprecatedParquetInputFormat"
  OUTPUTFORMAT "parquet.hive.DeprecatedParquetOutputFormat";
```

Impala 1.1.1 and higher can reuse Parquet data files created by Hive, without any action required.

As usual, make sure to upgrade the `impala-lzo-cdh4` package to the latest level at the same time as you upgrade the Impala server.

Incompatible Change Introduced in Impala 1.1

- The `REFRESH` statement now requires a table name; in Impala 1.0, the table name was optional. This syntax change is part of the internal rework to make `REFRESH` a true Impala SQL statement so that it can be called through the JDBC and ODBC APIs. `REFRESH` now reloads the metadata immediately, rather than marking it for update the next time any affected table is accessed. The previous behavior, where omitting the table name caused a refresh of the entire Impala metadata catalog, is available through the new `INVALIDATE_METADATA` statement. `INVALIDATE_METADATA` can be specified with a table name to affect a single table, or without a table name to affect the entire metadata catalog; the relevant metadata is reloaded the next time it is requested during the processing for a SQL statement. See [REFRESH Statement](#) and [INVALIDATE_METADATA Statement](#) for the latest details about these statements.

Incompatible Changes Introduced in Impala 1.0

- If you use LZO-compressed text files, when you upgrade Impala to version 1.0, also update the `impala-lzo-cdh4` to the latest level. See [Using LZO-Compressed Text Files](#) for details.
- Cloudera Manager 4.5.2 and higher only supports Impala 1.0 and higher, and vice versa. If you upgrade to Impala 1.0 or higher managed by Cloudera Manager, you must also upgrade Cloudera Manager to version 4.5.2 or higher. If you upgrade from an earlier version of Cloudera Manager, and were using Impala, you must also upgrade Impala to version 1.0 or higher. The beta versions of Impala are no longer supported as of the release of Impala 1.0.

Incompatible Change Introduced in Version 0.7 of the Impala Beta Release

- The defaults for the `-nn` and `-nn_port` flags have changed and are now read from `core-site.xml`. Impala prints the values of `-nn` and `-nn_port` to the log when it starts. The ability to set `-nn` and `-nn_port` on the command line is deprecated in 0.7 and may be removed in Impala 0.8.

Incompatible Change Introduced in Version 0.6 of the Impala Beta Release

- Cloudera Manager 4.5 supports only version 0.6 of the Impala Beta Release. It does not support the earlier beta versions. If you upgrade your Cloudera Manager installation, you must also upgrade Impala to beta version 0.6. If you upgrade Impala to beta version 0.6, you must upgrade Cloudera Manager to 4.5.

Incompatible Change Introduced in Version 0.4 of the Impala Beta Release

- Cloudera Manager 4.1.3 supports only version 0.4 of the Impala Beta Release. It does not support the earlier beta versions. If you upgrade your Cloudera Manager installation, you must also upgrade Impala to beta version 0.4. If you upgrade Impala to beta version 0.4, you must upgrade Cloudera Manager to 4.1.3.

Incompatible Change Introduced in Version 0.3 of the Impala Beta Release

- Cloudera Manager 4.1.2 supports only version 0.3 of the Impala Beta Release. It does not support the earlier beta versions. If you upgrade your Cloudera Manager installation, you must also upgrade Impala to beta version 0.3. If you upgrade Impala to beta version 0.3, you must upgrade Cloudera Manager to 4.1.2.

Cloudera Distribution of Apache Kafka Incompatible Changes and Limitations

Flume shipped with CDH 5.6.0 can only send data to Kafka 2.0 and higher via unsecured transport

To take advantage of security additions in Kafka 2.0 and higher, upgrade to CDH 5.7.0.

Topic Blacklist Removed

The MirrorMaker **Topic blacklist** setting has been removed in Cloudera Distribution of Kafka 2.x and higher.

Avoid Data Loss Option Removed

The **Avoid Data Loss** option from earlier releases has been removed in Kafka 2.x in favor of automatically setting the following properties.

1. Producer settings

- acks=all
- retries=max integer
- max.block.ms=max long

2. Consumer setting

- auto.commit.enable=false

3. MirrorMaker setting

- abort.on.send.failure=true

Kite Incompatible Changes and Limitations

Kite in CDH has been rebased on the 1.0 release upstream. This breaks backward compatibility with existing APIs. The APIs are documented at <http://kitesdk.org/docs/1.0.0/apidocs/index.html>. For more information, see [What's New In CDH 5.4.x](#) on page 27.

Llama Incompatible Changes and Limitations



Note: Llama no longer supported:

Although Impala can be used together with YARN via simple configuration of Static Service Pools in Cloudera Manager, the use of the general-purpose component Llama for integrated resource management between Impala and YARN is no longer supported as of CDH 5.5 / Impala 2.3. Please contact your Cloudera account team if this impacts a previously deployed system.

The following changes have made in the Llama API to help solve synchronization problems:

- As of CDH 5.1, the Reserve API requires you to provide the `reservationId` as part of the request. In previous releases, the `reservationId` was auto-generated and returned to the user.

The `TLLamaAMReservationRequest` has an additional field called `reservation_id` which needs to be initialized to a UUID value. If you do not set this field, the request will result in an error with the error code set to `ErrorCode.RESERVATION_NO_ID_PROVIDED`.

- The Expand API now requires you to provide the `expansionId` as part of the request. In previous releases, the `expansionId` was auto-generated and returned to the user.

The `TlLlamaAMReservationExpansionRequest` has an additional field called `expansion_id` which needs to be initialized to a UUID value. If you do not set this field, the request will result in an error with the error code set to `ErrorCode.EXPANSION_NO_EXPANSION_ID_PROVIDED`

Apache Mahout Incompatible Changes and Limitations

CDH 5.1 introduces the following incompatible changes.

Minor changes in behavior:

- [MAHOUT-1368](#)

The `org.apache.mahout.math.stats.OnlineSummarizer` algorithm has changed, potentially leading to different results.

- [MAHOUT-1392](#)

The streaming `KMeans` implementation has changed in how it outputs centroids when run outside of a Hadoop cluster.

- [MAHOUT-1565](#)

Major Developer API changes

- [MAHOUT-1296](#):

The following implementations have been removed:

- Clustering:

- `DirichletMeanShift`
- `MinHash`
- `Eigencuts` in `o.a.m.clustering.spectral.eigencuts`

- Classification:

- `WinnowPerceptron`
- Frequent Pattern Mining

- Collaborative Filtering:

- All recommenders in `o.a.m.cf.taste.impl.recommender.knn`
- `TreeClusteringRecommender` in `o.a.m.cf.taste.impl.recommender`
- `SlopeOne` implementations in `o.a.m..cf.taste.hadoop.slopeone` and `o.a.m.cf.taste.impl.recommender.slopeone`
- Distributed pseudo recommender in `o.a.m.cf.taste.hadoop.pseudo`

- [MAHOUT-1362](#):

The `examples/bin/build-reuters.sh` script has been removed.

Minor Developer API changes

- [MAHOUT-1280](#):

Some SSVD support code, such as `UpperTriangularMatrix`, has been moved to `mahout-math`.

- [MAHOUT-1363](#):

Scala-related math code has been moved into an `org.apache.mahout.math.scalabindings` sub-package.

Minor packaging changes

- [MAHOUT-1382](#):

Move to Guava 16.

- [MAHOUT-1364](#):

Update to require Lucene 4.6.1.

Apache Oozie Incompatible Changes and Limitations

The following incompatible changes occurred between CDH 4 and CDH 5:

- [OOZIE-1680](#) - By default, at submission time, Oozie will now reject any coordinators whose frequency is faster than 5 minutes. This check can be disabled by setting the `oozie.service.coord.check.maximum.frequency` property to false in `oozie-site.xml`; however, Cloudera does not recommend you disable this check or submit coordinators with frequencies greater than 5 minutes. Doing so can lead to unintended behavior and additional system stress.
- The procedure to install the Oozie Sharelib has changed.
- The Oozie Sharelib should be updated to the one provided with the CDH 5 package. See [Configuring Oozie](#).
- The Oozie database schema has changed and must be upgraded. See [Configuring Oozie](#) for more details. To configure Oozie using Cloudera Manager see [Managing Oozie](#).
- An Oozie client running CDH 4 will not work with an Oozie server running CDH 5 when obtaining coordinator job information. Make sure you update all the Oozie clients ([OOZIE-1482](#)).
- In CDH 4, subworkflows inherit all JAR files from their parent workflow by default. In CDH 5, this has changed so that subworkflows do not inherit JAR files by default, because the latter is actually the correct behavior. Cloudera recommends that you rework workflows and subworkflows to remove any reliance on inheriting JAR files from a parent workflow. However, setting `oozie.wf.subworkflow.classpath.inheritance` in `job.properties` or `oozie.subworkflow.classpath.inheritance` to true in `oozie-site.xml` will restore the old behavior. For more details, see the [Sub-workflow Action documentation](#)

As of CDH 5.2.0, a new [Hive 2 Action](#) allows Oozie to run HiveServer2 scripts. Using the Hive Action with HiveServer2 is now deprecated; you should switch to the new Hive 2 Action as soon as possible.

CDH 5.4.0 introduces sharelib packaging changes: the sharelib was previously shipped as a pair of tarballs, `oozie-sharelib-yarn.tar.gz` and `oozie-sharelib-mr1.tar.gz`. As of CDH 5.4.0, it is shipped as a pair of directories, `oozie-sharelib-yarn` and `oozie-sharelib-mr1`. They are still installed in `/usr/lib/oozie/` in a packages distribution, and in `/lib/oozie` in a parcels distribution.

Apache Pig Incompatible Changes and Limitations

- Apache Pig has been upgraded from version 0.11 to 0.12.
- A custom UDF must return the schema as a tuple with exactly one field.
- Added the `In`, `CASE` and `Assert` keywords; They cannot be used as variables or UDF names any more.

Cloudera Search Incompatible Changes and Limitations

General Limitations of Cloudera Search

- Cloudera Search supports one instance of the Solr service on each host in a cluster. Using multiple Solr instances on a host is not supported.
- **Converting existing file-based Sentry authorization policy files to permissions in the Sentry service does not support preserving case-sensitive role or group names**

The file-based model allows case-sensitive role names. During conversion, all roles and groups are converted to lower case.

- If a policy-file conversion will change the case of roles or groups, a warning is presented. Policy conversion can proceed, but if you have enabled document-level security and use role names as your tokens, you must re-index using the new lower case role names after conversion is complete.

- If a policy-file conversion will change the case of roles or groups, creating a name collision, an error occurs and conversion cannot occur. In such a case, you must eliminate the collisions before proceeding. For example, you could rename or delete all but one of the names that cause a collision.

Incompatible Changes Between Cloudera Search for CDH 5.8 and Previous Versions of Cloudera Search

- **Converting existing file-based Sentry authorization policy files to permissions in the Sentry service does not support preserving case-sensitive role or group names**

The file-based model allows case-sensitive role names. During conversion, all roles and groups are converted to lower case.

- If a policy-file conversion will change the case of roles or groups, a warning is presented. Policy conversion can proceed, but if you have enabled document-level security and use role names as your tokens, you must re-index using the new lower case role names after conversion is complete.
- If a policy-file conversion will change the case of roles or groups, creating a name collision, an error occurs and conversion cannot occur. In such a case, you must eliminate the collisions before proceeding. For example, you could rename or delete all but one of the names that cause a collision.

Incompatible Changes Between Cloudera Search for CDH 5.5 and Previous Versions of Cloudera Search

- **Using MapReduceIndexerTool with configurations that require an updateRequestProcessorChain may fail unless an alternate configuration is specified**

With Search for CDH 5.5, the MapReduceIndexerTool uses a default `solrconfig.xml` that is appropriate for most collection configurations. With this configuration, the MapReduceIndexerTool can index data even if Sentry is enabled. This default configuration does not include any `updateRequestProcessorChains`; if your configuration requires an `updateRequestProcessorChain`, you can configure the MapReduceIndexerTool to use the configuration from ZooKeeper by specifying `--use-zk-solrconfig.xml` or from local disk by specifying `--solr-home-dir`.

Incompatible Changes Between Cloudera Search for CDH 5.4 and Previous Versions of Cloudera Search

- **HDFS locality metrics are disabled by default**

HDFS locality metrics are disabled by default in CDH 5.4.8 and higher because they can generate numerous HDFS calls, negatively affecting performance. Performance degradation is more common in production-scale environments that include rapidly changing indexes. To re-enable these metrics, add the following to the directory config in `solrconfig.xml`:

```
<directoryFactory name="DirectoryFactory"
class="${solr.directoryFactory:org.apache.solr.core.HdfsDirectoryFactory}">
  <bool name="solr.hdfs.locality.metrics.enabled">true</bool>
</directoryFactory>
```

- **CloudSolrServer and LBHttpSolrServer no longer declare MalformedURLException as thrown from their constructors**

As a result of this change, compilation failures against the 4.10.3 Solr libraries may fail. To avoid this issue, make relevant source code changes, such as removing catch phrases related to `MalformedURLException`, and then recompile the application.

Related JIRA: Solr-5555

- **The solrJ client JavaBinCodec serializes unknown objects differently**

Starting with Search for CDH 5.4.0, Search moves from Solr 4.4 to Solr 4.1.0. With Solr 4.4, JavaBinCodec serialized unknown Java objects as `obj.toString()`. In Solr 4.10.0, JavaBinCodec serializes unknown Java objects as `obj.getClass().getName() + ':' + obj.toString()`.

As a result, the same objects may produce different results when serialized with CDH 5.4 and higher compared with objects serialized with CDH 5.3 and lower.

- **Parsing using schema.xml creates an init error when <dynamicField/> declarations include default or required attributes**

In previous releases, these attributes were ignored. If init errors occur when upgrading with an existing schema.xml, remove the default or required attributes. After removing these attributes, Search functions as it did before upgrading.

Related JIRA: SOLR-5227.

- **Indexing documents with terms that exceed Lucene's MAX_TERM_LENGTH registers errors**

In previous releases, terms that exceeded the length limit were silently ignored. To make Search function as it did in previous releases—silently ignoring longer terms—use solr.LengthFilterFactory in all of your Analyzers.

Related JIRA: LUCENE-5472.

- **The fieldType configuration docValuesFormat="Disk" is no longer supported**

If your schema.xml contains fieldTypes using docValuesFormat="Disk", modify the file to remove the docValuesFormat attribute and optimize your index to rewrite to the default codec. Make these changes before upgrading to CDH 5.4.

Related JIRA: LUCENE-5761.

- **UpdateRequestExt has been removed**

Use UpdateRequest instead.

Related JIRA: SOLR-4816.

- **Parsing schema.xml registers errors when multiple values exist where only a single value is permitted.**

With previous releases, when multiple values existed where only a single value was permitted, one value was silently chosen. In CDH 5.4, if multiple values exist where only a single value is supported, configuration parsing fails. The extra values must be removed.

Related JIRAs: SOLR-4953, SOLR-5108.

Incompatible Changes Between Cloudera Search for CDH 5.2 and Cloudera Search for CDH 5.3

Some packaging changes were made that have consequences for CrunchIndexerTool start-up scripts. If those startup scripts include the following line:

```
export myDriverJar=$(find $myDriverJarDir -maxdepth 1 -name \
'*.jar' ! -name '*-job.jar' ! -name '*-sources.jar')
```

That line in those scripts should be changed as follows:

```
export myDriverJar=$(find $myDriverJarDir -maxdepth 1 -name \
'search-crunch-*.jar' ! -name '*-job.jar' ! -name '*-sources.jar')
```

Incompatible changes between Cloudera Search for CDH 5 Beta 2 and Older Versions of Cloudera Search

The following incompatible changes occurred between Cloudera Search for CDH 5 beta 2 and older versions of Cloudera Search, including both lower versions of Cloudera Search for CDH 5 and Cloudera Search 1.x:

- Supported values for the --reducers option of the MapReduceIndexer tool change with the release of Search for CDH 5 beta 2. To use one reducer per output shard, 0 is used in Search 1.x and Search for CDH 5 beta 1. With the release of Search for CDH 5 beta 2, -2 is used for one reducer per output shard. Because of this change, commands using --reducers 0 that were written for previous Search releases do not continue to work in the same way after upgrading to Search for CDH 5 beta 2. After upgrading to Search for CDH 5 beta 2, using --reducers 0 results in an exception stating that zero is an illegal value.

Apache Sentry Incompatible Changes

- CDH 5.1 introduces a new privilege model in Sentry. This introduces a backward incompatible change for Impala. Creating a new object now requires the ALL privilege on the parent object. For example, creating a database now requires server-level privileges (previously needed database-level) and creating a table requires database-level privileges (previously needed table-level).
- Upgrading Sentry from a release **earlier than CDH 5.2 to CDH 5.2 or later** entails a schema upgrade to the Sentry database.
- As of CDH 5.3, MSCK REPAIR TABLE now requires ALL privileges for the table (previously this statement required ALL privileges on the parent database for the table).

Apache Spark Incompatible Changes and Limitations

- If you have uploaded the Spark assembly JAR file to HDFS, you must upload the new version of the file each time you upgrade Spark to a new *minor* CDH release (for example, any CDH 5.2, 5.3, 5.4, or 5.5 release). You may also need to modify the configured path for the file; see [CDH 5.2](#).
- **CDH 5.5**
 - Dynamic allocation is enabled by default but is not compatible with streaming. For streaming jobs, if you specify `--num-executors`, then dynamic allocation is implicitly disabled. To be safe, you can explicitly disable dynamic allocation using: `spark.dynamicAllocation.enabled = false`
 - The CDH 5.5 version of Spark 1.5 differs from the Apache Spark 1.5 release in using Akka version 2.2.3, the version used by Spark 1.1 and CDH 5.2. Apache Spark 1.5 uses Akka version 2.3.11.
- **CDH 5.4**
 - The CDH 5.4 version of Spark 1.3 differs from the Apache Spark 1.3 release in using Akka version 2.2.3, the version used by Spark 1.1 and CDH 5.2. Apache Spark 1.3 uses Akka version 2.3.4.
- **CDH 5.3**
 - Spark 1.2, on which CDH 5.3 is based, does not expose a transitive dependency on the Guava library. As a result, projects that use Guava but do not explicitly add it as a dependency will need to be modified: the dependency must be added to the project and also packaged with the job.
 - The CDH 5.3 version of Spark 1.2 differs from the Apache Spark 1.2 release in using Akka version 2.2.3, the version used by Spark 1.1 and CDH 5.2. Apache Spark 1.2 uses Akka version 2.3.4.
- **CDH 5.2**
 - The configured paths for `spark.eventLog.dir`, `spark.history.fs.logDirectory`, and the `SPARK_JAR` environment variable have changed in a way that may not be backward-compatible. By default, those paths now refer to the local filesystem. To make sure everything works as before, modify the paths as follows:
 - For HDFS, if this is not a federated cluster, prepend `hdfs:` to the path.
 - For HDFS in a federated cluster, prepend `viewfs:` to the path.Alternatively, you can prepend the value of `fs.defaultFS`, set in `core-site.xml` in the HDFS configuration.
 - The following changes may affect existing applications:
 - The default for I/O compression is now Snappy (changed from LZF).
 - PySpark now performs external spilling during aggregations.
 - The following Spark-related artifacts are no longer published as part of the Cloudera repository:
 - `spark-assembly`: The `spark-assembly` JAR is used internally by Spark distributions when running Spark applications and should not be referenced directly. Instead, projects should add dependencies for those parts of the Spark project that are being used, for example, `spark-core`.
 - `spark-yarn`
 - `spark-tools`
 - `spark-examples`
 - `spark-repl`

- **CDH 5.1**

- Before you can run Spark in standalone mode, you must set the `spark.master` property in `/etc/spark/conf/spark-defaults.conf`, as follows:

```
spark.master=spark://MASTER_IP:MASTER_PORT
```

where `MASTER_IP` is the IP address of the host the Spark master is running on and `MASTER_PORT` is the port.

This setting means that all jobs will run in standalone mode by default; you can override the default on the command line.

- Includes changes that will enable Spark to avoid breaking compatibility in the future. As a result, most applications will require a recompile to run against Spark 1.0, and some will require changes in source code. The details are as follows:
 - This release has two changes in the core Scala API:
 - The `cogroup` and `groupByKey` operators now return an `Iterator` over their values instead of a `Seq`. This change means that the set of values corresponding to a particular key need not all reside in memory at the same time.
 - `SparkContext.jarOfClass` now returns `Option[String]` instead of `Seq[String]`.
 - Spark Java APIs have been updated to accommodate Java 8 lambdas. See [Migrating from pre-1.0 Versions of Spark](#) for more information.

Apache Sqoop Incompatible Changes and Limitations

Upgrading Sqoop 2 from an earlier release to CDH 5.2.0 and later entails a schema upgrade to the repository database; see [Upgrading Sqoop 2 from an Earlier CDH 5 Release](#).

Apache Whirr Incompatible Changes and Limitations

The Apache Software Foundation has voted to terminate the Whirr project. Whirr is deprecated in CDH 5 and will be removed altogether in a future release.

Apache ZooKeeper Incompatible Changes and Limitations

There are no known incompatible changes between CDH 4 and CDH 5.

Known Issues in CDH 5

Operating System Known Issues

Random JVM Hangs on RHEL 6.6 (Kernels 3.14-3.17)

User Java processes—HBase RegionServers, HDFS DataNodes—can deadlock and hang for no apparent reason.

Symptom:

A Java application or daemon is ‘stuck’, making no progress. Confusingly, a `jstack`, or attaching a debugger, can unblock the process.

Workaround:

Upgrade the operating system to one with kernel version 3.18 or later; for example, RHEL 6.6z and later. For information on checking your RHEL kernel version, see <http://www.cyberciti.biz/faq/centos-redhat-rhel-6-kernel-version/>. In RHEL,

CDH 5 Release Notes

the issue has been fixed recently in kernel-2.6.32-504.16.2.el6 update (April 21, <https://rhn.redhat.com/errata/RHSA-2015-0864.html>). The following example shows how to check for the fix:

```
% rpm -qp --changelog kernel-2.6.32-504.16.2.el6.x86_64.rpm | grep 'Ensure  
get_futex_key_refs() always implies a barrier'
```

Other distributions running kernel versions 3.14-3.17 may be affected.

For more information, see <https://groups.google.com/forum/#topic/mechanical-sympathy/QbmpZxp6C64>

Leap-Second Events



Note: The last leap-second event will occur on June 30, 2015.

In general, the handling of leap-second events is tied to the time synchronization methods of Red Hat Linux releases and versions of Java: <https://access.redhat.com/articles/15145>.

- Oracle made Java more agnostic to the system clock progression, and less susceptible to the kernel mishandling a leap-second event.
- Amazon EC2 users can review the following release announcement:
<https://aws.amazon.com/blogs/aws/look-before-you-leap-the-coming-leap-second-and-aws/>
- This issue does not affect Ubuntu, SLES, or Debian.

Impact:

Internal Cloudera testing has not shown direct issues with the leap-second event; however, Cloudera recommends following Red Hat and Oracle guidance noted below.

Immediate action required:

For RedHat releases:

RHEL 5: Ensure that tzdata-2015a-1.el5 or later is installed.

RHEL 6: Ensure that tzdata-2015a-1.el6 or later is installed.

RHEL 7: Ensure that tzdata-2015a-1.el7 or later is installed.

For Oracle Java:

Oracle recommends upgrading to Java 7 for this event.

If you are running a version of Java 6, you need to update to the latest release of Java 7.

If you are running a version of Java 7 lower than 7u60, you need to update to the latest release.

Java 8 is unaffected by this issue.

Performance Known Issues



Important: For best practices, and solutions to known performance problems, see [Optimizing Performance in CDH](#).

Install and Upgrade Known Issues

Flume Kafka client incompatible changes in CDH5.8

Due to the change of offset storage from ZooKeeper to Kafka in the CDH5.8 Flume Kafka client, data might not be consumed by the Flume agents, or might be duplicated (if `kafka.auto.offset.reset=smallest`) during an upgrade to CDH5.8.

Bug: [TSB-173](#)

Workaround: See [Upgrading to CDH 5.8 When Using the Flume Kafka Client](#)

Upgrades to CDH 5.4.1 from Releases Earlier than 5.4.0 May Fail

Problem: Because of a change in the implementation of the NameNode metadata upgrade mechanism, upgrading to CDH 5.4.1 from a version lower than 5.4.0 can take an inordinately long time. In a cluster with NameNode high availability (HA) configured and a large number of edit logs, the upgrade can fail, with errors indicating a timeout in the pre-upgrade step on JournalNodes.

What to do:

To avoid the problem: Do not upgrade to CDH 5.4.1; upgrade to CDH 5.4.2 instead.

If you experience the problem: If you have already started an upgrade and seen it fail, contact Cloudera Support. This problem involves no risk of data loss, and manual recovery is possible.

If you have already completed an upgrade to CDH 5.4.1, or are installing a new cluster: In this case you are not affected and can continue to run CDH 5.4.1.

Potential job failures during YARN rolling upgrades to CDH 5.3.4

Problem: A MapReduce security fix introduced a compatibility issue that results in job failures during YARN rolling upgrades from CDH 5.3.3 to CDH 5.3.4.

Release affected: CDH 5.3.4

Release containing the fix: CDH 5.3.5

Workarounds: You can use any one of the following workarounds for this issue:

- Upgrade to CDH 5.3.5.
- Restart any jobs that might have failed during the upgrade.
- Explicitly set the version of MapReduce to be used so it is picked on a per-job basis.

1. Update the YARN property, **MR Application Classpath** (`mapreduce.application.classpath`), either in Cloudera Manager or in the `mapred-site.xml` file. Remove all existing values and add a new entry:
`<parcel-path>/lib/hadoop-mapreduce/*`, where `<parcel-path>` is the absolute path to the parcel installation. For example, the default installation path for the CDH 5.3.3 parcel would be:
`/opt/cloudera/parcels/CDH-5.3.3-1.cdh5.3.3.p0.5/lib/hadoop-mapreduce/*`.
2. Wait until jobs submitted with the above client configuration change have run to completion.
3. Upgrade to CDH 5.3.4.
4. Update the **MR Application Classpath** (`mapreduce.application.classpath`) property to point to the new CDH 5.3.4 parcel.

Do not delete the old parcel until after all jobs submitted prior to the upgrade have finished running.

NameNode Incorrectly Reports Missing Blocks During Rolling Upgrade

Problem: During a rolling upgrade to any of the releases listed below, the NameNode may report missing blocks after rolling back multiple DataNodes. This is caused by a race condition with block reporting between the DataNode and the NameNode. No permanent data loss occurs, but data can be unavailable for up to six hours before the problem corrects itself.

Releases affected: CDH 5.0.6, 5.1.5, 5.2.5, 5.3.3, 5.4.1, 5.4.2

What to do:

To avoid the problem: Cloudera advises skipping the affected releases and installing a release containing the fix. For example, do not upgrade to CDH 5.4.2; upgrade to CDH 5.4.3 instead.

The releases containing the fix are: CDH 5.3.4, 5.4.3

CDH 5 Release Notes

If you have already completed an upgrade to an affected release, or are installing a new cluster: You can continue to run the release, or upgrade to a release that is not affected.

No in-place upgrade to CDH 5 from CDH 4

Cloudera fully supports upgrade from Cloudera Enterprise 4 and CDH 4 to Cloudera Enterprise 5. Upgrade requires uninstalling the CDH 4 packages before installing CDH 5 packages.

CDH 4 and Cloudera Manager 4 End of Maintenance

Cloudera Manager version 4 and CDH 4 have reached End of Maintenance (EOM) on August 9, 2015. Cloudera will not support or provide patches for any of the Cloudera Manager version 4 or CDH 4 releases after that date.

Upgrading to CDH 5.4 or later requires an HDFS upgrade

Upgrading to CDH 5.4.0 or later from an earlier CDH 5 release requires an HDFS upgrade, and upgrading from a release earlier than CDH 5.2.0 requires additional steps. See [Upgrading from an Earlier CDH 5 Release to the Latest Release](#) for further information. See also [What's New In CDH 5.4.x](#) on page 27.

Upgrading from CDH 4 requires an HDFS upgrade

Upgrading from CDH 4 requires an HDFS upgrade. See [Upgrading from CDH 4 to CDH 5](#) for further information. See also [What's New In CDH 5.4.x](#) on page 27.

CDH 5 requires JDK 1.7

JDK 1.6 is not supported on any CDH 5 release, but before CDH 5.4.0, CDH libraries have been compatible with JDK 1.6. As of CDH 5.4.0, CDH libraries are no longer compatible with JDK 1.6 and **applications using CDH libraries must use JDK 1.7**.

In addition, you must upgrade your cluster to a supported version of JDK 1.7 before upgrading to CDH 5. See [Upgrading to Oracle JDK 1.7 before Upgrading to CDH 5](#) for instructions.

Extra step needed on Ubuntu Trusty if you add the Cloudera repository

If you install or upgrade CDH on Ubuntu Trusty using the command line, and add the Cloudera repository yourself (rather than using the "1-click Install" method) you need to perform an additional step to ensure that you get the CDH version of ZooKeeper, rather than the version that is bundled with Trusty.

No upgrade directly from CDH 3 to CDH 5

You must upgrade to CDH 4, then to CDH 5. See the [CDH 4 documentation](#) for instructions on upgrading from CDH 3 to CDH 4.

Upgrading hadoop-kms from 5.2.x and 5.3.x releases fails on SLES

Upgrading hadoop-kms fails on SLES when you try to upgrade an existing version from 5.2.x releases earlier than 5.2.4, and from 5.3.x releases earlier than 5.3.2.

After upgrading from a release earlier than CDH 4.6, you may see reports of corrupted files

Some older versions of CDH do not handle DataNodes with a large number of blocks correctly. The problem exists on versions 4.6, 4.7, 4.8, 5.0, and 5.1. The symptom is that the NameNode Web UI and the `fsck` command incorrectly report missing blocks, even when those blocks are present.

The cause of the problem is that if the DataNode attempts to send a block report that is larger than the maximum RPC buffer size, the NameNode rejects the report. This prevents the NameNode from becoming aware of the blocks on the affected DataNodes. The maximum buffer size is controlled by the `ipc.maximum.data.length` property, which defaults to 64 MB.

This problem does not exist in CDH 4.5 and earlier because there is no maximum RPC buffer size in these versions. Starting in CDH 5.2, DataNodes now send individual block reports for each storage volume, which mitigates the problem.

Bug: [HADOOP-9676](#)

Severity: Medium

Workaround: Immediately after upgrading, increase the value of `ipc.maximum.data.length`; Cloudera recommends doubling the default value, from 64 MB to 128 MB:

```
<property>
  <name>ipc.maximum.data.length</name>
  <value>134217728</value>
</property>
```

- In a Cloudera Manager installation, set this property in the `hdfs_service_config_safety_valve`.
- In a command-line-only installation, add and set this property in `core-site.xml`.

After setting `ipc.maximum.data.length`, restart the NameNode(s).

Must build native libraries when installing from tarballs

When installing Hadoop from Cloudera tarballs, you must build your own native libraries. The tarballs do not include libraries that are built for the different distributions and architectures.

Apache Flume Known Issues

Flume Kafka client incompatible changes in CDH5.8

Due to the change of offset storage from ZooKeeper to Kafka in the CDH5.8 Flume Kafka client, data might not be consumed by the Flume agents, or might be duplicated (if `kafka.auto.offset.reset=smallest`) during an upgrade to CDH5.8.

Bug: [TSB-173](#)

Workaround: See [Upgrading to CDH 5.8 When Using the Flume Kafka Client](#)

Hive sink support

Flume does not provide a native sink that stores the data that can be directly consumed by Hive.

Bug: [FLUME-1008](#)

Workaround: None

Fast Replay does not work with encrypted File Channel

If an encrypted file channel is set to use fast replay, the replay will fail and the channel will fail to start.

Bug: [FLUME-1885](#)

Workaround: Disable fast replay for the encrypted channel by setting `use-fast-replay` to false.

Apache Hadoop Known Issues

Deprecated Properties

In Hadoop 2.0.0 and later, a number of Hadoop and HDFS properties have been deprecated. (The change dates from Hadoop 0.23.1, on which the Beta releases of CDH 4 were based). A list of deprecated properties and their replacements can be found at

<https://archive.cloudera.com/cdh5/cdh/5/hadoop/hadoop-project-dist/hadoop-common/DeprecatedProperties.html>.

HDFS

DiskBalancer Occasionally Emits False Error Messages

Diskbalancer occasionally emits false error messages. For example:

```
2016-08-03 11:01:41,788 ERROR org.apache.hadoop.hdfs.server.datanode.DiskBalancer:
Disk Balancer is not enabled.
```

CDH 5 Release Notes

You can safely ignore this error message if you are not using DiskBalancer.

Workaround: Use the following command against all DataNodes to suppress DiskBalancer logs:

```
hadoop daemonlog -setlevel <host:port> org.apache.hadoop.hdfs.server.datanode.DiskBalancer  
FATAL
```

Another workaround is to suppress the warning by setting the log level of DiskBalancer to FATAL. Add the following to log4j.properties (DataNode Logging Advanced Configuration Snippet (Safety Valve)) and restart your DataNodes:

```
log4j.logger.org.apache.hadoop.hdfs.server.datanode.DiskBalancer = FATAL
```

Upgrade Requires an HDFS Upgrade

Upgrading from any release earlier than CDH 5.2.0 to CDH 5.2.0 or later requires an HDFS Upgrade.

Optimizing HDFS Encryption at Rest Requires Newer openssl Library on Some Systems

CDH 5.3 implements the **Advanced Encryption Standard New Instructions** (AES-NI), which provide substantial performance improvements. To get these improvements, you need a recent version of `libcrypto.so` on HDFS and MapReduce client hosts that is, any host from which you originate HDFS or MapReduce requests. Many OS versions have an older version of the library that does not support AES-NI.

See [HDFS Transparent Encryption](#) in the *Encryption* section of the *Cloudera Security* guide for instructions for obtaining the right version.

Other HDFS Encryption Known Issues

Potentially Incorrect Initialization Vector Calculation in HDFS Encryption

A mathematical error in the calculation of the Initialization Vector (IV) for encryption and decryption in HDFS could cause data to appear corrupted when read. The IV is a 16-byte value input to encryption and decryption ciphers. The calculation of the IV implemented in HDFS was found to be subtly different from that used by Java and OpenSSL cryptographic routines. The result is that data could possibly appear to be corrupted when it is read from a file inside an Encryption Zone.

Fortunately, the probability of this occurring is extremely small. For example, the maximum size of a file in HDFS is 64 TB. This enormous file would have a 1-in-4-million chance of hitting this condition. A more typically sized file of 1 GB would have a roughly 1-in-274-billion chance of hitting the condition.

Workaround: If you are using the experimental HDFS encryption feature in CDH 5.2, upgrade to CDH 5.3 and verify the integrity of all files inside an Encryption Zone.

DistCp between unencrypted and encrypted locations fails

By default, DistCp compares checksums provided by the filesystem to verify that data was successfully copied to the destination. However, when copying between unencrypted and encrypted locations, the filesystem checksums will not match since the underlying block data is different.

Workaround: Specify the `-skipcrccheck` and `-update` distcp flags to avoid verifying checksums.

Cannot move encrypted files to trash

With HDFS encryption enabled, you cannot move encrypted files or directories to the trash directory.

Bug: [HDFS-6767](#)

Workaround: To remove encrypted files/directories, use the following command with the `-skipTrash` flag specified to bypass trash.

```
rm -r -skipTrash /testdir
```

If you install CDH using packages, HDFS NFS gateway works out of the box only on RHEL-compatible systems

Because of a bug in native versions of `portmap/rpcbind`, the HDFS NFS gateway does not work out of the box on SLES, Ubuntu, or Debian systems if you install CDH from the command-line, using packages. It does work on supported versions of RHEL-compatible systems on which `rpcbind-0.2.0-10.el6` or later is installed, and it does work if you use Cloudera Manager to install CDH, or if you start the gateway as root.

Bug: [731542](#) (Red Hat), [823364](#) (SLES), [594880](#) (Debian)

Workarounds and caveats:

- On Red Hat and similar systems, make sure `rpcbind-0.2.0-10.el6` or later is installed.
- On SLES, Debian, and Ubuntu systems, do one of the following:
 - Install CDH using Cloudera Manager; or
 - As of CDH 5.1, start the NFS gateway as root; or
 - [Start the NFS gateway without using packages](#); or
 - You can use the gateway by running `rpcbind` in insecure mode, using the `-i` option, but keep in mind that this allows anyone from a remote host to bind to the portmap.

HDFS does not currently provide ACL support for the HDFS gateway

Bug: [HDFS-6949](#)

No error when changing permission to 777 on .snapshot directory

Snapshots are read-only; running `chmod 777` on the `.snapshot` directory does not change this, but does not produce an error (though other illegal operations do).

Bug: [HDFS-4981](#)

Workaround:

None

Snapshot operations are not supported by ViewFileSystem

Bug: None

Workaround:

None

Snapshots do not retain directories' quotas settings

Bug: [HDFS-4897](#)

Workaround:

None

Permissions for `dfs.namenode.name.dir` incorrectly set.

Hadoop daemons should set permissions for the `dfs.namenode.name.dir` (or `dfs.name.dir`) directories to `drwx-----` (700), but in fact these permissions are set to the file-system default, usually `drwxr-xr-x` (755).

Bug: [HDFS-2470](#)

Workaround:

Use `chmod` to set permissions to 700.

hadoop fsck -move does not work in a cluster with host-based Kerberos

Bug: None

Workaround:

Use `hadoop fsck -delete`

HttpFS cannot get delegation token without prior authenticated request.

A request to obtain a delegation token cannot initiate an SPNEGO authentication sequence; it must be accompanied by an authentication cookie from a prior SPNEGO authentication sequence.

Bug: [HDFS-3988](#)

Workaround:

Make another WebHDFS request (such as `GETHOMEDIR`) to initiate an SPNEGO authentication sequence and then make the delegation token request.

CDH 5 Release Notes

DistCp does not work between a secure cluster and an insecure cluster in some cases

See the upstream bug reports for details.

Bug: [HDFS-7037](#), [HADOOP-10016](#), [HADOOP-8828](#)

Workaround: None

Using DistCp with Hftp on a secure cluster using SPNEGO requires that the `dfs.https.port` property be configured

In order to DistCp using Hftp from a secure cluster using SPNEGO, you must configure the `dfs.https.port` property on the client to use the HTTP port (50070 by default).

Bug: [HDFS-3983](#)

Workaround: Configure `dfs.https.port` to use the HTTP port on the client

Non-HA DFS Clients do not attempt reconnects

This problem means that streams cannot survive a NameNode restart or network interruption that lasts longer than the time it takes to write a block.

Bug: [HDFS-4389](#)

DataNodes may become unresponsive to block creation requests

DataNodes may become unresponsive to block creation requests from clients when the directory scanner is running.

Bug: [HDFS-7489](#)

Workaround: Disable the directory scanner by setting `dfs.datanode.directoryscan.interval` to -1.

The active NameNode will not accept an fsimage sent from the standby during rolling upgrade

The result is that the NameNodes fail to checkpoint until the upgrade is finalized.



Note:

Rolling upgrade is supported only for clusters managed by Cloudera Manager; you cannot do a rolling upgrade in a command-line-only deployment.

Bug: [HDFS-7185](#)

Workaround: None.

On a DataNode with a large number of blocks, the block report may exceed the maximum RPC buffer size

Bug: None

Workaround: Increase the value `ipc.maximum.data.length` in `hdfs-site.xml`:

```
<property>
  <name>ipc.maximum.data.length</name>
  <value>268435456</value>
</property>
```

DistCp to S3a fails due to integer overflow in retry timer

Writing to S3 under high load can cause `com.amazonaws.AmazonClientException: Unable to complete transfer: timeout value is negative.`

Bug: [HADOOP-12267](#)

Workaround: Reduce the load to S3 by reducing the number of reducers or mappers.

MapReduce, YARN

Unsupported Features

The following features are not currently supported:

- **FileSystemRMStateStore:** Cloudera recommends you use `ZKRMStateStore` (ZooKeeper-based implementation) to store the ResourceManager's internal state for recovery on restart or failover. Cloudera does not support the use of `FileSystemRMStateStore` in production.
- **ApplicationTimelineSever (also known as Application History Server):** Cloudera does not support `ApplicationTimelineServer` v1. `ApplicationTimelineServer` v2 is under development and Cloudera does not currently support it.
- **Scheduler Reservations:** Scheduler reservations are currently at an experimental stage, and Cloudera does not support their use in production.
- **Scheduler node-labels:** Node-labels are currently experimental with CapacityScheduler. Cloudera does not support their use in production.

Hadoop Pipes should not be used in secure clusters

Hadoop Pipes should not be used in secure clusters. A shared password used by the framework for parent-child communications in the clear. A malicious user could intercept that password and potentially use it to access private data in a running application.

After Moving YARN JobHistory Server to a Different Host, the URLs Are Broken

After moving the JobHistory Server to a new host, the URLs listed for the JobHistory Server on the ResourceManager web UI still point to the old JobHistory Server. This affects existing jobs only. New jobs started after the move are not affected.

Workaround: For any existing jobs that have the incorrect JobHistory Server URL, there is no option other than to allow the jobs to roll off the history over time. For new jobs, make sure that all clients have the updated `mapred-site.xml` that references the correct JobHistory Server.

Starting an unmanaged ApplicationMaster may fail

Starting a custom Unmanaged ApplicationMaster may fail due to a race in getting the necessary tokens.

Bug: [YARN-1577](#)

Workaround: Try to get the tokens again; the custom unmanaged ApplicationMaster should be able to fetch the necessary tokens and start successfully.

Job movement between queues does not persist across ResourceManager restart

CDH 5 adds the capability to move a submitted application to a different scheduler queue. This queue placement is not persisted across ResourceManager restart or failover, which resumes the application in the original queue.

Bug: [YARN-1558](#)

Workaround: After ResourceManager restart, re-issue previously issued move requests.

No JobTracker becomes active if both JobTrackers are migrated to other hosts

If JobTrackers in an High Availability configuration are shut down, migrated to new hosts, then restarted, no JobTracker becomes active. The logs show a `Mismatched address` exception.

Bug: None

Workaround: After shutting down the JobTrackers on the original hosts, and before starting them on the new hosts, delete the ZooKeeper state using the following command:

```
$ zkCli.sh rmr /hadoop-ha/<logical name>
```

Hadoop Pipes may not be usable in an MRv1 Hadoop installation done through tarballs

Under MRv1, MapReduce's C++ interface, Hadoop Pipes, may not be usable with a Hadoop installation done through tarballs unless you build the C++ code on the operating system you are using.

Bug: None

Workaround: Build the C++ code on the operating system you are using. The C++ code is present under `src/c++` in the tarball.

CDH 5 Release Notes

Task-completed percentage may be reported as slightly under 100% in the web UI, even when all of a job's tasks have successfully completed.

Bug: None

Workaround: None

Oozie workflows will not be recovered in the event of a JobTracker failover on a secure cluster

Delegation tokens created by clients (via `JobClient#getDelegationToken()`) do not persist when the JobTracker [fails over](#). This limitation means that Oozie workflows will not be recovered successfully in the event of a failover on a secure cluster.

Bug: None

Workaround: Re-submit the workflow.

Encrypted shuffle in MRv2 does not work if used with LinuxContainerExecutor and encrypted web UIs.

In MRv2, if the `LinuxContainerExecutor` is used (usually as part of Kerberos security), and `hadoop.ssl.enabled` is set to `true` (See [Configuring Encrypted Shuffle, Encrypted Web UIs, and Encrypted HDFS Transport](#)), then the encrypted shuffle does not work and the submitted job fails.

Bug: [MAPREDUCE-4669](#)

Workaround: Use encrypted shuffle with Kerberos security without encrypted web UIs, or use encrypted shuffle with encrypted web UIs without Kerberos security.

Link from ResourceManager to Application Master does not work when the Web UI over HTTPS feature is enabled.

In MRv2 (YARN), if `hadoop.ssl.enabled` is set to `true` (use HTTPS for web UIs), then the link from the ResourceManager to the running MapReduce Application Master fails with an HTTP Error 500 because of a PKIX exception.

A job can still be run successfully, and, when it finishes, the link to the job history does work.

Bug: [YARN-113](#)

Workaround: Do not use encrypted web UIs.

Hadoop client JARs do not provide all the classes needed for clean compilation of client code

The compile does succeed, but you may see warnings as in the following example:

```
$ javac -cp '/usr/lib/hadoop/client/*' -d wordcount_classes WordCount.java
org/apache/hadoop/fs/Path.class(org/apache/hadoop/fs:Path.class): warning: Cannot find
annotation method 'value()'
in type 'org.apache.hadoop.classification.InterfaceAudience.LimitedPrivate': class file
for org.apache.hadoop.classification.InterfaceAudience not found
1 warning
```



Note: This means that the example at the bottom of the page on managing Hadoop API dependencies (see "Using the CDH 4 Maven Repository" under [CDH Version and Packaging Information](#)) will produce a similar warning.

Bug:

Workaround: None

The ulimits setting in `/etc/security/limits.conf` is applied to the wrong user if security is enabled.

Bug: <https://issues.apache.org/jira/browse/DAEMON-192>

Anticipated Resolution: None

Workaround: To increase the `ulimits` applied to DataNodes, you must change the `ulimit` settings for the root user, not the `hdfs` user.

Must set `yarn.resourcemanager.scheduler.address` to routable host:port when submitting a job from the ResourceManager

When you submit a job from the ResourceManager, `yarn.resourcemanager.scheduler.address` must be set to a real, routable address, not the wildcard 0.0.0.0.

Bug: None

Workaround: Set the address, in the form `host : port`, either in the client-side configuration, or on the command line when you submit the job.

Amazon S3 copy may time out

The Amazon S3 filesystem does not support renaming files, and performs a copy operation instead. If the file to be moved is very large, the operation can time out because S3 does not report progress to the TaskTracker during the operation.

Bug: [MAPREDUCE-972](#)

Workaround: Use `-Dmapred.task.timeout=15000000` to increase the MR task timeout.

Task Controller Changed from DefaultTaskController to LinuxTaskController

In CDH 5, the MapReduce task controller is changed from `DefaultTaskController` to `LinuxTaskController`. The new task controller has different directory ownership requirements which can cause jobs to fail. You can switch back to `DefaultTaskController` by adding the following to the MapReduce Advanced Configuration Snippet if you use Cloudera Manager, or directly to `mapred-default.xml` otherwise.

```
<property>
  <name>mapreduce.tasktracker.taskcontroller</name>
  <value>org.apache.hadoop.mapred.DefaultTaskController</value>
</property>
```

Out-of-memory errors may occur with Oracle JDK 1.8

The total JVM memory footprint for JDK8 can be larger than that of JDK7 in some cases. This may result in out-of-memory errors.

Bug: None

Workaround: Increase max default heap size (`-Xmx`). In the case of MapReduce, for example, increase **Reduce Task Maximum Heap Size** in Cloudera Manager (`mapred.reduce.child.java.opts`, or `mapreduce.reduce.java.opts` for YARN) to avoid out-of-memory errors during the shuffle phase.

`hadoop-test.jar` has been renamed to `hadoop-test-mr1.jar`

As of CDH 5.4.0, `hadoop-test.jar` has been renamed to `hadoop-test-mr1.jar`. This JAR file contains the `mrbench`, `TestDFSIO`, and `nnbench` tests.

Bug: None

Workaround: None.

Jobs in pool with DRF policy will not run if root pool is FAIR

If a child pool using DRF policy has a parent pool using Fairshare policy, jobs submitted to the child pool do not run.

Bug: [YARN-4212](#)

Workaround: Change parent pool to use DRF.

Large TeraValidate data sets can fail with MapReduce

In a cluster using MR1 as the MapReduce service (JobTracker/TaskTracker), TeraValidate fails when run over large TeraGen/TeraSort data sets (1TB and larger) with an `IndexOutOfBoundsException`. Smaller data sets do not show this issue.

Bug: [MAPREDUCE-6481](#)

CDH 5 Release Notes

Workaround: Use YARN as the MapReduce service.

FairScheduler might not Assign Containers

Under certain circumstances, turning on Fair Scheduler Assign Multiple Tasks (`yarn.scheduler.fair.assignmultiple`) causes the scheduler to stop assigning containers to applications. Possible symptoms are that running applications show no progress, and new applications do not start, staying in an Assigned state, despite the availability of free resources on the cluster.

Bug: [YARN-4477](#)

Workaround: Turn off Fair Scheduler Assign Multiple Tasks (`yarn.scheduler.fair.assignmultiple`) and restart the ResourceManager.

Jobs with encrypted spills do not recover if the AM goes down

The fix to [CVE-2015-1776](#) leads to not having enough information to recover a job should the Application Master fail. Releases with this security fix cannot tolerate Application Master failures.

Bug: [MAPREDUCE-6638](#)

Workaround: None. Fix to come in a later release.

FairScheduler: AMs can consume all vCores leading to a livelock

When using FAIR policy with the FairScheduler, Application Masters can consume all vCores which may lead to a livelock.

Bug: [YARN-4866](#)

Workaround: Use Dominant Resource Fairness (DRF) instead of FAIR; or make sure that the cluster has enough vCores in proportion to the memory.

MapReduce jobs might fail during a rolling upgrade to or from CDH 5.6.0

Cloudera recommends that you avoid doing rolling upgrades to CDH 5.6.0.

Bug: None.

Workaround: Restart failed jobs.

Apache HBase Known Issues

Except for version-specific sections, these issues affect each release of CDH.

Known Issues in CDH 5.9.0

Known Issues In CDH 5.7.0

Unsupported Features of Apache HBase 1.2

- Although Apache HBase 1.2 allows replication of `hbase:meta`, this feature is not supported by Cloudera and should not be used on CDH clusters until further notice.
- The [FIFO compaction policy](#) has not been thoroughly tested and is not supported in CDH 5.7.0.
- Although Apache HBase 1.2 adds a new permissive mode to allow mixed secure and insecure clients, this feature is not supported by Cloudera and should not be used on CDH clusters until further notice.

The ReplicationCleaner process can abort if its connection to ZooKeeper is inconsistent.

Bug: [HBASE-15234](#)

If the connection with ZooKeeper is inconsistent, the ReplicationCleaner may abort, and the following event will be logged by the HMaster:

```
WARN org.apache.hadoop.hbase.replication.master.ReplicationLogCleaner: Aborting
ReplicationLogCleaner
because Failed to get list of replicators
```

Unprocessed WALs will accumulate.

Workaround: Restart the HMaster occasionally. The ReplicationCleaner will restart if necessary and process the unprocessed WALs.

IntegrationTestReplication fails if replication does not finish before the verify phase begins.

Bug: None.

During IntegrationTestReplication, if the verify phase starts before the replication phase finishes, the test will fail because the target cluster does not contain all of the data. If the HBase services in the target cluster does not have enough memory, long garbage-collection pauses might occur.

Workaround: Use the -t flag to set the timeout value before starting verification.

Cloudera has tested the performance impact of using HDFS encryption with HBase. The overall overhead of HDFS encryption on HBase performance is in the range of 3 to 4% for both read and update workloads. Scan performance has not been thoroughly tested.

Known Issues in CDH 5.6.0

The ReplicationCleaner process can abort if its connection to ZooKeeper is inconsistent.

Bug: [HBASE-15234](#)

If the connection with ZooKeeper is inconsistent, the ReplicationCleaner may abort, and the following event will be logged by the HMaster:

```
WARN org.apache.hadoop.hbase.replication.master.ReplicationLogCleaner: Aborting
ReplicationLogCleaner
because Failed to get list of replicators
```

Unprocessed WALs will accumulate.

Workaround: Restart the HMaster occasionally. The ReplicationCleaner will restart if necessary and process the unprocessed WALs.

ExportSnapshot or DistCp operations may fail on the Amazon s3a:// protocol.

Bug: None.

ExportSnapshot or DistCP operations may fail on on AWS when using certain JDK 8 versions, due to an incompatibility between the AWS Java SDK 1.9.x and the joda-time date-parsing module.

Workaround: Use joda-time 2.8.1 or higher, which is included in AWS Java SDK 1.10.1 or higher.

Reverse scans do not work when Bloom blocks or leaf-level inode blocks are present.

Bug: [HBASE-14283](#)

Because the seekBefore() method calculates the size of the previous data block by assuming that data blocks are contiguous, and HFile v2 and higher store Bloom blocks and leaf-level inode blocks with the data, reverse scans do not work when Bloom blocks or leaf-level inode blocks are present when HFile v2 or higher is used.

Workaround: None.

Known Issues In CDH 5.5.1

The ReplicationCleaner process can abort if its connection to ZooKeeper is inconsistent.

Bug: [HBASE-15234](#)

CDH 5 Release Notes

If the connection with ZooKeeper is inconsistent, the ReplicationCleaner may abort, and the following event will be logged by the HMaster:

```
WARN org.apache.hadoop.hbase.replication.master.ReplicationLogCleaner: Aborting  
ReplicationLogCleaner  
because Failed to get list of replicators
```

Unprocessed WALs will accumulate.

Workaround: Restart the HMaster occasionally. The ReplicationCleaner will restart if necessary and process the unprocessed WALs.

Extra steps must be taken when upgrading from CDH 4.x to CDH 5.5.1.

The fix for TSB 2015-98 disables legacy object serialization. This will cause direct upgrades on HBase clusters from CDH 4.x to CDH 5.5.1 to fail if one of the workarounds below is not used.

Bug: [HBASE-14799](#)

Workaround: Use one of the following workarounds to upgrade from CDH 4.x to CDH 5.1:

- Upgrade to a CDH 5 version prior to CDH 5.5.1, and then upgrade from that version to CDH 5.5.1, or
- Set the `hbase.allow.legacy.object.serialization` to `true` in the Advanced Configuration Snippet for `hbase-site.xml` if using Cloudera Manager, or directly in `hbase-site.xml` on an unmanaged cluster. Upgrade your cluster to CDH 5.5.1. Remove the `hbase.allow.legacy.object.serialization` property or set it to `false` after migration is complete.

Known Issues In CDH 5.5.0

Data may not be replicated to worker cluster if multiwal multiplicity is set to greater than 1.

Bug: [HBASE-13703](#), [HBASE-6617](#), [HBASE-14501](#)

Workaround: Do not use multiwal > 1 with replication. If you use multiwal > 1, do not use replication.

An operating-system level tuning issue in RHEL7 causes 30% latency regressions.

Bug: None

Severity: Medium

Workaround: Avoid using RHEL 7 if you have a latency-critical workload. For a cached workload, consider tuning the C-state (power-saving) behavior of your CPUs.

A RegionServer under extreme duress due to back-to-back garbage collection combined with heavy load on HDFS can lock up while attempting to append to the WAL.

The RegionServer appears operational to ZooKeeper, and continues to host regions, but cannot complete any writes. The most obvious symptom is that all writes to regions on this RegionServer time out, and the RegionServer log shows no progress other than queuing of flushes that never complete. Log messages such as the following may occur:

```
124028 2015-11-14 05:54:48,659 WARN org.apache.hadoop.hbase.util.Sleeper: We slept  
42911ms instead of 3000ms,  
    this is likely due to a long garbage collecting pause and it's usually bad, see  
    http://hbase.#  
124029 2015-11-14 05:54:48,659 WARN org.apache.hadoop.hbase.util.JvmPauseMonitor: Detected  
    pause in JVM or  
    host machine (eg GC): pause of approximately 41110ms
```

```
1806 2015-11-14 04:58:09,952 INFO org.apache.hadoop.hbase.regionserver.wal.FSHLog:  
    Slow sync cost: 2734 ms, current pipeline:  
    [DatanodeInfoWithStorage[10.17.198.17:20002,DS-56e2cf88-f267-43a8-b964-b29858#  
1807 2015-11-14 04:58:09,952 INFO org.apache.hadoop.hbase.regionserver.wal.FSHLog:
```

```
Slow sync cost: 2963 ms, current pipeline:  
[DatanodeInfoWithStorage[10.17.198.17:20002,DS-56e2cf88-f267-43a8-b964-b29858#]
```

Bug: [HBASE-14374](#)

Workaround: Restart the RegionServer. To avoid the problem, adjust garbage-collection settings, give the RegionServer more RAM, and reduce the load on HDFS.

Known Issues In CDH 5.4 and Higher

The ReplicationCleaner process can abort if its connection to ZooKeeper is inconsistent.

Bug: [HBASE-15234](#)

If the connection with ZooKeeper is inconsistent, the ReplicationCleaner may abort, and the following event will be logged by the HMaster:

```
WARN org.apache.hadoop.hbase.replication.master.ReplicationLogCleaner: Aborting  
ReplicationLogCleaner  
because Failed to get list of replicators
```

Unprocessed WALs will accumulate.

Workaround: Restart the HMaster occasionally. The ReplicationCleaner will restart if necessary and process the unprocessed WALs.

Increments and CheckAnd* operations are much slower in CDH 5.4 and higher (since HBase 1.0.0) than in CDH 5.3 and earlier.

This is due to the unification of mvcc and sequenceid done in HBASE-8763.

Bug: [HBASE-14460](#)

Workaround: None

Known Issues In CDH 5.3 and Higher

Export to Azure Blob Storage (the wasb:// or wasbs:// protocol) is not supported.

Bug: [HADOOP-12717](#)

CDH 5.3 and higher supports Azure Blob Storage for some applications. However, a null pointer exception occurs when you specify a wasb:// or wasbs:// location in the --copy-to option of the ExportSnapshot command or as the output directory (the second positional argument) of the Export command.

Workaround: None.

Some HBase Features Not Supported in CDH 5

The following features, introduced upstream in HBase, are not supported in CDH 5:

- Visibility labels
- Transparent server-side encryption
- Stripe compaction
- Distributed log replay

UnknownScannerException Messages After Upgrade

HBase clients may throw exceptions like the following after an HBase upgrade:

```
org.apache.hadoop.hbase.UnknownScannerException:  
org.apache.hadoop.hbase.UnknownScannerException: Name: 10092964, already closed?
```

CDH 5 Release Notes

In this upgrade scenario, these messages are caused by restarting the RegionServer during the upgrade. Restart the HBase client to stop seeing the exceptions. The log message has been improved in CDH 5.8.0 and higher.

HBase moves to Protoc 2.5.0

This change may cause JAR conflicts with applications that have older versions of `protobuf` in their Java classpath.

Bug: None

Workaround: Update applications to use Protoc 2.5.0.

Write performance may be a little slower in CDH 5 than in CDH 4

Bug: None

Workaround: None.

Must explicitly add permissions for `owner` users before upgrading from 4.1.x

In CDH 4.1.x, an HBase table could have an owner. The `owner` user had full administrative permissions on the table (`RWXCA`). These permissions were implicit (that is, they were not stored explicitly in the HBase `acl` table), but the code checked them when determining if a user could perform an operation.

The `owner` construct was removed as of CDH 4.2.0, and the code now relies exclusively on entries in the `acl` table. Since table owners do not have an entry in this table, their permissions are removed on upgrade from CDH 4.1.x to CDH 4.2.0 or later.

Bug: None

Anticipated Resolution: None; use workaround

Workaround: Add permissions for `owner` users before upgrading from CDH 4.1.x. You can automate the task of making the owner users' implicit permissions explicit, using code similar to the following. This snippet is intended only to give you an idea of how to proceed; it may not compile and run as it stands.

```
PERMISSIONS = 'RWXCA'

tables.each do |t|
  table_name = t.getNameAsString
  owner = t.getOwnerString
  LOG.warn( "Granting " + owner + " with
            " + PERMISSIONS + " for
            table " + table_name)
  user_permission = UserPermission. new(owner.to_java_bytes, table_name.to_java_bytes,
                                         nil, nil, PERMISSIONS.to_java_bytes)
  protocol.grant(user_permission)
end
```

Change in default splitting policy from `ConstantSizeRegionSplitPolicy` to `IncreasingToUpperBoundRegionSplitPolicy` may create too many splits

This affects you only if you are upgrading from CDH 4.1 or earlier.

Split size is the number of regions that are on this server that all are part of the same table, squared, times the region flush size or the maximum region split size, whichever is smaller. For example, if the flush size is 128MB, then on first flush we will split, making two regions that will split when their size is $2 * 2 * 128\text{MB} = 512\text{MB}$. If one of these regions splits, there are three regions and now the split size is $3 * 3 * 128\text{MB} = 1152\text{MB}$, and so on until we reach the configured maximum file size, and then from then, we'll use that.

This new default policy could create many splits if you have many tables in your cluster.

This default split size has also changed - from 64MB to 128MB; and the region eventual split size, `hbase.hregion.max.filesize`, is now 10GB (it was 1GB).

Bug: None

Anticipated Resolution: None; use workaround

Workaround: If find you are getting too many splits, either go back to the old split policy or increase the `hbase.hregion.memstore.flush.size`.

In a cluster where the HBase directory in HDFS is encrypted, an IOException can occur if the BulkLoad staging directory is not in the same encryption zone as the HBase root directory.

If you have encrypted the HBase root directory (`hbase.rootdir`) and you attempt a BulkLoad where the staging directory is in a different encryption zone from the HBase root directory, you may encounter errors such as:

```
org.apache.hadoop.ipc.RemoteException(java.io.IOException):
  /tmp/output/f/5237a8430561409bb641507f0c531448 can't be moved into an encryption zone.
```

There are three different directories involved in BulkLoad operations, any of which will cause a similar error if it is not in the same encryption zone as the HBase root directory:

- The location where the output of the HBase export is dumped
- The HBase staging directory, which defaults to `/tmp/hbase-staging`, and is configured using the configuration key `hbase.bulkload.staging.dir`
- The final destination, which is usually `/hbase`

Bug: None

Anticipated Resolution: None; use workaround

Workaround: Configure each of the three locations involved in a BulkLoad operation to be in the same encryption zone. It may be necessary to manually copy the exported files from the HBase export location to a directory within `/hbase` before attempting the LoadIncremental step of the BulkLoad procedure, and remove the copied files after the BulkLoad has completed. In the interim, extra storage space will be used.

In a nonsecure cluster, MapReduce over HBase does not properly handle splits in the BulkLoad case

You may see errors because of:

- Missing permissions on the directory that contains the files to bulk load
- Missing ACL rights for the table/families

Bug: None

Anticipated Resolution: None; use workaround

Workaround: In a nonsecure cluster, execute BulkLoad as the `hbase` user.



Note: For important information about configuration that is required for BulkLoad in a secure cluster as of CDH 4.3, see the [Apache HBase Incompatible Changes and Limitations](#) on page 87 subsection under Incompatible Changes in these Release Notes.

User-provided coprocessors not supported

Cloudera does not provide support for user-provided custom coprocessors of any kind.

Bug: [HBASE-6427](#)

Workaround: None

Custom constraints coprocessors (HBASE-4605) not supported

The constraints coprocessor feature provides a framework for constraints and requires you to add your own custom code. Cloudera does not support user-provided custom code, and hence does not support this feature.

Bug: [HBASE-4605](#)

Workaround: None

CDH 5 Release Notes

Pluggable split key policy (HBASE-5304) not supported

Cloudera supports the two split policies that are supplied and tested: `ConstantSizeSplitPolicy` and `PrefixSplitKeyPolicy`. The code also provides a mechanism for custom policies that are specified by adding a class name to the `HTableDescriptor`. Custom code added via this mechanism must be provided by the user. Cloudera does not support user-provided custom code, and hence does not support this feature.

Bug: [HBASE-5304](#)

Workaround: None

HBase may not tolerate HDFS root directory changes

While HBase is running, do not stop the HDFS instance running under it and restart it again with a different root directory for HBase.

Bug: None

Workaround: None

AccessController postOperation problems in asynchronous operations

When security and Access Control are enabled, the following problems occur:

- If a `Delete Table` fails for a reason other than missing permissions, the access rights are removed but the table may still exist and may be used again.
- If `hbaseAdmin.modifyTable()` is used to delete column families, the rights are not removed from the Access Control List (ACL) table. The `postOperation` is implemented only for `postDeleteColumn()`.
- If `Create Table` fails, full rights for that table persist for the user who attempted to create it. If another user later succeeds in creating the table, the user who made the failed attempt still has the full rights.

Bug: [HBASE-6992](#)

Workaround: None

Native library not included in tarballs

The native library that enables RegionServer page pinning on Linux is not included in tarballs. This could impair performance if you install HBase from tarballs.

Bug: None

Workaround: None

Apache Hive Known Issues



Note: As of CDH 5, HCatalog is part of Apache Hive; HCatalog known issues are included [below](#).

Built-in `version()` function is not supported

Cloudera does not currently support the built-in `version()` function.

`EXPORT` and `IMPORT` commands fail for tables or partitions with data residing on Amazon S3

The `EXPORT` and `IMPORT` commands fail when the data resides on the Amazon S3 filesystem because the default Hive configuration restricts which file systems can be used for these statements.

Bug: None.

Resolution: Use workaround.

Workaround: Add S3 to the list of supported filesystems for `EXPORT` and `IMPORT` by setting the following property in HiveServer2 Advanced Configuration Snippet (Safety Valve) for `hive-site.xml` in Cloudera Manager (select **Hive** service > **Configuration** > **HiveServer2**):

```
<property>
<name>hive.exim.uri.scheme.whitelist</name>
<value>hdfs,pfile,s3a</value>
</property>
```

Hive queries on MapReduce 1 cannot use Amazon S3 when Cloudera Manager External Account feature is used

Hive queries that read or write data to Amazon S3 and use the Cloudera Manager External Account feature for S3 credential management do not work with MapReduce 1 (MRv1) because it is deprecated on CDH.

Bug: None.

Resolution: Use workaround.

Workaround: Migrate your cluster from MRv1 to MRv2. See [Migrating from MapReduce \(MRv1\) to MapReduce \(MRv2\)](#).

ALTER PARTITION does not work on Amazon S3 or between S3 and HDFS

Cloudera recommends that you do not use `ALTER PARTITION` on S3 or between S3 and HDFS.

Bug: None.

Hive cannot drop encrypted databases in cascade if trash is enabled

Bug: [HIVE-11418](#).

Workaround: Remove each table using the `PURGE` keyword (`DROP TABLE table PURGE`). After all tables are removed, remove the empty database (`DROP DATABASE database`).

Hive upgrade from CDH 5.0.5 fails on Debian 7.0 if a Sentry 5.0.x release is installed

Upgrading Hive from CDH 5.0.5 to CDH 5.4, 5.3 or 5.2 fails with the following error if a Sentry version later than 5.0.4 and earlier than 5.1.0 is installed. You will see an error such as the following:

```
: error processing
/var/cache/apt/archives/hive_0.13.1+cdh5.2.0+221-1.cdh5.2.0.p0.32~precise-cdh5.2.0_all.deb
      (--unpack): trying to overwrite '/usr/lib/hive/lib/commons-lang-2.6.jar', which
is also
      in package sentry 1.2.0+cdh5.0.5
```

This is because of a conflict involving `commons-lang-2.6.jar`.

Bug: None.

Workaround: Upgrade Sentry first and then upgrade Hive. Upgrading Sentry deletes all the JAR files that Sentry has installed under `/usr/lib/hive/lib` and installs them under `/usr/lib/sentry/lib` instead.

Hive ACID is not supported

Hive [ACID](#) is an experimental feature and Cloudera does not currently support it.

Hive creates an invalid table if you specify more than one partition with `alter table`

Hive (in all known versions from 0.7) allows you to configure multiple partitions with a single `alter table` command, but the configuration it creates is invalid for both Hive and Impala.

Bug: None

Resolution: Use workaround.

Workaround:

Correct results can be obtained by configuring each partition with its own alter table command in either Hive or Impala .For example, the following:

```
ALTER TABLE page_view ADD PARTITION (dt='2008-08-08', country='us') location
'/path/to/us/part080808' PARTITION
(dt='2008-08-09', country='us') location '/path/to/us/part080809';
```

should be replaced with:

```
ALTER TABLE page_view ADD PARTITION (dt='2008-08-08', country='us') location
'/path/to/us/part080808';
ALTER TABLE page_view ADD PARTITION (dt='2008-08-09', country='us') location
'/path/to/us/part080809';
```

PostgreSQL 9.0+ requires additional configuration

The Hive metastore will not start if you use a version of PostgreSQL later than 9.0 in the default configuration. You will see output similar to this in the log:

```
Caused by: javax.jdo.JDODataStoreException: Error executing JDOQL query
"SELECT \"THIS\".\"TBL_NAME\" AS NUCORDER0 FROM \"TBLS\" \"THIS\" LEFT OUTER JOIN \"DBS\"
\"THIS_DATABASE_NAME\" ON \"THIS\".\"DB_ID\" = \"THIS_DATABASE_NAME\".\"DB_ID\"
WHERE \"THIS_DATABASE_NAME\".\"NAME\" = ? AND (LOWER(\"THIS\".\"TBL_NAME\") LIKE ? ESCAPE '\\'
) ORDER BY NUCORDER0 " : ERROR: invalid escape string
Hint: Escape string must be empty or one character..
NestedThrowables:
org.postgresql.util.PSQLException: ERROR: invalid escape string
    Hint: Escape string must be empty or one character.
    at
org.datanucleus.jdo.NucleusJDOHelper.getJDOExceptionForNucleusException(NucleusJDOHelper.java:313)

    at org.datanucleus.jdo.JDOQuery.execute(JDOQuery.java:252)
    at org.apache.hadoop.hive.metastore.ObjectStore.getTable(ObjectStore.java:759)
    ... 28 more
Caused by: org.postgresql.util.PSQLException: ERROR: invalid escape string
    Hint: Escape string must be empty or one character.
    at
org.postgresql.core.v3.QueryExecutorImpl.receiveErrorResponse(QueryExecutorImpl.java:2096)

    at org.postgresql.core.v3.QueryExecutorImpl.processResults(QueryExecutorImpl.java:1829)

    at org.postgresql.core.v3.QueryExecutorImpl.execute(QueryExecutorImpl.java:257)
    at org.postgresql.jdbc2.AbstractJdbc2Statement.execute(AbstractJdbc2Statement.java:510)

    at
org.postgresql.jdbc2.AbstractJdbc2Statement.executeUpdate(AbstractJdbc2Statement.java:386)

    at
org.postgresql.jdbc2.AbstractJdbc2Statement.executeQuery(AbstractJdbc2Statement.java:271)

    at
org.apache.commons.dbcp.DelegatingPreparedStatement.executeQuery(DelegatingPreparedStatement.java:96)

    at
org.apache.commons.dbcp.DelegatingPreparedStatement.executeQuery(DelegatingPreparedStatement.java:96)

    at
org.datanucleus.store.rdbms.SQLController.executeStatementQuery(SQLController.java:457)

    at org.datanucleus.store.rdbms.query.legacy.SQLEvaluator.evaluate(SQLEvaluator.java:123)

    at
org.datanucleus.store.rdbms.query.legacy.JDOQLQuery.performExecute(JDOQLQuery.java:288)

    at org.datanucleus.store.query.Query.executeQuery(Query.java:1657)
    at org.datanucleus.store.rdbms.query.legacy.JDOQLQuery.executeQuery(JDOQLQuery.java:245)

    at org.datanucleus.store.query.Query.executeWithArray(Query.java:1499)
```

```
at org.datanucleus.jdo.JDOQuery.execute(JDOQuery.java:243)
... 29 more
```

The problem is caused by a backward-incompatible change in the default value of the `standard_conforming_strings` property. Versions up to PostgreSQL 9.0 defaulted to `off`, but starting with version 9.0 the default is `on`.

Bug: None

Resolution: Use workaround.

Workaround: As the administrator user, use the following command to turn `standard_conforming_strings` off:

```
ALTER DATABASE <hive_db_name> SET standard_conforming_strings = off;
```

Queries spawned from MapReduce jobs in MRv1 fail if `mapreduce.framework.name` is set to `yarn`

Queries spawned from MapReduce jobs fail in MRv1 with a null pointer exception (NPE) if `/etc/mapred/conf/mapred-site.xml` has the following:

```
<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
</property>
```

Bug: None

Resolution: Use workaround

Workaround: Remove the `mapreduce.framework.name` property from `mapred-site.xml`.

Commands run against an Oracle backed Metastore may fail

Commands run against an Oracle-backed Metastore fail with error:

```
javax.jdo.JDODataStoreException Incompatible data type for column TBLS.VIEW_EXPANDED_TEXT
: was CLOB (datastore),
but type expected was LONGVARCHAR (metadata). Please check that the type in the datastore
and the type specified in the MetaData are consistent.
```

This error may occur if the metastore is run on top of an Oracle database with the configuration property `datanucleus.validateColumns` set to true.

Bug: None

Workaround: Set `datanucleus.validateColumns=false` in the `hive-site.xml` configuration file.

Hive, Pig, and Sqoop 1 fail in MRv1 tarball installation because `/usr/bin/hbase` sets `HADOOP_MAPRED_HOME` to MR2

This problem affects tarball installations only.

Bug: None

Resolution: Use workaround

Workaround: If you are using MRv1, edit the following line in `/etc/default/hadoop` from

```
export HADOOP_MAPRED_HOME=/usr/lib/hadoop-mapreduce
```

to

```
export HADOOP_MAPRED_HOME=/usr/lib/hadoop-0.20-mapreduce
```

In addition, `/usr/lib/hadoop-mapreduce` must not exist in `HADOOP_CLASSPATH`.

Hive Web Interface not supported

Cloudera no longer supports the Hive Web Interface because of inconsistent upstream maintenance of this project.

Bug: [DISTRO-77](#)

Resolution: Use workaround

Workaround: Use Hue and Beeswax instead of the Hive Web Interface.

Hive may need additional configuration to make it work in an Federated HDFS cluster

Hive jobs normally move data from a temporary directory to a warehouse directory during execution. Hive uses `/tmp` as its temporary directory by default, and users usually configure `/user/hive/warehouse/` as the warehouse directory. Under Federated HDFS, `/tmp` and `/user` are configured as ViewFS mount tables, and so the Hive job will actually try to move data between two ViewFS mount tables. Federated HDFS does not support this, and the job will fail with the following error:

```
Failed with exception Renames across Mount points not supported
```

Bug: None

Resolution: No software fix planned; use the workaround.

Workaround: Modify `/etc/hive/conf/hive-site.xml` to allow the temporary directory and warehouse directory to use the same ViewFS mount table. For example, if the warehouse directory is `/user/hive/warehouse`, add the following property to `/etc/hive/conf/hive-site.xml` so both directories use the ViewFS mount table for `/user`.

```
<property>
  <name>hive.exec.scratchdir</name>
  <value>/user/${user.name}/tmp</value>
</property>
```

Cannot create archive partitions with external HAR (Hadoop Archive) tables

`ALTER TABLE ... ARCHIVE PARTITION` is not supported on external tables.

Bug: None

Workaround: None

Setting `hive.optimize.skewjoin` to true causes long running queries to fail

Bug: None

Workaround: None

JDBC - `executeUpdate` does not returns the number of rows modified

Contrary to the documentation, method `executeUpdate` always returns zero.

Workaround: None

Hive Auth (Grant/Revoke>Show Grant) statements do not support fully qualified table names (`default.tab1`)

Bug: None

Workaround: Switch to the database before granting privileges on the table.

Object types Server and URI are not supported in "SHOW GRANT ROLE `roleName` on OBJECT `objectName`"

Bug: None

Workaround: Use `SHOW GRANT ROLE roleName` to list all privileges granted to the role.

Kerberized HS2 with LDAP Authentication Fails in Multi-domain LDAP Case

In CDH 5.7, Hive introduced a feature to support HS2 with Kerberos plus LDAP authentication; but it broke compatibility with multi-domain LDAP cases on CDH 5.7.x and C5.8.x versions.

Bug: [HIVE-13590](#).

Workaround: None.

HCatalog Known Issues



Note: As of CDH 5, HCatalog is part of Apache Hive.

Hive's DECIMAL data type cannot be mapped to Pig via HCatalog

HCatalog does recognize the DECIMAL data type.

Bug: none

Workaround: None

Job submission using WebHCatalog might not work correctly

Bug: none

Resolution: Use workaround.

Workaround: Cloudera recommends using the Oozie REST interface to submit jobs, as it's a more mature and capable tool.

WebHCatalog does not work in a Kerberos-secured Federated cluster

Bug: none

Resolution: None planned.

Workaround: None

With Encrypted HDFS, 'drop database if exists <db_name> cascade' fails

Hive cannot drop encrypted databases in cascade if trash is enabled.

Bug: [HIVE-11418](#)

Workaround: Remove each table, using the PURGE keyword (DROP TABLE table PURGE). After all tables are removed, remove the empty database (DROP DATABASE database).

Hive External LDAP Configuration Requires Full Distinguished Name

This problem affects OpenLDAP only.

Due to a change in search and bind authentication, Hive users authenticating to external LDAP without the distinguishedName (dn) attribute may encounter errors.

Bug: [HIVE-7193](#)

Workaround: Set distinguishedName attribute to its full value.

Creating external Hive tables on an empty S3 bucket may result in NullPointerException

This bug only occurs on a completely empty s3 bucket.

Bug: None.

Workaround: Create any file in the bucket first.

CDH 5 Release Notes

Hive on Spark (HoS)

Hive on Spark throws exception for multi-insert with join

A multi-insert combined with a join query with Hive on Spark (HoS) sometimes throws an exception. It occurs only when multiple parts of the resultant operator tree are executed on the same executor by Spark.

Bug: [HIVE-13300](#)

Workaround: Run inserts one at a time.

NullPointerException when spark session is reused to run a mapjoin

Some Hive on Spark (HoS) queries may fail with a `NullPointerException` if a Spark dependency is not set.

Bug: [HIVE-12616](#)

Workaround: Configure Hive to depend on the Spark (on YARN) service in Cloudera Manager.

Large Hive on Spark queries may fail in Spark tasks with ExecutorLostFailure

The root cause is `java.lang.OutOfMemoryError: Unable to acquire XX bytes of memory, got 0`. Spark executors can OOM because of a failure to correctly spill shuffle data from memory to disk.

Bug: None.

Workaround: Run this query using MapReduce.

Hue Known Issues

Installing Hue on Debian/Ubuntu may require manual restart

When you install or upgrade Hue from packages on Debian/Ubuntu, the system tries to restart the Hue service in between updating the apps. This causes Hue to start with an interface that could be missing some apps.

Affected Versions: CDH 5 Beta 2

Bug: None

Workaround: Restart Hue once the installation/upgrade is complete: `sudo service hue restart`.

Hue can hang or fail when SQLite database is overloaded

Hue can hang or fail because the SQLite database is overloaded, returning the error, `database is locked`.

Affected Versions: CDH 5.2.x and higher

Bug: None

Workaround: Do one of the following:

- Increase the timeout setting in `[desktop][[database]]` in the Hue configuration file, *OR*
- Install an external database. See [Connect to a Custom Database](#).

Importing Hue data to MySQL can cause columns to truncate

Importing Hue data to MySQL can cause columns to be truncated on import, displaying `Warning: Data truncated for column 'name' at row 1`

Affected Versions: CDH 4.6.x and higher

Bug: None

Workaround: In the `/etc/my.cnf` file, configure the database operation to fail rather than truncate data:

```
[mysqld]
sql_mode=STRICT_ALL_TABLES
```

Problems implementing Sqoop 2 version 1.99.5

There are currently two obstacles when implementing Sqoop2:

1. Occasionally the listings pages have Unknown as a title.

Workaround: Refresh the page.

2. Autocompletes do not work.

Workaround: None.

Hue does not support the Spark App

Hue does not currently support the Spark application.

Apache Impala (incubating) Known Issues

The following sections describe known issues and workarounds in Impala, as of the current production release. This page summarizes the most serious or frequently encountered issues in the current release, to help you make planning decisions about installing and upgrading. Any workarounds are listed here. The bug links take you to the Impala issues site, where you can see the diagnosis and whether a fix is in the pipeline.



Note: The online issue tracking system for Impala contains comprehensive information and is updated in real time. To verify whether an issue you are experiencing has already been reported, or which release an issue is fixed in, search on the [issues.cloudera.org JIRA tracker](#).

For issues fixed in various Impala releases, see [Fixed Issues in Apache Impala \(incubating\)](#) on page 306.

Impala Known Issues: Crashes and Hangs

These issues can cause Impala to quit or become unresponsive.

Setting BATCH_SIZE query option too large can cause a crash

Using a value in the millions for the BATCH_SIZE query option, together with wide rows or large string values in columns, could cause a memory allocation of more than 2 GB resulting in a crash.

Bug: [IMPALA-3069](#)

Severity: High

Resolution: Fixed in CDH 5.9.0 / Impala 2.7.0.

Malformed Avro data, such as out-of-bounds integers or values in the wrong format, could cause a crash when queried.

Bug: [IMPALA-3441](#)

Severity: High

Resolution: Fixed in CDH 5.9.0 / Impala 2.7.0 and CDH 5.8.2 / Impala 2.6.2.

Queries may hang on server-to-server exchange errors

The DataStreamSender::Channel::CloseInternal() does not close the channel on an error. This causes the node on the other side of the channel to wait indefinitely, causing a hang.

Bug: [IMPALA-2592](#)

Resolution: Fixed in CDH 5.7.0 / Impala 2.5.0.

Impalad is crashing if udf jar is not available in hdfs location for first time

If the JAR file corresponding to a Java UDF is removed from HDFS after the Impala CREATE FUNCTION statement is issued, the impalad daemon crashes.

Bug: [IMPALA-2365](#)

CDH 5 Release Notes

Resolution: Fixed in CDH 5.7.0 / Impala 2.5.0.

Impala Known Issues: Performance

These issues involve the performance of operations such as queries or DDL statements.

Slow DDL statements for tables with large number of partitions

DDL statements for tables with a large number of partitions might be slow.

Bug: <https://issues.cloudera.org/browse/IMPALA-1480> IMPALA-1480

Workaround: Run the DDL statement in Hive if the slowness is an issue.

Resolution: Fixed in CDH 5.7.0 / Impala 2.5.0.

Impala Known Issues: Usability

These issues affect the convenience of interacting directly with Impala, typically through the Impala shell or Hue.

Unexpected privileges in show output

Due to a timing condition in updating cached policy data from Sentry, the `SHOW` statements for Sentry roles could sometimes display out-of-date role settings. Because Impala rechecks authorization for each SQL statement, this discrepancy does not represent a security issue for other statements.

Bug: [IMPALA-3133](#)

Severity: High

Resolution: Fixes have been issued for some but not all CDH / Impala releases. Check the JIRA for details of fix releases.

Resolution: Fixed in CDH 5.8.0 / Impala 2.6.0 and CDH 5.7.1 / Impala 2.5.1.

Less than 100% progress on completed simple SELECT queries

Simple `SELECT` queries show less than 100% progress even though they are already completed.

Bug: [IMPALA-1776](#)

Unexpected column overflow behavior with INT datatypes

Impala does not return column overflows as `NULL`, so that customers can distinguish between `NULL` data and overflow conditions similar to how they do so with traditional database systems. Impala returns the largest or smallest value in the range for the type. For example, valid values for a `tinyint` range from -128 to 127. In Impala, a `tinyint` with a value of -200 returns -128 rather than `NULL`. A `tinyint` with a value of 200 returns 127.

Bug: [IMPALA-3123](#)

Impala Known Issues: JDBC and ODBC Drivers

These issues affect applications that use the JDBC or ODBC APIs, such as business intelligence tools or custom-written applications in languages such as Java or C++.

ImpalaODBC: Can not get the value in the `SQLGetData(m-x th column)` after the `SQLBindCol(m th column)`

If the ODBC `SQLGetData` is called on a series of columns, the function calls must follow the same order as the columns. For example, if data is fetched from column 2 then column 1, the `SQLGetData` call for column 1 returns `NULL`.

Bug: [IMPALA-1792](#)

Workaround: Fetch columns in the same order they are defined in the table.

Impala Known Issues: Security

These issues relate to security features, such as Kerberos authentication, Sentry authorization, encryption, auditing, and redaction.

Kerberos tickets must be renewable

In a Kerberos environment, the `impalad` daemon might not start if Kerberos tickets are not renewable.

Workaround: Configure your KDC to allow tickets to be renewed, and configure `krb5.conf` to request renewable tickets.

Impala Known Issues: Resources

These issues involve memory or disk usage, including out-of-memory conditions, the spill-to-disk feature, and resource management features.

Impala catalogd heap issues when upgrading to 5.7

The default heap size for Impala `catalogd` has changed in 5.7 and higher:

- Before 5.7, by default `catalogd` was using the JVM's default heap size, which is the smaller of 1/4th of the physical memory or 32 GB.
- Starting with CDH 5.7.0, the default `catalogd` heap size is 4 GB.

For example, on a host with 128GB physical memory this will result in `catalogd` heap decreasing from 32GB to 4GB. This can result in out-of-memory errors in `catalogd` and leading to query failures.

Severity: High

Workaround: Increase the `catalogd` memory limit as follows.

For schemas with large numbers of tables, partitions, and data files, the `catalogd` daemon might encounter an out-of-memory error. To increase the memory limit for the `catalogd` daemon:

1. Check current memory usage for the `catalogd` daemon by running the following commands on the host where that daemon runs on your cluster:

```
jcmsg catalogd_pid VM.flags
jmap -heap catalogd_pid
```

2. Decide on a large enough value for the `catalogd` heap. You express it as an environment variable value as follows:

```
JAVA_TOOL_OPTIONS="-Xmx8g"
```

3. On systems managed by Cloudera Manager, include this value in the configuration field **Java Heap Size of Catalog Server in Bytes** (Cloudera Manager 5.7 and higher), or **Impala Catalog Server Environment Advanced Configuration Snippet (Safety Valve)** (prior to Cloudera Manager 5.7). Then restart the Impala service.
4. On systems not managed by Cloudera Manager, put this environment variable setting into the startup script for the `catalogd` daemon, then restart the `catalogd` daemon.
5. Use the same `jcmsg` and `jmap` commands as earlier to verify that the new settings are in effect.

Breakpad minidumps can be very large when the thread count is high

The size of the breakpad minidump files grows linearly with the number of threads. By default, each thread adds 8 KB to the minidump size. Minidump files could consume significant disk space when the daemons have a high number of threads.

Bug: [IMPALA-3509](#)

Severity: High

Workaround: Add `--minidump_size_limit_hint_kb=size` to set a soft upper limit on the size of each minidump file. If the minidump file would exceed that limit, Impala reduces the amount of information for each thread from 8

CDH 5 Release Notes

KB to 2 KB. (Full thread information is captured for the first 20 threads, then 2 KB per thread after that.) The minidump file can still grow larger than the “hinted” size. For example, if you have 10,000 threads, the minidump file can be more than 20 MB.

Parquet scanner memory increase after IMPALA-2736

The initial release of sometimes has a higher peak memory usage than in previous releases while reading Parquet files.

addresses the issue IMPALA-2736, which improves the efficiency of Parquet scans by up to 2x. The faster scans may result in a higher peak memory consumption compared to earlier versions of Impala due to the new column-wise row materialization strategy. You are likely to experience higher memory consumption in any of the following scenarios:

- Very wide rows due to projecting many columns in a scan.
- Very large rows due to big column values, for example, long strings or nested collections with many items.
- Producer/consumer speed imbalances, leading to more rows being buffered between a scan (producer) and downstream (consumer) plan nodes.

Bug: [IMPALA-3662](#)

Severity: High

Workaround: The following query options might help to reduce memory consumption in the Parquet scanner:

- Reduce the number of scanner threads, for example: `set num_scanner_threads=30`
- Reduce the batch size, for example: `set batch_size=512`
- Increase the memory limit, for example: `set mem_limit=64g`

Process mem limit does not account for the JVM's memory usage

Some memory allocated by the JVM used internally by Impala is not counted against the memory limit for the `impalad` daemon.

Bug: [IMPALA-691](#)

Workaround: To monitor overall memory usage, use the `top` command, or add the memory figures in the Impala web UI `/memz` tab to JVM memory usage shown on the `/metrics` tab.

Fix issues with the legacy join and agg nodes using `--enable_partitioned_hash_join=false` and `--enable_partitioned_aggregation=false`

Bug: [IMPALA-2375](#)

Workaround: Transition away from the “old-style” join and aggregation mechanism if practical.

Resolution: Fixed in CDH 5.7.0 / Impala 2.5.0.

Impala Known Issues: Correctness

These issues can cause incorrect or unexpected results from queries. They typically only arise in very specific circumstances.

Incorrect assignment of NULL checking predicate through an outer join of a nested collection.

A query could return wrong results (too many or too few NULL values) if it referenced an outer-joined nested collection and also contained a null-checking predicate (`IS NULL`, `IS NOT NULL`, or the `<=>` operator) in the `WHERE` clause.

Bug: [IMPALA-3084](#)

Severity: High

Resolution: Fixed in CDH 5.9.0 / Impala 2.7.0.

Incorrect result due to constant evaluation in query with outer join

An OUTER JOIN query could omit some expected result rows due to a constant such as FALSE in another join clause. For example:

```
explain SELECT 1 FROM alltypestiny a1
    INNER JOIN alltypesagg a2 ON a1.smallint_col = a2.year AND false
    RIGHT JOIN alltypes a3 ON a1.year = a3.bigint_col;
+-----+
| Explain String
+-----+
| Estimated Per-Host Requirements: Memory=1.00KB VCores=1
|
| 00:EMPTYSET
+-----+
```

Bug: [IMPALA-3094](#)

Severity: High

Resolution:

Workaround:

Incorrect assignment of an inner join On-clause predicate through an outer join.

Impala may return incorrect results for queries that have the following properties:

- There is an INNER JOIN following a series of OUTER JOINS.
- The INNER JOIN has an On-clause with a predicate that references at least two tables that are on the nullable side of the preceding OUTER JOINs.

The following query demonstrates the issue:

```
select 1 from functional.alltypes a left outer join
    functional.alltypes b on a.id = b.id left outer join
    functional.alltypes c on b.id = c.id right outer join
    functional.alltypes d on c.id = d.id inner join functional.alltypes e
    on b.int_col = c.int_col;
```

The following listing shows the incorrect EXPLAIN plan:

```
+-----+
| Explain String
+-----+
| Estimated Per-Host Requirements: Memory=480.04MB VCores=4
|
14:EXCHANGE [UNPARTITIONED]
|
08:NESTED LOOP JOIN [CROSS JOIN, BROADCAST]
|
--13:EXCHANGE [BROADCAST]
|
| 04:SCAN HDFS [functional.alltypes e]
|     partitions=24/24 files=24 size=478.45KB
|
07:HASH JOIN [RIGHT OUTER JOIN, PARTITIONED]
| hash predicates: c.id = d.id
| runtime filters: RF000 <- d.id
|
--12:EXCHANGE [HASH(d.id)]
|
| 03:SCAN HDFS [functional.alltypes d]
|     partitions=24/24 files=24 size=478.45KB
```

```

06:HASH JOIN [LEFT OUTER JOIN, PARTITIONED]
|   hash predicates: b.id = c.id
|   other predicates: b.int_col = c.int_col      <--- incorrect placement; should be at
node 07 or 08
|   runtime filters: RF001 <- c.int_col

--11:EXCHANGE [HASH(c.id)]

02:SCAN HDFS [functional.alltypes c]
|   partitions=24/24 files=24 size=478.45KB
|   runtime filters: RF000 -> c.id

05:HASH JOIN [RIGHT OUTER JOIN, PARTITIONED]
|   hash predicates: b.id = a.id
|   runtime filters: RF002 <- a.id

--10:EXCHANGE [HASH(a.id)]

00:SCAN HDFS [functional.alltypes a]
|   partitions=24/24 files=24 size=478.45KB

09:EXCHANGE [HASH(b.id)]

01:SCAN HDFS [functional.alltypes b]
|   partitions=24/24 files=24 size=478.45KB
|   runtime filters: RF001 -> b.int_col, RF002 -> b.id
+-----+

```

Bug: [IMPALA-3126](#)

Severity: High

Workaround: High

For some queries, this problem can be worked around by placing the problematic ON clause predicate in the WHERE clause instead, or changing the preceding OUTER JOINS to INNER JOINS (if the ON clause predicate would discard NULLs). For example, to fix the problematic query above:

```

select 1 from functional.alltypes a
  left outer join functional.alltypes b
    on a.id = b.id
  left outer join functional.alltypes c
    on b.id = c.id
  right outer join functional.alltypes d
    on c.id = d.id
  inner join functional.alltypes e
where b.int_col = c.int_col

+-----+
| Explain String
+-----+
Estimated Per-Host Requirements: Memory=480.04MB VCores=4

14:EXCHANGE [UNPARTITIONED]

08:NESTED LOOP JOIN [CROSS JOIN, BROADCAST]
|   --13:EXCHANGE [BROADCAST]
|       04:SCAN HDFS [functional.alltypes e]
|           partitions=24/24 files=24 size=478.45KB

07:HASH JOIN [RIGHT OUTER JOIN, PARTITIONED]
|   hash predicates: c.id = d.id
|   other predicates: b.int_col = c.int_col      <--- correct assignment
|   runtime filters: RF000 <- d.id

--12:EXCHANGE [HASH(d.id)]

03:SCAN HDFS [functional.alltypes d]
+-----+

```

```

    partitions=24/24 files=24 size=478.45KB
06:HASH JOIN [LEFT OUTER JOIN, PARTITIONED]
  hash predicates: b.id = c.id
  --11:EXCHANGE [HASH(c.id)]
  |
  02:SCAN HDFS [functional.alltypes c]
    partitions=24/24 files=24 size=478.45KB
    runtime filters: RF000 -> c.id

05:HASH JOIN [RIGHT OUTER JOIN, PARTITIONED]
  hash predicates: b.id = a.id
  runtime filters: RF001 <- a.id
  --10:EXCHANGE [HASH(a.id)]
  |
  00:SCAN HDFS [functional.alltypes a]
    partitions=24/24 files=24 size=478.45KB

09:EXCHANGE [HASH(b.id)]
01:SCAN HDFS [functional.alltypes b]
  partitions=24/24 files=24 size=478.45KB
  runtime filters: RF001 -> b.id
+-----+

```

Impala may use incorrect bit order with BIT_PACKED encoding

Parquet BIT_PACKED encoding as implemented by Impala is LSB first. The parquet standard says it is MSB first.

Bug: [IMPALA-3006](#)

Severity: High, but rare in practice because BIT_PACKED is infrequently used, is not written by Impala, and is deprecated in Parquet 2.0.

BST between 1972 and 1995

The calculation of start and end times for the BST (British Summer Time) time zone could be incorrect between 1972 and 1995. Between 1972 and 1995, BST began and ended at 02:00 GMT on the third Sunday in March (or second Sunday when Easter fell on the third) and fourth Sunday in October. For example, both function calls should return 13, but actually return 12, in a query such as:

```

select
  extract(from_utc_timestamp(cast('1970-01-01 12:00:00' as timestamp), 'Europe/London'),
  "hour") summer70start,
  extract(from_utc_timestamp(cast('1970-12-31 12:00:00' as timestamp), 'Europe/London'),
  "hour") summer70end;

```

Bug: [IMPALA-3082](#)

Severity: High

`parse_url()` returns incorrect result if @ character in URL

If a URL contains an @ character, the `parse_url()` function could return an incorrect value for the hostname field.

Bug: <https://issues.cloudera.org/browse/IMPALA-1170> IMPALA-1170

Resolution: Fixed in CDH 5.7.0 / Impala 2.5.0 and CDH 5.5.4 / Impala 2.3.4.

`% escaping does not work correctly when occurs at the end in a LIKE clause`

If the final character in the RHS argument of a `LIKE` operator is an escaped \% character, it does not match a % final character of the LHS argument.

Bug: [IMPALA-2422](#)

CDH 5 Release Notes

ORDER BY rand() does not work.

Because the value for `xrand()` is computed early in a query, using an `ORDER BY` expression involving a call to `xrand()` does not actually randomize the results.

Bug: [IMPALA-397](#)

Duplicated column in inline view causes dropping null slots during scan

If the same column is queried twice within a view, `NULL` values for that column are omitted. For example, the result of `COUNT(*)` on the view could be less than expected.

Bug: [IMPALA-2643](#)

Workaround: Avoid selecting the same column twice within an inline view.

Resolution: Fixed in CDH 5.7.0 / Impala 2.5.0, CDH 5.5.2 / Impala 2.3.2, and CDH 5.4.10 / Impala 2.2.10.

Incorrect assignment of predicates through an outer join in an inline view.

A query involving an `OUTER JOIN` clause where one of the table references is an inline view might apply predicates from the `ON` clause incorrectly.

Bug: [IMPALA-1459](#)

Resolution: Fixed in CDH 5.7.0 / Impala 2.5.0, CDH 5.5.2 / Impala 2.3.2, and CDH 5.4.9 / Impala 2.2.9.

Crash: impala::Coordinator::ValidateCollectionSlots

A query could encounter a serious error if includes multiple nested levels of `INNER JOIN` clauses involving subqueries.

Bug: [IMPALA-2603](#)

Incorrect assignment of On-clause predicate inside inline view with an outer join.

A query might return incorrect results due to wrong predicate assignment in the following scenario:

1. There is an inline view that contains an outer join
2. That inline view is joined with another table in the enclosing query block
3. That join has an On-clause containing a predicate that only references columns originating from the outer-joined tables inside the inline view

Bug: [IMPALA-2665](#)

Resolution: Fixed in CDH 5.7.0 / Impala 2.5.0, CDH 5.5.2 / Impala 2.3.2, and CDH 5.4.9 / Impala 2.2.9.

Wrong assignment of having clause predicate across outer join

In an `OUTER JOIN` query with a `HAVING` clause, the comparison from the `HAVING` clause might be applied at the wrong stage of query processing, leading to incorrect results.

Bug: [IMPALA-2144](#)

Resolution: Fixed in CDH 5.7.0 / Impala 2.5.0.

Wrong plan of NOT IN aggregate subquery when a constant is used in subquery predicate

A `NOT IN` operator with a subquery that calls an aggregate function, such as `NOT IN (SELECT SUM(...))`, could return incorrect results.

Bug: [IMPALA-2093](#)

Resolution: Fixed in CDH 5.7.0 / Impala 2.5.0 and CDH 5.5.4 / Impala 2.3.4.

Impala Known Issues: Metadata

These issues affect how Impala interacts with metadata. They cover areas such as the metastore database, the `COMPUTE STATS` statement, and the Impala catalogd daemon.

Catalogd may crash when loading metadata for tables with many partitions, many columns and with incremental stats

Incremental stats use up about 400 bytes per partition for each column. For example, for a table with 20K partitions and 100 columns, the memory overhead from incremental statistics is about 800 MB. When serialized for transmission across the network, this metadata exceeds the 2 GB Java array size limit and leads to a catalogd crash.

Bugs: [IMPALA-2647](#), [IMPALA-2648](#), [IMPALA-2649](#)

Workaround: If feasible, compute full stats periodically and avoid computing incremental stats for that table. The scalability of incremental stats computation is a continuing work item.

[Can't update stats manually via alter table after upgrading to CDH 5.2](#)

Bug: [IMPALA-1420](#)

Workaround: On CDH 5.2, when adjusting table statistics manually by setting the numRows, you must also enable the Boolean property STATS_GENERATED_VIA_STATS_TASK. For example, use a statement like the following to set both properties with a single ALTER TABLE statement:

```
ALTER TABLE table_name SET TBLPROPERTIES('numRows'='new_value',
    'STATS_GENERATED_VIA_STATS_TASK' = 'true');
```

Resolution: The underlying cause is the issue [HIVE-8648](#) that affects the metastore in Hive 0.13. The workaround is only needed until the fix for this issue is incorporated into a CDH release.

[Impala Known Issues: Interoperability](#)

These issues affect the ability to interchange data between Impala and other database systems. They cover areas such as data types and file formats.

[DESCRIBE FORMATTED gives error on Avro table](#)

This issue can occur either on old Avro tables (created prior to Hive 1.1 / CDH 5.4) or when changing the Avro schema file by adding or removing columns. Columns added to the schema file will not show up in the output of the DESCRIBE FORMATTED command. Removing columns from the schema file will trigger a NullPointerException.

As a workaround, you can use the output of SHOW CREATE TABLE to drop and recreate the table. This will populate the Hive metastore database with the correct column definitions.



Warning: Only use this for external tables, or Impala will remove the data files. In case of an internal table, set it to external first:

```
ALTER TABLE table_name SET TBLPROPERTIES('EXTERNAL'='TRUE');
```

(The part in parentheses is case sensitive.) Make sure to pick the right choice between internal and external when recreating the table. See [Overview of Impala Tables](#) for the differences between internal and external tables.

Severity: High

Deviation from Hive behavior: Impala does not do implicit casts between string and numeric and boolean types.

Anticipated Resolution: None

Workaround: Use explicit casts.

Deviation from Hive behavior: Out of range values float/double values are returned as maximum allowed value of type (Hive returns NULL)

Impala behavior differs from Hive with respect to out of range float/double values. Out of range values are returned as maximum allowed value of type (Hive returns NULL).

Workaround: None

Configuration needed for Flume to be compatible with Impala

For compatibility with Impala, the value for the Flume HDFS Sink `hdfs.writeFormat` must be set to `Text`, rather than its default value of `Writable`. The `hdfs.writeFormat` setting must be changed to `Text` before creating data files with Flume; otherwise, those files cannot be read by either Impala or Hive.

Resolution: This information has been requested to be added to the upstream Flume documentation.

Avro Scanner fails to parse some schemas

Querying certain Avro tables could cause a crash or return no rows, even though Impala could `DESCRIBE` the table.

Bug: [IMPALA-635](#)

Workaround: Swap the order of the fields in the schema specification. For example, `["null", "string"]` instead of `["string", "null"]`.

Resolution: Not allowing this syntax agrees with the Avro specification, so it may still cause an error even when the crashing issue is resolved.

Impala BE cannot parse Avro schema that contains a trailing semi-colon

If an Avro table has a schema definition with a trailing semicolon, Impala encounters an error when the table is queried.

Bug: [IMPALA-1024](#)

Severity: Remove trailing semicolon from the Avro schema.

Fix decompressor to allow parsing gzips with multiple streams

Currently, Impala can only read gzipped files containing a single stream. If a gzipped file contains multiple concatenated streams, the Impala query only processes the data from the first stream.

Bug: [IMPALA-2154](#)

Workaround: Use a different gzip tool to compress file to a single stream file.

Resolution: Fixed in CDH 5.7.0 / Impala 2.5.0.

Impala incorrectly handles text data when the new line character \n\r is split between different HDFS block

If a carriage return / newline pair of characters in a text table is split between HDFS data blocks, Impala incorrectly processes the row following the `\n\r` pair twice.

Bug: [IMPALA-1578](#)

Workaround: Use the Parquet format for large volumes of data where practical.

Resolution: Fixed in CDH 5.8.0 / Impala 2.6.0.

Invalid bool value not reported as a scanner error

In some cases, an invalid `BOOLEAN` value read from a table does not produce a warning message about the bad value. The result is still `NONE` as expected. Therefore, this is not a query correctness issue, but it could lead to overlooking the presence of invalid data.

Bug: [IMPALA-1862](#)

Incorrect results with basic predicate on CHAR typed column.

When comparing a `CHAR` column value to a string literal, the literal value is not blank-padded and so the comparison might fail when it should match.

Bug: [IMPALA-1652](#)

Workaround: Use the `RPAD()` function to blank-pad literals compared with `CHAR` columns to the expected length.

Impala Known Issues: Limitations

These issues are current limitations of Impala that require evaluation as you plan how to integrate Impala into your data management workflow.

Impala does not support running on clusters with federated namespaces

Impala does not support running on clusters with federated namespaces. The `impalad` process will not start on a node running such a filesystem based on the `org.apache.hadoop.fs.viewfs.ViewFs` class.

Bug: [IMPALA-77](#)

Anticipated Resolution: Limitation

Workaround: Use standard HDFS on all Impala nodes.

Impala Known Issues: Miscellaneous / Older Issues

These issues do not fall into one of the above categories or have not been categorized yet.

A failed CTAS does not drop the table if the insert fails.

If a `CREATE TABLE AS SELECT` operation successfully creates the target table but an error occurs while querying the source table or copying the data, the new table is left behind rather than being dropped.

Bug: [IMPALA-2005](#)

Workaround: Drop the new table manually after a failed `CREATE TABLE AS SELECT`.

Casting scenarios with invalid/inconsistent results

Using a `CAST()` function to convert large literal values to smaller types, or to convert special values such as `NaN` or `Inf`, produces values not consistent with other database systems. This could lead to unexpected results from queries.

Bug: [IMPALA-1821](#)

Support individual memory allocations larger than 1 GB

The largest single block of memory that Impala can allocate during a query is 1 GiB. Therefore, a query could fail or Impala could crash if a compressed text file resulted in more than 1 GiB of data in uncompressed form, or if a string function such as `group_concat()` returned a value greater than 1 GiB.

Bug: [IMPALA-1619](#)

Resolution: Fixed in CDH 5.9.0 / Impala 2.7.0 and CDH 5.8.3 / Impala 2.6.3.

Impala Parser issue when using fully qualified table names that start with a number

A fully qualified table name starting with a number could cause a parsing error. In a name such as `db.571_market`, the decimal point followed by digits is interpreted as a floating-point number.

Bug: [IMPALA-941](#)

Workaround: Surround each part of the fully qualified name with backticks (`\``).

Impala should tolerate bad locale settings

If the `LC_*` environment variables specify an unsupported locale, Impala does not start.

Bug: [IMPALA-532](#)

Workaround: Add `LC_ALL="C"` to the environment settings for both the Impala daemon and the Statestore daemon. See [Modifying Impala Startup Options](#) for details about modifying these environment settings.

Resolution: Fixing this issue would require an upgrade to Boost 1.47 in the Impala distribution.

Log Level 3 Not Recommended for Impala

The extensive logging produced by log level 3 can cause serious performance overhead and capacity issues.

Workaround: Reduce the log level to its default value of 1, that is, `GLOG_v=1`. See [Setting Logging Levels](#) for details about the effects of setting different logging levels.

Cloudera Distribution of Apache Kafka Known Issues

Flume Kafka client incompatible changes in CDH5.8

Due to the change of offset storage from ZooKeeper to Kafka in the CDH5.8 Flume Kafka client, data might not be consumed by the Flume agents, or might be duplicated (if `kafka.auto.offset.reset=smallest`) during an upgrade to CDH5.8.

Bug: [TSB-173](#)

Workaround: See [Upgrading to CDH 5.8 When Using the Flume Kafka Client](#)

Apache Mahout Known Issues

The Scala module with Mahout does not support Java 8

Bug: None

Workaround: None

Apache Oozie Known Issues

After enabling HA, Oozie may fail to start due to "NoSuchFieldError: EXTERNAL_PROPERTY

This issue happens in rare cases. Due to an incompatibility with the version of Jackson used by Oozie and Hive, and depending on the order that jars are loaded into Oozie's classpath, Oozie may fail to start.

Affected Versions: CDH5.x and higher.

Bug: [HIVE-1640](#)

Workaround: If using parcels:

1. Delete or move `/opt/cloudera/parcels/CDH/lib/oozie/libserver/hive-exec.jar` and `/opt/cloudera/parcels/CDH/lib/oozie/libtools/hive-exec.jar`.
2. Download `hive-exec-<cdh version>-core.jar` from the Cloudera repo and put it in `/opt/cloudera/parcels/CDH/lib/oozie/libserver/` and `/opt/cloudera/parcels/CDH/lib/oozie/libtools/`.
3. Download `kryo-2.22.jar` from the maven repo and put it in `/opt/cloudera/parcels/CDH/lib/oozie/libserver/` and `/opt/cloudera/parcels/CDH/lib/oozie/libtools/`.



Note: If using packages, use `/usr/lib/oozie` instead of `/opt/cloudera/parcels/CDH/lib/oozie/` in the above paths.

Oozie Web Console returns 500 error when Oozie server runs on JDK 8u75 or higher

The Oozie Web Console returns a 500 error when the Oozie server is running on JDK 8u75 and higher. The Oozie server still functions, and you can use the Oozie command line, REST API, Java API, or the Hue Oozie Dashboard to review status of those jobs.

Affected Versions: CDH5.x and higher.

Bug: [OOZIE-2365](#)

Workaround: Use an earlier version of Java 8 or use the Hue Oozie Dashboard.

Oozie jobs fail (gracefully) on secure YARN clusters when JobHistory server is down

If the JobHistory server is down on a YARN (MRv2) cluster, Oozie attempts to submit a job, by default, three times. If the job fails, Oozie automatically puts the workflow in a SUSPEND state.

Affected Versions: CDH 5 Beta 1 and higher.

Workaround: When the JobHistory server is running again, use the `resume` command to tell Oozie to continue the workflow from the point at which it left off.

Oozie does not start when `oozie.email.smtp.auth` is disabled

If you enable SLA integration, and `oozie.email.smtp.auth` is disabled, Oozie throws a `NullPointerException` and fails to start.

Affected Versions: C5.5.1 and lower.

Bug: [OOZIE-2365](#)

Workaround: In Cloudera Manager, configure **Oozie Server Advanced Configuration Snippet (Safety Valve) for `oozie-site.xml`** as follows:

```
<property>
  <name>oozie.email.smtp.password</name>
  <value>none</value>
</property>
<property>
  <name>oozie.email.smtp.username</name>
  <value>none</value>
</property>
```

Oozie works with MapReduce or YARN, but not both

The Oozie server works with a MapReduce (MRv1) cluster or a YARN (MRv2) cluster, but not both at the same time.

Workaround: Use two different Oozie servers.

Apache Parquet Known Issues

Parquet file writes run out of memory if (number of partitions) times (block size) exceeds available memory

The Parquet output writer allocates one block for each table partition it is processing and writes partitions in parallel. The MapReduce or YARN task will run out of memory if (number of partitions) times (Parquet block size) is greater than the available memory.

Bug: None

Workaround: None; if necessary, reduce the number of partitions in the table.

parquet-thrift cannot read Parquet data written by Hive

`parquet-thrift` cannot read Parquet data written by Hive, and `parquet-avro` will show an additional record level in lists named `array_element`.

Bug: [PARQUET-113](#)

Workaround: None; arrays written by `parquet-avro` or `parquet-thrift` cannot currently be read by `parquet-hive`.

Apache Pig Known Issues

Hive, Pig, and Sqoop 1 fail in MRv1 tarball installation because `/usr/bin/hbase` sets `HADOOP_MAPRED_HOME` to MRv2

This problem affects tarball installations only.

Bug: None

Resolution: Use workaround.

Workaround: If you are using MRv1, edit the following line in `/etc/default/hadoop` from

```
export HADOOP_MAPRED_HOME=/usr/lib/hadoop-mapreduce
```

to

```
export HADOOP_MAPRED_HOME=/usr/lib/hadoop-0.20-mapreduce
```

In addition, /usr/lib/hadoop-mapreduce must not exist in HADOOP_CLASSPATH.

Pig fails to read Parquet file (created with Hive) with a complex field if schema not specified explicitly

Bug: None

Workaround: Provide the schema of the fields in the `LOAD` statement.

Cloudera Search Known Issues

The current release includes the following known limitations:

Collection Creation No Longer Supports Automatically Selecting A Configuration If Only One Exists

Before CDH 5.5.0, a collection could be created without specifying a configuration. If no `-c` value was specified, then:

- If there was only one configuration, that configuration was chosen.
- If the collection name matched a configuration name, that configuration was chosen.

Search for CDH 5.5.0 includes multiple built-in configurations. As a result, there is no longer a case in which only one configuration can be chosen by default.

To avoid this issue, explicitly specify the collection configuration to use by passing `-c configName` to `solrctl collection --create`.

Title: Configuration Templates Are Not Automatically Created on Existing SolrCloud Deployments

Configuration templates are only created when Solr is initialized. As a result, templates are not automatically available on existing SolrCloud deployments, even after upgrading to CDH 5.5.0.

If you do not need to retain information in your SolrCloud cluster, you can reinitialize solr using Cloudera Manager or using `solrctl init`. If you need to retain information in your cluster, there is no automated way to create the templates, but you can access the templates on a host with Solr installed at `/usr/lib/solr/templateName` for packages and `/opt/cloudera/parcels/CDH/lib/solr/templateName` for parcels. The templates can be uploaded to ZooKeeper using `instancedir` commands such as `solrctl instancedir --create templateName /path/to/templateName`, although they are not protected by ZooKeeper ACLs.

Solr ZooKeeper ACLs Are Not Automatically Applied to Existing ZNodes

As of CDH 5.4, in Kerberos-enabled environments, ZooKeeper ACLs restrict access to Solr metadata stored in ZooKeeper to the solr user. This metadata cannot be modified by other users. These ACLs that limit access to the solr user are only applied automatically to new znodes.

This protection is not automatically applied to existing deployments.

To enable Solr ZooKeeper ACLs without retaining the existing cluster's Solr state, remove the solr znodes and reinitialize solr.

To remove solr znodes and reinitialize solr:

1. Using the `zookeeper-client`, enter the command `rmr /solr`.
2. Reinitialize Solr:

- Select **Initialize Solr** in Cloudera Manager *OR*
- Use `solrctl init`

To enable Solr ZooKeeper ACLs while retaining the existing cluster's Solr state, manually modify the existing znode's ACL information. For example, using `zookeeper-client`, run the command `setAcl [path]`

`sasl:solr:cdrwa,world:anyone:r`. This grants the solr user ownership of the specified path. Run this command for /solr and every znode under /solr except for the configuration znodes under and including /solr/configs.

HBase Indexer ACLs Are Not Automatically Applied to Existing ZNodes

As of CDH 5.4, in Kerberos-enabled environments, ZooKeeper ACLs restrict access to Lily HBase Indexer metadata stored in ZooKeeper to hbase user. This metadata cannot be modified by other users. These ACLs that limit access to the hbase user are only applied automatically to new znodes.

This protection is not automatically applied to existing deployments.

To enable Lily HBase Indexer ACLs without retaining the existing cluster's Lily HBase Indexer state, turn off the Lily HBase Indexer, remove the hbase-indexer znodes, and then restart the Lily HBase Indexer.

To remove hbase-indexer znodes and reinitialize Lily HBase Indexer:

1. In Cloudera Manager, click



to the right of the Lily HBase Indexer service and select **Stop**.

2. Using the zookeeper-client, enter the command `rmr /ngdata`.

3. In Cloudera Manager, click



to the right of the Lily HBase Indexer service and select **Start**.

The Lily HBase Indexer automatically creates all required znodes when it is started.

To enable Lily HBase Indexer while retaining the existing HBase-Indexer state, manually modify the existing znode's ACL information. For example, using zookeeper-client, run the command `setAcl [path]sasl:hbbase:cdrwa,world:anyone:r`. This grants the hbase user ownership of every znode under /ngdata (inclusive of /ngdata).



Note: This operation is not recursive, so creating a simple script may be helpful.

CrunchIndexerTool which includes Spark indexer requires specific input file format specifications

If the `--input-file-format` option is specified with CrunchIndexerTool then its argument must be `text`, `avro`, or `avroParquet`, rather than a fully qualified class name.

Previously deleted empty shards may reappear after restarting the leader host

It is possible to be in the process of deleting a collection when hosts are shut down. In such a case, when hosts are restarted, some shards from the deleted collection may still exist, but be empty.

Workaround: To delete these empty shards, manually delete the folder matching the shard. On the hosts on which the shards exist, remove folders under `/var/lib/solr/` that match the collection and shard. For example, if you had an empty shard 1 and empty shard 2 in a collection called `MyCollection`, you might delete all folders matching `/var/lib/solr/MyCollection{1,2}_replica*/`.

The quickstart.sh file does not validate ZooKeeper and the NameNode on some operating systems

The `quickstart.sh` file uses the `timeout` function to determine if ZooKeeper and the NameNode are available. To ensure this check can be complete as intended, the `quickstart.sh` determines if the operating system on which the script is running supports `timeout`. If the script detects that the operating system does not support `timeout`, the script continues without checking if the NameNode and ZooKeeper are available. If your environment is configured properly or you are using an operating system that supports `timeout`, this issue does not apply.

Workaround: This issue only occurs in some operating systems. If timeout is not available, a warning is displayed, but the quickstart continues and final validation is always done by the MapReduce jobs and Solr commands that are run by the quickstart.

Field value class guessing and Automatic schema field addition are not supported with the MapReduceIndexerTool nor the HBaseMapReduceIndexerTool

The MapReduceIndexerTool and the HBaseMapReduceIndexerTool can be used with a Managed Schema created via NRT indexing of documents or via the Solr Schema API. However, neither tool supports adding fields automatically to the schema during ingest.

Workaround: Define the schema before running the MapReduceIndexerTool or HBaseMapReduceIndexerTool. In non-schemaless mode, define in the schema using the schema.xml file. In schemaless mode, either define the schema using the Solr Schema API or index sample documents using NRT indexing before invoking the tools. In either case, Cloudera recommends that you verify that the schema is what you expect using the List Fields API command.

The “Browse” and “Spell” Request Handlers are not enabled in schemaless mode

The “Browse” and “Spell” Request Handlers require certain fields be present in the schema. Since those fields cannot be guaranteed to exist in a Schemaless setup, the “Browse” and “Spell” Request Handlers are not enabled by default.

Workaround: If you require the “Browse” and “Spell” Request Handlers, add them to the solrconfig.xml configuration file. Generate a non-schemaless configuration to see the usual settings and modify the required fields to fit your schema.

Using Solr with Sentry may consume more memory than required

The sentry-enabled solrconfig.xml.secure configuration file does not enable the hdfs global block cache. This does not cause correctness issues, but it can greatly increase the amount of memory that solr requires.

Workaround: Enable the hdfs global block cache, by adding the following line to solrconfig.xml.secure under the directoryFactory element:

```
<str name="solr.hdfs.blockcache.global">${solr.hdfs.blockcache.global: true}</str>
```

Enabling blockcache writing may result in unusable indexes

It is possible to create indexes with solr.hdfs.blockcache.write.enabled set to true. Such indexes may appear corrupt to readers, and reading these indexes may irrecoverably corrupt indexes. Blockcache writing is disabled by default.

Workaround: Do not enable blockcache writing.

Solr fails to start when Trusted Realms are added for Solr into Cloudera Manager

Cloudera Manager generates name rules with spaces as a result of entries in the Trusted Realms, which do not work with Solr. This causes Solr to not start.

Workaround: Do not use the Trusted Realm field for Solr in Cloudera Manager. To write your own name rule mapping, add an environment variable SOLR_AUTHENTICATION_KERBEROS_NAME_RULES with the mapping. See the [Cloudera Manager Security Guide](#) for more information.

Lily HBase batch indexer jobs fail to launch

A symptom of this issue is an exception similar to the following:

```
Exception in thread "main" java.lang.IllegalAccessError: class com.google.protobuf.ZeroCopyLiteralByteString cannot access its superclass com.google.protobuf.LiteralByteString
  at java.lang.ClassLoader.defineClass1(Native Method)
  at java.lang.ClassLoader.defineClass(ClassLoader.java:792)
```

```

at java.security.SecureClassLoader.defineClass(SecureClassLoader.java:142)
at java.net.URLClassLoader.defineClass(URLClassLoader.java:449)
at java.net.URLClassLoader.access$100(URLClassLoader.java:71)
at java.net.URLClassLoader$1.run(URLClassLoader.java:361)
at java.net.URLClassLoader$1.run(URLClassLoader.java:355)
at java.security.AccessController.doPrivileged(Native Method)
at java.net.URLClassLoader.findClass(URLClassLoader.java:354)
at java.lang.ClassLoader.loadClass( ClassLoader.java:424)
at java.lang.ClassLoader.loadClass( ClassLoader.java:357)
at org.apache.hadoop.hbase.protobuf.ProtobufUtil.toScan(ProtobufUtil.java:818)
at
org.apache.hadoop.hbase.mapreduce.TableMapReduceUtil.convertScanToString(TableMapReduceUtil.java:433)

at
org.apache.hadoop.hbase.mapreduce.TableMapReduceUtil.initTableMapperJob(TableMapReduceUtil.java:186)

at
org.apache.hadoop.hbase.mapreduce.TableMapReduceUtil.initTableMapperJob(TableMapReduceUtil.java:147)

at
org.apache.hadoop.hbase.mapreduce.TableMapReduceUtil.initTableMapperJob(TableMapReduceUtil.java:270)

at
org.apache.hadoop.hbase.mapreduce.TableMapReduceUtil.initTableMapperJob(TableMapReduceUtil.java:100)

at
com.ngdata.hbaseindexer.mr.HBaseMapReduceIndexerTool.run(HBaseMapReduceIndexerTool.java:124)

at
com.ngdata.hbaseindexer.mr.HBaseMapReduceIndexerTool.run(HBaseMapReduceIndexerTool.java:64)

at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:70)
at
com.ngdata.hbaseindexer.mr.HBaseMapReduceIndexerTool.main(HBaseMapReduceIndexerTool.java:51)

at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)

at java.lang.reflect.Method.invoke(Method.java:606)
at org.apache.hadoop.util.RunJar.main(RunJar.java:212)

```

This is because of an optimization introduced in [HBASE-9867](#) that inadvertently introduced a classloader dependency. In order to satisfy the new classloader requirements, `hbase-protocol.jar` must be included in Hadoop's classpath. This can be resolved on a per-job launch basis by including it in the `HADOOP_CLASSPATH` environment variable when you submit the job.

Workaround: Run the following command before issuing Lily HBase MapReduce jobs. Replace the .jar file names and filepaths as appropriate.

```
$ export HADOOP_CLASSPATH=</path/to/hbase-protocol>.jar; hadoop jar <MyJob>.jar
<MyJobMainClass>
```

Users may receive limited error messages on requests in Sentry-protected environment.

Users submit requests which are received by a host. The host that receives the request may be different from the host with the relevant information. In such a case, Solr forwards the request to the appropriate host. Once the correct host receives the request, Sentry may deny access.

Because the request was forwarded, available information may be limited. In such a case, the user's client display the error message Server returned HTTP response code: 401 for URL: followed by the Solr machine reporting the error.

Workaround: For complete error information, review the contents of the Solr logs on the machine reporting the error.

CDH 5 Release Notes

Users with insufficient Solr permissions may receive a "Page Loading" message from the Solr Web Admin UI

Users who are not authorized to use the Solr Admin UI are not given page explaining that access is denied, and instead receive a web page that never finishes loading.

Workaround: None

Using MapReduceIndexerTool or HBaseMapReduceIndexerTool multiple times may produce duplicate entries in a collection.

Repeatedly running the MapReduceIndexerTool on the same set of input files can result in duplicate entries in the Solr collection. This occurs because the tool can only insert documents and cannot update or delete existing Solr documents.

Workaround: To avoid this issue, use HBaseMapReduceIndexerTool with zero reducers. This must be done without Kerberos.

Deleting collections may fail if hosts are unavailable.

It is possible to delete a collection when hosts that host some of the collection are unavailable. After such a deletion, if the previously unavailable hosts are brought back online, the deleted collection may be restored.

Workaround: Ensure all hosts are online before deleting collections.

Lily HBase Indexer is slow to index new data after restart.

After restarting the Lily HBase Indexer, you can add data to one of the HBase tables. There may be a delay of a few minutes before this newly added data appears in Solr. This delay only occurs with a first HBase addition after a restart. Similar subsequent additions are not subject to this delay.

Workaround: None

Some configurations for Lily HBase Indexers cannot be modified after initial creation.

Newly created Lily HBase Indexers define their configuration using the properties in /etc/hbase-solr/conf/hbase-indexer-site.xml. Therefore, if the properties in the hbase-indexer-site.xml file are incorrectly defined, new indexers do not work properly. Even after correcting the contents of hbase-indexer-site.xml and restarting the indexer service, old, incorrect content persists. This continues to create non-functioning indexers.

Workaround:



Warning: This workaround involves completing destructive operations that delete all of your other Lily HBase Indexers.

To resolve this issue:

1. Connect to each machine running the Lily HBase Indexer service and stop the indexer:

```
service hbase-solr-indexer stop
```



Note: You may need to stop the service on multiple machines.

2. For each indexer machine, modify the /etc/hbase-solr/conf/hbase-indexer-site.xml file to include valid settings.

3. Connect to the ZooKeeper machine, invoke the ZooKeeper CLI, and remove all contents of the /ngdata chroot:

```
$ /usr/lib/zookeeper/bin/zkCli.sh  
[zk: localhost:2181( CONNECTED ) 0] rmr /ngdata
```

4. Connect to each indexer machine and restart the indexer service.

```
service hbase-solr-indexer start
```

After restarting the client services, ZooKeeper is updated with the correct information stored on the updated clients.

Saving search results is not supported.

This version of Cloudera Search does not support the ability to save search results.

Workaround: None

HDFS Federation is not supported.

This version of Cloudera Search does not support HDFS Federation.

Workaround: None

Block Cache Metrics are not supported.

This version of Cloudera Search does not support block cache metrics.

Workaround: None

User with update access to the administrative collection can elevate the access.

Users are granted access to collections. Access to several collections can be simplified by aliasing a set of collections. Creating an alias requires update access to the administrative collection. Any user with update access to the administrative collection is granted query access to all collections in the resulting alias. This is true even if the user with update access to the administrative collection otherwise would be unable to query the other collections that have been aliased.

Workaround: None. Mitigate the risk by limiting the users with update access to the administrative collection.

Apache Sentry Known Issues

CREATE FUNCTION ... USING JAR does not work on Sentry-secured clusters

In a cluster without Sentry, a user is able to create a UDF using the CREATE FUNCTION ... USING <hdfs location> command in Hive, with a JAR located on HDFS. However, once Sentry is enabled, this command does not work even if the user is granted the ALL privilege to the URI on HDFS.

Affected Versions: CDH 5.7, 5.6, 5.5, 5.4

With Sentry enabled, only Hive admin users have access to YARN job logs

As a prerequisite of enabling Sentry, Hive impersonation is turned off, which means all YARN jobs are submitted to the Hive job queue, and are run as the hive user. This is an issue because the YARN History Server now has to block users from accessing logs for their own jobs, since their own usernames are not associated with the jobs. As a result, end users cannot access any job logs unless they can get sudo access to the cluster as the hdfs, hive or other admin users.

In CDH 5.8 (and higher), Hive overrides the default configuration, mapred.job.queuename, and places incoming jobs into the connected user's job queue, even though the submitting user remains hive. Hive obtains the relevant queue/username information for each job by using YARN's fair-scheduler.xml file.

Affected Versions: CDH 5.7 and lower

Fixed Versions: CDH 5.8

CDH 5 Release Notes

Moving a partitioned table to a new location on the filesystem does not affect ACLs set on the previous location

With HDFS/Sentry sync enabled, if you move a partitioned table to a new location on the filesystem using the `ALTER TABLE .. SET LOCATION` command, ACLs set on the previous location remain unchanged. This occurs irrespective of whether the table is managed by Sentry.

Bug: [SENTRY-1373](#)

Column-level privileges are not supported on Hive Metastore views

`GRANT` and `REVOKE` for column level privileges is not supported on Hive Metastore views.

Bug: [SENTRY-754](#)

`SELECT` privilege on all columns does not equate to `SELECT` privilege on table

Users who have been explicitly granted the `SELECT` privilege on all columns of a table, will *not* have the permission to perform table-level operations. For example, operations such as `SELECT COUNT (1)` or `SELECT COUNT (*)` will not work even if you have the `SELECT` privilege on all columns.

There is one exception to this. The `SELECT * FROM TABLE` command will work even if you do not have explicit table-level access.

Bug: [SENTRY-838](#)

The EXPLAIN SELECT operation works without table or column-level privileges

Users are able to run the `EXPLAIN SELECT` operation, exposing metadata for all columns, even for tables/columns to which they weren't explicitly granted access.

Bug: [SENTRY-849](#)

With HDFS sync enabled, unexpected directory permissions are set when the NameNode plugin cannot communicate with the Sentry Server

With HDFS-Sentry sync enabled, if the NameNode plugin is unable to communicate with the Sentry Service for a particular period of time (configurable by the `sentry.authorization-provider.cache-stale-threshold.ms` property), permissions for *all* directories under Sentry-managed path prefixes, irrespective of whether those file paths correspond to Hive warehouse objects, will be set to `hive:hive`.

Hive authorization (Grant/Revoke>Show) statements do not support fully qualified table names (`default.tab1`)

Bug: None

Workaround: Switch to the database before granting privileges on the table.

Object types Server and URI are not supported in `SHOW GRANT ROLE roleName on OBJECT objectName`

Bug: None

Workaround: Use `SHOW GRANT ROLE roleName` to list all privileges granted to the role.

Relative URI paths not supported by Sentry

Sentry supports only absolute (not relative) URI paths in permission grants. Although some early releases (for example, CDH 5.7.0) may not have raised explicit errors when relative paths were set, upgrading a system that uses relative paths causes the system to lose Sentry permissions.

Affected Versions: All versions. Relative paths are not supported in Sentry for permission grants.

Resolution: Revoke privileges that have been set using relative paths, and grant permissions using absolute paths before upgrading.

Absolute (Use this form)	Relative (Do not use this form)
<code>hdfs://absolute/path/</code>	<code>hdfs://relative/path</code>

Absolute (Use this form)	Relative (Do not use this form)
s3a://bucketname/	s3a://bucketname

Apache Spark Known Issues

[Apache Spark experimental features are not supported unless specifically identified as supported](#)

If an Apache Spark feature or API is identified as experimental, in general Cloudera does not provide support for it.

[Certain Spark Streaming features not supported](#)

- The `mapWithState` method is unsupported because it is a nascent unstable API

[Certain Spark SQL features not supported](#)

The following Spark SQL features are not supported:

- Thrift JDBC/ODBC server
- Spark SQL CLI

[Spark Dataset API not supported](#)

Cloudera does not support the Spark Dataset API.

[GraphX not supported](#)

Cloudera does not support GraphX.

[SparkR not supported](#)

Cloudera does not support SparkR.

[Scala 2.11 not supported](#)

Cloudera does not support Spark on Scala 2.11 because it is binary incompatible, and also not yet full-featured.

[Spark Streaming cannot consume from secure Kafka till it starts using Kafka 0.9 Consumer API](#)

Bug: [SPARK-12177](#).

Workaround: None.

[Tables saved with the Spark SQL DataFrame.saveAsTable method are not compatible with Hive](#)

Writing a DataFrame directly to a Hive table creates a table that is not compatible with Hive; the metadata stored in the metastore can only be correctly interpreted by Spark. For example:

```
val hsc = new HiveContext(sc)
import hsc.implicits._
val df = sc.parallelize(data).toDF()
df.write.format("parquet").saveAsTable(tableName)
```

creates a table with this metadata:

```
inputFormat:org.apache.hadoop.mapred.SequenceFileInputFormat
outputFormat:org.apache.hadoop.hive.io.HiveSequenceFileOutputFormat
```

This is also occurs when using explicit schema, such as:

```
val schema = StructType(Seq(...))
val data = sc.parallelize(Seq(Row(...), ...))
val df = hsc.createDataFrame(data, schema)
df.write.format("parquet").saveAsTable(tableName)
```

Workaround: Explicitly create a Hive table to store the data. For example:

```
df.registerTempTable(tempName)
hsc.sql(s"""
CREATE TABLE $tableName (
// field definitions
)
STORED AS $format """)
hsc.sql(s"INSERT INTO TABLE $tableName SELECT * FROM $tempName")
```

Cannot create Parquet tables containing date fields in Spark SQL

If you create a Parquet table containing date field in Spark SQL, you see the following exception:

```
Exception in thread "main" org.apache.spark.sql.execution.QueryExecutionException:
FAILED: Execution Error, return code 1 from
org.apache.hadoop.hive.ql.exec.DDLTask.java.lang.UnsupportedOperationException: Parquet
does not support date.
at
org.apache.spark.sql.hive.client.ClientWrapper$$anonfun$runHive$1.apply(ClientWrapper.scala:433)
```

This is due to a limitation ([HIVE-6384](#)) in the version of Hive (1.1) included in CDH 5.5.0.

Spark SQL does not support the union type

Tables containing union fields cannot be read or created using Spark SQL.

Spark SQL does not respect size limit for the varchar type

Spark SQL treats `varchar` as string (that is, there no size limit). The observed behavior is that Spark reads and writes these columns as regular strings; if inserted values exceed the size limit, no error will occur. The data will be truncated when read from Hive, but not when read from Spark.

Bug: [SPARK-5918](#)

Spark SQL does not support the char type

Spark SQL does not support the `char` type (fixed-length strings). Like unions, tables with such fields cannot be created from or read by Spark.

Spark SQL does not support transactional tables

Spark SQL does not support Hive transactions ("ACID").

Spark SQL does not prevent you from writing key types not supported by Avro tables

Spark allows you to declare DataFrames with any key type. Avro supports only string keys and trying to write any other key type to an Avro table will fail.

Spark SQL does not support timestamp in Avro tables

Spark SQL does not support all 'ANALYZE TABLE COMPUTE STATISTICS' syntax

`ANALYZE TABLE <table name> COMPUTE STATISTICS NOSCAN` works. `ANALYZE TABLE <table name> COMPUTE STATISTICS` (without noscan) and `ANALYZE TABLE <table name> COMPUTE STATISTICS FOR COLUMNS` both return errors.

Spark SQL statements that can result in table partition metadata changes may fail

Because Spark does not have access to Sentry data, it may not know that a user has permissions to execute an operation and instead fail it. SQL statements that can result in table partition metadata changes, for example, "ALTER TABLE" or "INSERT", may fail.

Spark SQL does not respect Sentry ACLs when communicating with Hive metastore

Even if user is configured via Sentry to not have read permission to a Hive table, a Spark SQL job running as that user can still read the table's metadata directly from the Hive metastore.

Dynamic allocation and Spark Streaming

If you are using Spark Streaming, Cloudera recommends that you disable dynamic allocation by setting `spark.dynamicAllocation.enabled` to `false` when running streaming applications.

Spark uses Akka version 2.2.3

The CDH 5.5 version of Spark 1.5 differs from the Apache Spark 1.5 release in using Akka version 2.2.3, the version used by Spark 1.1 and CDH 5.2. Apache Spark 1.5 uses Akka version 2.3.11.

Spark standalone mode does not work on secure clusters

Workaround: On secure clusters, run Spark applications on YARN.

Apache Sqoop Known Issues

MySQL JDBC driver shipped with CentOS 6 systems does not work with Sqoop

CentOS 6 systems currently ship with version 5.1.17 of the MySQL JDBC driver. This version does not work correctly with Sqoop.

Bug: None

Resolution: Use workaround.

Workaround: Install version 5.1.31 of the JDBC driver, following directions in (Sqoop 1) or (Sqoop 2).

MS SQL Server "integratedSecurity" option unavailable in Sqoop

The `integratedSecurity` option is not available in the Sqoop CLI.

Bug: None

Resolution: None

Workaround: None

Sqoop 1

Hive, Pig, and Sqoop 1 fail in MRv1 tarball installation because `/usr/bin/hbase` sets `HADOOP_MAPRED_HOME` to `MR2`

This problem affects tarball installations only.

Bug: None

Resolution: Use workaround.

Workaround: If you are using MRv1, edit the following line in `/etc/default/hadoop` from

```
export HADOOP_MAPRED_HOME=/usr/lib/hadoop-mapreduce
```

to

```
export HADOOP_MAPRED_HOME=/usr/lib/hadoop-0.20-mapreduce
```

In addition, `/usr/lib/hadoop-mapreduce` must not exist in `HADOOP_CLASSPATH`.

Sqoop import into Hive Causes a Null Pointer Exception (NPE)

Bug: None

Workaround: Import the data into HDFS via Sqoop first and then import it into Hive from HDFS.

Sqoop 2

Sqoop 2 client cannot be used with a different version of the Sqoop 2 server

The Sqoop 2 client and server must be running the same CDH version.

Bug: None

Workaround: Make sure all Sqoop 2 components are running the same version of CDH.

[Sqoop 2 upgrade may fail if any job's source and destination links point to the same connector](#)

For example, the links for the job shown in the following output both point to generic-jdbc-connector:

```
sqoop:000> show job --all
1 job(s) to show:
Job with id 1 and name job1 (Enabled: true, Created by null at 5/13/15 3:05 PM, Updated
by null at 5/13/15 6:04 PM)
  Throttling resources
    Extractors:
    Loaders:
  From link: 1
    From database configuration
      Schema name: schemal
      Table name: tab1
      Table SQL statement:
      Table column names: col1
      Partition column name:
      Null value allowed for the partition column: false
      Boundary query:
      Incremental read
      Check column:
      Last value:
  To link: 2
    To database configuration
      Schema name: schema2
      Table name: tab2
      Table SQL statement:
      Table column names: col2
      Stage table name:
      Should clear stage table:

sqoop:000> show link --all
2 link(s) to show:
link with id 1 and name try1 (Enabled: true, Created by null at 5/13/15 2:59 PM, Updated
by null at 5/13/15 5:47 PM)
Using Connector generic-jdbc-connector with id 2
  Link configuration
    JDBC Driver Class: com.mysql.jdbc.Driver
    JDBC Connection String: jdbc:mysql://mysql.server/database
    Username: nvaidya
    Password:
    JDBC Connection Properties:
link with id 2 and name try2 (Enabled: true, Created by null at 5/13/15 3:01 PM, Updated
by null at 5/13/15 5:47 PM)
Using Connector generic-jdbc-connector with id 2
  Link configuration
    JDBC Driver Class: com.mysql.jdbc.Driver
    JDBC Connection String: jdbc:mysql://mysql.server/database
    Username: nvaidya
    Password:
    JDBC Connection Properties:
```

Bug: None

Workaround: Before upgrading, make sure no jobs have source and destination links that point to the same connector.

Apache ZooKeeper Known Issues

[Adding New ZooKeeper Servers Can Lead to Data Loss](#)

When the number of new ZooKeeper servers exceeds the number that already exist in the ZooKeeper service (for example, if you increase the number of servers from 1 to 3), and a Start command is immediately issued to the ZooKeeper service, the new servers can form a quorum, which causes data loss in existing servers.

Users of the following versions of Cloudera Manager are affected:

5.0.0–5.0.5, 5.1.0–5.1.4, 5.2.0–5.2.4, and 5.3.0–5.3.2

Workaround: If you use a version of Cloudera Manager listed above, upgrade to the next available maintenance release with the bug fix (within the minor version), or to Cloudera Manager 5.4.

[The ZooKeeper server cannot be migrated from version 3.4 to 3.3, then back to 3.4, without user intervention.](#)

Upgrading from 3.3 to 3.4 is supported, as is downgrading from 3.4 to 3.3. However, moving from 3.4 to 3.3 and back to 3.4 will fail. 3.4 is checking the datadir for acceptedEpoch and currentEpoch files and comparing these against the snapshot and log files contained in the same directory. These epoch files are new in 3.4.

As a result: 1) Upgrading from 3.3 to 3.4 is fine - the *Epoch files do not exist, and the server creates them. 2) Downgrading from 3.4 to 3.3 is also fine as version 3.3 ignores the *Epoch files. 3) Going from 3.4 to 3.3 then back to 3.4 fails because 3.4 sees invalid *Epoch files in the datadir; 3.3 will have ignored them, applying changes to the snapshot and log files without updating the *Epoch files.

Bug: [ZOOKEEPER-1149](#)

Anticipated Resolution: See workaround

Workaround: Delete the *Epoch files if this situation occurs — the version 3.4 server will recreate them as in case 1 above.

Issues Fixed in CDH 5



Note: For links to the detailed change lists that describe the bug fixes and improvements to all of the CDH 5 projects, including bug-fix reports for the corresponding upstream Apache projects, see the packaging section of [CDH Version and Packaging Information](#).

Issues Fixed in CDH 5.9.x

The following topics describe issues fixed in CDH 5.9.x, from newest to oldest release. You can also review [What's New In CDH 5.9.x](#) on page 14 or [Known Issues in CDH 5](#) on page 111.

Issues Fixed in CDH 5.9.0

CDH 5.9.0 fixes the following issues.

Apache Flume

Use SourceCounter for SyslogTcpSource

Bug: [FLUME-2797](#)

SyslogTcpSource uses a deprecated class. Use the newer SouceCounter class for SyslogTcpSource, and mark SyslogTcpSouce as deprecated.

SpillableMemoryChannel must start ChannelCounter

Bug: [FLUME-2844](#)

When using SpillableMemoryChannel, a bug causes the values of all metrics of channel component monitoring system to be zero.

CDH 5 Release Notes

Add localhost escape sequence to HDFS sink

Bug: [FLUME-2982](#)

The HDFS sink should just use localhost escape sequence instead of having to pass in a header and use the host interceptor.

Apache Hadoop

When removing a stored block, use stored BlockInfo for updating UnderReplicatedBlocks

During the removal of a stored block, BlockManager UnderReplicatedBlocks needs to be updated with the stored BlockInfo instead of the incoming Block, so that the Block with the correct generation stamp is used to calculate any pending replication that needs to be triggered.

Fix a race condition in MetricsSourceAdapter.updateJmxCache

Bug: [HADOOP-11361](#)

KMS Server should log exceptions before throwing

Bug: [HADOOP-13669](#)[HADOOP-13317](#)

When KMS throws an exception, it is not logged anywhere, and the exception message can only be seen from the client side, not the stacktrace

LdapGroupsMapping getPassword should not return null when IOException throws

Bug: [HADOOP-13353](#)

When IOException throws in getPassword(), getPassword() returns a null string. This causes setConf() to throw a java.lang.NullPointerException.

Add detailed logging in KMS for the authentication failure of a proxy user

Bug: [HADOOP-13526](#)

The log message from AuthenticationFilter.java indicates that the user was successfully authenticated. However, when the filter on DelegationTokenAuthenticationFilter is called, it hits an exception and there is no log message

UserGroupInformation created from a Subject incorrectly tries to renew the Kerberos ticket

Bug: [HADOOP-13558](#)

KMS should set UGI Configuration object properly

Bug: [HADOOP-13638](#)

The Configuration object in UGI in KMS server is not initialized properly, so it does not load core-site.xml from KMSConfiguration.KMS_CONFIG_DIR.

Throw helpful exception when DNS entry for JournalNode cannot be resolved

Bug: [HDFS-4210](#)

FSDirectory and FSNameSystem issues

Bug: [HDFS-7415](#) [HDFS-7420](#) [HDFS-7463](#) [HDFS-7478](#) [HDFS-7517](#) [HDFS-8269](#)

Schedule a block for scanning if its metadata file is corrupt

Bug: [HDFS-8224](#)

A block with corrupt metadata is not scheduled for scanning by BlockPoolSliceScanner, so it is not reported as corrupt.

DiskBalancer: Use SHA1 for plan ID

Bug: [HDFS-10559](#)

SHA1 should be used as the plan ID instead of Sha512 because much shorter and easier to handle.

Improve plan command help message

Bug: [HDFS-10567](#)

The help message for the plan command needs to be clarified and corrected.

PlanCommand#getThrsholdPercentage should not use throughput value

Bug: [HDFS-10600](#)

Namenode should use loginUser(hdfs) to generateEncryptedKey

Bug: [HDFS-10643](#)

HDFS namenode login user (hdfs) should always be used when talking to KMS to generateEncryptedKey for new file creation.

DataXceiver#run() should not log InvalidToken exception as an error

Bug: [HDFS-10760](#)

When the client has an expired token and refetches a new token, the DataNode logs an error.

Log DataNodes in the write pipeline

Bug: [HDFS-10822](#)

You cannot tell which DataNodes are involved in the write pipeline. A DEBUG trace should be added to print the list of DataNodes in the pipeline.

Reduce log level when network topology cannot find enough DataNodes

Bug: [HDFS-10963](#)

The warning-level log event should be reduced to help prevent unnecessary concern about the event.

Unnecessary INFO logging on DFSClients for InvalidToken

Bug: [HDFS-11012](#)

InvalidBlockTokenException should be changed to DEBUG to help prevent unnecessary concern about the event.

Potential memory leak in CryptoOutputStream

Bug: [MAPREDUCE-6628](#)

A potential memory leak in CryptoOutputStream.java allocates two direct byte buffers (inBuffer and outBuffer) that are freed when close() method is called.

Add progress log to JHS during startup

Bug: [MAPREDUCE-6718](#)

When the JHS starts up, it initializes the internal caches and storage through the HistoryFileManager. If a large number of finished jobs exist, this startup phase can last minutes without logging progress.

RMContainerAllocator sends container diagnostics event after corresponding completion event

Bug: [MAPREDUCE-6771](#)

Diagnostics information may never get into .jhst file, so when the job completes, the diagnostics information associated with the failed task attempts is empty.

yarn node -list -all fails if ResourceManager starts with decommissioned node

Bug: [YARN-4940](#)

Apache HBase

Superuser does not consider the keytab credentials

Bug: [HBASE-15622](#)

The superuser added by default (the process running HBase) does not take in consideration the keytab credential.

CDH 5 Release Notes

Do not cache unresolved addresses for connections

Bug: [HBASE-15856](#)

During periods where DNS is not working properly, caching can occur for connections to Master or RegionServers where the initial hostname resolution and the resolution is never reattempted. This causes clients to get UnknownHostException for any calls.

Overflows in AverageIntervalRateLimiter's refill() and getWaitInterval()

Bug: [HBASE-16699](#)

Mob compaction needs to clean up files in /hbase/mobdir/.tmp and /hbase/mobdir/.tmp/.bulkload when running into IO exceptions

Bug: [HBASE-16767](#)

Apache Hive

WebUI Elapsed Time may mislead users to mean query run time

Bug: [HIVE-13420](#)

When running queries from Hue, WebUI's elapsed time continuously increases.

Table not found error when trying to refresh column statistics in Metastore Manager

Bug: [HIVE-10007](#)

When you try to refresh column statistics in the Metastore Manager, it fails with the following error message:

```
Error while compiling statement: FAILED: SemanticException  
[Error 10001]: Line 1:157 Table not found ''
```

StorageBasedAuthorizationProvider requires write permission on table for SELECT

Bug: [HIVE-11901](#)

StorageBasedAuthorizationProvider requires write permission on table for SELECT statements.

Subquery inside a view will have the object in the subquery as the direct input

Bug: [HIVE-14805](#)

Hive query fails from HUE for some users

Bug: [HIVE-14784](#)

Operation logging is disabled automatically for the query if for some reason the parent directory (named after the Hive session ID) created when the session is established gets deleted.

Hive throws NumberFormatException with query with Null value

Bug: [HIVE-14715](#)

Hue

[editor] Improved autocomplete

Bug: [HUE-4039](#)

[editor] Improve result table scroll performance and fix header positioning

Bug: [HUE-4910](#)

[editor] Progress status and truncating warning when direct downloading results as Excel

Bug: [HUE-4438](#)

[fb] Allow to browse S3 (other filesystems)

Bug: [HUE-2915](#)

[editor] Export query result to S3

Bug: [HUE-4367](#)

[metastore] UX create table from a file, S3 or from a directory

Bug: [HUE-4425](#)

[core] Missing some security related response headers

Bug: [HUE-4372](#)

Wildcard Certificates not supported

Bug: None

Hue now supports wildcard certificates and certificates using Subject Alternative Name (SAN).

Cookie without HttpOnly flag set

Bug: None

The HttpOnly flag is now in the response header, preventing cookies from being accessed through a client side script.

Password reveal should be disabled

Bug: None

The Password Reveal button is now disabled on Internet Explorer.

Web server and platform version disclosure should be prevented

Bug: None

Web server and platform versions were removed from the response headers.

Apache Oozie

Oozie should mask any passwords in logs and REST interfaces

Bug: [OOZIE-1814](#)

The following passwords are currently visible in the instrumentation log, REST endpoints, WebUI, and CLI and should be masked:

```
javax.net.ssl.trustStorePassword
oozie.https.keystore.pass
HADOOP_CREDSTORE_PASSWORD
OOZIE_HTTPS_KEYSTORE_PASSWORD
OOZIE_HTTPS_TRUSTSTORE_PASSWORD
```

Cloudera Search

Crunchindexertool YARN job reports success even when indexing fails

When running a Spark indexer job that fails due to shards being down, Kerberos failures, or other problems), the YARN final status is still reported as SUCCEEDED when it should be failed.

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.9.0:

- [FLUME-2797](#) - Use SourceCounter for SyslogTcpSource
- [FLUME-2844](#) - SpillableMemoryChannel must start ChannelCounter
- [FLUME-2909](#) - Upgrade RAT to 0.11
- [FLUME-2920](#) - Kafka Channel Should Not Commit Offsets When Stopping
- [HADOOP-11031](#) - Design Document for Credential Provider API
- [HADOOP-11149](#) - Increase the timeout of TestZKFailoverController
- [HADOOP-11180](#) - Change log message "token.Token: Cannot find class for token kind kms-dt" to debug
- [HADOOP-11814](#) - Reformat hadoop-annotations, o.a.h. classification.tools

CDH 5 Release Notes

- [HADOOP-11872](#) - hadoop dfs command prints message about using "yarn jar" on Windows (branch-2 only)
- [HADOOP-12465](#) - Incorrect Javadoc in WritableUtils.java
- [HADOOP-12537](#) - S3A to support Amazon STS temporary credentials
- [HADOOP-12548](#) - Read s3a creds from a Credential Provider
- [HADOOP-12589](#) - Fix intermittent test failure of TestCopyPreserveFlag
- [HADOOP-12613](#) - TestFind.processArguments occasionally fails
- [HADOOP-12636](#) - Prevent ServiceLoader failure init for unused FileSystems
- [HADOOP-12723](#) - S3A: Add ability to plug in any AWS Credentials Provider
- [HADOOP-12800](#) - Copy docker directory from 2.8 to 2.7/2.6 repos to enable pre-commit Jenkins runs
- [HADOOP-12847](#) - Hadoop daemonlog should support https and SPNEGO for Kerberized cluster
- [HADOOP-12893](#) - Verify LICENSE.txt and NOTICE.txt
- [HADOOP-12928](#) - Update netty to 3.10.5.Final to sync with ZooKeeper
- [HADOOP-12991](#) - Conflicting default ports in DelegateToFileSystem
- [HADOOP-13098](#) - Dynamic LogLevel setting page should accept case-insensitive log level string
- [HADOOP-13103](#) - Group resolution from LDAP may fail on javax.naming.ServiceUnavailableException
- [HADOOP-13154](#) - S3AFileSystem printAmazonServiceException/printAmazonClientException appear copy and paste of AWS examples
- [HADOOP-13189](#) - FairCallQueue makes callQueue larger than the configured capacity
- [HADOOP-13192](#) - org.apache.hadoop.util.LineReader cannot handle multibyte delimiters correctly
- [HADOOP-13202](#) - Avoid possible overflow in org.apache.hadoop.util.bloom.BloomFilter#getNBytes
- [HADOOP-13228](#) - Add delegation token to the connection in DelegationTokenAuthenticator
- [HADOOP-13254](#) - Create framework for configurable disk checkers
- [HADOOP-13290](#) - Appropriate use of generics in FairCallQueue
- [HADOOP-13297](#) - Add missing dependency in setting maven-remote-resource-plugin to fix builds
- [HADOOP-13298](#) - Fix the leftover L&N files in hadoop-build-tools/src/main/resources/META-INF/
- [HADOOP-13316](#) - Enforce Kerberos authentication for required ops in DelegationTokenAuthenticator
- [HADOOP-13350](#) - Additional fix to LICENSE and NOTICE
- [HADOOP-13351](#) - TestDFSClientSocketSize buffer size tests are flaky
- [HADOOP-13362](#) - DefaultMetricsSystem leaks the source name when a source unregisters
- [HADOOP-13380](#) - TestBasicDiskValidator should not write data to /tmp
- [HADOOP-13395](#) - Enhance TestKMSAudit
- [HADOOP-13443](#) - KMS should check the type of underlying keyprovider of KeyProviderExtension before falling back to default
- [HADOOP-13461](#) - NPE in KeyProvider.rollNewVersion
- [HADOOP-13476](#) - CredentialProviderFactory fails at class loading from libhdfs (JNI)
- [HADOOP-13494](#) - ReconfigurableBase can log sensitive information
- [HADOOP-13526](#) - Add detailed logging in KMS for the authentication failure of proxy user
- [HADOOP-13558](#) - UserGroupInformation created from a Subject incorrectly tries to renew the Kerberos ticket
- [HADOOP-13579](#) - Fix source-level compatibility after HADOOP-11252
- [HADOOP-13638](#) - KMS should set UGI's Configuration object properly
- [HDFS-2580](#) - NameNode#main(...) can make use of GenericOptionsParser
- [HDFS-4176](#) - EditLogTailer should call rollEdits with a timeout
- [HDFS-7415](#) - Move FSNameSystem.resolvePath() to FSDirectory
- [HDFS-7420](#) - Delegate permission checks to FSDirectory
- [HDFS-7463](#) - Simplify FSNamesystem#getBlockLocationsUpdateTimes
- [HDFS-7478](#) - Move org.apache.hadoop.hdfs.server.namenode.NNConf to FSNamesystem
- [HDFS-7517](#) - Remove redundant non-null checks in FSNamesystem#getBlockLocations
- [HDFS-7934](#) - Update RollingUpgrade rollback documentation: should use bootstrapstandby for standby NN
- [HDFS-8101](#) - DFSClient use of non-constant DFSCConfigKeys pulls in WebHDFS classes at runtime
- [HDFS-8224](#) - Schedule a block for scanning if its metadata file is corrupt

- [HDFS-8269](#) - getBlockLocations() does not resolve the .reserved path and generates incorrect edit logs when updating the atime
- [HDFS-8521](#) - Add VisibleForTesting annotation to BlockPoolSlice#selectReplicaToDelete
- [HDFS-8600](#) - TestWebHdfsFileSystemContract.testGetFileBlockLocations fails in branch-2.7
- [HDFS-8608](#) - Merge HDFS-7912 to trunk and branch-2 (track BlockInfo instead of Block in UnderReplicatedBlocks and PendingReplicationBlocks)
- [HDFS-8633](#) - Fix setting of dfs.datanode.readahead.bytes in hdfs-default.xml to match DFSConfigKeys
- [HDFS-8682](#) - Should not remove decommissioned node while calculating the number of live/dead decommissioned nodes
- [HDFS-8780](#) - Fetching live/dead datanode list with arg true for removeDecommissionNode, returns list with decom node
- [HDFS-9033](#) - dfsadmin -metasave prints "NaN" for cache used%
- [HDFS-9048](#) - DistCp documentation is out-of-date
- [HDFS-9413](#) - getContentSummary() on standby should throw StandbyException
- [HDFS-9530](#) - ReservedSpace is not cleared for abandoned Blocks
- [HDFS-9533](#) - seen_txid in the shared edits directory is modified during bootstrapping
- [HDFS-9601](#) - NNThroughputBenchmark.BlockReportStats should handle NotReplicatedYetException on adding block
- [HDFS-9696](#) - Garbage snapshot records linger forever
- [HDFS-9765](#) - TestBlockScanner#testVolumelIteratorWithCaching fails intermittently
- [HDFS-9781](#) - FsDatasetImpl#getBlockReports can occasionally throw NullPointerException
- [HDFS-10225](#) - DataNode hot swap drives should disallow storage type changes
- [HDFS-10245](#) - Fix the findbugs warnings in branch-2.7
- [HDFS-10335](#) - Mover\$Processor#chooseTarget() always chooses the first matching target storage group
- [HDFS-10336](#) - TestBalancer failing intermittently because of not resetting UserGroupInformation completely
- [HDFS-10347](#) - Namenode report bad block method doesn't log the bad block or datanode
- [HDFS-10458](#) - getFileEncryptionInfo should return quickly for non-encrypted cluster
- [HDFS-10474](#) - hftp copy fails when file name with Chinese+special char in branch-2
- [HDFS-10623](#) - Remove unused import of httpclient.HttpConnection from TestWebHdfsTokens
- [HDFS-10625](#) - VolumeScanner to report why a block is found bad
- [HDFS-10641](#) - TestBlockManager#testBlockReportQueueing fails intermittently
- [HDFS-10653](#) - Optimize conversion from path string to components
- [HDFS-10688](#) - BPServiceActor may run into a tight loop for sending block report when hitting IOException
- [HDFS-10691](#) - FileDistribution fails in hdfs oiv command due to ArrayIndexOutOfBoundsException
- [HDFS-10693](#) - metaSave should print blocks, not LightWeightHashSet
- [HDFS-10703](#) - HA NameNode Web UI should show last checkpoint time
- [HDFS-10716](#) - In Balancer, the target task should be removed when its size < 0
- [HDFS-10879](#) - TestEncryptionZonesWithKMS#testReadWrite fails intermittently
- [HDFS-10962](#) - TestRequestHedgingProxyProvider is flaky
- [HDFS-10963](#) - Reduce log level when network topology cannot find enough datanodes
- [MAPREDUCE-6242](#) - Progress report log is incredibly excessive in application master
- [MAPREDUCE-6259](#) - IllegalArgumentException due to missing job submit time
- [MAPREDUCE-6374](#) - Distributed Cache File visibility should check permission of full path
- [MAPREDUCE-6533](#) - testDetermineCacheVisibilities of TestClientDistributedCacheManager is broken
- [MAPREDUCE-6628](#) - Potential memory leak in CryptoOutputStream
- [MAPREDUCE-6633](#) - AM should retry map attempts if the reduce task encounters compression related errors
- [MAPREDUCE-6641](#) - TestTaskAttempt fails in trunk
- [MAPREDUCE-6652](#) - Add configuration property to prevent JHS from loading jobs with a task count greater than X
- [MAPREDUCE-6718](#) - Add progress log to JHS during startup
- [MAPREDUCE-6724](#) - Single shuffle to memory must not exceed Integer#MAX_VALUE

CDH 5 Release Notes

- [MAPREDUCE-6741](#) - Add MR support to redact job conf properties
- [MAPREDUCE-6751](#) - Add debug log message when splitting is not possible due to unsplittable compression
- [MAPREDUCE-6771](#) - RMContainerAllocator sends container diagnostics event after corresponding completion event
- [YARN-3212](#) - RMNode State Transition Update with DECOMMISSIONING state
- [YARN-3226](#) - UI changes for decommissioning node
- [YARN-3445](#) - Cache runningApps in RMNode for getting running apps on given Nodeld
- [YARN-4556](#) - TestFifoScheduler.testResourceOverCommit fails
- [YARN-4568](#) - Fix message when NodeManager runs into errors initializing the recovery directory
- [YARN-4702](#) - FairScheduler: Allow setting maxResources for ad hoc queues
- [YARN-4940](#) - yarn node -list -all failed if RM start with decommissioned node
- [YARN-5024](#) - TestContainerResourceUsage#testUsageAfterAMRestartWithMultipleContainers random failure
- [YARN-5343](#) - TestContinuousScheduling#testSortedNodes fails intermittently
- [YARN-5434](#) - Add -client|server argument for graceful decommission
- [YARN-5483](#) - Optimize RMAppAttempt#pullJustFinishedContainers
- [YARN-5549](#) - AMLauncher#createAMContainerLaunchContext() should not log the command to be launched indiscriminately
- [YARN-5566](#) - Client-side NM graceful decom is not triggered when jobs finish
- [YARN-5655](#) - TestContainerManagerSecurity#testNMTokens is asserting
- [HBASE-15396](#) - Enhance mapreduce.TableSplit to add encoded region name
- [HBASE-15856](#) - Don't cache unresolved addresses for connections
- [HBASE-15889](#) - String case conversions are locale-sensitive, used without locale
- [HBASE-15891](#) - Closeable resources potentially not getting closed if exception is thrown
- [HBASE-15946](#) - Eliminate possible security concerns in Store File metrics
- [HBASE-16023](#) - Fastpath for the FIFO rpcscheduler
- [HBASE-16035](#) - Nested AutoCloseables might not all get closed
- [HBASE-16096](#) - Backport. Cleanly remove replication peers from ZooKeeper
- [HBASE-16140](#) - Bump owasp.esapi from 2.1.0 to 2.1.0.1
- [HBASE-16294](#) - hbck reporting "No HDFS region dir found" for replicas
- [HBASE-16379](#) - [replication] Minor improvement to replication/copy_tables_desc.rb
- [HBASE-16450](#) - Shell tool to dump replication queues
- [HBASE-16699](#) - Overflows in AverageIntervalRateLimiter's refill() and getWaitInterval()
- [HBASE-16767](#) - Mob compaction needs to clean up files in /hbase/mobdir/.tmp and /hbase/mobdir/.tmp/.bulkload when running into IO exceptions
- [HIVE-4570](#) - Add more information to GetOperationStatus in HiveServer2 when query is still executing
- [HIVE-6758](#) - Beeline doesn't work with -e option when started in background
- [HIVE-9013](#) - Hive set command exposes metastore db password
- [HIVE-9302](#) - Beeline add commands to register local jdbc driver names and jars
- [HIVE-9570](#) - Investigate test failure on union_view
- [HIVE-9657](#) - Use new parquet Types API builder to construct data types
- [HIVE-10190](#) - CBO: AST mode checks for TABLESAMPLE with AST.toString().contains
- [HIVE-10485](#) - Create md5 UDF
- [HIVE-10624](#) - Update the initial script to make beeline bucked cli as default and allow user choose old hive cli by env
- [HIVE-10684](#) - Fix the unit test failures for HIVE-7553 after HIVE-10674 removed the binary jar files
- [HIVE-10705](#) - Update tests for HIVE-9302 after removing binaries
- [HIVE-10722](#) - External table creation with msck in Hive can create unusable partition
- [HIVE-10755](#) - Rework on HIVE-5193 to enhance the column oriented table access
- [HIVE-10824](#) - Need to update start script changes in .cmd files
- [HIVE-10904](#) - Use beeline-log4j.properties for migrated CLI [beeline-cli Branch]

- [HIVE-10965](#) - Direct SQL for stats fails in 0-column case
- [HIVE-11028](#) - Tez: table self join and join with another table fails with IndexOutOfBoundsException
- [HIVE-11226](#) - BeeLine-Cli: support hive.cli.prompt in new CLI
- [HIVE-11236](#) - BeeLine-Cli: use the same output format as old CLI in the new CLI
- [HIVE-11280](#) - Support executing script file from hdfs in new CLI [Beeline-CLI branch]
- [HIVE-11316](#) - Use datastructure that doesnt duplicate any part of string for ASTNode::toStringTree()
- [HIVE-11336](#) - Support initial file option for new CLI [beeline-cli branch]
- [HIVE-11352](#) - Avoid the double connections with 'e' option [beeline-cli branch]
- [HIVE-11375](#) - Broken processing of queries containing NOT (x IS NOT NULL and x <> 0)
- [HIVE-11490](#) - Lazily call ASTNode::toStringTree() after tree modification
- [HIVE-11624](#) - Beeline-cli: support hive.cli.print.header in new CLI [beeline-cli branch]
- [HIVE-11637](#) - Support hive.cli.print.current.db in new CLI [beeline-cli branch]
- [HIVE-11640](#) - Shell command doesn't work for new CLI [Beeline-CLI branch]
- [HIVE-11717](#) - nohup mode is not supported for new hive CLI
- [HIVE-11746](#) - Connect command should not to be allowed from user [beeline-cli branch]
- [HIVE-11796](#) - CLI option is not updated when executing the initial files [beeline-cli]
- [HIVE-11943](#) - Set old CLI as the default Client when using Hive script
- [HIVE-11944](#) - Address the review items on HIVE-11778
- [HIVE-11990](#) - Loading data inpath from a temporary table dir fails on Windows
- [HIVE-12018](#) - beeline --help doesn't return to original prompt
- [HIVE-12080](#) - Support auto type widening (int->bigint & float->double) for Parquet table
- [HIVE-12083](#) - HIVE-10965 introduces thrift error if partNames or colNames are empty
- [HIVE-12215](#) - Exchange partition does not show outputs field for post/pre execute hooks
- [HIVE-12246](#) - Orc FileDump fails with Missing CLI jar
- [HIVE-12259](#) - Command containing semicolon is broken in Beeline
- [HIVE-12345](#) - Followup for HIVE-9013 : Hidden conf vars still visible through beeline
- [HIVE-12475](#) - Parquet schema evolution within array<struct<>> does not work
- [HIVE-12590](#) - Repeated UDAFs with literals can produce incorrect result
- [HIVE-12721](#) - Add UUID built in function
- [HIVE-12785](#) - View with union type and UDF to the struct is broken
- [HIVE-12834](#) - Fix to accept the arrow keys in BeeLine CLI
- [HIVE-12983](#) - Provide a builtin function to get Hive version
- [HIVE-12987](#) - Add metrics for HS2 active users and SQL operations
- [HIVE-13058](#) - Add session and operation_log directory deletion messages
- [HIVE-13093](#) - Hive metastore does not exit on start failure
- [HIVE-13151](#) - Clean up UGI objects in FileSystem cache for transactions
- [HIVE-13198](#) - Authorization issues with cascading views
- [HIVE-13237](#) - Select parquet struct field with upper case throws NPE
- [HIVE-13381](#) - Timestamp and date should have precedence in type hierarchy than string group
- [HIVE-13420](#) - Clarify HS2 WebUI Query 'Elapsed Time'
- [HIVE-13462](#) - HiveResultSetMetaData.getPrecision() fails for NULL columns
- [HIVE-13502](#) - Beeline doesnt support session parameters in JDBC URL as documentation states
- [HIVE-13620](#) - Merge llap branch work to master
- [HIVE-13625](#) - Hive Prepared Statement when executed with escape characters in parameter fails
- [HIVE-13645](#) - Beeline needs null-guard around hiveVars and hiveConfVars read
- [HIVE-13670](#) - Improve Beeline connect/reconnect semantics
- [HIVE-13783](#) - Display a secondary prompt on beeline for multi-line statements
- [HIVE-13788](#) - hive msck listpartitions need to make use of directSQL instead of datanucleus
- [HIVE-13953](#) - Issues in HiveLockObject equals method
- [HIVE-13964](#) - Add a parameter to beeline to allow a properties file to be passed in

CDH 5 Release Notes

- [HIVE-13984](#) - Use multi-threaded approach to listing files for msck
- [HIVE-13987](#) - Clarify current error shown when HS2 is down
- [HIVE-13997](#) - Insert overwrite directory doesn't overwrite existing files
- [HIVE-14001](#) - Beeline doesn't give out an error when takes either "-e" or "-f" in command instead of both
- [HIVE-14013](#) - Describe table doesn't show unicode properly
- [HIVE-14037](#) - java.lang.ClassNotFoundException for the jar in hive.reloadable.aux.jars.path in mapreduce
- [HIVE-14074](#) - RELOAD FUNCTION should update dropped functions
- [HIVE-14075](#) - BeeLine.java.orig was accidentally committed during HIVE-14001 patch
- [HIVE-14085](#) - Allow type widening primitive conversion on hive/parquet tables
- [HIVE-14090](#) - JDOExceptions thrown by the Metastore have their full stack trace returned to clients
- [HIVE-14135](#) - Beeline output not formatted correctly for large column widths
- [HIVE-14137](#) - Hive on Spark throws FileAlreadyExistsException for jobs with multiple empty tables
- [HIVE-14149](#) - Joda Time causes an AmazonS3Exception on Hadoop3.0.0
- [HIVE-14151](#) - Use of USE_DEPRECATED_CLI environment variable does not work
- [HIVE-14153](#) - Beeline: beeline history doesn't work on Hive2
- [HIVE-14207](#) - Strip HiveConf hidden params in webui conf
- [HIVE-14215](#) - Displaying inconsistent CPU usage data with MR execution engine
- [HIVE-14226](#) - Invalid check on an ASTNode#toStringTree in CalcitePlanner
- [HIVE-14267](#) - HS2 open_operations metrics not decremented when an operation gets timed out
- [HIVE-14270](#) - Write temporary data to HDFS when doing inserts on tables located on S3
- [HIVE-14294](#) - HiveSchemaConverter for Parquet doesn't translate TINYINT and SMALLINT into proper Parquet types
- [HIVE-14296](#) - Session count is not decremented when HS2 clients do not shutdown cleanly
- [HIVE-14342](#) - Beeline output is garbled when executed from a remote shell
- [HIVE-14360](#) - Starting BeeLine after using !save, there is an error logged: "Error setting configuration: conf"
- [HIVE-14383](#) - SparkClientImpl should pass principal and keytab to spark-submit instead of calling kinit explicitly
- [HIVE-14426](#) - Extensive logging on info level in WebHCat
- [HIVE-14513](#) - Enhance custom query feature in LDAP atn to support resultset of ldap groups
- [HIVE-14588](#) - Add S3 credentials to the hidden configuration variable supported on HIVE-14207
- [HIVE-14715](#) - Hive throws NumberFormatException with query with Null value
- [HIVE-14743](#) - ArrayIndexOutOfBoundsException - HBASE-backed views' query with JOINs
- [HIVE-14784](#) - Operation logs are disabled automatically if the parent directory does not exist
- [HIVE-14799](#) - Query operation not thread safe during its cancellation
- [HIVE-14805](#) - Subquery inside a view will have the object in the subquery as the direct input
- [HIVE-14848](#) - PROPOSEDS3 creds added to a hidden list by HIVE-14588 are not working on MR jobs
- [HIVE-14889](#) - Beeline leaks sensitive environment variables of HiveServer2 when you type set
- [HUE-2645](#) - [oozie] Better Spark action UX
- [HUE-2975](#) - [aws] Add S3 UploadFileHandler and implement upload to S3
- [HUE-3065](#) - [desktop] Fix test_redact_statements
- [HUE-3065](#) - [desktop] Allow searching of raw SQL in saved queries
- [HUE-3227](#) - [connector] Java snippet
- [HUE-3278](#) - [core] Migrate Google Analytics from ga.js to analytics.js and use https if possible
- [HUE-3308](#) - [spark] Migrate Livy to external repository
- [HUE-3324](#) - [editor] Create initial json parser
- [HUE-3325](#) - [editor] Generate the parser using the makefile
- [HUE-3326](#) - [editor] Introduce the cursor and add suggestions for possible statements
- [HUE-3327](#) - [editor] Suggest tables after select
- [HUE-3328](#) - [editor] Follow case in keyword suggestions
- [HUE-3329](#) - [editor] Include database references with table suggestions
- [HUE-3348](#) - [indexer] Update deprecated field in schema.xml

- [HUE-3394](#) - [notebook] Jar submit button is always disabled
- [HUE-3647](#) - [editor] SELECT statement that starts with a comment cannot be saved to table
- [HUE-3667](#) - [sentry] Support TListSentryPrivilegesByAuthRequest v2
- [HUE-3764](#) - [editor] When exporting results to a new table, the new table is not shown
- [HUE-3765](#) - [editor] Graphing with many columns is not readable
- [HUE-3786](#) - [editor] Search box and foreach binding in the snippet DB list dropdown
- [HUE-3788](#) - [home] No way to middle click to open document in a new tab
- [HUE-3795](#) - [editor] Scroll lock a row to enable comparison with other rows
- [HUE-3797](#) - [oozie] API to create workflow for Hive documents
- [HUE-3799](#) - [editor] Impala logs have a new line
- [HUE-3831](#) - [oozie] Show operations inside the HDFS fs graph node for external workflows
- [HUE-3838](#) - [oozie] Deprecate jar_path copy and use oozie.libpath instead for actions with jars
- [HUE-3842](#) - [core] HTTP 500 while emptying Hue 3.9 trash directory
- [HUE-3885](#) - [metastore] Add check to prevent space in table name in both create table wizards
- [HUE-3896](#) - [editor] Add search on column list result set
- [HUE-3897](#) - [metastore] Add column search on the table page
- [HUE-3913](#) - [editor] Drag and drop table names is flaky
- [HUE-3921](#) - [fb] Clicking outside of an empty upload window does not close it
- [HUE-3941](#) - [search] Use sentry base solr config for examples when Sentry is on
- [HUE-3954](#) - [search] Editing record more than once will error because of old version id
- [HUE-3956](#) - [sentry] Browse to table does not load the object in v1
- [HUE-3959](#) - [core] Double blurry login popup on /security/solr page
- [HUE-3960](#) - [editor] Implement virtual rendering of the result table
- [HUE-3981](#) - [core] Restyle the login page a little
- [HUE-3997](#) - [core] Add useradmin feature that resets axes lockouts
- [HUE-3998](#) - [oozie] Blacklisting oozie prevents hue launch
- [HUE-4002](#) - [editor] Add parser support for all generic SQL test cases from the old autocomplete
- [HUE-4003](#) - [editor] Add parser support for HDFS path completion
- [HUE-4004](#) - [editor] Make all parser tests pass from the previous autocomplete
- [HUE-4008](#) - [core] Expandable and fixed close icon for error popup
- [HUE-4014](#) - [editor] Weird scroll and menu part hidden when a few rows
- [HUE-4016](#) - [editor] Add configuration option to enable the new autocomplete
- [HUE-4017](#) - [editor] Make ace editor suggestions from the new autocomplete
- [HUE-4018](#) - [editor] Error when canceling a query that is slow being submitted
- [HUE-4020](#) - [editor] Canceling query at wrong timing breaks the editor
- [HUE-4022](#) - [doc] Promote docker image to get ramped up quicker
- [HUE-4024](#) - [editor] The new autocomplete should suggest identifiers and aliases
- [HUE-4025](#) - [editor] Add keyword completion based on the current parser capabilities
- [HUE-4027](#) - [editor] New autocomplete should support functions with descriptors and types
- [HUE-4032](#) - [editor] Identify locations of databases, tables, columns and functions
- [HUE-4037](#) - [doc] Update docker image with 3.10
- [HUE-4040](#) - [editor] US map blinks when hovering texas
- [HUE-4041](#) - [editor] Auto select axes value if there is only one possible choice
- [HUE-4043](#) - [editor] Bubble up error about queries with \u2002 fails
- [HUE-4044](#) - [editor] Copy pasted query with \u2002 fails
- [HUE-4048](#) - [editor] Clicking on .. of a file settings deletes the snippet header and the row
- [HUE-4051](#) - [hive] Have lighter Impala and Hive check configs call than list DBs
- [HUE-4054](#) - [oozie] Improve of bundle inputs
- [HUE-4055](#) - [oozie] Upgrade oozie spark action to support spark-action:0.2 schema version
- [HUE-4056](#) - [home] Click on trash icons errors

- [HUE-4058](#) - [home] Rename a directory
- [HUE-4059](#) - [editor] Lost the green bar when the query finishes on query execution
- [HUE-4062](#) - [editor] Formatting action add extra spaces after group by and skip LIMIT
- [HUE-4063](#) - [home] 'done' button on wizard redirects to old home2
- [HUE-4064](#) - [editor] Format creation and update date on the table details popover
- [HUE-4066](#) - [editor] Put a toggle all columns in the result column list
- [HUE-4067](#) - [assist] Fix issue with truncated columns of the last assist entry
- [HUE-4068](#) - [metastore] Format date and file size of table stats
- [HUE-4068](#) - [metastore] Show owners and last time instead of files and size
- [HUE-4069](#) - [metastore] Add scroll to top on sample and column pages of a table
- [HUE-4070](#) - [editor] Uncheck all freezes the browser with many columns
- [HUE-4071](#) - [metastore] Table columns should sorted on their initial table order
- [HUE-4072](#) - [editor] Focus on input search box when opening result column filtering
- [HUE-4073](#) - [editor] Scroll to column in result set does not update the horizontal scroll bar
- [HUE-4074](#) - [editor] Clicking on table in assist sometimes expand or insert the name
- [HUE-4076](#) - [editor] Column value autocomplete does not filter duplicates
- [HUE-4077](#) - [editor] Decimal not supported in bar charts
- [HUE-4079](#) - [editor] Amount of 2M is missing a zero in the Y axis of bar charts
- [HUE-4080](#) - [editor] Delete queries from editor does not send to the trash
- [HUE-4081](#) - [core] Skip idle session timeout relogin popup on running jb jobs call when idle session timeout is disabled
- [HUE-4083](#) - [editor] Bar charts gets all of the same color
- [HUE-4085](#) - [editor] Fixed result col headers when cell is wide
- [HUE-4088](#) - [oozie] Make workflow action parameter dragabble
- [HUE-4089](#) - [editor] Keyboard shortcut for a new query and save query
- [HUE-4096](#) - [oozie] Fix unit tests for HUE-4054
- [HUE-4097](#) - [core] Still print error message when Oracle connector is not configured
- [HUE-4099](#) - [editor] Allow searching of raw SQL in query history
- [HUE-4100](#) - [core] First user login checks are not displayed
- [HUE-4101](#) - [editor] Hide search input when input is empty and it loses focus
- [HUE-4102](#) - [editor] Move session settings to a context panel
- [HUE-4103](#) - [core] Upgrade Font Awesome to 4.6.3
- [HUE-4104](#) - [core] Introduce Roboto as a main font
- [HUE-4105](#) - [editor] Make the editor 'editorType' a ko variable
- [HUE-4106](#) - [editor] Move schedule edition to the context panel
- [HUE-4107](#) - [desktop] Add a 'managed' field to Document2
- [HUE-4109](#) - [oozie] First pass at refactoring Coordinator editor to be pluggable
- [HUE-4110](#) - [editor] Skeleton of running schedule actions
- [HUE-4111](#) - [editor] Toggle to show or hide the column search field on result is always on
- [HUE-4112](#) - [oozie] Automatically persist the workflow of a scheduled query
- [HUE-4115](#) - [oozie] Autoatically persist the coordinator of a scheduled query
- [HUE-4116](#) - [editor] Loads the coordinator of the query
- [HUE-4118](#) - [fb] Context menu can be hidden by bottom bar
- [HUE-4120](#) - [core] Update NVD3 color palette
- [HUE-4121](#) - [oozie] Improve UX of entering args to distcp action in workflow editor
- [HUE-4122](#) - [editor] Examples should have isSaved set to true
- [HUE-4123](#) - [core] Verify id parameter values in GET requests
- [HUE-4127](#) - [oozie] SLA can not be enabled for a coordinator
- [HUE-4130](#) - [core] Triple amount of lines in the /logs page
- [HUE-4131](#) - [assist] Search on the query editors should be case insensitive

- [HUE-4133](#) - [search] Queries don't handle field names that contain spaces
- [HUE-4135](#) - [editor] Automatically update on save the variables of a scheduled query
- [HUE-4136](#) - [fb] Fix normpath and parent path logic for aws fs
- [HUE-4137](#) - [editor] Open editor, create query and save it, click on Saved queries will have one item
- [HUE-4138](#) - [editor] Last modified time of a saved query is not in the correct timezone
- [HUE-4140](#) - [editor] Do not reload saved query list when opening a saved query
- [HUE-4141](#) - [oozie] Graph breaks for external workflows when there is more than one kill node
- [HUE-4142](#) - [aws] Enable S3 browse option in filebrowser
- [HUE-4143](#) - [editor] Snippet configuration layout is broken
- [HUE-4144](#) - [editor] Add search functionality to snippet DB selection
- [HUE-4147](#) - [editor] Older queries after upgrade do not have any document1
- [HUE-4149](#) - [editor] Integrated progress report of MR jobs
- [HUE-4151](#) - [oozie] Avoid dashboard page crash when coordinator does not exist in Oozie
- [HUE-4152](#) - [oozie] Managed document property blocks access to workflows of coordinators
- [HUE-4155](#) - [home] Improve document drag to select
- [HUE-4156](#) - [oozie] Do not list the managed workflows in the coordinator editor
- [HUE-4157](#) - [aws] Fix aws config_check
- [HUE-4158](#) - [oozie] Do not lose existing values when renaming variables
- [HUE-4160](#) - [editor] Do not create multiple coordinator for a new saved query
- [HUE-4161](#) - [editor] Autocomplete tables and columns in update statements
- [HUE-4163](#) - [aws] Implement S3 mkdir operation
- [HUE-4166](#) - [metastore] Sample data dialog can't display very long table name properly
- [HUE-4167](#) - [aws] Implement S3 rmtree operation
- [HUE-4170](#) - [aws] Fix S3 object timestamp
- [HUE-4173](#) - [editor] Enable autocomplete when there are backticks in table and database names
- [HUE-4174](#) - [editor] Fix nested type autocompletion of arrays, maps and structs
- [HUE-4175](#) - [editor] Improved autocompletion after WHERE
- [HUE-4176](#) - [editor] Autocomplete field values
- [HUE-4180](#) - [editor] JS error on help icon
- [HUE-4183](#) - [fb] Offer two File Browser links when S3 is configured
- [HUE-4184](#) - [editor] Set the result legend section as fixed
- [HUE-4186](#) - [editor] Create a batch_oozie connector
- [HUE-4187](#) - [editor] Integrate batch submission in the UI
- [HUE-4190](#) - [oozie] Option to return coordinator id instead of redirect in coordinator submission popup
- [HUE-4191](#) - [oozie] Prevent 500 error when listing a workflow which coordinator is none
- [HUE-4192](#) - [editor] Document2 matching query does not exist when saving
- [HUE-4193](#) - [search] Toggling on all the fields of a grid widgets breaks
- [HUE-4194](#) - [editor] Autocomplete CREATE DATABASE
- [HUE-4195](#) - [editor] Last digit of multiquery can be hidden
- [HUE-4196](#) - [editor] Catch open notebook errors
- [HUE-4197](#) - [editor] Autocomplete for DROP TABLE, DATABASE and SCHEMA
- [HUE-4198](#) - [metastore] Sample fetch breaks the page when it fails with error 1
- [HUE-4199](#) - [editor] Make the editor 'editorMode' a ko variable
- [HUE-4200](#) - [editor] Autocompletion of SHOW statements
- [HUE-4203](#) - [editor] Column search on resultset is not empty, do a scroll to any column will hide the search input
- [HUE-4204](#) - [editor] Add autocompletion support for CAST functions
- [HUE-4205](#) - [doc2] Ignore history dependencies on import
- [HUE-4206](#) - [notebook] Ignore history and other calls in notebook mode
- [HUE-4207](#) - [doc2] Delete history command to clean up old history or orphaned history
- [HUE-4210](#) - [editor] Don't revert to Hive on browser refresh of new Impala query

- [HUE-4211](#) - [editor] Autoexpand query editor sometimes bugs the result headers
- [HUE-4212](#) - [hive] Also provide if the job is started or finished
- [HUE-4216](#) - [editor] The editor type title blinks at page load and shows all the different types
- [HUE-4217](#) - [editor] Autocomplete for DESCRIBE
- [HUE-4218](#) - [oozie] Refresh parameters of a d&dropped action
- [HUE-4219](#) - [oozie] Add quick link to selected d&dropped query
- [HUE-4220](#) - [oozie] Saving a workflow make the save button blink
- [HUE-4223](#) - [oozie] No indication that \${} can be used as parameters
- [HUE-4225](#) - [editor] Colors are flipped when sorting bar results in reverse order
- [HUE-4228](#) - [search] Used to be able to toggle the field list of grid widget
- [HUE-4229](#) - [oozie] js error when selecting another query in the hive document action
- [HUE-4230](#) - [core] Finer version number logging
- [HUE-4232](#) - [notebook] Bulk operation button is gone
- [HUE-4234](#) - [editor] Improved JOIN autocompletion
- [HUE-4236](#) - [editor] Changing fields of a query we chart should reselect default values for the axis
- [HUE-4239](#) - [liboauth] Does not handle "next" urls, always sending user back to homepage
- [HUE-4240](#) - [filebrowser] Remove unused sortby param returned from listdir
- [HUE-4241](#) - [editor] Schedule are showing up in home
- [HUE-4242](#) - [editor] Add a message when query submission takes time
- [HUE-4243](#) - [aws] Only raise aws warning if default account configs are provided
- [HUE-4244](#) - [indexer] First skeleton on scalable indexer
- [HUE-4245](#) - [editor] Autocomplete subqueries in FROM and WHERE
- [HUE-4246](#) - [libsaml] SAML Integration Doesn't Check Forwarded Protocol
- [HUE-4247](#) - [batch] Support for batch pyspark or spark
- [HUE-4248](#) - [liboauth] OAuth integration Redirect URL is Assumed to be HTTP
- [HUE-4250](#) - [editor] Add configurable timeout for autocompletion API calls
- [HUE-4251](#) - [editor] Explain on incorrect query should not error
- [HUE-4262](#) - [editor] Improve performances of the table fixed header
- [HUE-4266](#) - [indexer] Add geo, host, grok, split operations
- [HUE-4268](#) - [core] Do not remove the table fixed rows on plugin redraw
- [HUE-4269](#) - [editor] Improve sample popover
- [HUE-4270](#) - [fb] Enable S3 on the file pickers
- [HUE-4271](#) - [editor] New autocomplete should support UDAFs and numeric expressions
- [HUE-4275](#) - [core] Avoid filechooser modals to scroll to the action buttons on modal open
- [HUE-4276](#) - [editor] Autocomplete around arithmetic operations and support BETWEEN, EXISTS, RLIKE and REGEX
- [HUE-4277](#) - [fb] Add endpoint that returns enabled status for each filesystem
- [HUE-4278](#) - [useradmin] Provide basic useradmin API
- [HUE-4279](#) - [editor] Shrink back the editor if the query is not big enough and it hasn't been resized manually
- [HUE-4281](#) - [editor] Fail gracefully and return an error when trying to import a wrong/corrupted JSON file
- [HUE-4285](#) - [editor] Fix Scatter plot display
- [HUE-4290](#) - [editor] Style the coordinator history on the context panel
- [HUE-4291](#) - [metastore] Do not use smallints in the create table from file wizard
- [HUE-4296](#) - [core] Need info message if kt_renwer exits due to no hue_keytab
- [HUE-4298](#) - [security] Offer ability to resize the Impersonate the user drop down in Hue security app
- [HUE-4301](#) - [editor] Autocomplete for Impala extract function
- [HUE-4302](#) - [editor] New autocomplete should suggest DISTINCT or ALL where appropriate in aggregate functions
- [HUE-4303](#) - [useradmin] Reintroduce '?' icon displaying description of each field
- [HUE-4304](#) - [core] Pie slices shrink after wrong timing of resize of the pie
- [HUE-4306](#) - [editor] The new autocomplete should complete CASE functions
- [HUE-4307](#) - [editor] Goto line number shorter conflicts with highlighting the address bar contents in Safari

- [HUE-4308](#) - [editor] UX lock rows keep adding rows when clicking on the same
- [HUE-4309](#) - [editor] UX lock rows should tooltip about clicking to lock and delete icon on hover
- [HUE-4310](#) - [fb] File chooser picker modal should have action buttons fixed to the bottom
- [HUE-4311](#) - [fb] Add boto http_socket_timeout configuration
- [HUE-4312](#) - [editor] Load query history, loses editor type when clicking on new query
- [HUE-4313](#) - [editor] The new autocomplete should accept non-reserved keywords for DBs, tables, columns etc
- [HUE-4315](#) - [editor] Error when sharing a query with R permission
- [HUE-4317](#) - [jb] Logs on attempt page do not make links to job id
- [HUE-4320](#) - [editor] New autocomplete should make suggestions based on type in value expressions
- [HUE-4322](#) - [impala] Better error message when user already downloaded all the impala resultset cache
- [HUE-4324](#) - [editor] Batch API get_jobs() should return the job ids
- [HUE-4325](#) - [editor] Toggle the execute action depending on if last execution was batch or not
- [HUE-4326](#) - [editor] Keep snippet DB list in sync like the assist
- [HUE-4327](#) - [editor] Disable batch submit by default when oozie is not setup
- [HUE-4328](#) - [editor] File chooser in 'In HDFS (large file)' always open on /
- [HUE-4329](#) - [editor] Multi query does not go past statement 1
- [HUE-4331](#) - [editor] Log call should return only the additional ones after the first call
- [HUE-4332](#) - [indexer] Add Hue logs as a supported file type
- [HUE-4334](#) - [editor] Update status to error when canceling batch job
- [HUE-4335](#) - [editor] New autocomplete should handle date types
- [HUE-4337](#) - [editor] The autocomplete should suggest backticked values when applicable
- [HUE-4340](#) - [indexer] Fix double matching and set required field default to false
- [HUE-4341](#) - [indexer] Use info level logging for morphlines, add error handling for invalid dates
- [HUE-4342](#) - [editor] DB prefix autocomplete can be lost
- [HUE-4343](#) - [editor] Improved autocompletion in select list
- [HUE-4344](#) - [fb] Buffer reading parquet file and limit in size
- [HUE-4345](#) - [aws] Fix Actions when viewing an S3 file in filebrowser
- [HUE-4347](#) - [core] Rename tests_convert.py to follow test conventions
- [HUE-4348](#) - [indexer] Refactor format types such that all information is in one place
- [HUE-4351](#) - [editor] Show number of columns in the query result
- [HUE-4352](#) - [indexer] Fix typo for grok dictionary file exists check
- [HUE-4354](#) - [core] Warn than SQLite is the cause of the 'Database locked' error
- [HUE-4355](#) - [indexer] Add i18n for arguments and extension based file type guessing
- [HUE-4356](#) - [fb] File chooser picker does not select S3 tab when opening an existing path
- [HUE-4357](#) - [oozie] Return the launcher task logs when submitting Oozie batch job
- [HUE-4358](#) - [meta] Split metadata services into two
- [HUE-4360](#) - [meta] Read configuration from a properties file
- [HUE-4363](#) - [core] Do not rely on import hadoop config before loading core
- [HUE-4366](#) - [metastore] Add an option to create a table from a file in an external location
- [HUE-4367](#) - [editor] Export query result to S3
- [HUE-4371](#) - [editor] Enable the new autocomplete by default
- [HUE-4372](#) - [core] Backport Django Security middleware or add HTTP security headers
- [HUE-4373](#) - [home] Exporting / Importing queries should handle associated workflows and coordinator
- [HUE-4374](#) - [editor] No autocomplete appearing in ORDER BY with an operation
- [HUE-4376](#) - [indexer] Added assist panel and filechooser
- [HUE-4377](#) - [editor] Autocomplete should support variable references like '\${var}' in statements
- [HUE-4378](#) - [core] File chooser with multiple options overflow the modals
- [HUE-4379](#) - [editor] Autocomplete should support CREATE TABLE statements
- [HUE-4380](#) - [assist] Only enable navigator search when configured
- [HUE-4384](#) - [indexer] Add support for combined apache log files

CDH 5 Release Notes

- [HUE-4385](#) - [indexer] Stop the indexer from dropping text fields that are under 100 characters long
- [HUE-4386](#) - [editor] Make it possible to turn autocomplete on or off
- [HUE-4390](#) - [core] Do not re-create a new user for each request in demo mode
- [HUE-4391](#) - [editor] Blacklisting Oozie should not break the editor
- [HUE-4392](#) - [fb] S3 buckets should have a different icon
- [HUE-4393](#) - [fb] S3 bucket should not have any file upload or file creation action
- [HUE-4394](#) - [fb] Home button redirects to HDFS even if on S3
- [HUE-4395](#) - [fb] S3 delete bucket implementation
- [HUE-4396](#) - [fb] Delete button should be 'delete forever' all the time
- [HUE-4397](#) - [fb] Hide upload archive in S3
- [HUE-4398](#) - [fb] API should check if bucket already exists in the region and bubble up the error
- [HUE-4399](#) - [fb] Renaming a key (non-bucket directory) creates an extraneous empty file
- [HUE-4400](#) - [fb] Disable chmod and summary for S3 for now
- [HUE-4401](#) - [fb] Do not return any date for bucket and directory objects
- [HUE-4402](#) - [fb] S3 copy action is not using the picker and fails
- [HUE-4403](#) - [aws] Close file after read using fast=True
- [HUE-4407](#) - [assist] Tables and views checkboxes can't be clicked
- [HUE-4410](#) - [editor] Fix editor size issue for large pasted queries
- [HUE-4413](#) - [dbms] Security: Full XSS in DBQuery editor
- [HUE-4414](#) - [editor] Allow to send queries even after closing the session
- [HUE-4415](#) - [editor] Disable "Save" button in query export unless output is filled
- [HUE-4416](#) - [editor] Do not save the coordinator constants in the json model
- [HUE-4417](#) - [aws] Switch from s3 to s3a to enable direct save to S3
- [HUE-4418](#) - [editor] Sessions tab doesn't display properly once the Settings is expanded
- [HUE-4419](#) - [editor] The new completer should suggest columns from sub-queries
- [HUE-4421](#) - [editor] New saved query should be schedulable
- [HUE-4423](#) - [editor] Persist the coordinator ID of a submitted schedule
- [HUE-4424](#) - [librdbms] Support more than one database
- [HUE-4425](#) - [metastore] UX create table from a file, S3 or from a directory
- [HUE-4426](#) - [editor] The new completer should work with all SQL dialects
- [HUE-4427](#) - [metastore] Filepicker should allow directories when selecting external table
- [HUE-4429](#) - [oozie] Generate unique workspace for managed workflows
- [HUE-4430](#) - [editor] Batch execution often fails to show any output
- [HUE-4431](#) - [editor] Batch mode does not always update the query
- [HUE-4432](#) - [editor] The new completer should be more forgiving when editing incomplete UDFs in select list
- [HUE-4433](#) - [editor] The new completer should handle analytical functions
- [HUE-4434](#) - [editor] Fixed header has a wrong offset in fullscreen when the editor is manually resized
- [HUE-4435](#) - [test] Add list_modules subcommand to `hue test`
- [HUE-4436](#) - [hive] Fix test_install_examples tests to point at the database configured in BeeswaxSampleProvide
- [HUE-4437](#) - [editor] The completer should handle multiple db references in table primaries for Impala
- [HUE-4438](#) - [editor] Progress status and truncating warning when direct downloading results as Excel
- [HUE-4439](#) - [notebook] Fix notebook TestHiveserver2ApiWithHadoop.test_explain
- [HUE-4440](#) - [editor] Assist slightly scrolls horizontally
- [HUE-4441](#) - [editor] Set result menu as fixed
- [HUE-4442](#) - [editor] Improve column list scroll bars
- [HUE-4443](#) - [security] Automatically default secure_ssl_redirect to True when HTTPS is setup
- [HUE-4444](#) - [beeswax] test_analyze_table_and_read_statistics doesn't test correctly with Hive 2
- [HUE-4445](#) - [fb] Return correct result response on file upload
- [HUE-4448](#) - [hbase] Change test_list_tables hbase test to query each cluster name
- [HUE-4449](#) - [indexer] Add ruby log to indexer

- [HUE-4450](#) - [desktop] Fix HTTPS termination using X-Forwarded-Proto header
- [HUE-4451](#) - [editor] Add tooltips to result set legend
- [HUE-4452](#) - [metadata] Fix optimizer view name and update navigator API version to v9
- [HUE-4454](#) - [security] Disclosure of Web Server Information
- [HUE-4455](#) - [security] secure_content_security_policy breaks the editor and GA
- [HUE-4457](#) - [editor] Column legend horizontal scroll bar are missing until you scroll down
- [HUE-4459](#) - [indexer] Add morphline generation tests for each file format
- [HUE-4461](#) - [metastore] If create external table is selected, input path needs to be a directory
- [HUE-4462](#) - [editor] Status and progress report when downloading large Excel files
- [HUE-4463](#) - [editor] Result column search on column types too
- [HUE-4464](#) - [editor] Result column quick scroll does not scroll with a search
- [HUE-4465](#) - [notebook] Bump the darkness of the non selected icons
- [HUE-4466](#) - [security] Deliver csrfToken cookie with secure bit set if possible
- [HUE-4467](#) - [indexer] Add field operation morphline generation tests for each operation
- [HUE-4469](#) - [editor] Can have lot of white spaces below the query
- [HUE-4470](#) - [editor] Snippet properties for File filechooser always starts at /
- [HUE-4471](#) - [indexer] Add geoip operation automatically to hue format
- [HUE-4474](#) - [indexer] Let user set an existing field to unique id
- [HUE-4475](#) - [editor] Lines are not expanded when loading a long query from the cookie
- [HUE-4476](#) - [editor] Add RESET to the Hive highlight
- [HUE-4478](#) - [editor] Style the editor settings panel
- [HUE-4479](#) - [indexer] Allow indexing multiple files on the same collection
- [HUE-4480](#) - [meta] Filter SQL objects in the backend
- [HUE-4481](#) - [editor] Show a message when the Excel results have been truncated
- [HUE-4483](#) - [core] Do not use backticks in the ini files
- [HUE-4484](#) - [editor] Add column name and type to the title of column browser
- [HUE-4485](#) - [editor] Scroll to assist entries from search result
- [HUE-4486](#) - [editor] Add back sorting to the result table
- [HUE-4487](#) - [editor] Result table search
- [HUE-4488](#) - [editor] Show result row detail view
- [HUE-4489](#) - [editor] The new autocomplete should allow white space between UDF name and parenthesis
- [HUE-4490](#) - [oozie] Generate workflow from Java snippet
- [HUE-4491](#) - [fb] Enable creating buckets on S3
- [HUE-4492](#) - [indexer] Migrate to the oozie library to submit the java indexer job
- [HUE-4494](#) - [indexer] Add a load from JSON function to Create Index Wizard
- [HUE-4495](#) - [assist] Switching DB does not display the tables
- [HUE-4498](#) - [security] Fixed Content Security Policy blocks PDF in HBase app
- [HUE-4499](#) - [editor] Gradient map not always showing on the editor
- [HUE-4501](#) - [indexer] Add basic hive CSV file format outline to indexer
- [HUE-4502](#) - [editor] Fix issue with column type precision
- [HUE-4503](#) - [fb] Disable timestamp for S3 buckets and folders
- [HUE-4505](#) - [editor] Resultset search is too transparent
- [HUE-4506](#) - [editor] Provide more information why query couldn't be saved
- [HUE-4507](#) - [editor] Cannot see column names in the gradient map visualization
- [HUE-4508](#) - [editor] Search in result set is issuing a js error
- [HUE-4511](#) - [editor] Field scroll does not scroll or grey the column
- [HUE-4512](#) - [editor] Header and column not always aligned
- [HUE-4513](#) - [editor] Use weights for keyword autocompletion
- [HUE-4514](#) - [beeswax] Fix test_explain_query and test_explain_query_i18n for default execution engine Tez
- [HUE-4516](#) - [indexer] Switching file types doesn't update fields properly in the indexer wizard

CDH 5 Release Notes

- [HUE-4518](#) - [editor] Use different icon for the column list
- [HUE-4519](#) - [fb] Check for user permissions first before uploading
- [HUE-4521](#) - [editor] Hide/show column doesn't work on fixed rows
- [HUE-4522](#) - [beeswax] Remove max_rows limit for HiveServer2 get_configuration
- [HUE-4523](#) - [editor] Enable resizing of the column list panel
- [HUE-4524](#) - [hadoop] Fix test_yarn_ssl_validate which assumes default YARN_CLUSTER is defined
- [HUE-4526](#) - [oozie] API to generate sequential workflow with sequential document actions
- [HUE-4527](#) - [indexer] Date fields that aren't required can crash the indexing job
- [HUE-4530](#) - [indexer] Only warn when the collection name is empty
- [HUE-4531](#) - [fb] Enable deleting buckets on S3
- [HUE-4534](#) - [indexer] Fix syslog grokking for smart indexer
- [HUE-4535](#) - [editor] Add autocomplete for CREATE and ALTER statements
- [HUE-4536](#) - [editor] Load query generates JS error
- [HUE-4538](#) - [search] Field detail popup cannot always be closed
- [HUE-4539](#) - [search] Remove field collapse icon
- [HUE-4540](#) - [jb] Lower max number of jobs displayed on the main page
- [HUE-4542](#) - [editor] Column filter should be 100% of the column list width
- [HUE-4544](#) - [indexer] Extend Parquet morphline to support generic fields
- [HUE-4548](#) - [editor] Sample popup tab seems off
- [HUE-4549](#) - [core] Missing file for the collect static plugin
- [HUE-4550](#) - [fb] Fix file chooser for S3 after S3A migration
- [HUE-4551](#) - [editor] The new autocomplete should support CTEs (WITH) and UNION
- [HUE-4552](#) - [editor] Improve performance on checking/unchecked all fields
- [HUE-4553](#) - [oozie] The dashboard table scrolls horizontally
- [HUE-4554](#) - [editor] Status checking can go in infinite loop in batch mode
- [HUE-4555](#) - [core] Avoid opening file chooser on enter key
- [HUE-4556](#) - [core] Fix coverage failure due to missing source files
- [HUE-4559](#) - [indexer] Enable Select Fields button when a path has been entered into wizard
- [HUE-4560](#) - [core] Add permissions controls to authorize access to S3 across all components
- [HUE-4562](#) - [editor] Switching to marker map makes the page freeze
- [HUE-4563](#) - [impala] Handle case for non-string partition keys or values
- [HUE-4564](#) - [core] Log stderr on failure to coerce password from script
- [HUE-4565](#) - [core] Update natural sort algorithm
- [HUE-4566](#) - [metastore] Query created by Browse Data action should have a limit and semi-colon
- [HUE-4567](#) - [aws] Read AWS credentials from configured script files
- [HUE-4568](#) - [editor] Export a query to an index
- [HUE-4569](#) - [editor] If we don't have a query history select the saved query tab
- [HUE-4570](#) - [editor] Autocomplete popup disappears even when calling it when typing
- [HUE-4572](#) - [core] Fix HueDataTable with multiple instances on the same page
- [HUE-4573](#) - [editor] The autocomplete should support carriage return in statements
- [HUE-4574](#) - [editor] Enable autocompletion after '-' without whitespace
- [HUE-4575](#) - [core] Add notice to hue.ini regarding the custom maps and security
- [HUE-4576](#) - [editor] Fix autocompletion before OR in value expressions
- [HUE-4577](#) - [editor] The new autocomplete should support CREATE, DROP and SET ROLE
- [HUE-4579](#) - [editor] The new autocomplete should support Hive INSERT statements
- [HUE-4580](#) - [editor] The new autocomplete should support ANALYZE TABLE
- [HUE-4581](#) - [editor] The new autocomplete should support DROP FUNCTION
- [HUE-4582](#) - [editor] The new autocomplete should fully support LOAD
- [HUE-4584](#) - [editor] The new autocomplete should support DELETE
- [HUE-4586](#) - [editor] The new autocomplete should support GROUPING SETS, CUBE and ROLLUP

- [HUE-4587](#) - [editor] The new autocomplete should completely support EXPLAIN
- [HUE-4588](#) - [editor] The new autocomplete should support SET
- [HUE-4589](#) - [editor] The new autocomplete should completely support COMPUTE and DROP STATS
- [HUE-4590](#) - [editor] The new autocomplete should support INVALIDATE METADATA
- [HUE-4591](#) - [editor] The new autocomplete should support REFRESH
- [HUE-4598](#) - [editor] Empty table list sometimes the first time
- [HUE-4600](#) - [editor] Result table headers breaks on sort
- [HUE-4601](#) - [fb] Creating a bucket with a not allowed name on S3 will HTTP 500
- [HUE-4603](#) - [fb] Disable bucket move and copy
- [HUE-4604](#) - [editor] Locked rows does not go well with headers
- [HUE-4605](#) - [fb] Move and copy in S3 have the HDFS path autocompleters in the popup
- [HUE-4609](#) - [editor] Resizing the window should resize also the column list/grid
- [HUE-4610](#) - [editor] Avoid trembling of fixed headers on query editing
- [HUE-4611](#) - [editor] Export to S3 should not say HDFS
- [HUE-4614](#) - [editor] Query result search box always present when switching tabs
- [HUE-4618](#) - [editor] Query selection highlighting does not stay
- [HUE-4621](#) - [core] On loaddata failure, print stdout to logs
- [HUE-4623](#) - [editor] ON and ASC and DESC are not being highlighted on Impala
- [HUE-4624](#) - [editor] Doing enter in the query result search box could go to the next match
- [HUE-4625](#) - [fb] S3 filechooser popup do not display the parent directory as
- [HUE-4626](#) - [fb] S3 in the filechooser do not show up with FF
- [HUE-4628](#) - [useradmin] Improve and unify help icons
- [HUE-4631](#) - [home] DB transaction failing because of atomic block on home page
- [HUE-4632](#) - [oozie] Log which type of oozie action are created in workflows
- [HUE-4633](#) - [notebook] Disable by default Oozie integration until 3.12
- [HUE-4634](#) - [editor] The autocomplete should allow some errors in the select list
- [HUE-4637](#) - [editor] Scroll on column list is extremly slow on Linux
- [HUE-4638](#) - [editor] Infinite scroll blinks the fixed legend when fetching the next 100 batch
- [HUE-4640](#) - [editor] Fix for DB prefix autocomplete when typing table names after "? FROM db."
- [HUE-4641](#) - [indexer] GeoIP will stop indexing if an IP is missing from it's lookup table
- [HUE-4642](#) - [editor] Ace autocomplete should honour the weights for partial matching
- [HUE-4643](#) - [editor] Autocomplete is silent for the second argument of concat
- [HUE-4644](#) - [editor] Don't refresh the assist 5 seconds after loading a query
- [HUE-4645](#) - [core] Disable inline display of SVG files
- [HUE-4646](#) - [core] The initial welcome wizard redirects to the old home
- [HUE-4647](#) - [home] Better align new document dropdown
- [HUE-4648](#) - [security] The File ACLs tree doesn't render if there's a file in the HDFS root
- [HUE-4649](#) - [sqoop] SqoopResource got broken after change in Resource
- [HUE-4650](#) - [fb] Selecting S3A from filechooser is not displaying the S3 filesystem
- [HUE-4651](#) - [fb] Prevent inline display of non authorized mime types
- [HUE-4652](#) - [core] Add settings for django-axes AXES_BEHIND_REVERSE_PROXY and AXES_REVERSE_PROXY_HEADER
- [HUE-4657](#) - [indexer] A failed grok match attempt will stop indexing
- [HUE-4658](#) - [doc] Update release 3.11
- [HUE-4659](#) - [editor] The new autocomplete should merge columns from multiple tables when suggesting columns
- [HUE-4660](#) - [editor] Ace autocomplete spinner disappears too quick
- [HUE-4661](#) - [core] Demo backend should keep the admin user logged in
- [HUE-4665](#) - [editor] Autocompletion of lateral views should not have parenthesis around the column aliases
- [HUE-4666](#) - [search] Use the actual Hue user for impersonating in non secure mode
- [HUE-4667](#) - [editor] Geo plotting of states is case sensitive
- [HUE-4668](#) - [fb] S3 rename directory raises IOError

CDH 5 Release Notes

- [HUE-4670](#) - [metastore] Allow headers to be ignored/removed when creating external table
- [HUE-4671](#) - [fb] S3 create bucket should allow underscores in name
- [HUE-4673](#) - [fb] S3 bucket names should automatically lowercase them
- [HUE-4675](#) - [metastore] Assist not loaded and pointing to the source on the create table from a file page
- [HUE-4676](#) - [editor] Table header is not XSS safe
- [HUE-4678](#) - [editor] The label on marker maps is not updated to reflect what you choose on the left side dropdown
- [HUE-4682](#) - [editor] Scroll to a col, reexec the query, the greyed column is still there
- [HUE-4683](#) - [editor] Export result question mark icon does nothing
- [HUE-4684](#) - [editor] No placeholder showing up in new editors
- [HUE-4686](#) - [fb] Bubble up error when trying to create directory in / in filechooser
- [HUE-4688](#) - [metastore] Authorize the selection of a directory in the create table from a file wizard
- [HUE-4689](#) - [fb] D&D a file into a directory in S3 gets 'RenameFormFormSet' object has no attribute 'data'
- [HUE-4691](#) - [fb] S3 D&D is allowed after a page refresh but not after opening a directory
- [HUE-4696](#) - [editor] JS .top error on certain reload/execute combos
- [HUE-4697](#) - [editor] The autocomplete should suggest select list aliases
- [HUE-4699](#) - [editor] Autocomplete fix for decimal values
- [HUE-4701](#) - [editor] Older saved queries throw "'NoneType' object has no attribute 'update_data'"
- [HUE-4703](#) - Fixing[yarn] Correct username is not used on hard failover
- [HUE-4704](#) - [security] Fixed Arbitrary host header accepted in Hue
- [HUE-4705](#) - [editor] Possible XSS when hovering on saved query description
- [HUE-4706](#) - [core] Importing documents should ignore reserved directories
- [HUE-4707](#) - [aws] Enable non-US region support for AWS S3
- [HUE-4708](#) - [aws] Add test_rename for directory rename in S3
- [HUE-4709](#) - [editor] Search does not load cells when navigating
- [HUE-4713](#) - [oozie] Spark example should use yarn client mode and show how to add dependencies
- [HUE-4718](#) - [core] jHueScrollUp pollutes the DOM with more than one scroll up anchor
- [HUE-4719](#) - [editor] Search disappears on load of new records
- [HUE-4720](#) - [oozie] Drag & Drop saved Spark app into a workflow
- [HUE-4724](#) - [editor] Closing the table search should un-highlight the selected cell
- [HUE-4725](#) - [editor] The autocomplete should handle Impala complex types in the table list
- [HUE-4727](#) - [fb] Raise warning if CHECKACCESS is not supported
- [HUE-4732](#) - [editor] Autocomplete should work inside parenthesized value expressions
- [HUE-4733](#) - [editor] Autocomplete doesn't support arithmetic operations in in value list
- [HUE-4734](#) - [editor] Autocomplete should trigger after '!' when autocomplete as you type is enabled
- [HUE-4735](#) - [editor] The autocomplete should also suggest tables with '.' when suggesting columns from multiple tables
- [HUE-4736](#) - [editor] The autocomplete throws JS error
- [HUE-4737](#) - [core] Upgrade boto to 2.42.0
- [HUE-4742](#) - [core] Data sample popup HueDataTable renders weirdly on FF
- [HUE-4747](#) - [editor] Download form should be submitted to a new tab otherwise the snippet gets closed
- [HUE-4753](#) - [editor] The autocomplete doesn't merge columns correctly
- [HUE-4754](#) - [editor] Increase default autocomplete API timeout to 5 seconds
- [HUE-4758](#) - [editor] The autocomplete should suggest keywords at the start of a statement
- [HUE-4759](#) - [editor] Don't suggest tables in the select list when there is a FROM clause
- [HUE-4760](#) - [core] Fix error with indexer config when search is disabled
- [HUE-4763](#) - [editor] Update the editor keyword highlight rules for Hive and Impala
- [HUE-4767](#) - [editor] Limit the autocomplete length before and after the cursor
- [HUE-4768](#) - [editor] Autocomplete is being called multiple times when the cursor is on top of '*' in a select statement
- [HUE-4769](#) - [editor] The autocomplete should support INSERT OVERWRITE DIRECTORY
- [HUE-4773](#) - [editor] Autocomplete spins forever when completing subqueries containing asterisk

- [HUE-4787](#) - [editor] Fixing Marker map tiles are not showing up
- [HUE-4801](#) - [core] Do not hide the username in small resolution
- [HUE-4804](#) - [search] Download function of HTML widget breaks the display
- [HUE-4809](#) - [oozie] Only add trustore paths when they are actually existing
- [HUE-4810](#) - [core] Fix tests by setting data to valid JSON type
- [HUE-4813](#) - [core] Disable collecting referrer URL in Google Analytics
- [HUE-4823](#) - [fb] Display exception when Hue is configured with invalid S3 access key
- [HUE-4827](#) - [fb] Pointing to a wrong S3 region breaks with no information
- [HUE-4871](#) - [useradmin] An unprivileged user can enumerate users
- [HUE-4882](#) - [editor] Redraw fixed header on column toggling
- [HUE-4891](#) - [useradmin] An unprivileged user can list document items
- [HUE-4910](#) - [editor] Improve result table scroll performance and fix header positioning
- [HUE-4916](#) - [core] Truncate last name to 30 chars on ldap import
- [HUE-4917](#) - [fb] S3 upload progress of large files is not consistent
- [HUE-4921](#) - [oozie] Workflow dashboard page should not fail when properties have unicode characters
- [HUE-4928](#) - [oozie] New Spark action can't be added in workflows
- [HUE-4935](#) - [aws] Enable support for AWS security token
- [HUE-4938](#) - [editor] Fix issue where the map crashes the browser for large coordinate values
- [HUE-4941](#) - [editor] Fixed Content Security Policy directive blocks an image when navigating on marker map
- [HUE-4968](#) - [oozie] Remove access to /oozie/import_wokflow when v2 is enabled
- [HUE-4994](#) - [oozie] Consider default path for decision nodes in dashboard graph
- [HUE-4996](#) - [editor] Nested types can not be extended in assist
- [HUE-4997](#) - [impala] get_partitions is not being closed
- [HUE-4998](#) - [impala] get_configuration is not being closed
- [HUE-4999](#) - [impala] Set default Impala idle_session_timeout to 1 hour
- [HUE-5007](#) - [core] Get ride of bad utf8 character when doing make-locale
- [HUE-5007](#) - [core] Update localization for de, es, fr, ja, ko, zh
- [HUE-5014](#) - [core] Prevent copying a Document folder into itself which then creates a recursive depth issue
- [HUE-5040](#) - [notebook] Keep the progress bar to 99% orange until the status of query is READY
- [HUE-5041](#) - [editor] Hue export large file to HDFS doesn't work on non-default database
- [HUE-5050](#) - [core] Logout fails for local login when multiple backends are used
- [IMPALA-1112](#) - Remove some unnecessary code from cross-compilation
- [IMPALA-1240](#) - add back spilling sort now that sorter is not flaky
- [IMPALA-1440](#) - test for insert mem limit
- [IMPALA-3018](#) - Address various small memory allocation related bugs
- [IMPALA-1619](#) - Support 64-bit allocations
- [IMPALA-1633](#) - GetOperationStatus should set errorMessage and sqlState
- [IMPALA-1671](#) - Print time and link to coordinator web UI once query is submitted in shell
- [IMPALA-1683](#) - Allow REFRESH on a single partition
- [IMPALA-2347](#) - Reuse metastore client connections in Catalog
- [IMPALA-2459](#) - Implement next_day date/time UDF
- [IMPALA-2700](#) - ASCII NUL characters are doubled on insert into text tables
- [IMPALA-2767](#) - Web UI call to force expire sessions
- [IMPALA-2878](#) - Fix Base64Decode error and remove duplicate codes
- [IMPALA-2885](#) - ScannerContext::Stream objects should be owned by ScannerContext
- [IMPALA-2979](#) - Fix scheduling on remote hosts
- [IMPALA-3018](#) - Don't return NULL on zero length allocations
- [IMPALA-3063](#) - Separate join inversion from join ordering
- [IMPALA-3084](#) - Cache the sequence of table ref and materialized tuple ids during analysis
- [IMPALA-3181](#) - Add noexcept to some functions

CDH 5 Release Notes

- [IMPALA-3201](#) - buffer pool header only
- [IMPALA-3206](#) - Enable codegen for AVRO_DECIMAL
- [IMPALA-3210](#) - last/first_value() support for IGNORE NULLS
- [IMPALA-3223](#) - Remove boost multiprecision in thirdparty
- [IMPALA-3225](#) - Add script to push from gerrit to ASF
- [IMPALA-3227](#) - generate test TPC data sets during data load
- [IMPALA-3253](#) - Modify gen_build_version.sh to always output the right version
- [IMPALA-3336](#) - qgen: do not randomly generate query options
- [IMPALA-3376](#) - Extra definition level when writing Parquet files
- [IMPALA-3418](#) - The Impala FE project relies on Z-tools snapshot builds
- [IMPALA-3442](#) - Replace '>' with '>>' in template decls
- [IMPALA-3449](#) - Kudu deploy.py should find clusters by displayName
- [IMPALA-3454](#) - Kudu deletes may fail if subqueries are used
- [IMPALA-3470](#) - DecompressorTest is flaky
- [IMPALA-3491](#) - Merge test_hbase_metadata.py into compute_stats.py. Use unique db fixture
- [IMPALA-3501](#) - ee tests: detect build type and support different timeouts based on the same
- [IMPALA-3507](#) - update binutils version to fix slow linking
- [IMPALA-3521](#) - Impalad should communicate with the statestore after binding to the hs2 and besswax ports
- [IMPALA-3530](#) - Clean up test_ddl.py. Part 1
- [IMPALA-3567](#) - Part 1: groundwork to make Join build sides DataSinks
- [IMPALA-3575](#) - Add retry to backend connection request and rpc timeout
- [IMPALA-3587](#) - Get rid of not_default_fs skip marker
- [IMPALA-3600](#) - Add missing admission control tests
- [IMPALA-3606](#) - Fix Java NPE when trying to add an existing partition
- [IMPALA-3611](#) - track unused Disk IO buffer memory
- [IMPALA-3627](#) - Clean up RPC structures in ImpalaInternalService
- [IMPALA-3632](#) - Add script for running cpclean over the BE code
- [IMPALA-3647](#) - track runtime filter memory in separate tracker
- [IMPALA-3656](#) - Hitting DCHECK/CHECK does not write minidumps
- [IMPALA-3664](#) - S3A test_keys_do_not_work fails
- [IMPALA-3674](#) - Lazy materialization of LLVM module bitcode
- [IMPALA-3677](#) - Write minidump on SIGUSR1
- [IMPALA-3682](#) - Don't retry unrecoverable socket creation errors
- [IMPALA-3687](#) - Prefer Avro field name during schema reconciliation
- [IMPALA-3715](#) - Include total usage of JVM memory
- [IMPALA-3715](#) - Include more info by default in Impala debug memz webpage
- [IMPALA-3716](#) - Add Memory Tab in query's Details page
- [IMPALA-3727](#) - Change microbenchmarks to use percentile-based reporting
- [IMPALA-3729](#) - batch_size=1 coverage for avro scanner
- [IMPALA-3734](#) - C++11 - Replace boost::shared_ptr with std:: equivalent
- [IMPALA-3736](#) - Move Impala HTTP handlers to a separate class
- [IMPALA-3737](#) - Local filesystem build failed loading custom schemas
- [IMPALA-3751](#) - fix clang build errors and warnings
- [IMPALA-3753](#) - Disable create table test for old aggs and joins
- [IMPALA-3756](#) - Fix wrong argument type in HiveStringsTest
- [IMPALA-3757](#) - Add missing lock in RuntimeProfile::ComputeTimeInProfile
- [IMPALA-3762](#) - Download Python requirements before they are needed
- [IMPALA-3763](#) - download_requirements fixes
- [IMPALA-3764](#) - fuzz test HDFS scanners and fix parquet bugs found
- [IMPALA-3767](#) - bootstrap_virtualenv fails to find cython distribution

- [IMPALA-3774](#) - fix download_requirements for older Python versions
- [IMPALA-3778](#) - Fix ASF packaging build
- [IMPALA-3779](#) - Disable cache pool reader thread when HDFS isn't running
- [IMPALA-3780](#) - avoid many small reads past end of block
- [IMPALA-3786](#) - Remove "Cloudera" from impalad webpage title
- [IMPALA-3790](#) - AC tests timeout in codecov coverage builds
- [IMPALA-3799](#) - Make MAX_SCAN_RANGE_LENGTH accept formatted quantities
- [IMPALA-3806](#) - remove a few modern shell idioms to improve RHEL5 support
- [IMPALA-3817](#) - Ensure filter hash function is the same on all hardware
- [IMPALA-3839](#) - Fix race condition in impala_cluster.py
- [IMPALA-3843](#) - Update warning for non-SSSE3 CPUs
- [IMPALA-3845](#) - Split up hdfs-parquet-scanner.cc into more files/components
- [IMPALA-3852](#) - Remove Derby and Shiro FE dependencies
- [IMPALA-3854](#) - Fix use-after-free in HdfsTextScanner::Close()
- [IMPALA-3856](#) - Fix BinaryPredicate normalization for Kudu
- [IMPALA-3857](#) - KuduScanNode race on returning "optional" threads
- [IMPALA-3864](#) - qgen: reduce likelihood of create_query() exceptions
- [IMPALA-3866](#) - consistent user-facing terminology for scratch dirs
- [IMPALA-3881](#) - Add DataTables 1.10.12 to www/
- [IMPALA-3886](#) - Improve log of pip_download.py
- [IMPALA-3892](#) - qgen: always run Impala with -convert_legacy_hive_parquet_utc_timestamps=true
- [IMPALA-3905](#) - Add HdfsScanner::GetNext() interface and implementation for Parquet
- [IMPALA-3906](#) - Materialize implicitly referenced IR functions
- [IMPALA-3914](#) - SKIP_TOOLCHAIN_BOOTSTRAP skips Python package downloads
- [IMPALA-3918](#) - Remove Cloudera copyrights and add ASF license header
- [IMPALA-3923](#) - fix overflow in BufferedTupleStream::GetRows()
- [IMPALA-3924](#) - Ubuntu16 support
- [IMPALA-3936](#) - BufferedBlockMgr fixes for Pin() while write in flight
- [IMPALA-3939](#) - Data loading may fail on tpch kudu
- [IMPALA-3943](#) - Do not throw scan errors for empty Parquet files
- [IMPALA-3946](#) - fix MemPool integrity issues with empty chunks
- [IMPALA-3952](#) - Clear scratch batch mem pool if Open() failed
- [IMPALA-3953](#) - Fixes for KuduScanNode BE test failure
- [IMPALA-3954](#) - Add unique_database to scanner test
- [IMPALA-3957](#) - Test failure in S3 build: TestLoadData.test_load
- [IMPALA-3964](#) - Fix crash when a count(*) is performed on a nested collection
- [IMPALA-3969](#) - stress test: add option to set common query options
- [IMPALA-3972](#) - Improve display of /varz page
- [IMPALA-3992](#) - bad shell error message when running nonexistent file
- [IMPALA-3223](#) - Fix the include path for Hadoop's header files
- [IMPALA-3253](#) - Modify gen_build_version.sh to always output the right version
- [IMPALA-3832](#) - Merge "fix DCHECK calling AddDetail() on an OK status" into cdh5-trunk
- [IMPALA-3832](#) - fix DCHECK calling AddDetail() on an OK status
- [IMPALA-3905](#) - Fixes for the modified HdfsScanner interface
- [OOZIE-1814](#) - Oozie should mask any passwords in logs and REST interfaces
- [OOZIE-2244](#) - Oozie should mask passwords in the logs when logging command arguments
- [OOZIE-2349](#) - Method getCoordJobInfo(String jobId, String filter, int offset, int length, boolean desc) is not present in LocalOozieClientCoord
- [OOZIE-2362](#) - SQL injection in BulkJPAExecutor
- [OOZIE-2402](#) - oozie-setup.sh sharelib create takes a long time on large clusters

CDH 5 Release Notes

- [OOZIE-2447](#) - Illegal character 0x0 oozie client
- [OOZIE-2462](#) - When calling ParamChecker.notNull() in CoordActionsIgnoreXCommand.java, 'Action' should be passed instead of 'Action cannot be null'
- [OOZIE-2467](#) - Oozie can shutdown itself on long GC pause
- [OOZIE-2475](#) - Oozie does not cleanup action dir of killed actions
- [OOZIE-2476](#) - When one of the action from fork fails with transient error, WF never joins
- [OOZIE-2493](#) - TestDistcpMain deletes action.xml from wrong filesystem
- [OOZIE-2515](#) - Duplicate information for 'Changing endtime/pausetime of a Bundle Job' in CommandLineTool wiki
- [OOZIE-2516](#) - Update web service documentation for jobs API
- [OOZIE-2539](#) - Incorrect property key is used for 'hive log4j configuration file for execution mode'
- [OOZIE-2541](#) - Possible resource leak in Hive2Credentials
- [OOZIE-2542](#) - Option to disable OpenJPA BrokerImpl finalization
- [OOZIE-2546](#) - Improperly closed resources in OozieDBCLI
- [OOZIE-2548](#) - Flaky test TestZKLocksService.testLockReaper
- [OOZIE-2550](#) - Flaky tests in TestZKUUIDService.java
- [OOZIE-2551](#) - Feature request: epoch timestamp generation
- [OOZIE-2552](#) - Update ActiveMQ version for security and other fixes
- [OOZIE-2553](#) - Cred tag is required for all actions in the workflow even if an action does not require it
- [OOZIE-2556](#) - TestAbandonedCoordChecker.testCatchupJob is flaky
- [OOZIE-2574](#) - Oozie to support replication-enabled mysql urls
- [OOZIE-2577](#) - Flaky tests TestCoordActionInputCheckXCommand.testTimeout and testTimeoutWithException
- [OOZIE-2578](#) - Oozie example distcp job fails to run within an encrypted zone with checksum match error
- [OOZIE-2579](#) - Bulk kill tests in TestBulkWorkflowXCommand might fail because of a race condition
- [OOZIE-2581](#) - Oozie should reset SecurityManager in finally block
- [OOZIE-2587](#) - Disable SchedulerService on certain tests
- [OOZIE-2603](#) - Give thread pools a meaningful name in CallableQueueService and SchedulerService
- [OOZIE-2615](#) - Flaky tests TestCoordActionsKillXCommand.testActionKillCommandActionNumbers and testActionKillCommandData
- [OOZIE-2623](#) - Oozie should use a dummy OutputFormat
- [OOZIE-2632](#) - Provide database dump/load functionality to make database migration easier
- [OOZIE-2660](#) - Create documentation for DB Dump/Load functionality
- [PIG-2949](#) - JsonLoader only reads arrays of objects
- [PIG-3413](#) - JsonLoader fails the pig job in case of malformed json input
- [PIG-3619](#) - Update piggybank to include XPath function
- [PIG-3664](#) - Piggy Bank XPath UDF can't be called
- [PIG-3730](#) - Performance issue in SelfSpillBag
- [PIG-3970](#) - Increase PermGen size, tests ran out of memory
- [PIG-4355](#) - Piggybank: XPath cant handle namespace in xpath, nor can it return more than one match
- [PIG-4787](#) - CHD:14987: BackportLog JSONLoader exception while parsing records
- [SENTRY-1201](#) - Sentry ignores database prefix for MSCK statement
- [SENTRY-1208](#) - Make HOST implied in privileges if not specified explicitly
- [SENTRY-1228](#) - SimpleFileProviderBackend error message, remove missing spaces
- [SENTRY-1230](#) - Add filesystem tests to test Sentry with user data storage on S3
- [SENTRY-1231](#) - Sentry doesn't secure index location uri, when do "CREATE INDEX LOCATION "/uri"
- [SENTRY-1252](#) - grantServerPrivilege and revokeServerPrivilege should treat * and ALL as synonyms when action is not explicitly specified
- [SENTRY-1253](#) - SentryShellKafka incorrectly sets component as "KAFKA"
- [SENTRY-1265](#) - Sentry service should not require a TGT
- [SENTRY-1269](#) - Converter vs Convertor - spelling inconsistent
- [SENTRY-1292](#) - Reorder DBModelAction EnumSet

- [SENTRY-1293](#) - Avoid converting string permission to Privilege object
- [SENTRY-1299](#) - Add a test case to verify SentryStore#verifySentryStoreSchema works
- [SENTRY-1311](#) - Improve usability of URI privileges by supporting mixed use of URLs with and without scheme
- [SENTRY-1320](#) - Truncate table db_name.table_name fails
- [SENTRY-1334](#) - backport[column level privileges] test and add test for CTAS and Create View AS SELECT (cross databases cases)
- [SENTRY-1345](#) - ACLS on table folder disappear after insert for unpartitioned tables
- [SENTRY-1376](#) - Fix alter property case correctly - Deletes ACLS on the table
- [SENTRY-1378](#) - Login fails for a secure Sentry Web UI
- [SENTRY-1416](#) - kafka-sentry tool service name default is different from KafkaSentryAuthorizer default service name
- [SENTRY-1429](#) - TestHDFSIntegration improvements
- [SENTRY-1454](#): Fix intermittent timeout issue for TestHDFSIntegration
- [SENTRY-1447](#) - When s3 is configured as HDFS defaultFS and Hive Warehouse Dir, need to fix some e2e test failures
- [SENTRY-1450](#) - Have privilege converter set by Kafka binding
- [SENTRY-1453](#) - Enable passing sentry client cache configs from Kafka conf
- [SENTRY-1233](#) - Logging improvements for SentryConfigToolSolr.
- [SENTRY-1119](#) - Allow data engines to obtain the ActionFactory directly from the configuration, instead of having hardcoded component-specific classes. This will allow external data engines to integrate with Sentry easily.
- [SENTRY-1229](#) - Added a basic configurable cache to SentryGenericProviderBackend.
- [SOLR-5750](#) - Backup/Restore API for SolrCloud
- [SOLR-5922](#) - Add support for adding core properties to SolrJ Collection Admin Request calls
- [SOLR-6637](#) - Solr should have a way to restore a core
- [SOLR-6761](#) - Ability to ignore commit and optimize requests from clients when running in SolrCloud mode
- [SOLR-7294](#) - Migrate API fails with 'Invalid status request: notfoundretried 6times' message
- [SOLR-7374](#) - Backup/Restore should provide a param for specifying the directory implementation it should use
- [SOLR-7766](#) - Support creation of a coreless collection
- [SOLR-8449](#) - Multiple restores on the same core does not work
- [SOLR-9055](#) - Make collection backup/restore extensible
- [SOLR-9242](#) - Collection level backup/restore should provide a param for specifying the repository implementation it should use
- [SOLR-9269](#) - Ability to create/delete/list snapshots for a solr core
- [SOLR-9310](#) - PeerSync fails on a node restart due to IndexFingerPrint mismatch
- [SOLR-9326](#) - Ability to create/delete/list snapshots for a solr collection
- [SOLR-9441](#) - Solr collection backup on HDFS can only be manipulated by the Solr process owner
- [SPARK-10372](#) - [CORE] Basic test framework for entire spark scheduler
- [SPARK-12009](#) - [YARN] Avoid reallocating YARN container when driver wants to stop all executors
- [SPARK-12941](#) - [SQL][MASTER] Spark-SQL JDBC Oracle dialect fails to map string datatypes to Oracle VARCHAR datatype
- [SPARK-13904](#) - [SCHEDULER] Add support for pluggable cluster manager
- [SPARK-14881](#) - [PYTHON] [SPARKR] pyspark and sparkR shell default log level should match spark-shell/Scala
- [SPARK-15714](#) - [CORE] Fix flaky o.a.s.scheduler.BlacklistIntegrationSuite
- [SPARK-15754](#) - [YARN] Not letting the credentials containing HDFS delegation tokens to be added in current user credential
- [SPARK-15865](#) - [CORE] Blacklist should not result in job hanging with less than 4 executors
- [SPARK-16711](#) - YarnShuffleService doesn't re-init properly on YARN rolling upgrade
- [SPARK-17171](#) - [WEB UI] DAG will list all partitions in the graph
- [SPARK-17433](#) - YarnShuffleService doesn't handle moving credentials levelDb
- [SPARK-17611](#) - [YARN][TEST] Make shuffle service test acutally test authentication
- [SPARK-17644](#) - [CORE] Do not add failedStages when abortStage for fetch failure
- [SPARK-17696](#) - [CORE] Partial backport of to branch-1.6

CDH 5 Release Notes

- [SQOOP-2938](#) - Mainframe import module extension to support data sets on tape
- [SQOOP-2952](#) - Fixing bug
- [SQOOP-3021](#) - ClassWriter fails if a column name contains a backslash character
- [ZOOKEEPER-2405](#) - getTGT() in Login.java mishandles confidential information
- [ZOOKEEPER-2477](#) - Documentation should refer to Java CLI shell and not C CLI shell

Issues Fixed in CDH 5.8.x

The following topics describe issues fixed in CDH 5.8.x, from newest to oldest release. You can also review [What's New In CDH 5.8.x](#) on page 16 or [Known Issues in CDH 5](#) on page 111.

Issues Fixed in CDH 5.8.3

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.8.3:

- [FLUME-2797](#) - Use SourceCounter for SyslogTcpSource
- [FLUME-2844](#) - SpillableMemoryChannel must start ChannelCounter
- [HADOOP-12548](#) - Read s3a credentials from a Credential Provider
- [HADOOP-13353](#) - LdapGroupsMapping getPassword should not return null when IOException throws
- [HADOOP-13526](#) - Add detailed logging in KMS for the authentication failure of proxy user
- [HADOOP-13558](#) - UserGroupInformation created from a Subject incorrectly tries to renew the Kerberos ticket
- [HADOOP-13579](#) - Fix source-level compatibility after HADOOP-11252
- [HADOOP-13638](#) - KMS should set UGI's Configuration object properly
- [HDFS-7415](#) - Move FSNameSystem.resolvePath() to FSDirectory
- [HDFS-7420](#) - Delegate permission checks to FSDirectory
- [HDFS-7463](#) - Simplify FSNamesystem#getBlockLocationsUpdateTimes
- [HDFS-7478](#) - Move org.apache.hadoop.hdfs.server.namenode.NNConf to FSNamesystem
- [HDFS-7517](#) - Remove redundant non-null checks in FSNamesystem#getBlockLocations
- [HDFS-8224](#) - Schedule a block for scanning if its metadata file is corrupt
- [HDFS-8269](#) - getBlockLocations() does not resolve the .reserved path and generates incorrect edit logs when updating the atime
- [HDFS-9601](#) - NNThroughputBenchmark.BlockReportStats should handle NotReplicatedYetException on adding block.
- [HDFS-9781](#) - FsDatasetImpl#getBlockReports can occasionally throw NullPointerException
- [HDFS-10641](#) - TestBlockManager#testBlockReportQueueing fails intermittently
- [HDFS-10879](#) - TestEncryptionZonesWithKMS#testReadWrite fails intermittently
- [HDFS-10962](#) - TestRequestHedgingProxyProvider fails intermittently
- [HDFS-10963](#) - Reduce log level when network topology cannot find enough datanodes
- [MAPREDUCE-6628](#) - Potential memory leak in CryptoOutputStream
- [MAPREDUCE-6641](#) - TestTaskAttempt fails in trunk
- [MAPREDUCE-6718](#) - Add progress log to JHS during startup
- [MAPREDUCE-6771](#) - RMContainerAllocator sends container diagnostics event after corresponding completion event
- [YARN-4940](#) - yarn node -list -all fails if RM starts with decommissioned node
- [HBASE-15856](#) - Addendum Fix UnknownHostException import in MetaTableLocator
- [HBASE-15856](#) - Do not cache unresolved addresses for connections
- [HBASE-16294](#) - hbck reporting "No HDFS region dir found" for replicas
- [HBASE-16699](#) - Overflows in AverageIntervalRateLimiter's refill() and getWaitInterval()
- [HBASE-16767](#) - Mob compaction needs to clean up files in /hbase/mobdir/.tmp and /hbase/mobdir/.tmp/.bulkload when running into IO exceptions
- [HIVE-9570](#) - Investigate test failure on union_view.q
- [HIVE-10965](#) - Direct SQL for stats fails in 0-column case

- [HIVE-12083](#) - HIVE-10965 introduces thrift error if partNames or colNames are empty
- [HIVE-12475](#) - Parquet schema evolution within array<struct> does not work
- [HIVE-12785](#) - View with union type and UDF to the struct is broken
- [HIVE-13058](#) - Add session and operation_log directory deletion messages
- [HIVE-13198](#) - Authorization issues with cascading views
- [HIVE-13237](#) - Select parquet struct field with upper case throws NPE
- [HIVE-13620](#) - Merge llap branch work to master
- [HIVE-13625](#) - Hive Prepared Statement when executed with escape characters in parameter fails
- [HIVE-13645](#) - Beeline needs null-guard around hiveVars and hiveConfVars read
- [HIVE-14296](#) - Session count is not decremented when HS2 clients do not shutdown cleanly
- [HIVE-14383](#) - SparkClientImpl should pass principal and keytab to spark-submit instead of calling kinit explicitly
- [HIVE-14715](#) - Hive throws NumberFormatException with query with Null value
- [HIVE-14743](#) - ArrayIndexOutOfBoundsException - HBASE-backed views' query with JOINs
- [HIVE-14784](#) - Operation logs are disabled automatically if the parent directory does not exist.
- [HIVE-14805](#) - Subquery inside a view will have the object in the subquery as the direct input
- [HUE-4064](#) - Format creation and update date on the table details popover
- [HUE-4138](#) - Last modified time of a saved query is not in the correct timezone
- [HUE-4141](#) - Graph breaks for external workflows when there is more than one kill node
- [HUE-4804](#) - Download function of HTML widget breaks the display
- [HUE-4809](#) - Add trustore parameters only if SSL is turned on
- [HUE-4809](#) - Only add trustore paths when they are actually existing
- [HUE-4810](#) - Fix tests by setting data to valid JSON type
- [HUE-4871](#) - An unprivileged user can enumerate users
- [HUE-4891](#) - An unprivileged user can list document items
- [HUE-4916](#) - Truncate last name to 30 chars on ldap import
- [HUE-4968](#) - Remove access to /oozie/import_wokflow when v2 is enabled
- [HUE-4994](#) - Consider default path for decision nodes in dashboard graph
- [HUE-5041](#) - Hue export large file to HDFS does not work on non-default database
- [IMPALA-1619](#) - Support 64-bit allocations
- [IMPALA-3687](#) - Prefer Avro field name during schema reconciliation
- [IMPALA-3751](#) - Fix clang build errors and warnings
- [IMPALA-4135](#) - Thrift threaded server times-out connections during high load
- [IMPALA-4170](#) - Fix identifier quoting in COMPUTE INCREMENTAL STATS
- [IMPALA-4180](#) - Synchronize accesses to RuntimeState::reader_contexts_
- [IMPALA-4196](#) - Cross compile bit-byte-functions
- [IMPALA-4237](#) - Fix materialization of 4-byte decimals in data source scan node
- [OOZIE-1814](#) - Oozie should mask any passwords in logs and REST interfaces
- [SOLR-9310](#) - PeerSync fails on a node restart due to IndexFingerPrint mismatch
- [SPARK-12009](#) - Avoid reallocating YARN container when driver wants to stop all Executors
- [SPARK-12392](#) - Optimize a location order of broadcast blocks by considering preferred local hosts
- [SPARK-12941](#) - Spark-SQL JDBC Oracle dialect fails to map string datatypes to Oracle VARCHAR datatype mapping
- [SPARK-12941](#) - Spark-SQL JDBC Oracle dialect fails to map string datatypes to Oracle VARCHAR datatype
- [SPARK-13328](#) - Poor read performance for broadcast variables with dynamic resource allocation
- [SPARK-16625](#) - General data types to be mapped to Oracle
- [SPARK-16711](#) - YarnShuffleService doesn't re-init properly on YARN rolling upgrade
- [SPARK-17171](#) - DAG will list all partitions in the graph
- [SPARK-17433](#) - YarnShuffleService doesn't handle moving credentials levelDb
- [SPARK-17611](#) - Make shuffle service test really test authentication
- [SPARK-17644](#) - Do not add failedStages when abortStage for fetch failure
- [SPARK-17696](#) - Partial backport of to branch-1.6.

CDH 5 Release Notes

- [SQOOP-3021](#) - ClassWriter fails if a column name contains a backslash character

Issues Fixed in CDH 5.8.2

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.8.2:

- [FLUME-1899](#) - Make SpoolDir work with subdirectories
- [FLUME-2652](#) - Documented transaction handling semantics incorrect in developer guide.
- [FLUME-2901](#) - Document Kerberos setup for Kafka channel
- [FLUME-2910](#) - AsyncHBaseSink: Failure callbacks should log the exception that caused them
- [FLUME-2913](#) - Don't strip SLF4J from imported classpaths
- [FLUME-2918](#) - Speed up TaildirSource on directories with many files
- [FLUME-2922](#) - Sync SequenceFile.Writer before calling hflush
- [FLUME-2923](#) - Bump asynchbase version to 1.7.0
- [FLUME-2934](#) - Document new cachePatternMatching option for TaildirSource
- [FLUME-2935](#) - Bump java target version to 1.7
- [FLUME-2948](#) - docs: Fix parameters on Replicating Channel Selector example
- [FLUME-2954](#) - Make raw data appearing in log messages explicit
- [FLUME-2963](#) - FlumeUserGuide: Fix error in Kafka Source properties table
- [FLUME-2972](#) - Handle offset migration in the new Kafka Channel
- [FLUME-2975](#) - docs: Fix NetcatSource example
- [FLUME-2982](#) - Add localhost escape sequence to HDFS sink
- [FLUME-2983](#) - Handle offset migration in the new Kafka Source
- [HADOOP-8436](#) - NPE in getLocalPathForWrite (path, conf) when the required context item is not configured
- [HADOOP-8437](#) - getLocalPathForWrite should throw IOException for invalid paths
- [HADOOP-8934](#) - Shell command ls should include sort options (Jonathan Allen via aw)
- [HADOOP-8934](#) - Shell command ls should include sort options
- [HADOOP-10048](#) - LocalDirAllocator should avoid holding locks while accessing the filesystem
- [HADOOP-10971](#) - Add -C flag to make `hadoop fs -ls` print filenames only
- [HDFS-10512](#) - VolumeScanner can terminate due to NPE in DataNode.reportBadBlocks.
- [HADOOP-11361](#) - Fix a race condition in MetricsSourceAdapter.updateJmxCache.
- [HADOOP-11469](#) - KMS should skip default.key.acl and whitelist.key.acl when loading key acl.
- [HADOOP-11901](#) - BytesWritable fails to support 2G chunks due to integer overflow
- [HADOOP-12252](#) - LocalDirAllocator should not throw NPE with empty string configuration
- [HADOOP-12609](#) - Fix intermittent failure of TestDecayRpcScheduler.
- [HADOOP-12659](#) - Incorrect usage of config parameters in token manager of KMS
- [HADOOP-12963](#) - Allow using path style addressing for accessing the s3 endpoint.
- [HADOOP-13079](#) - Add -q option to ls to print ? instead of non-printable characters
- [HADOOP-13132](#) - Handle ClassCastException on AuthenticationException in LoadBalancingKMSClientProvider
- [HADOOP-13155](#) - Implement TokenRenewer to renew and cancel delegation tokens in KMS
- [HADOOP-13251](#) - Authenticate with Kerberos credentials when renewing KMS delegation token
- [HADOOP-13255](#) - KMSClientProvider should check and renew tgt when doing delegation token operations.
- [HADOOP-13263](#) - Reload cached groups in background after expiry.
- [HADOOP-13270](#) - BZip2CompressionInputStream finds the same compression marker twice in corner case, causing duplicate data blocks
- [HADOOP-13381](#) - KMS clients should use KMS Delegation Tokens from current UGI
- [HADOOP-13437](#) - KMS should reload whitelist and default key ACLs when hot-reloading [HADOOP-13457](#) - Remove hardcoded absolute path for shell executable.
- [HADOOP-13487](#) - Hadoop KMS should load old delegation tokens from Zookeeper on startup
- [HDFS-4210](#) - Throw helpful exception when DNS entry for JournalNode cannot be resolved
- [HDFS-6434](#) - Default permission for creating file should be 644 for WebHdfs/HttpFS

- [HDFS-7597](#) - DelegationTokenIdentifier should cache the TokenIdentifier to UGI mapping
- [HDFS-8581](#) - ContentSummary on / skips further counts on yielding lock
- [HDFS-8829](#) - Make SO_RCVBUF and SO_SNDBUF size configurable for DataTransferProtocol sockets and allow configuring auto-tuning
- [HDFS-8897](#) - Balancer should handle fs.defaultFS trailing slash in HA
- [HDFS-9085](#) - Show renewer information in DelegationTokenIdentifier#toString
- [HDFS-9137](#) - DeadLock between DataNode#refreshVolumes and BPOfferService#registrationSucceeded.
- [HDFS-9141](#) - Thread leak in Datanode#refreshVolumes.
- [HDFS-9259](#) - Make SO_SNDBUF size configurable at DFSClient side for hdfs write scenario.
- [HDFS-9276](#) - Failed to Update HDFS Delegation Token for long running application in HA mode
- [HDFS-9365](#) - Balancer does not work with the HDFS-6376 HA setup.
- [HDFS-9461](#) - DiskBalancer: Add Report Command
- [HDFS-9466](#) - TestShortCircuitCache#testDataXceiverCleansUpSlotsOnFailure is flaky
- [HDFS-9700](#) - DFSClient and DFSOutputStream should set TCP_NODELAY on sockets for DataTransferProtocol
- [HDFS-9732](#) - , Improve DelegationTokenIdentifier.toString() for better logging
- [HDFS-9805](#) - Add server-side configuration for enabling TCP_NODELAY for DataTransferProtocol and default it to true
- [HDFS-9906](#) - Remove spammy log spew when a datanode is restarted.
- [HDFS-9939](#) - Increase DecompressorStream skip buffer size
- [HDFS-9958](#) - BlockManager#createLocatedBlocks can throw NPE for corruptBlocks on failed storages
- [HDFS-10270](#) - TestJMXGet: testNameNode() fails
- [HDFS-10381](#) - , DataStreamer DataNode exclusion log message should be warning.
- [HDFS-10403](#) - DiskBalancer: Add cancel command
- [HDFS-10457](#) - DataNode should not auto-format block pool directory if VERSION is missing.
- [HDFS-10481](#) - HTTPFS server should correctly impersonate as end user to open file
- [HDFS-10500](#) - Diskbalancer: Print out information when a plan is not generated
- [HDFS-10501](#) - DiskBalancer: Use the default datanode port if port is not provided
- [HDFS-10516](#) - Fix bug when warming up EDEK cache of more than one encryption zone
- [HDFS-10517](#) - DiskBalancer: Support help command
- [HDFS-10525](#) - Fix NPE in CacheReplicationMonitor#rescanCachedBlockMap
- [HDFS-10541](#) - Diskbalancer: When no actions in plan, error message says "Plan was generated more than 24 hours ago"
- [HDFS-10544](#) - Balancer doesn't work with IPFailoverProxyProvider.
- [HDFS-10552](#) - DiskBalancer "-query" results in NPE if no plan for the node
- [HDFS-10559](#) - DiskBalancer: Use SHA1 for Plan ID
- [HDFS-10567](#) - Improve plan command help message
- [HDFS-10588](#) - False alarm in datanode log - ERROR - Disk Balancer is not enabled
- [HDFS-10598](#) - DiskBalancer does not execute multi-steps plan
- [HDFS-10600](#) - PlanCommand#getThrsholdPercentage should not use throughput value.
- [HDFS-10643](#) - Namenode should use loginUser(hdfs) to generateEncryptedKey
- [HDFS-10681](#) - DiskBalancer: query command should report Plan file path apart from PlanID.
- [HDFS-10822](#) - Log DataNodes in the write pipeline. John Zhuge via Lei Xu
- [MAPREDUCE-4784](#) - TestRecovery occasionally fails
- [MAPREDUCE-6359](#) - In RM HA setup, Cluster tab links populated with AM hostname instead of RM
- [MAPREDUCE-6442](#) - Stack trace is missing when error occurs in client protocol provider's constructor Contributed by Chang Li.
- [MAPREDUCE-6473](#) - Revert "Revert "Job submission can take a long time during Cluster initialization
- [MAPREDUCE-6473](#) - Revert "Job submission can take a long time during Cluster initialization
- [MAPREDUCE-6473](#) - Job submission can take a long time during Cluster initialization
- [MAPREDUCE-6670](#) - TestJobListCache#testEviction sometimes fails on Windows with timeout

CDH 5 Release Notes

- [MAPREDUCE-6680](#) - JHS UserLogDir scan algorithm sometime could skip directory with update in CloudFS (Azure FileSystem, S3, etc)
- [MAPREDUCE-6738](#) - TestJobListCache.testAddExisting failed intermittently in slow VM testbed
- [MAPREDUCE-6761](#) - Regression when handling providers - invalid configuration ServiceConfiguration causes Cluster initialization failure
- [YARN-2605](#) - [RM HA] Rest api endpoints doing redirect incorrectly.
- [YARN-2977](#) - Fixed intermittent TestNMClient failure.
- [YARN-4411](#) - RMAppAttemptImpl#createApplicationAttemptReport throws IllegalArgumentException
- [YARN-4459](#) - container-executor should only kill process groups
- [YARN-4866](#) - FairScheduler: AMs can consume all vcores leading to a livelock when using FAIR policy.
- [YARN-4878](#) - Expose scheduling policy and max running apps over JMX for Yarn queues.
- [YARN-4989](#) - TestWorkPreservingRMRestart#testCapacitySchedulerRecovery fails intermittently
- [YARN-5048](#) - DelegationTokenRenewer#skipTokenRenewal may throw NPE
- [YARN-5077](#) - Fix FSLeafQueue#getFairShare() for queues with zero fairshare.
- [YARN-5107](#) - TestContainerMetrics fails.
- [YARN-5272](#) - Handle queue names consistently in FairScheduler.
- [YARN-5608](#) - TestAMRMClient.setup() fails with ArrayOutOfBoundsException
- [HBASE-14644](#) - Region in transition metric is broken -- addendum
- [HBASE-14644](#) - Region in transition metric is broken
- [HBASE-14818](#) - user_permission does not list namespace permissions
- [HBASE-14963](#) - Remove use of Guava Stopwatch from HBase client code
- [HBASE-15465](#) - userPermission returned by getUserPermission() for the selected namespace does not have namespace set
- [HBASE-15496](#) - Throw RowTooBigException only for user scan/get
- [HBASE-15621](#) - Suppress Hbase SnapshotHFile cleaner error messages when a snapshot is going on
- [HBASE-15683](#) - Min latency in latency histograms are emitted as Long.MAX_VALUE
- [HBASE-15698](#) - Increment TimeRange not serialized to server
- [HBASE-15746](#) - Remove extra RegionCoprocessor preClose() in RSRpcServices#closeRegion
- [HBASE-15808](#) - Reduce potential bulk load intermediate space usage and waste
- [HBASE-15872](#) - Split TestWALProcedureStore
- [HBASE-15873](#) - ACL for snapshot restore / clone is not enforced
- [HBASE-15925](#) - provide default values for hadoop compat module related properties that match default hadoop profile.
- [HBASE-16034](#) - Fix ProcedureTestingUtility#LoadCounter.setMaxProclId()
- [HBASE-16056](#) - Procedure v2 - fix master crash for FileNotFoundException
- [HBASE-16093](#) - Fix splits failed before creating daughter regions leave meta inconsistent
- [HBASE-16135](#) - PeerClusterZnode under rs of removed peer may never be deleted
- [HBASE-16194](#) - Should count in MSLAB chunk allocation into heap size change when adding duplicate cells
- [HBASE-16195](#) - Should not add chunk into chunkQueue if not using chunk pool in HeapMemStoreLAB
- [HBASE-16207](#) - can't restore snapshot without "Admin" permission
- [HBASE-16227](#) - [Shell] Column value formatter not working in scans. Tested : manually using shell.
- [HBASE-16284](#) - Unauthorized client can shutdown the cluster
- [HBASE-16288](#) - HFile intermediate block level indexes might recurse forever creating multi TB files.
- [HBASE-16319](#) - Fix TestCacheOnWrite after HBASE-16288.
- [HBASE-16317](#) - revert all ESAPI changes
- [HBASE-16318](#) - fail build while rendering velocity template if dependency license isn't in whitelist.
- [HBASE-16318](#) - consistently use the correct name for 'Apache License, Version 2.0'
- [HBASE-16321](#) - ensure no findbugs-jsr305
- [HBASE-16340](#) - exclude Xerces implementation jars from coming in transitively.
- [HBASE-16360](#) - TableMapReduceUtil addHBaseDependencyJars has the wrong class name for PrefixTreeCodec

- [HIVE-7443](#) - Fix HiveConnection to communicate with Kerberized Hive JDBC server and alternative JDks
- [HIVE-10007](#) - Support qualified table name in analyze table compute statistics for columns
- [HIVE-10728](#) - deprecate unix_timestamp(void) and make it deterministic (Sergey Shelukhin, reviewed by Ashutosh Chauhan(Also include the unit tests by HIVE-10932 : Unit test udf_nondeterministic failure due to
- [HIVE-11243](#) - Changing log level in Utilities.getBaseWork
- [HIVE-11432](#) - Hive macro give same result for different arguments
- [HIVE-11487](#) - Add getNumPartitionsByFilter api in metastore api
- [HIVE-11747](#) - Unnecessary error log is shown when executing a "INSERT OVERWRITE LOCAL DIRECTORY" cmd in the embedded mode
- [HIVE-11827](#) - STORED AS AVRO fails SELECT COUNT(*) when empty
- [HIVE-11901](#) - StorageBasedAuthorizationProvider requires write permission on table for SELECT statements
- [HIVE-11980](#) - Follow up on HIVE-11696, exception is thrown from CTAS from the table with table-level serde is Parquet while partition-level serde is JSON
- [HIVE-12277](#) - Hive macro results on macro_duplicate.q different after adding ORDER BY
- [HIVE-12556](#) - Ctrl-C in beeline doesn't kill Tez query on HS2
- [HIVE-12635](#) - Hive should return the latest hbase cell timestamp as the row timestamp value
- [HIVE-13043](#) - Reload function has no impact to function registry
- [HIVE-13090](#) - Hive metastore crashes on NPE with ZooKeeperTokenStore
- [HIVE-13372](#) - Hive Macro overwritten when multiple macros are used in one column
- [HIVE-13462](#) - HiveResultSetMetaData.getPrecision() fails for NULL columns
- [HIVE-13590](#) - Kerberized HS2 with LDAP auth enabled fails in multi-domain LDAP case
- [HIVE-13704](#) - Don't call DistCp.execute() instead of DistCp.run()
- [HIVE-13736](#) - View's input/output formats are TEXT by default.
- [HIVE-13749](#) - Memory leak in Hive Metastore
- [HIVE-13884](#) - Disallow queries in HMS fetching more than a configured number of partitions
- [HIVE-13932](#) - Hive SMB Map Join with small set of LIMIT failed with NPE
- [HIVE-13953](#) - Issues in HiveLockObject.equals method
- [HIVE-13991](#) - Union All on view fail with no valid permission on underneath table
- [HIVE-14006](#) - Hive query with UNION ALL fails with ArrayIndexOutOfBoundsException.
- [HIVE-14015](#) - SMB MapJoin failed for Hive on Spark when kerberized
- [HIVE-14055](#) - directSql - getting the number of partitions is broken
- [HIVE-14098](#) - Logging task properties, and environment variables might contain passwords
- [HIVE-14118](#) - Make the alter partition exception more meaningful
- [HIVE-14187](#) - JDOPersistenceManager objects remain cached if MetaStoreClient#close is not called
- [HIVE-14209](#) - Add some logging info for session and operation management
- [HIVE-14436](#) - Hive 1.2.1/Hitting "ql.Driver: FAILED: IllegalArgumentException Error: , expected at the end of 'decimal(9)" after enabling hive.optimize.skewjoin and with MR engine
- [HIVE-14457](#) - Partitions in encryption zone are still trashed though an exception is returned
- [HIVE-14519](#) - Multi insert query bug
- [HIVE-14538](#) - beeline throws exceptions with parsing hive config when using !sh statement
- [HIVE-14697](#) - Can not access kerberized HS2 Web UI
- [HUE-2689](#) - Sub-workflow submitted from coordinator gets parent workflow graph
- [HUE-2971](#) - Some links of a Fork can point to deleted nodes
- [HUE-3842](#) - HTTP 500 while emptying Hue 3.9 trash directory
- [HUE-3908](#) - [useradmin] Ignore (objectclass=*) filter when searching for LDAP users
- [HUE-3988](#) - Support schemaless collections
- [HUE-3999](#) - list_oozie_workflow page shouldn't break incase of bad json from oozie
- [HUE-4005](#) - Remove oozie.coord.application.path from properties when rerunning workflow
- [HUE-4006](#) - Create new deployment directory when coordinator or bundle is copied
- [HUE-4007](#) - Fix deployment_dir for the bundle in oozie example fixtures

CDH 5 Release Notes

- [HUE-4019](#) - Always fetch the logs on check status
- [HUE-4019](#) - Do not blank error on query with good syntax but invalid query
- [HUE-4021](#) - [libsolr] Allow customization of the Solr path in ZooKeeper
- [HUE-4023](#) - [useradmin] update AuthenticationForm to allow activated users to login
- [HUE-4078](#) - Drag & Drop hive queries shows queries from the trash
- [HUE-4087](#) - Unable to kill jobs with Resource Manager HA enabled
- [HUE-4092](#) - Can't type any / in the HDFS ACLs path input
- [HUE-4119](#) - Change list jobs call to POST
- [HUE-4129](#) - Long running query getting terminated when leaving the editor
- [HUE-4134](#) - [liboozie] Avoid logging truststore credentials
- [HUE-4145](#) - Older queries after upgrade do not provide direct save
- [HUE-4146](#) - Older saved queries defaults to default' DB
- [HUE-4148](#) - Improve import testing of beeswax queries to notebook format
- [HUE-4153](#) - Report last seen progress when running impala query
- [HUE-4164](#) - The ApiHelper should treat any negative status in the response as an error
- [HUE-4177](#) - Horizontal scroll in FF (Chrome fine) with touch pad is extremely slow
- [HUE-4201](#) - Add warning about max limit of cells before truncation in the export / download query result
- [HUE-4202](#) - Enable offset param for fetching jobbrowser logs
- [HUE-4215](#) - Reset API_CACHE on logout
- [HUE-4224](#) - 'Did you know' on home page is gone
- [HUE-4227](#) - Fix unittest for MR API Cache
- [HUE-4238](#) - Ignore history docs in find_jobs_with_no_doc during sync documents
- [HUE-4238](#) - Ignore history docs in find_jobs_with_no_doc during sync documents
- [HUE-4252](#) - Handle 307 redirect from YARN upon standby failover
- [HUE-4252](#) - Handle 307 redirect from YARN upon standby failover
- [HUE-4253](#) - Prompt for variables just once per variable name
- [HUE-4258](#) - Close and pool Spark History Server connections
- [HUE-4265](#) - Bring back the show preview in the assist
- [HUE-4300](#) - Avoid double file listing call on folder search
- [HUE-4321](#) - Batch submit of SQL show USE the correct DB
- [HUE-4333](#) - Properly reset API_CACHE on failover
- [HUE-4346](#) - Query History disappeared after upgrade to 3.10
- [HUE-4353](#) - Typing in the search bar always redirect to the end of the input
- [HUE-4362](#) - List more oozie workflow parameters on the workflow dashboard page
- [HUE-4364](#) - Handle files with carriage return in create table from a file
- [HUE-4365](#) - No information surfaced when LOAD data from Create table from file fails
- [HUE-4375](#) - Horizontal scrollbar can be hidden under the first fixed column
- [HUE-4383](#) - Trashed queries are showing up in the list of saved queries
- [HUE-4406](#) - Fails to start if Hive/Impala Not Installed
- [HUE-4409](#) - Main right scrollbar does not scroll when on the very right of the screen
- [HUE-4411](#) - Enable scrolling past the end of the editor
- [HUE-4412](#) - Errors should scroll to the line AND the column too
- [HUE-4477](#) - Select All is not filtering out the non visible roles from the selection
- [HUE-4493](#) - Fix sync-workflow action when Workflow includes sub-workflow
- [HUE-4515](#) - Remove oozie.bundle.application.path from properties when rerunning workflow
- [HUE-4533](#) - Disable password reveal on IE
- [HUE-4537](#) - Fix database_logging in hue config so it logs debug database messages
- [HUE-4541](#) - fixing Hue job browser - Kerberos mutual authentication error in Hue
- [HUE-4564](#) - Log stderr on failure to coerce password from script
- [HUE-4616](#) - Only select the snippet DB when executing the first statement

- [HUE-4635](#) - Fix duration on jobs page for running jobs
- [HUE-4662](#) - fixing Hue - Wildcard Certificates not supported
- [HUE-4700](#) - Protect against setting XSS in old editor
- [HUE-4738](#) - Use Concurrency and Throttle values set in coordinator settings
- [HUE-4739](#) - fixed Jobbrowser tests which were failing after resource manager pool change
- [HUE-4766](#) - Replace illegal characters on CSV downloads
- [HUE-4781](#) - Fix export to hdfs to use download_cell_limit from beeswax.conf
- [HUE-4801](#) - When importing oozie documents and remapping UUIDs, data should be updated accordingly
- [HUE-4808](#) - Don't show the edit link for sub workflows when submitted outside Hue
- [IMPALA-1346](#) - /1590/2344: fix sorter buffer mgmt when spilling
- [IMPALA-3159](#) - impala-shell does not accept wildcard or SAN certificates
- [IMPALA-3344](#) - Simplify sorter and document/enforce invariants.
- [IMPALA-3441](#) - , IMPALA-3659: check for malformed Avro data
- [IMPALA-3499](#) - Split catalog update.
- [IMPALA-3628](#) - Fix cancellation from shell when security is enabled
- [IMPALA-3633](#) - cancel fragment if coordinator is gone
- [IMPALA-3646](#) - Handle corrupt RLE literal or repeat counts of 0.
- [IMPALA-3670](#) - fix sorter buffer mgmt bugs
- [IMPALA-3678](#) - Fix migration of predicates into union operands with an order by + limit.
- [IMPALA-3680](#) - Cleanup the scan range state after failed hdfs cache reads
- [IMPALA-3711](#) - Remove unnecessary privilege checks in getDbsMetadata().
- [IMPALA-3732](#) - handle string length overflow in avro files
- [IMPALA-3745](#) - parquet invalid data handling
- [IMPALA-3754](#) - fix TestParquet.test_corrupt_rle_counts flakiness
- [IMPALA-3772](#) - Fix admission control stress test.
- [IMPALA-3776](#) - fix 'describe formatted' for Avro tables
- [IMPALA-3820](#) - Handle linkage errors while loading Java UDFs in Catalog
- [IMPALA-3861](#) - Replace BetweenPredicates with their equivalent CompoundPredicate.
- [IMPALA-3915](#) - Register privilege and audit requests when analyzing resolved table refs.
- [IMPALA-3930](#) - Fix shuffle insert hint with constant partition exprs.
- [IMPALA-3940](#) - Fix getting column stats through views.
- [IMPALA-3965](#) - TSSLocketWithWildcardSAN.py not exported as part of impala-shell build lib
- [IMPALA-4020](#) - Handle external conflicting changes to HMS gracefully
- [IMPALA-4049](#) - fix empty batch handling NLJ build side
- [OOZIE-2068](#) - Configuration as part of sharelib
- [OOZIE-2314](#) - Unable to kill old instance child job by workflow or coord rerun by Launcher
- [OOZIE-2329](#) - Make handling yarn restarts configurable
- [OOZIE-2345](#) - Parallel job submission for forked actions
- [OOZIE-2347](#) - AmendRemove unnecessary new Configuration()/new jobConf() calls from oozie
- [OOZIE-2347](#) - amendments patch toRemove unnecessary new Configuration()/new jobConf() calls from oozie
- [OOZIE-2347](#) - Remove unnecessary new Configuration()/new jobConf() calls from oozie
- [OOZIE-2436](#) - Fork/join workflow fails with oozie.action.yarn.tag must not be null
- [OOZIE-2504](#) - Create a log4j.properties under HADOOP_CONF_DIR in Shell Action
- [OOZIE-2533](#) - Patch-1550 - workaround for
- [OOZIE-2555](#) - Oozie SSL enable setup does not return port for admin -servers
- [OOZIE-2567](#) - HCat connection is not closed while getting hcat cred
- [OOZIE-2589](#) - CompletedActionXCommand is hardcoded to wrong priority
- [OOZIE-2649](#) - Can't override sub-workflow configuration property if defined in parent workflow XML
- [OOZIE-2656](#) - OozieShareLibCLI uses op system username instead of Kerberos to upload jars
- [PIG-3807](#) - Pig creates wrong schema after dereferencing nested tuple fields with sorts

CDH 5 Release Notes

- [SENTRY-1201](#) - Sentry ignores database prefix for MSCK statement
- [SENTRY-1311](#) - Improve usability of URI privileges by supporting mixed use of URIs with and without scheme
- [SENTRY-1345](#) - Revert "ACLS on table folder disappear after insert for unpartitioned tables (Sravya Tirukkavalur, Reviewed by: Hao Hao and Anne Yu)"
- [SENTRY-1345](#) - ACLS on table folder disappear after insert for unpartitioned tables
- [SENTRY-1346](#) - add a test case into hdfs acl e2e suite to test a db.tbl wit out partition, can take more than certain number groups. (Anne Yu, reviewed by Haohao).
- [SOLR-6295](#) - Fix child filter query creation to never match parent docs in SolrExampleTests
- [SOLR-7280](#) - Missing test resources
- [SOLR-7280](#) - BackportLoad cores in sorted order and tweak coreLoadThread counts to improve cluster stability on restarts
- [SOLR-7866](#) - Harden code to prevent an unhandled NPE when trying to determine the max value of the version field.
- [SOLR-9091](#) - ZkController#publishAndWaitForDownStates logic is inefficient
- [SOLR-9236](#) - AutoAddReplicas will append an extra /tlog to the update log location on replica failover.
- [SPARK-8428](#) - Fix integer overflows in TimSort
- [SPARK-12339](#) - Added a null check that was removed in
- [SPARK-13242](#) - codegen fallback in case-when if there many branches
- [SPARK-14391](#) - Fix launcher communication test, take 2.
- [SPARK-14963](#) - Fix typo in YarnShuffleService recovery file name
- [SPARK-14963](#) - Using recoveryPath if NM recovery is enabled
- [SPARK-15165](#) - Introduce place holder for comments in generated code
- [SPARK-16106](#) - TaskSchedulerImpl should properly track executors added to existing hosts
- [SPARK-16505](#) - Optionally propagate error during shuffle service startup.
- [SQOOP-2561](#) - Special Character removal from Column name as avro data results in duplicate column and fails the import
- [SQOOP-2846](#) - Sqoop Export with update-key failing for avro data file
- [SQOOP-2906](#) - Optimization of AvroUtil.toAvroIdentifier
- [SQOOP-2920](#) - sqoop performance deteriorates significantly on wide datasets; sqoop 100% on cpu
- [SQOOP-2971](#) - OraOop does not close connections properly
- [SQOOP-2995](#) - Backward incompatibility introduced by Custom Tool options.
- [SQOOP-2999](#) - Sqoop ClassNotFoundException (org.apache.commons.lang3.StringUtils) is thrown when executing Oracle direct import map task

Issues Fixed in CDH 5.8.0

CDH 5.8.0 fixes the following issues.

Apache Flume

Flume fully compatible with Kafka 2.x

In release CDH 5.8.0, Flume is fully compatible with Kafka 2.x, including support for security features.

Apache HBase

Premature EOF detected in a WAL During Replication

Bug: NoneDuring the parsing of a write-ahead log (WAL) during replication, an `InvalidProtobufException` can occur while reading the source RegionServer WAL, if `EOF` (end-of-file) is incorrectly detected before the actual end of the file. HBase stops reading the WAL after the EOF, and does not parse any bytes which occur after the EOF, causing data loss.

To work around this problem, Cloudera has patched HBase. HBase in CDH 5.8.0 and higher detect whether unparsed bytes exist after the EOF, and if so, the WAL is reset and re-read from the beginning, to attempt a clean read-through.

In testing, a single reset has been sufficient to work around observed data loss. However, the above change will retry a given WAL file indefinitely. On each attempt, a log message such as this will be emitted at the `WARN` level:

```
Processing end of WAL file '{}'. At position {}, which is too far away from
reported file length {}. Restarting WAL reading
```

Additional log detail are emitted at the `TRACE` level about file offsets seen while handling recoverable errors.

Batch Get after Batch Put Does Not Fetch All Cells

Bug: [HBASE-15811](#)

A batch Get after a batch Put could fail to fetch cells that were written by the Get, resulting in a "read-your-writes" failure. This bug was exacerbated by high load on the client.

Read Replica Failure For PUT Operation During Region Transition

Bug: None

When the patch for HBASE-10794 was applied in CDH 5.4.4, a new bug was introduced, where, if the primary RegionServer becomes unavailable (for any reason, even a graceful shutdown), while a client is performing PUTs on that region, subsequent PUTs will fail.

Latency Metrics Inaccurate for MultiGet Operations

Bug: [HBASE-15673](#)

Latency values are written after each row is processed. However, if `MultiGet` is enabled, some rows are not counted in the metrics. This causes the metrics for the 50th, 75th, and 90th percentiles to be reported as 0.

Inconsistent Behavior Among DeleteColumnFamilyProcedure, CreateTableProcedure, and ModifyTableProcedure

Bug: [HBASE-15456](#)

If there is only one family in the table, `DeleteColumnFamilyProcedure` will fail. When `hbase.table.sanity.checks` is set to `false`, the HMaster logs a warning, but `CreateTableProcedure` and `ModifyTableProcedure` will now fail, where before they logged a warning, but succeeded. This makes the behavior of all three methods consistent.

Failed hbase-spark Bulk Loads Leave Files Behind

Bug: [HBASE-15271](#)

When using the bulk load helper provided by the `hbase-spark` module, output files are now written into temporary files and only made available when the executor has successfully completed. Previously, failed executors would leave files behind, and these files would be picked up by subsequent bulk load commands, and spurious copies of some cells were written.

Apache Hive

[HIVE-13217](#): Replication for HoS MapJoin small file needs to respect `dfs.replication.max`

[HIVE-13039](#): BETWEEN predicate is not functioning correctly with predicate PUSHDOWN on Parquet table

[HIVE-13065](#): Hive throws `NullPointerException` (NPE) when writing map type data to an HBase-backed table

[HIVE-13160](#): HS2 unable to load UDFs on startup when HMS is not ready

[HIVE-13243](#): Hive DROP TABLE on encryption zone fails for external tables

[HIVE-13302](#): Direct SQL: CAST to DATE doesn't work on Oracle

[HIVE-13115](#): MetaStore Direct SQL `getPartitions()` call fails when the columns schemas for a partition are NULL

[HIVE-10303](#): HIVE-9471 broke forward compatibility of ORC files

[HIVE-12706](#): Incorrect output from `from_utc_timestamp() / to_utc_timestamp` when local timezone has DST

[HIVE-10685](#): ALTER TABLE concatenate operator will cause duplicate data

CDH 5 Release Notes

- [HIVE-13500](#): Launching big queries fails with OutOfMemoryException
- [HIVE-13527](#): Using deprecated APIs in HBase client causes ZooKeeper connection leaks.
- [HIVE-12517](#): Beeline's use of failed connection(s) causes failures and leaks.
- [HIVE-13632](#): Hive failing on INSERT empty array into parquet table
- [HIVE-13285](#): Orc concatenation may drop old files from moving to final path
- [HIVE-13836](#): DbNotifications giving an error = Invalid state. Transaction has already started
- [HIVE-9499](#): `hive.limit.query.max.table.partition` makes queries fail on non-partitioned tables
- [HIVE-13462](#): HiveResultSetMetaData.getPrecision() fails for NULL columns
- [HIVE-11408](#): HiveServer2 is leaking ClassLoaders when add jar / temporary functions are used due to constructor caching in Hadoop ReflectionUtils
- [HIVE-12481](#): Occasionally "Request is a replay" will be thrown from HS2
- [HIVE-10698](#): Query on view results fails with "table not found error" if view is created with subquery alias (CTE)
- [HIVE-12941](#): Unexpected result when using MIN() on struct with NULL in first field
- [HIVE-13200](#): Aggregation functions returning empty rows on partitioned columns
- [HIVE-11054](#): Read error : Partition Varchar column cannot be cast to string
- [HIVE-13401](#): Kerberized HS2 with LDAP auth enabled fails kerberos/delegation token authentication
- [HIVE-13217](#): Some queries with UNION all fail when CBO is off
- [HIVE-11369](#): MapJoins in HiveServer2 fail when `jmxremote` is used
- [HIVE-13261](#): Can not compute column stats for partition when schema evolves

With Sentry enabled, only Hive admin users have access to YARN job logs

As a prerequisite of enabling Sentry, Hive impersonation is turned off, which means all YARN jobs are submitted to the Hive job queue, and are run as the `hive` user. This is an issue because the YARN History Server now has to block users from accessing logs for their own jobs, since their own usernames are not associated with the jobs. As a result, end users cannot access any job logs unless they can get `sudo` access to the cluster as the `hdfs`, `hive` or other admin users.

In CDH 5.8 (and higher), Hive overrides the default configuration, `mapred.job.queuename`, and places incoming jobs into the connected user's job queue, even though the submitting user remains `hive`. Hive obtains the relevant queue/username information for each job by using YARN's `fair-scheduler.xml` file.

Hue

Cannot query the customers table in Hue

Bug: [HUE-3040](#)

To query the `customers` table, users must re-create the parquet data for compatibility.

Cloudera Distribution of Apache Kafka

CDH 5.7 is not compatible with Cloudera Distribution of Apache Kafka 1.x

Cloudera Distribution of Apache Kafka 1.x is compatible with CDH 5.4+.

Apache Oozie

PySpark does not work from the Oozie Spark Action

Bug: [OOZIE-2482](#)

The Spark Action would typically fail with a message like, "key not found: SPARK_HOME," but other error messages were possible. After the fix, the Spark Action has the necessary changes to successfully run PySpark jobs. See [Oozie Spark Action Extension](#) for more details and an example. Cloudera makes the PySpark dependencies available.

Apache Sentry *Security*

Sentry does not check privileges on the URI used for the `CREATE INDEX LOCATION '/path'` command

Bug: [SENTRY-1231](#)

The `CREATE INDEX LOCATION '/path'` command would succeed even if a user did not have the required URI privileges for the `/path`.

Upgraded libthrift to version 0.9.3 due to a security vulnerability

For details on the security vulnerability in the Apache Thrift client libraries, see [THRIFT-3231](#).

Hive Binding

The TRUNCATE TABLE query fails

Bug: [SENTRY-1320](#)

Precondition checks expect only one child node in the ASTNode for a `TRUNCATE TABLE` query. However, in queries of the form, `TRUNCATE TABLE db_name.table_name;`, it is possible to have two child nodes. Previously, such queries would fail with an `IllegalArgumentException` error because they violated precondition checks.

INSERT OVERWRITE DIRECTORY command does not work correctly

Bug: [SENTRY-922](#)

The `INSERT OVERWRITE DIRECTORY` command would write table data into an HDFS directory (`hdfs://path/`), even if privileges are granted only for the local directory (`file://path/`).

INSERT INTO no longer requires URI privilege on partition locations

Bug: [SENTRY-1095](#)

The `INSERT INTO` Hive command adds location information to the partition description. Usually if location information is included, you must ensure that the user has privileges on the corresponding URI. However, in this case, since the partition locations are under the table directory and can be easily generated, these requirements have been relaxed.

Change default value of `sentry.hive.server`

Bug: [SENTRY-1112](#)

The default value for `sentry.hive.server` was changed from `server1` to an empty string.

Sentry Service

Sentry's Oracle upgrade scripts fails with ORA-00955

Bug: [SENTRY-1066](#)

Sentry upgrade scripts for Oracle would fail with error, ORA-00955, because during the upgrade, the script inadvertently creates an index with the same name as the constraint being dropped. The script will now run `DROP INDEX` before it adds the constraint again and completes the schema upgrade successfully.

grantServerPrivilege() and revokeServerPrivilege() should treat '*' and 'ALL' as synonyms

Bug: [SENTRY-1252](#)

The `grantServerPrivilege()` and `revokeServerPrivilege()` methods should treat `*` and `ALL` as synonyms when an action is not explicitly specified. Previously, if `grantServerPrivilege()` was called without an action, and followed up with a `revokeServerPrivilege()` invocation with an action such as `ALL`, the server-level privilege would not be revoked. This fix only applies to privileges that are granted after upgrading to CDH 5.8.

Sentry Debugging

Error in Hive Metastore Plugin (`renameAuthzObject`) log messages

Bug: [SENTRY-1169](#)

The `renameAuthzObject` plugin prints log messages with old path names in place of new path names.

Apache ZooKeeper

Upgrade Netty Due to Security Vulnerabilities

Bug: [ZOOKEEPER-2450](#)

Netty was upgraded from version 3.2.2 to 3.10.5 to resolve security vulnerabilities.

Fix Privacy Violation in Login.java

Bug: [ZOOKEEPER-2405](#)

In `Login.java`, `getTGT()` was logging confidential information in DEBUG mode. After the fix, only principals are logged.

Issues Fixed in CDH 5.7.x

The following topics describe issues fixed in CDH 5.7.x, from newest to oldest release. You can also review [What's New In CDH 5.7.x](#) on page 18 or [Known Issues in CDH 5](#) on page 111.

Issues Fixed in CDH 5.7.5

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.7.5:

- [HADOOP-10300](#) - Allowed deferred sending of call responses
- [HADOOP-12483](#) - Maintain wrapped SASL ordering for postponed IPC responses
- [HADOOP-13317](#) - Add logs to KMS server-side to improve supportability
- [HADOOP-13558](#) - UserGroupInformation created from a Subject incorrectly tries to renew the Kerberos ticket
- [HADOOP-13638](#) - KMS should set UGI's Configuration object properly
- [HADOOP-13669](#) - KMS Server should log exceptions before throwing
- [HADOOP-13693](#) - Remove the message about HTTP OPTIONS in SPNEGO initialization message from kms audit log
- [HDFS-4176](#) - EditLogTailer should call rollEdits with a timeout
- [HDFS-6962](#) - ACLs inheritance conflict with umaskmode
- [HDFS-7413](#) - Some unit tests should use NameNodeProtocols instead of FSNameSystem
- [HDFS-7964](#) - Add support for async edit logging
- [HDFS-8224](#) - Schedule a block for scanning if its metadata file is corrupt
- [HDFS-8709](#) - Clarify automatic sync in FSEditLog#logEdit
- [HDFS-9038](#) - DFS reserved space is erroneously counted towards non-DFS used
- [HDFS-10178](#) - Permanent write failures can happen if pipeline recoveries occur for the first packet
- [HDFS-10609](#) - Uncaught InvalidEncryptionKeyException during pipeline recovery can abort downstream applications
- [HDFS-10641](#) - TestBlockManager#testBlockReportQueueing fails intermittently
- [HDFS-10722](#) - Fix race condition in TestEditLog#testBatchedSyncWithClosedLogs
- [HDFS-10760](#) - DataXceiver#run() should not log InvalidToken exception as an error
- [HDFS-10879](#) - TestEncryptionZonesWithKMS#testReadWrite fails intermittently
- [HDFS-10962](#) - TestRequestHedgingProxyProvider fails intermittently
- [HDFS-11012](#) - Unnecessary INFO logging on DFSClients for InvalidToken
- [MAPREDUCE-6633](#) - AM should retry map attempts if the reduce task encounters compression related errors
- [MAPREDUCE-6718](#) - Add progress log to JHS during startup
- [MAPREDUCE-6728](#) - Give fetchers hint when ShuffleHandler rejects a shuffling connection

- [MAPREDUCE-6771](#) - RMContainerAllocator sends container diagnostics event after corresponding completion event
- [YARN-4004](#) - container-executor should print output of Docker logs if the Docker container exits with non-0 exit status
- [YARN-4017](#) - container-executor overuses PATH_MAX
- [YARN-4245](#) - Generalize config file handling in container-executor
- [YARN-4255](#) - container-executor does not clean up Docker operation command files
- [YARN-4723](#) - NodesListManager\$UnknownNodeID ClassCastException
- [YARN-4940](#) - YARN node -list -all fails if RM starts with decommissioned node
- [YARN-5704](#) - Provide configuration knobs to control enabling/disabling new/work in progress features in container-executor
- [HBASE-16294](#) - hbck reporting "No HDFS region dir found" for replicas
- [HBASE-16699](#) - Overflows in AverageIntervalRateLimiter's refill() and getWaitInterval()
- [HBASE-16767](#) - Mob compaction needs to clean up files in /hbase/mobdir/.tmp and /hbase/mobdir/.tmp/.bulkload when running into IO exceptions
- [HIVE-10384](#) - BackportRetryingMetaStoreClient does not retry wrapped TTransportExceptions
- [HIVE-12077](#) - MSCK Repair table should fix partitions in batches
- [HIVE-12475](#) - Parquet schema evolution within array<struct>> does not work
- [HIVE-12785](#) - View with union type and UDF to the struct is broken
- [HIVE-13058](#) - Add session and operation_log directory deletion messages
- [HIVE-13198](#) - Authorization issues with cascading views
- [HIVE-13237](#) - Select parquet struct field with upper case throws NPE
- [HIVE-13429](#) - Tool to remove dangling scratch dir
- [HIVE-13997](#) - Insert overwrite directory does not overwrite existing files
- [HIVE-14313](#) - Test failure TestMetaStoreMetrics.testConnections
- [HIVE-14421](#) - FS.deleteOnExit holds references to _tmp_space.db files
- [HIVE-14762](#) - Add logging while removing scratch space
- [HIVE-14784](#) - Operation logs are disabled automatically if the parent directory does not exist
- [HIVE-14799](#) - Query operations are not thread safe during cancellation
- [HIVE-14805](#) - Subquery inside a view will have the object in the subquery as the direct input
- [HIVE-14810](#) - Fix failing test: TestMetaStoreMetrics.testMetaDataCounts
- [HIVE-14817](#) - Shutdown the SessionManager timeoutChecker thread properly upon shutdown
- [HIVE-14839](#) - Improve the stability of TestSessionManagerMetrics
- [HUE-3860](#) - Fix unittest beeswax.tests.test_hiveserver2_jdbc_url
- [HUE-3905](#) - Reset beeswax.conf params in beeswax.tests:test_hiveserver2_jdbc_url
- [HUE-4201](#) - Add warning about max limit of cells before truncation in the download query result
- [HUE-4662](#) - Fixed: Wildcard Certificates not supported
- [HUE-4739](#) - Fixed Jobbrowser tests which were failing after resource manager pool change
- [HUE-4916](#) - Truncate last name to 30 chars on ldap import
- [HUE-4968](#) - Remove access to /oozie/import_wokflow when v2 is enabled
- [HUE-5042](#) - Unable to kill jobs after Resource Manager failover
- [HUE-5050](#) - Logout fails for local login when multiple backends are used
- [HUE-5161](#) - Speed up roles rendering
- [HUE-5163](#) - Speed up initial page rendering
- [IMPALA-1619](#) - Support 64-bit allocations
- [IMPALA-1740](#) - Add support for skip.header.line.count
- [IMPALA-3458](#) - Fix table creation to test insert with header lines
- [IMPALA-3949](#) - Log the error message in FileSystemUtil.copyToLocal()
- [IMPALA-4037](#) - Fx locking during query cancellation
- [IMPALA-4076](#) - Fix runtime filter sort compare method

CDH 5 Release Notes

- [IMPALA-4099](#) - Fix the error message while loading UDFs with no JARs
- [IMPALA-4120](#) - Incorrect results with LEAD() analytic function
- [IMPALA-4135](#) - Thrift threaded server times-out connections during high load
- [IMPALA-4170](#) - Fix identifier quoting in COMPUTE INCREMENTAL STATS
- [IMPALA-4196](#) - Cross compile bit-byte functions
- [IMPALA-4237](#) - Fix materialization of 4 byte decimals in data source scan node
- [IMPALA-4246](#) - SleepForMs() utility function has undefined behavior for > 1s
- [OOZIE-1814](#) - Oozie should mask any passwords in logs and REST interfaces
- [SOLR-9310](#) - PeerSync fails on a node restart due to IndexFingerPrint mismatch
- [SPARK-12009](#) - Avoid re-allocating YARN container when driver wants to stop all Executors
- [SPARK-12392](#) - Optimize a location order of broadcast blocks by considering preferred local hosts
- [SPARK-12941](#) - Spark-SQL JDBC Oracle dialect fails to map string datatypes to Oracle VARCHAR datatype mapping
- [SPARK-13328](#) - Poor read performance for broadcast variables with dynamic resource allocation
- [SPARK-16625](#) - General data types to be mapped to Oracle
- [SPARK-16711](#) - YarnShuffleService does not re-init properly on YARN rolling upgrade
- [SPARK-17171](#) - DAG lists all partitions in the graph
- [SPARK-17433](#) - YarnShuffleService does not handle moving credentials levelDb
- [SPARK-17611](#) - Make shuffle service test really test authentication
- [SPARK-17644](#) - Do not add failedStages when abortStage for fetch failure
- [SPARK-17696](#) - Partial backport of to branch-1.6.
- [SQOOP-2952](#) - Row key not added into column family using --hbase-bulkload
- [SQOOP-2986](#) - Add validation check for --hive-import and --incremental lastmodified
- [SQOOP-3021](#) - ClassWriter fails if a column name contains a backslash character

Issues Fixed in CDH 5.7.4

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.7.4:

- [FLUME-2797](#) - Use SourceCounter for SyslogTcpSource
- [FLUME-2844](#) - SpillableMemoryChannel must start ChannelCounter
- [HADOOP-8436](#) - NPE In getLocalPathForWrite (path, conf) when the required context item is not configured
- [HADOOP-8437](#) - getLocalPathForWrite should throw IOException for invalid paths
- [HADOOP-10048](#) - LocalDirAllocator should avoid holding locks while accessing the filesystem
- [HADOOP-11469](#) - KMS should skip default.key.acl and whitelist.key.acl when loading key acl.
- [HADOOP-12252](#) - LocalDirAllocator should not throw NPE with empty string configuration
- [HADOOP-12548](#) - Read s3a credentials from a Credential Provider
- [HADOOP-12609](#) - Fix intermittent failure of TestDecayRpcScheduler.
- [HADOOP-13270](#) - BZip2CompressionInputStream finds the same compression marker twice in corner case, causing duplicate data blocks
- [HADOOP-13353](#) - LdapGroupsMapping getPassword should not return null when IOException is thrown
- [HADOOP-13437](#) - KMS should reload whitelist and default key ACLs when hot-reloading
- [HADOOP-13487](#) - Hadoop KMS should load old delegation tokens from Zookeeper on startup
- [HADOOP-13526](#) - Add detailed logging in KMS for the authentication failure of proxy user
- [HADOOP-13579](#) - Fix source-level compatibility after HADOOP-11252
- [HDFS-4210](#) - Throw helpful exception when DNS entry for JournalNode cannot be resolved
- [HDFS-7415](#) - Move FSNameSystem.resolvePath() to FSDirectory
- [HDFS-7420](#) - Delegate permission checks to FSDirectory
- [HDFS-7463](#) - Simplify FSNamesystem#getBlockLocationsUpdateTimes
- [HDFS-7478](#) - Move org.apache.hadoop.hdfs.server.namenode.NNConf to FSNamesystem
- [HDFS-7517](#) - Remove redundant non-null checks in FSNamesystem#getBlockLocations

- [HDFS-8269](#) - getBlockLocations() does not resolve the .reserved path and generates incorrect edit logs when updating the atime
- [HDFS-8897](#) - Balancer should handle fs.defaultFS trailing slash in HA
- [HDFS-9198](#) - Coalesce IBR processing in the NameNode.
- [HDFS-9781](#) - FsDatasetImpl#getBlockReports can occasionally throw NullPointerException
- [HDFS-9906](#) - Remove unhelpful log entries when restarting a datanode
- [HDFS-9958](#) - BlockManager#createLocatedBlocks can throw NPE for corruptBlocks on failed storages
- [HDFS-10270](#) - TestJMXGet: testNameNode() fails
- [HDFS-10457](#) - DataNode should not auto-format block pool directory if VERSION is missing
- [HDFS-10544](#) - Balancer does not work with IPFailoverProxyProvider
- [HDFS-10643](#) - Namenode should use loginUser(hdfs) to generateEncryptedKey
- [HDFS-10822](#) - Log DataNodes in the write pipeline
- [MAPREDUCE-4784](#) - TestRecovery occasionally fails
- [MAPREDUCE-6359](#) - In RM HA setup, Cluster tab links populated with AM hostname instead of RM
- [MAPREDUCE-6514](#) - Fixed MapReduce ApplicationMaster to properly updated resources ask after ramping down of all reducers avoiding job hangs
- [MAPREDUCE-6628](#) - Potential memory leak in CryptoOutputStream
- [MAPREDUCE-6670](#) - TestJobListCache#testEviction sometimes fails on Windows with timeout
- [MAPREDUCE-6680](#) - JHS UserLogDir scan algorithm sometimes could skip directory with update in CloudFS (Azure FileSystem, S3, and so on)
- [MAPREDUCE-6684](#) - High contention on scanning of user directory under immediate_done in Job History Server
- [MAPREDUCE-6738](#) - TestJobListCache.testAddExisting failed intermittently in slow VM testbed
- [MAPREDUCE-6761](#) - Regression when handling providers - invalid configuration ServiceConfiguration causes Cluster initialization failure
- [YARN-2977](#) - Fixed intermittent TestNMClient failure
- [YARN-4989](#) - TestWorkPreservingRMRestart#testCapacitySchedulerRecovery fails intermittently
- [YARN-5608](#) - TestAMRMClient.setup() fails with ArrayOutOfBoundsException
- [HBASE-15856](#) - Fix UnknownHostException import in MetaTableLocator
- [HBASE-15856](#) - Do not cache unresolved addresses for connections
- [HBASE-16194](#) - Should count in MSLAB chunk allocation into heap size change when adding duplicate cells
- [HBASE-16195](#) - Should not add chunk into chunkQueue if not using chunk pool in HeapMemStoreLAB
- [HBASE-16284](#) - Unauthorized client can shut down the cluster
- [HBASE-16317](#) - Revert all ESAPI changes
- [HBASE-16318](#) - Fail build while rendering velocity template if dependency license is not in whitelist
- [HBASE-16318](#) - Consistently use the correct name for "Apache License, Version 2.0"
- [HBASE-16321](#) - Ensure no findbugs-jsr305
- [HBASE-16340](#) - Exclude Xerces implementation jars from coming in transitively
- [HBASE-16360](#) - TableMapReduceUtil addHBaseDependencyJars has the wrong class name for PrefixTreeCodec
- [HIVE-9570](#) - Investigate test failure on union_view.q
- [HIVE-10007](#) - Support qualified table name in analyze table compute statistics for columns
- [HIVE-10728](#) - Deprecate unix_timestamp(void) and make it deterministic
- [HIVE-11901](#) - StorageBasedAuthorizationProvider requires write permission on table for SELECT statements
- [HIVE-12556](#) - Ctrl-C in Beeline does not kill Tez query on HS2
- [HIVE-13160](#) - HS2 unable to load UDFs on startup when HMS is not ready
- [HIVE-13620](#) - Merge llap branch work to master
- [HIVE-13645](#) - Beeline needs null-guard around hiveVars and hiveConfVars read
- [HIVE-14296](#) - Session count is not decremented when HS2 clients do not shutdown cleanly
- [HIVE-14436](#) - Hive 1.2.1/Hitting "ql.Driver: FAILED: IllegalArgumentException Error"
- [HIVE-14519](#) - Multi insert query bug
- [HIVE-14538](#) - Beeline throws exceptions with parsing Hive configuration when using !sh statement

CDH 5 Release Notes

- [HIVE-14715](#) - Hive throws NumberFormatException with query with Null value
- [HIVE-14743](#) - ArrayIndexOutOfBoundsException - HBASE-backed views query with JOINS
- [HUE-2689](#) - Sub-workflow submitted from coordinator gets parent workflow graph
- [HUE-4541](#) - Fixing Hue job browser - Kerberos mutual authentication error in Hue
- [HUE-4635](#) - Fix duration on jobs page for running jobs
- [HUE-4804](#) - Download function of HTML widget breaks the display
- [HUE-4808](#) - Do not show the edit link for sub-workflows when submitted outside Hue
- [HUE-4809](#) - Add truststore parameters only if SSL is turned on
- [HUE-4809](#) - Only add truststore paths when they actually exist
- [IMPALA-3081](#) - Increase memory limit for TestWideRow
- [IMPALA-3311](#) - Fix string data coming out of aggs in subplans
- [IMPALA-3575](#) - Add retry to back end connection request and rpc timeout
- [IMPALA-3678](#) - Fix migration of predicates into union operands with an order by + limit.
- [IMPALA-3682](#) - Do not retry unrecoverable socket creation errors
- [IMPALA-3687](#) - Fix test failure introduced by backporting
- [IMPALA-3687](#) - Prefer Avro field name during schema reconciliation
- [IMPALA-3820](#) - Handle linkage errors while loading Java UDFs in Catalog
- [IMPALA-3930](#) - Fix shuffle insert hint with constant partition exprs
- [IMPALA-3940](#) - Fix getting column stats through views
- [IMPALA-4020](#) - Handle external conflicting changes to HMS gracefully
- [IMPALA-4049](#) - Fix empty batch handling NLJ build side
- [OOZIE-2068](#) - Configuration as part of sharelib
- [OOZIE-2347](#) - Remove unnecessary new Configuration()/new jobConf() calls from Oozie
- [OOZIE-2555](#) - Oozie SSL enable setup does not return port for admin -servers
- [OOZIE-2567](#) - HCat connection is not closed while getting hcat credentials
- [OOZIE-2589](#) - CompletedActionXCommand is hardcoded to wrong priority
- [OOZIE-2649](#) - Cannot override sub-workflow configuration property if defined in parent workflow XML
- [PIG-3807](#) - Pig creates wrong schema after dereferencing nested tuple fields with sorts
- [SPARK-8428](#) -Fix integer overflows in TimSort
- [SPARK-12339](#) - Added a null check that was removed in
- [SPARK-13242](#) - codegen fallback when there many branches

Issues Fixed in CDH 5.7.3

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.7.3:

- [FLUME-2821](#) - KafkaSourceUtil Can Log Passwords at Info remove logging of security related data in older releases.
- [FLUME-2913](#) - Don't strip SLF4J from imported classpaths.
- [FLUME-2922](#) - Sync SequenceFile.Writer before calling hflush
- [HADOOP-8751](#) - NPE in Token.toString() when Token is constructed using null identifier.
- [HADOOP-11361](#) - Fix a race condition in MetricsSourceAdapter.updateJmxCache.
- [HADOOP-11901](#) - BytesWritable fails to support 2G chunks due to integer overflow.
- [HADOOP-12659](#) - Incorrect usage of configuration parameters in token manager of KMS.
- [HADOOP-13263](#) - Reload cached groups in background after expiry.
- [HADOOP-13381](#) - KMS clients should use KMS Delegation Tokens from current UGI.
- [HADOOP-13457](#) - Remove hardcoded absolute path for shell executable.
- [HDFS-6434](#) - Default permission for creating file should be 644 for WebHdfs/HttpFS.
- [HDFS-7597](#) - DelegationTokenIdentifier should cache the TokenIdentifier to UGI mapping.
- [HDFS-8008](#) - Support client-side back off when the datanodes are congested.
- [HDFS-9276](#) - Failed to Update HDFS Delegation Token for long running application in HA mode.
- [HDFS-9466](#) - TestShortCircuitCache#testDataXceiverCleansUpSlotsOnFailure is flaky.

- [HDFS-9939](#) - Increase DecompressorStream skip buffer size.
- [HDFS-10512](#) - VolumeScanner may terminate due to NPE in DataNode.reportBadBlocks.
- [MAPREDUCE-6442](#) - Stack trace is missing when error occurs in client protocol provider's constructor.
- [MAPREDUCE-6473](#) - Job submission can take a long time during Cluster initialization.
- [MAPREDUCE-6675](#) - TestJobImpl.testUnusableNode failed
- [YARN-4459](#) - container-executor should only kill process groups.
- [YARN-4784](#) - Fairscheduler: defaultQueueSchedulingPolicy should not accept FIFO.
- [YARN-4866](#) - FairScheduler: AMs can consume all vcores leading to a livelock when using FAIR policy.
- [YARN-4878](#) - Expose scheduling policy and max running apps over JMX for Yarn queues.
- [YARN-5077](#) - Fix FSLeafQueue#getFairShare() for queues with zero fairshare.
- [YARN-5272](#) - Handle queue names consistently in FairScheduler.
- [HBASE-14963](#) - Remove use of Guava Stopwatch from HBase client code.
- [HBASE-15621](#) - Suppress Hbase SnapshotHFile cleaner error messages when a snapshot is going on.
- [HBASE-15808](#) - Reduce potential bulk load intermediate space usage and waste.
- [HBASE-16135](#) - PeerClusterZnode under rs of removed peer may never be deleted
- [HBASE-16207](#) - can't restore snapshot without "Admin" permission
- [HBASE-16227](#) - [Shell] Column value formatter not working in scans. Tested : manually using shell.
- [HBASE-16288](#) - HFile intermediate block level indexes might recurse forever creating multi TB files.
- [HBASE-16319](#) - Fix TestCacheOnWrite after HBASE-16288.
- [HIVE-11432](#) - Hive macro gives same result for different arguments.
- [HIVE-11487](#) - Add getNumPartitionsByFilter api in metastore api.
- [HIVE-11980](#) - Follow up on HIVE-11696, exception is thrown from CTAS from the table with table-level serde is Parquet while partition-level serde is JSON.
- [HIVE-12277](#) - Hive macro results on macro_duplicate.q different after adding ORDER BY.
- [HIVE-12635](#) - Hive should return the latest HBase cell timestamp as the row timestamp value.
- [HIVE-13043](#) - Reload function has no impact to function registry.
- [HIVE-13090](#) - Hive metastore crashes on NPE with ZooKeeperTokenStore.
- [HIVE-13372](#) - Hive Macro overwritten when multiple macros are used in one column.
- [HIVE-13749](#) - Memory leak in Hive Metastore.
- [HIVE-13884](#) - Disallow queries in HMS fetching more than a configured number of partitions
- [HIVE-14055](#) - directSql - getting the number of partitions is broken.
- [HIVE-14187](#) - JDOPersistenceManager objects remain cached if MetaStoreClient#close is not called.
- [HIVE-14209](#) - Add some logging info for session and operation management.
- [HIVE-14298](#) - NPE could be thrown in HMS when an ExpressionTree could not be made from a filter.
- [HIVE-14359](#) - Hive on Spark might fail in HS2 with LDAP authentication in a kerberized cluster.
- [HIVE-14457](#) - Partitions in encryption zone are still trashed though an exception is returned.
- [HUE-3481](#) - [assist] Do not sort the columns by name, instead use the creation order.
- [HUE-3842](#) - [core] HTTP 500 while emptying Hue 3.9 trash directory.
- [HUE-3845](#) - [sentry] Sometimes see group as editable on role section.
- [HUE-3880](#) - [core] Add importlib directly for Python 2.6.
- [HUE-3988](#) - [search] Support schemaless collections.
- [HUE-3999](#) - [oozie] list_oozie_workflow page should not break in case of bad json from oozie.
- [HUE-4265](#) - [beeswax] Bring back show preview in the assist.
- [HUE-4300](#) - [fb] Avoid double file listing call on folder search.
- [HUE-4333](#) - [core] Properly reset API_CACHE on failover.
- [HUE-4477](#) - [security] Select All is not filtering out the non visible roles from the selection .
- [HUE-4493](#) - [oozie] Fix sync-workflow action when Workflow includes sub-workflow.
- [HUE-4515](#) - [oozie] Remove oozie.bundle.application.path from properties when rerunning workflow.
- [IMPALA-3711](#) - Remove unnecessary privilege checks in getDbsMetadata().
- [IMPALA-3915](#) - Register privilege and audit requests when analyzing resolved table refs.

CDH 5 Release Notes

- [OOZIE-2391](#) - spark-opts value in workflow.xml is not parsed properly.
- [OOZIE-2537](#) - SqoopMain does not set up log4j properly.
- [SOLR-7280](#) - BackportLoad cores in sorted order and tweak coreLoadThread counts to improve cluster stability on restarts.
- [SOLR-9236](#) - AutoAddReplicas will append an extra /tlog to the update log location on replica failover.
- [SPARK-14963](#) - [YARN] Using recoveryPath if NM recovery is enabled.
- [SPARK-16505](#) - [YARN] Optionally propagate error during shuffle service startup.
- [SQOOP-2561](#) - Special Character removal from Column name as avro data results in duplicate column and fails the import.
- [SQOOP-2906](#) - Optimization of AvroUtil.toAvroIdentifier.
- [SQOOP-2971](#) - OraOop does not close connections properly.
- [SQOOP-2995](#) - Backward incompatibility introduced by Custom Tool options.

Issues Fixed in CDH 5.7.2

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.7.2:

- [FLUME-1899](#) - Make SpoolDir work with subdirectories
- [FLUME-2910](#) - AsyncHBaseSink: Failure callbacks should log the exception that caused them
- [FLUME-2918](#) - Speed up TaildirSource on directories with many files
- [HADOOP-8934](#) - Shell command ls should include sort options
- [HADOOP-10971](#) - Add -C flag to make `hadoop fs -ls` print filenames only
- [HADOOP-11409](#) - FileContext.getFileContext can stack overflow if default fs misconfigured
- [HADOOP-11432](#) - Fix SymlinkBaseTest#testCreateLinkUsingPartQualPath2.
- [HADOOP-12787](#) - KMS SPNEGO sequence does not work with WebHDFS
- [HADOOP-12841](#) - Update s3-related properties in core-default.xml.
- [HADOOP-12901](#) - Add warning log when KMSClientProvider cannot create a connection to the KMS server.
- [HADOOP-12963](#) - Allow using path style addressing for accessing the S3 endpoint.
- [HADOOP-13079](#) - Add -q option to ls to print ? instead of non-printable characters
- [HADOOP-13132](#) - Handle ClassCastException on AuthenticationException in LoadBalancingKMSClientProvider
- [HADOOP-13155](#) - Implement TokenRenewer to renew and cancel delegation tokens in KMS
- [HADOOP-13251](#) - Authenticate with Kerberos credentials when renewing KMS delegation token
- [HADOOP-13255](#) - KMSClientProvider should check and renew TGT when doing delegation token operations
- [HDFS-8581](#) - ContentSummary on / skips further counts on yielding lock
- [HDFS-8829](#) - Make SO_RCVBUF and SO_SNDBUF size configurable for DataTransferProtocol sockets and allow configuring auto-tuning
- [HDFS-9085](#) - Show renewer information in DelegationTokenIdentifier#toString
- [HDFS-9259](#) - Make SO_SNDBUF size configurable at DFSClient side for hdfs write scenario.
- [HDFS-9365](#) - Balancer does not work with the HDFS-6376 HA setup.
- [HDFS-9405](#) - Warmup NameNode EDEK caches in background thread
- [HDFS-9700](#) - DFSClient and DFSOutputStream should set TCP_NODELAY on sockets for DataTransferProtocol
- [HDFS-9732](#) - Improve DelegationTokenIdentifier.toString() for better logging
- [HDFS-9805](#) - Add server-side configuration for enabling TCP_NODELAY for DataTransferProtocol and default it to true
- [HDFS-10360](#) - DataNode may format directory and lose blocks if current/VERSION is missing
- [HDFS-10381](#) - DataStreamer DataNode exclusion log message should be warning
- [HDFS-10396](#) - Using -diff option with DistCp may get "Comparison method violates its general contract" exception
- [HDFS-10481](#) - HTTPFS server should correctly impersonate as end user to open file
- [HDFS-10516](#) - Fix bug when warming up EDEK cache of more than one encryption zone
- [HDFS-10525](#) - Fix NPE in CacheReplicationMonitor#rescanCachedBlockMap
- [MAPREDUCE-6558](#) - multibyte delimiters with compressed input files generate duplicate records

- [MAPREDUCE-6577](#) - MR AM unable to load native library without MR_AM_ADMIN_USER_ENV set
- [MAPREDUCE-6635](#) - Unsafe long to int conversion in UncompressedSplitLineReader and IndexOutOfBoundsException
- [MAPREDUCE-6701](#) - Application Master log unavailable when clicking JobHistory's AM logs link
- [YARN-2605](#) - [RM HA] Rest API endpoints doing redirect incorrectly.
- [YARN-4812](#) - TestFairScheduler#testContinuousScheduling fails intermittently.
- [YARN-4916](#) - Revert "TestNMProxy.tesNMProxyRPCRetry fails
- [YARN-5048](#) - DelegationTokenRenewer#skipTokenRenewal may throw NPE
- [HBASE-11625](#) - Reading datablock throws "Invalid HFile block magic" and can not switch to hdfs checksum
- [HBASE-13532](#) - Make UnknownScannerException less scary by giving more information in the exception string.
- [HBASE-14644](#) - Region in transition metric is broken
- [HBASE-14818](#) - user_permission does not list namespace permissions
- [HBASE-15236](#) - Inconsistent cell reads over multiple bulk-loaded HFiles
- [HBASE-15439](#) - getMaximumAllowedTimeBetweenRuns in ScheduledChore ignores the TimeUnit
- [HBASE-15465](#) - userPermission returned by getUserPermission() for the selected namespace does not have namespace set
- [HBASE-15496](#) - Throw RowTooBigException only for user scan/get
- [HBASE-15698](#) - Increment TimeRange not serialized to server
- [HBASE-15746](#) - Remove extra RegionCoprocessor preClose() in RSRpcServices#closeRegion
- [HBASE-15791](#) - Improve javadoc around ScheduledChore
- [HBASE-15811](#) - Batch Get after batch Put does not fetch all cells
- [HBASE-15872](#) - Split TestWALProcedureStore
- [HBASE-15873](#) - ACL for snapshot restore / clone is not enforced
- [HBASE-15925](#) - provide default values for hadoop compat module related properties that match default hadoop profile.
- [HBASE-16034](#) - Fix ProcedureTestingUtility#LoadCounter.setMaxProcl()
- [HBASE-16056](#) - Procedure v2 - fix master crash for FileNotFoundException
- [HBASE-16093](#) - Fix splits failed before creating daughter regions leave meta inconsistent
- [HIVE-7443](#) - Fix HiveConnection to communicate with Kerberized Hive JDBC server and alternative JDKs
- [HIVE-9486](#) - Use session classloader instead of application loader
- [HIVE-9499](#) - hive.limit.query.max.table.partition makes queries fail on non-partitioned tables
- [HIVE-10685](#) - Alter table concatenate oparetor will cause duplicate data
- [HIVE-10925](#) - Non-static threadlocals in metastore code can potentially cause memory leak
- [HIVE-11031](#) - ORC concatenation of old files can fail while merging column statistics
- [HIVE-11243](#) - Changing log level in Utilities.getBaseWork
- [HIVE-11747](#) - Unnecessary error log is shown when executing a "INSERT OVERWRITE LOCAL DIRECTORY" cmd in the embedded mode
- [HIVE-11827](#) - STORED AS AVRO fails SELECT COUNT(*) when empty
- [HIVE-12742](#) - NULL table comparison within CASE does not work as previous hive versions
- [HIVE-12958](#) - Make embedded Jetty server more configurable
- [HIVE-13285](#) - ORC concatenation may drop old files from moving to final path
- [HIVE-13462](#) - HiveResultSetMetaData.getPrecision() fails for NULL columns
- [HIVE-13590](#) - Kerberized HS2 with LDAP auth enabled fails in multi-domain LDAP case
- [HIVE-13736](#) - View's input/output formats are TEXT by default.
- [HIVE-13932](#) - Hive SMB Map Join with small set of LIMIT failed with NPE
- [HIVE-13953](#) - Issues in HiveLockObject equals method
- [HIVE-13991](#) - Union All on view fail with no valid permission on underneath table
- [HIVE-14006](#) - Hive query with UNION ALL fails with ArrayIndexOutOfBoundsException.
- [HIVE-14015](#) - SMB MapJoin failed for Hive on Spark when kerberized
- [HIVE-14098](#) - Logging task properties and environment variables might contain passwords
- [HIVE-14118](#) - Make the alter partition exception more meaningful

CDH 5 Release Notes

- [HUE-2678](#) - [jobbrowser] Read Spark job data from Spark History Server API
- [HUE-3197](#) - [oozie] Decision node support in external Workflow graph
- [HUE-3520](#) - [jb] Use impersonation to access JHS if security is enabled
- [HUE-3521](#) - [core] Provide a force_username_uppercase option
- [HUE-3526](#) - [useradmin] Fix LDAP tests for force_username_uppercase
- [HUE-3688](#) - [oozie] Fix TestEditor.test_workflow_dependencies unit test
- [HUE-3700](#) - [core] Support force_username_lowercase and ignore_username_case for all Auth backends
- [HUE-3802](#) - [oozie] Fix HS2 action on SSL enabled cluster
- [HUE-3805](#) - [oozie] Add support for oozie schema 0.4 in dashboard graph for external workflows
- [HUE-3808](#) - [core] Offer to live turn on/off debug level
- [HUE-3821](#) - [pig] Logs are never returned on running script
- [HUE-3822](#) - [pig] Display logs when found
- [HUE-3861](#) - [core] Upgrade Django Axes to 1.5
- [HUE-3866](#) - [core] Hue CPU reaches ~100% usage while uploading files with SSL to HTTPFS/WebHDFS
- [HUE-3908](#) - [useradmin] Ignore (objectclass=*) filter when searching for LDAP users
- [HUE-3923](#) - [core] Simplify force debug logic option
- [HUE-4005](#) - [oozie] Remove oozie.coord.application.path from properties when rerunning workflow
- [HUE-4006](#) - [oozie] Create new deployment directory when coordinator or bundle is copied
- [HUE-4007](#) - [oozie] Fix deployment_dir for the bundle in oozie example fixtures
- [HUE-4021](#) - [libsolr] Allow customization of the Solr path in ZooKeeper
- [HUE-4023](#) - [useradmin] update AuthenticationForm to allow activated users to login
- [HUE-4061](#) - [jb] Job attempt logs not appearing for running jobs
- [HUE-4087](#) - [jobbrowser] Unable to kill jobs with Resource Manager HA enabled
- [HUE-4092](#) - [security] Can't type any / in the HDFS ACLs path input
- [HUE-4113](#) - [Pig] Hue breaks when user has only access to pig app
- [HUE-4134](#) - [liboozie] Avoid logging truststore credentials
- [HUE-4202](#) - [jb] Enable offset param for fetching jobbrowser logs
- [HUE-4215](#) - [yarn] Reset API_CACHE on logout
- [HUE-4227](#) - [yarn] Fix unittest for MR API Cache
- [HUE-4238](#) - [doc2] Ignore history docs in find_jobs_with_no_doc during sync documents
- [HUE-4252](#) - [core] Handle 307 redirect from YARN upon standby failover
- [HUE-4258](#) - [jb] Close and pool Spark History Server connections
- [IMPALA-1928](#) - Fix Thrift client transport wrapping order
- [IMPALA-2660](#) - Respect auth_to_local configs from hdfs configs
- [IMPALA-3276](#) - Consistently handle pin failure in BTS::PrepareForRead()
- [IMPALA-3369](#) - Add ALTER TABLE SET COLUMN STATS statement.
- [IMPALA-3441](#) - Impala should not crash for invalid avro serialized data
- [IMPALA-3499](#) - Split catalog update
- [IMPALA-3502](#) - Fix race in the coordinator while updating filter routing table
- [IMPALA-3633](#) - Cancel fragment if coordinator is gone
- [IMPALA-3732](#) - Handle string length overflow in Avro files
- [IMPALA-3745](#) - Corrupt encoded values in parquet files can cause crashes
- [IMPALA-3751](#) - Fix clang build errors and warnings
- [IMPALA-3754](#) - Fix TestParquet.test_corrupt_rle_counts flakiness
- [OOZIE-2314](#) - Unable to kill old instance child job by workflow or coord rerun by Launcher
- [OOZIE-2329](#) - Make handling yarn restarts configurable
- [OOZIE-2330](#) - Spark action should take the global jobTracker and nameNode configs by default and allow file and archive elements
- [OOZIE-2345](#) - Parallel job submission for forked actions
- [OOZIE-2436](#) - Fork/join workflow fails with oozie.action.yarn.tag must not be null

- [OOZIE-2481](#) - Add YARN_CONF_DIR in the Shell action
- [OOZIE-2504](#) - Create a log4j.properties under HADOOP_CONF_DIR in Shell Action
- [OOZIE-2511](#) - SubWorkflow missing variable set from option if config-default is present in parent workflow
- [OOZIE-2533](#) - Oozie Web UI gives Error 500 with Java 8u91
- [SENTRY-1175](#) - Improve usability of URI privileges when granting URIs
- [SENTRY-1201](#) - Sentry ignores database prefix for MSCK statement
- [SENTRY-1252](#) - grantServerPrivilege and revokeServerPrivilege should treat "*" and "ALL" as synonyms when action is not explicitly specified
- [SENTRY-1265](#) - Sentry service should not require a TGT as it is not talking to other kerberos services as a client
- [SENTRY-1292](#) - Reorder DBModelAction EnumSet
- [SENTRY-1293](#) - Avoid converting string permission to Privilege object
- [SENTRY-1311](#) - Improve usability of URI privileges by supporting mixed use of URIs with and without scheme
- [SENTRY-1320](#) - truncate table db_name.table_name fails
- [SOLR-7178](#) - OverseerAutoReplicaFailoverThread compares Integer objects using ==
- [SOLR-8451](#) - We should not call method.abort in HttpSolrClient and HttpSolrCall#remoteQuery should not close streams
- [SOLR-8497](#) - Merge indexes should mark its directories as done rather than keep them around in the directory cache.
- [SOLR-8691](#) - Cache index fingerprints per searcher
- [SOLR-9053](#) - Upgrade commons-fileupload to 1.3.1, fixing a potential vulnerability
- [SPARK-13278](#) - [CORE] Launcher fails to start with JDK 9 EA
- [SPARK-14391](#) - [LAUNCHER] Fix launcher communication test
- [SPARK-15067](#) - [YARN] YARN executors are launched with fixed perm gen size
- [SPARK-15165](#) - [SPARK-15205] [SQL] Introduce place holder for comments in generated code
- [SQOOP-2846](#) - Sqoop Export with update-key failing for avro data file
- [SQOOP-2864](#) - ClassWriter chokes on column names containing double quotes
- [SQOOP-2920](#) - Sqoop performance deteriorates significantly on wide datasets; sqoop 100% on CPU

Issues Fixed in CDH 5.7.1

CDH 5.7.1 fixes the following issues.

Apache HBase

The ReplicationCleaner process can abort if its connection to ZooKeeper is inconsistent

Bug: [HBASE-15234](#)

If the connection with ZooKeeper is inconsistent, the ReplicationCleaner may abort, and the following event is logged by the HMaster:

```
WARN org.apache.hadoop.hbase.replication.master.ReplicationLogCleaner: Aborting
ReplicationLogCleaner
because Failed to get list of replicators
```

Unprocessed WALS accumulate.

The seekBefore() method calculates the size of the previous data block by assuming that data blocks are contiguous, and HFile v2 and higher store Bloom blocks and leaf-level INode blocks with the data. As a result, reverse scans do not work when Bloom blocks or leaf-level INode blocks are present when HFile v2 or higher is used.

Workaround: Restart the HMaster occasionally. The ReplicationCleaner restarts if necessary and process the unprocessed WALS.

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.7.1:

- [AVRO-1781](#) - Schema.parse is not thread safe

- [FLUME-2781](#) - Kafka Channel with parseAsFlumeEvent=true should write data as is, not as flume events
- [FLUME-2891](#) - Revert FLUME-2712 and FLUME-2886
- [FLUME-2897](#) - AsyncHBase sink NPE when Channel.getTransaction() fails
- [HADOOP-7139](#) - Allow appending to existing SequenceFiles
- [HADOOP-7817](#) - RawLocalFileSystem.append() should give FSDataOutputStream with accurate .getPos()
- [HADOOP-11321](#) - copyToLocal cannot save a file to an SMB share unless the user has Full Control permissions
- [HADOOP-11687](#) - Ignore x-* and response headers when copying an Amazon S3 object
- [HADOOP-12668](#) - Support excluding weak Ciphers in HttpServer2 through ssl-server.conf
- [HADOOP-12825](#) - Log slow name resolutions
- [HADOOP-12954](#) - Add a way to change hadoop.security.token.service.use_ip
- [HADOOP-12972](#) - Lz4Compressor#getLibraryName returns the wrong version number
- [HDFS-3519](#) - Checkpoint upload may interfere with a concurrent saveNamespace
- [HDFS-6520](#) - hdfs fsck passes invalid length value when creating BlockReader
- [HDFS-7600](#) - Refine hdfs admin classes to reuse common code
- [HDFS-8142](#) - DistributedFileSystem encryption zone commands should resolve relative paths
- [HDFS-8211](#) - DataNode UUID is always null in the JMX counter.
- [HDFS-8496](#) - Calling stopWriter() with FsDatasetImpl lock held may block other threads
- [HDFS-8855](#) - Webhdfs client leaks active NameNode connections
- [HDFS-9549](#) - TestCacheDirectives#testExceedsCapacity is flaky
- [HDFS-9589](#) - Block files which have been hardlinked should be duplicated before the DataNode appends to them
- [HDFS-9949](#) - Add a test case to ensure that the DataNode does not regenerate its UUID when a storage directory is cleared
- [HDFS-10223](#) - peerFromSocketAndKey performs SASL exchange before setting connection timeouts
- [HDFS-10267](#) - Extra "synchronized" on FsDatasetImpl#recoverAppend and FsDatasetImpl#recoverClose
- [HDFS-10324](#) - Trash directory in an encryption zone should be pre-created with correct permissions
- [HDFS-10344](#) - DistributedFileSystem#getTrashRoots should skip encryption zone that does not have .Trash
- [MAPREDUCE-4785](#) - TestMRApp occasionally fails
- [MAPREDUCE-6297](#) - Task Id of the failed task in diagnostics should link to the task page
- [MAPREDUCE-6333](#) - TestEvents,TestAMWebServicesTasks,TestAppController are broken due to MAPREDUCE-6297
- [MAPREDUCE-6384](#) - Add the last reporting reducer info for too many fetch failure diagnostics
- [MAPREDUCE-6388](#) - Remove deprecation warnings from JobHistoryServer classes
- [MAPREDUCE-6485](#) - Create a new task attempt with failed map task priority if in-progress attempts are unassigned
- [MAPREDUCE-6513](#) - MR job got hanged forever when one NM unstable for some time
- [MAPREDUCE-6535](#) - TaskID default constructor results in NPE on toString()
- [MAPREDUCE-6580](#) - Test failure: TestMRJobsWithProfiler
- [YARN-2871](#) - TestRMRestart#testRMRestartGetApplicationList sometime fails in trunk
- [YARN-3104](#) - Fixed RM to not generate new AMRM tokens on every heartbeat between rolling and activation
- [YARN-3493](#) - RM fails to come up with error "Failed to load/recover state" when mem settings are changed
- [YARN-3695](#) - ServerProxy (NMProxy, etc.) shouldn't retry forever for non network exception
- [YARN-4168](#) - Fixed a failing test TestLogAggregationService.testLocalFileDeletionOnDiskFull
- [YARN-4414](#) - Nodemanager connection errors are retried at multiple levels
- [YARN-4579](#) - Allow DefaultContainerExecutor container log directory permissions to be configurable
- [YARN-4629](#) - Distributed shell breaks under strong security
- [YARN-4639](#) - Remove dead code in TestDelegationTokenRenewer added in YARN-3055
- [YARN-4717](#) - TestResourceLocalizationService.testPublicResourceInitializesLocalDir fails Intermittently due to IllegalArgumentException from cleanup
- [YARN-4795](#) - ContainerMetrics drops records
- [YARN-4916](#) - TestNMProxy.testNMProxyRPCRetry fails
- [HBASE-15234](#) - Don't abort ReplicationLogCleaner on ZooKeeper errors

- [HBASE-15271](#) - Spark bulk load should write to temporary location and then rename on success
- [HBASE-15349](#) - Update surefire version to 2.19.1
- [HBASE-15405](#) - Fix PE logging and wrong defaults in help message
- [HBASE-15456](#) - CreateTableProcedure/ModifyTableProcedure needs to fail when there is no family in table descriptor
- [HBASE-15479](#) - No more garbage or beware of autoboxing
- [HBASE-15582](#) - SnapshotManifestV1 too verbose when there are no regions
- [HBASE-15591](#) - ServerCrashProcedure not yielding
- [HBASE-15592](#) - Print Procedure WAL content
- [HBASE-15622](#) - Superusers does not consider the keytab credentials
- [HBASE-15622](#) - Superusers does not consider the keytab credentials
- [HBASE-15673](#) - Fix latency metrics for multiGet. - Also fixes some stuff in help text
- [HBASE-15707](#) - ImportTSV bulk output does not support tags with hfile.format.version=3
- [HIVE-6099](#) - Multi insert does not work properly with distinct count
- [HIVE-10303](#) - HIVE-9471 broke forward compatibility of ORC files
- [HIVE-10313](#) - Literal Decimal ExprNodeConstantDesc should contain value of HiveDecimal instead of String
- [HIVE-10396](#) - moredecimal_precision2.q test is failing on trunk
- [HIVE-10636](#) - CASE comparison operator rotation optimization
- [HIVE-11054](#) - Handle varchar/char partition columns in vectorization
- [HIVE-11097](#) - HiveInputFormat uses String.startsWith to compare splitPath and PathToAliases
- [HIVE-11369](#) - Mapjoins in HiveServer2 fail when jmxremote is used
- [HIVE-11408](#) - HiveServer2 is leaking ClassLoaders when add jar / temporary functions are used due to constructor caching in Hadoop ReflectionUtils
- [HIVE-11427](#) - Location of temporary table for CREATE TABLE SELECT broken by HIVE-7079
- [HIVE-11590](#) - AvroDeserializer is very chatty
- [HIVE-11919](#) - Hive Union Type Mismatch
- [HIVE-12481](#) - Occasionally "Request is a replay" will be thrown from HS2
- [HIVE-12506](#) - SHOW CREATE TABLE command creates a table that does not work for RCFile format
- [HIVE-12568](#) - Provide an option to specify network interface used by Spark remote client [Spark Branch]
- [HIVE-12616](#) - NullPointerException when spark session is reused to run a mapjoin
- [HIVE-12706](#) - Incorrect output from from_utc_timestamp()/to_utc_timestamp when local timezone has DST
- [HIVE-12941](#) - Unexpected result when using MIN() on struct with NULL in first field
- [HIVE-13082](#) - Enable constant propagation optimization in query with left semi join
- [HIVE-13082](#) - Enable constant propagation optimization in query with left semi join
- [HIVE-13115](#) - MetaStore Direct SQL getPartitions call fail when the columns schemas for a partition are null
- [HIVE-13200](#) - Aggregation functions returning empty rows on partitioned columns
- [HIVE-13217](#) - Replication for HoS mapjoin small file needs to respect dfs.replication.max
- [HIVE-13243](#) - Hive drop table on encryption zone fails for external tables
- [HIVE-13251](#) - hive can't read the decimal in AVRO file generated from previous version
- [HIVE-13286](#) - Query ID is being reused across queries
- [HIVE-13295](#) - Improvement to LDAP search queries in HS2 LDAP Authenticator
- [HIVE-13300](#) - Hive on spark throws exception for multi-insert with join
- [HIVE-13302](#) - direct SQL: cast to date doesn't work on Oracle
- [HIVE-13376](#) - HoS emits too many logs with application state
- [HIVE-13401](#) - Kerberized HS2 with LDAP auth enabled fails kerberos/delegation token authentication
- [HIVE-13410](#) - PerfLog metrics scopes not closed if there are exceptions on HS2
- [HIVE-13500](#) - Fix OOM with explain output being logged
- [HIVE-13527](#) - Using deprecated APIs in HBase client causes zookeeper connection leaks
- [HIVE-13530](#) - Hive on Spark throws Kryo exception in some cases
- [HIVE-13570](#) - Some queries with Union all fail when CBO is off

CDH 5 Release Notes

- [HIVE-13585](#) - Add counter metric for direct sql failures
- [HIVE-13632](#) - CDH39911Hive failing on insert empty array into parquet table
- [HIVE-13657](#) - Spark driver stderr logs should appear in hive client logs
- [HUE-3171](#) - Fix vertical resize handle for queries with long descriptions
- [HUE-3171](#) - Long descriptions doesn't wrap and headers from table follows with horizontal scroll
- [HUE-3221](#) - Styling on column stats popup leaks on the tables page
- [HUE-3293](#) - Prevent document matching query error when going one home 1
- [HUE-3293](#) - Fix mis-switching to new home page when new editor is on
- [HUE-3293](#) - Move new editor flag to desktop
- [HUE-3303](#) - PostgreSQL requires data update and alter table operations in separate transactions
- [HUE-3310](#) - Prevent browsing job designs by API
- [HUE-3334](#) - Update test, now se send empty query instead of error
- [HUE-3334](#) - Skip checking for multi queries if there is no semi colon
- [HUE-3350](#) - Reverse browsing link to use the correct version of the editor
- [HUE-3398](#) - Filter out sessions with empty guid or secret key
- [HUE-3434](#) - Logs of finished Oozie workflow are not displayed
- [HUE-3436](#) - Retain old dependencies when saving a workflow
- [HUE-3437](#) - PamBackend does not honor ignore_username_case
- [HUE-3459](#) - Put stat popover on top
- [HUE-3459](#) - Fixed Flexbox for IE10
- [HUE-3459](#) - Use fixed positioning for assist panel
- [HUE-3459](#) - Revert sticky assist
- [HUE-3459](#) - Fix issue with single panel in metastore and new editor
- [HUE-3459](#) - Clear the height interval on update
- [HUE-3459](#) - Assist doesn't stretch to the end of the page in the old editors
- [HUE-3471](#) - Set the assist database on design update
- [HUE-3471](#) - Assist does not show the DB from the saved query
- [HUE-3476](#) - Clear any running intervals after closing the stats popover
- [HUE-3480](#) - Impala refresh pop-over won't close after assist action while open
- [HUE-3506](#) - Limit length of comments on table page
- [HUE-3511](#) - Reduce flickering of action icons when moving the pointer across several entries
- [HUE-3523](#) - Modify find_jobs_with_no_doc method to exclude jobs with no name
- [HUE-3528](#) - Call correct metrics api to avoid 500 error
- [HUE-3543](#) - Timeout prevents refreshing of the Assist tables/dbs
- [HUE-3594](#) - Smarter DOM based XSS filter on hashes
- [HUE-3601](#) - Spinner not positioned correctly
- [HUE-3613](#) - Empty div elements are added when scrolling the DB assist panel
- [HUE-3614](#) - Scrolling on assist in old editor also scrolls the editor
- [HUE-3637](#) - Avoid decode errors on attribute values
- [HUE-3650](#) - Notify of caught errors in the watch logs process
- [HUE-3651](#) - Upgrade Moment.js
- [HUE-3704](#) - Force enable notebook permissions
- [HUE-3716](#) - Add gen-py paths to hue.pth
- [HUE-3725](#) - 'SparkJob' object has no attribute 'amHostHttpAddress'
- [HUE-3731](#) - Send database on Impala refresh with invalidate
- [HUE-3741](#) - Display field validation errors on create table wizard
- [HUE-3800](#) - Job attempt logs not appearing for some Oozie jobs
- [HUE-3819](#) - Make the upload and create icons not disappear under 1180px
- [IMPALA-2076](#) - Correct execution time tracking for DataStreamSender
- [IMPALA-2502](#) - Don't redundantly repartition grouping aggregations

- [IMPALA-2892](#) - Buffered-tuple-stream-ir.cc is not cross-compiled
- [IMPALA-3133](#) - Wrong privileges after a REVOKE ALL ON SERVER statement
- [IMPALA-3139](#) - Fix drop table statement to not drop views and vice versa
- [IMPALA-3141](#) - Send dummy filters when filter production is disabled
- [IMPALA-3194](#) - Allow queries materializing scalar type columns in RC/sequence files
- [IMPALA-3220](#) - Skip logging empty ScannerContext's stream in parse error
- [IMPALA-3236](#) - Increase timeout for runtime filter tests
- [IMPALA-3238](#) - Avoid log spam for very large hash tables
- [IMPALA-3245](#), [IMPALA-3305](#): Fix crash with global filters when NUM_NODES=1
- [IMPALA-3269](#) - Remove authz checks on default table location in CTAS queries
- [IMPALA-3285](#) - Fix ASAN failure in webserver-test
- [IMPALA-3317](#) - Fix crash in sorter when spilling zero-length strings
- [IMPALA-3334](#) - Fix some bugs in query options' parsing.
- [IMPALA-3367](#) - Ensure runtime filters tests run on 3 nodes
- [IMPALA-3378](#), [IMPALA-3379](#): fix various JNI issues
- [IMPALA-3385](#) - Fix crashes on accessing error_log
- [IMPALA-3395](#) - Old HT filter code uses wrong expr type
- [IMPALA-3396](#) - Fix ConcurrentTimerCounter unit test "TimerCounterTest" failure.
- [IMPALA-3412](#) - Fix CHAR codegen crash in tuple comparator
- [IMPALA-3420](#) - Set IMPALA_THRIFT_VERSION patch level to +4
- [KITE-1108](#) - Add optional retry feature to loadSolr morphline command
- [KITE-1114](#) - Kite CLI json-import HDFS temp file path not multiuser safe
- [OOZIE-2429](#) - TestEventGeneration test is flakey
- [OOZIE-2466](#) - Repeated failure of TestMetricsInstrumentation.testSamplers
- [OOZIE-2486](#) - TestSLAEventsGetForFilterJPAExecutor is flakey
- [OOZIE-2490](#) - Oozie can't set hadoop.security.token.service.use_ip
- [SENTRY-922](#) - INSERT OVERWRITE DIRECTORY permission not working correctly
- [SENTRY-1112](#) - Change default value of "sentry.hive.server" to empty string
- [SENTRY-1164](#) - testCaseSensitivity test failure on a real cluster and also a minor improvements to testConcurrentClients to run locally. (Anne Yu, reviewed by Haohao).
- [SENTRY-1169](#) - MetastorePlugin#renameAuthzObject log message prints oldpathname as newpathname
- [SENTRY-1184](#) - Clean up HMSPaths.renameAuthzObject
- [SENTRY-1190](#) - IMPORT TABLE silently fails if Sentry is enabled
- [SOLR-6631](#) - DistributedQueue spinning on calling zookeeper getChildren()
- [SOLR-6879](#) - Have an option to disable autoAddReplicas temporarily for all collections.
- [SOLR-7493](#) - Requests aren't distributed evenly if the collection isn't present locally. Merges r1683946 and r1683948 from trunk.
- [SOLR-8551](#) - Make collection deletion more robust.
- [SOLR-8683](#) - Tune down stream closed logging
- [SOLR-8720](#) - ZkController#publishAndWaitForDownStates should use #publishNodeAsDown.
- [SOLR-8771](#) - Multi-threaded core shutdown creates executor per core
- [SOLR-8855](#) - The HDFS BlockDirectory should not clean up its cache on shutdown.
- [SOLR-8856](#) - Do not cache merge or 'read once' contexts in the hdfs block cache.
- [SOLR-8857](#) - HdfsUpdateLog does not use configured or new default number of version buckets and is hard coded to 256.
- [SOLR-8869](#) - Optionally disable printing field cache entries in SolrFieldCacheMBean
- [SPARK-4452](#) - Shuffle data structures can starve others on the same thread for memory
- [SPARK-12614](#) - Don't throw non fatal exception from ask
- [SPARK-13622](#) - Issue creating level db for YARN shuffle service
- [SPARK-14242](#) - Avoid copy in compositeBuffer for frame decoder

CDH 5 Release Notes

- [SPARK-14290](#) - Avoid significant memory copy in Netty's tran...
- [SPARK-14363](#) - Fix executor OOM due to memory leak in the Sorter
- [SPARK-14477](#) - Allow custom mirrors for downloading artifacts in build/mvn
- [SPARK-14679](#) - Fix UI DAG visualization OOM.
- [SQOOP-2847](#) - Sqoop --incremental + missing parent --target-dir reports success with no data

Issues Fixed in CDH 5.7.0

CDH 5.7.0 fixes the following issues.

Apache Flume

TailDirSource throws FileNotFoundException if ~/ .flume directory is not created already

Bug: [FLUME-2773](#)

This fix ensures that any missing parent directories in the `positionFile` path (either default or user given input) are always created.

flume_env script should handle JVM parameters like -javaagent -agentpath -agentlib

Bug: [FLUME-2763](#)

This fix enables the `flume_env` script to handle JVM parameters such as `-javaagent -agentpath` and `-agentlib`.

Kafka channel timeout property is overridden by default value

Bug: [FLUME-2734](#)

When the Kafka channel timeout property is passed to the Kafka consumer internally, it does not work as expected. It is overridden by the default value or the value specified by the `.timeout` property, which is undocumented. Now the `kafka.consumer.timeout.ms` value specified in the configuration takes effect like other Kafka consumer properties.

Apache Hadoop

ReplicationMonitor can infinitely loop in BlockPlacementPolicyDefault#chooseRandom()

Bug: [HDFS-4937](#)

When a large number of nodes are removed by refreshing node lists, the network topology is updated and the replication monitor thread may get stuck in the `while` loop of `chooseRandom()`.

Clean up temporary files after fsimage transfer failures

Bug: [HDFS-7373](#)

When an `fsimage` (or checkpoint) transfer fails, a temporary file is left in each storage directory. If the namespace is large, these files can take up a large amount of space.

Lease recovery should return true if the lease can be released and the file can be closed

Bug: [HDFS-8576](#)

`FSNamesystem#recoverLease` should return `true` when a lease is recovered both explicitly and implicitly—that is, when a lease recovery is successful and the file is closed, and also when a file is closed and the lease is released *without* a recovery.

fsck does not list correct file path when bad replicas or blocks are in a snapshot

Bug: [HDFS-9231](#)

When blocks are corrupt in a snapshot, the `fsck` command lists the original directory and not the snapshot directory. This happens even when the original file is deleted. The specific commands are `fsck -list-corruptfileblocks` and `fsck -list-corruptfileblocks -includeSnapshots`.

Make DataStreamer#block thread safe and verify generationStamp in commitBlock

Bug: [HDFS-9289](#)

When the client calls `updatePipeline`, a block might commit with an old `generationStamp`, causing replicas to look corrupt.

Delayed heartbeat processing causes storm of subsequent heartbeats

Bug: [HDFS-9305](#)

The NameNode usually handles DataNode heartbeats quickly, but can be delayed for various reasons, such as a long garbage collection or lock contention. After the NameNode recovers, the DataNode sends a storm of heartbeat messages in a tight loop which, in a big cluster, can overload the NameNode and make cluster recovery difficult.

FSImage may get corrupted after deleting snapshot

Bug: [HDFS-9406](#)

When deleting a snapshot that contains the last record of a given INode, the fsimage may become corrupt because the create list of the snapshot diff in the previous snapshot and the child list of the parent `INodeDirectory` are not cleaned.

Apache HBase

See also [Known Issues In CDH 5.7.0](#) on page 122.

Potential data loss after a RegionServerAbortedException

Bug: [HBASE-13895](#)

If the master attempts to assign a region while handling a RegionServer abort, the returned `RegionServerAbortedException` is handled as though the region had been cleanly taken offline, so the new assignment is allowed to proceed. If the region is opened in its new location before WAL replay has completed, the replayed edits are ignored, or are later played back on top of new edits that happened after the region was opened. In either case, data can be lost.

Workaround: None.

Data loss can occur if a table has more than 2,147,483,647 columns

Bug: [HBASE-15133](#)

Data loss can occur if a table has more than 2,147,483,647 (`Integer.MAX_INT`) columns, because some key variable types are INT rather than LONG.

Workaround: Adjust your schema to use fewer than `Integer.MAX_INT` columns.

Delete operations that occur during a region merge may be eclipsed by new Put operations

Bug: [HBASE-13938](#)

The master's timestamp is not used when sending `hbase:meta` edits on region merges, so correct ordering of new region additions and old region deletes is not assured and data loss can occur if edits are applied in the wrong order.

Workaround: None.

Conflicts between HBase Balancer and hbase:meta reassignment

Bug: [HBASE-14536](#)

If `hbase:meta` is assigned to a RegionServer that becomes unavailable, and the HBase balancer has scheduled but not completed a plan to move `hbase:meta` to a different RegionServer, the `hbase:meta` becomes unassigned.

Workaround: None.

Regions can fail to transition in a write-heavy cluster with a small number of read handlers

Bug: [HBASE-13635](#)

On a write-heavy cluster configured with a small number of read handlers, all requests that are not mutations are sent to the read handlers, including `ReportRegionInTransition` requests. If these requests time out, the RegionServer is assumed to be unavailable, and the regions cannot transition correctly.

Workaround: None.

In a secured environment, when a RegionServer is stopped, znodes may not be cleaned up correctly

Bug: [HBASE-14581](#)

When a RegionServer process is stopped, the `zkcli` command is invoked to delete its znodes. In a secure cluster, the `zkcli` command does not authenticate to ZooKeeper and the deletion fails. This problem occurs because the `REGIONSERVER_OPTS` environment variable is not correctly passed when invoking the `zkcli` command.

Workaround: None.

Delays in RegionServer responses can cause a region to be closed indefinitely

Bug: [HBASE-14407](#)

Handling of region assignment by the master has a flaw when RegionServer responses are delayed due to network delays, system load, or other reasons. This flaw can cause the master to close a region indefinitely.

Workaround: Restart the RegionServer to force the region to be reassigned.

When a RegionServer crashes, replication peers can crash due to inode exhaustion from old WALS

Bug: [HBASE-14621](#)

The fix for [HBASE-12865](#) ensures that `loadWALsFromQueues` attempts a retry when the replication source version is changed while loading the replication queue. However, the fix introduced a bug in `ReplicationLogCleaner` that causes an infinite loop when a RegionServer crashes. As a result, old WALs are not cleaned up. In a cluster under high load, the inode limit on the replication peer RegionServer can be exhausted, causing the RegionServer to crash.

Workaround: None.

When a RegionServer crashes, cell-level visibility tags may be lost during WAL replay

Bug: [HBASE-15218](#)

When reading cells after a RegionServer crash, the `KeyValueCodec` and the `WallCellCodec` both use `NoTagsKeyValue`, which does not preserve visibility tags.

Workaround: None.

Column is not deleted if you do not pass the visibility label

Bug: [HBASE-14761](#)

If a column was created or modified with a visibility label, and you attempt to delete it without passing the visibility label, the column is not deleted. It is not visible using a `Scan` operation, but is visible using a raw `Scan`, and is marked with `deleteColumn`.

Workaround: None.

If multiple users are configured with the role `hbase.superuser`, an attempt to connect to a secure ZooKeeper instance fails

Bug: [HBASE-14425](#)

The `hbase.superuser` configuration option is a comma-separated list of users. A bug in the code to connect to a secure ZooKeeper causes the list to be evaluated as a single value, so a list of multiple users fails because no username matches the comma-separated list.

Workaround: Only specify a single user in the `hbase.superuser` configuration option.

Region split request audits are performed against the `hbase` user instead of the requesting user

Bug: [HBASE-14475](#)

When checking whether the requesting user has permission to perform a region split, the `hbase` user's permissions are checked instead of those of the requesting user. Due to this bug, `CREATE` is sufficient for the split, rather than `CREATE` and `ADMIN`. Because `CREATE` permissions are also sufficient for flushes and compactions, this issue is not severe in most environments.

Workaround: None.

Incorrect timestamp checking causes unpredictable deletes with VisibilityScanDeleteChecker.

Bug: [HBASE-13635](#)

Incorrect timestamp checking when `VisibilityScanDeleteChecker` is used causes unpredictable results when deleting cells. In some cases, the timestamp is deleted but the cell contents are not deleted. In other cases, a request to delete an entire row or to delete a version results in only a single cell being deleted.

Workaround: None.

A BulkLoad of an HFile with tags that requires splits does not preserve the tags

Bug: [HBASE-15035](#)

When an HFile is created with cell tags and is imported into HBase using a bulk load, the tags are present as expected when the HFile is loaded into a single region. However, if the bulk load spans multiple regions, the original HFile is automatically split into a set of HFiles corresponding to each of the regions the original HFile covers. Tags, including ACLs, TTLs, and MOB pointers, are not copied to the split files.

Workaround: None.

Restoring a snapshot from a table in a user-defined namespace causes a URISyntaxException

Bug: [HBASE-14578](#)

A table in a user-defined namespace uses a colon between the namespace and the table name (for instance, `example_ns:users`). This colon is interpreted incorrectly when restoring from a snapshot.

Workaround: None.

The list_snapshots HBase shell command shows all snapshots, regardless of the user's permission to view them

Bug: [HBASE-12552](#)

A user with no privileges to interact with a snapshot can list the snapshot using the `list_snapshots` HBase shell command.

Workaround: None.

ExportSnapshot or DistCp operations may fail on the Amazon s3a:// protocol

Bug: None.

`ExportSnapshot` or `DistCp` operations may fail on AWS when using certain JDK 8 versions, due to an incompatibility between AWS Java SDK 1.9.x and the `joda-time` date-parsing module.

Workaround: Use `joda-time` 2.8.1 or higher, which is included in AWS Java SDK 1.10.1 or higher.

If HDFS is restarted while HBase is running, WALs being replicated may not close correctly

Bug: [HBASE-15019](#)

The RegionServer receiving the replicated WALs has no mechanism to be notified to perform a recovery if HDFS is restarted on the source cluster.

Workaround: Restart the RegionServer to force the master to trigger the lease recovery during WAL splitting.

The permissions of the .top/ directory are not explicitly set during LoadIncrementalHFiles operations

Bug: [HBASE-14005](#)

Permissions are not explicitly set on the `.top/` directory created during `LoadIncrementalHFiles`. The permissions should be set the same as the `.bottom/` and `.tmp/` directories.

Workaround: None.

Nonfatal errors in the FSHLog subsystem are incorrectly logged as fatal errors

Bug: [HBASE-14042](#)

CDH 5 Release Notes

If an `IOException` causes a log roll to be requested, it is logged as a fatal event, although it should be logged as a warning.

Workaround: None.

FuzzyRowFilter may omit some rows if multiple fuzzy keys are present

Bug: [HBASE-14269](#)

If you use the `FuzzyRowFilter` for Scan operations, and you have multiple fuzzy keys, some rows may be omitted from the `RowTracker`.

Workaround: None.

The prefix-tree module is not automatically included in MapReduce jobs

Bug: [HBASE-15152](#)

JARs for `prefix-tree` module are not automatically included in `YarnChildren` processes. This causes a `ClassNotFoundException`.

Workaround: Manually add the `prefix-tree` JARs to the classpath if needed.

Values of some metrics may appear to be negative

Bug: [HBASE-12961](#)

Some metric value are stored in integers, and cannot accommodate real-world values. This causes metric values to appear to be negative.

Workaround: None.

The HBase Shell cannot handle Scan filters which contain non-UTF8 characters

Bug: [HBASE-15032](#)

The HBase Shell incorrectly handles filter strings which contain non-UTF8 characters.

Workaround: None.

Reverse scans do not work when Bloom blocks or leaf-level inode blocks are present

Bug: [HBASE-14283](#)

Because the `seekBefore()` method calculates the size of the previous data block by assuming that data blocks are contiguous, and HFile v2 and higher store Bloom blocks and leaf-level inode blocks with the data, reverse scans do not work when Bloom blocks or leaf-level inode blocks are present when HFile v2 or higher is used.

Workaround: None.

Apache Hive

Fix regression in bind and search logic for Hive external LDAP authentication

Bug: [HIVE-12885](#)

Fixes a regression in LDAP bind and search authentication from CDH 5.5.0.

Some queries using LEFT SEMI JOIN fail with IndexOutOfBoundsException

Bug: [HIVE-13082](#)

Some queries using `LEFT SEMI JOIN` fail with `IndexOutOfBoundsException`. Constant propagation optimization for these queries is now enabled.

BETWEEN predicate is not functioning correctly with predicate pushdown on Parquet tables

Bug: [HIVE-13039](#)

`BETWEEN` becomes exclusive in Parquet table when predicate pushdown (PPD) is enabled for Parquet tables, leading to potential incorrect results.

Performance degradation when running Hive queries against wide tables with Sentry enabled

Bug: [SENTRY-1007](#)

Fixes a performance regression due to inefficient authorization checks in the Sentry Hive binding for Hive tables that are wide (more than 100 columns).

Optionally cancel queries after configurable timeout waiting on compilation lock

Bug: [HIVE-12431](#)

Adds a new configuration option, `hive.server2.compile.lock.timeout`, that cancels queries if they are waiting for the compile lock for more than this amount of time. This applies only to queries waiting on compilation and does not cancel queries that are being compiled. By default, the timeout is unlimited.

java.io.DeleteOnExitHook leaks memory on long-running HiveServer2 Instances

Bug: [HIVE-11768](#)

The Hive plugin hook, `java.io.DeleteOnExitHook`, leaks memory on long-running HiveServer2 Instances. Over time this may lead to `OutOfMemoryError` (OOM) and service crashes. This fixes the memory leak by ensuring all resources are always cleaned up after processing an event.

Hue

The Hive Sample Table, customer, Cannot be Queried

Bug: [HUE-3040](#)

Users can now query the customer table in the Hive editor.

Apache Impala (incubating)

For the list of Impala fixed issues, see [Issues Fixed in Impala for CDH 5.7.0](#) on page 312.

See also [Apache Impala \(incubating\) Known Issues](#) on page 135 for issues that are known but not resolved yet.

MapReduce

MapReduce Rolling Upgrades To and From CDH 5.6.0 Fail

Bug: [CDH-38587](#)

Users can now safely use rolling upgrade from releases CDH 5.6.0 and lower to CDH 5.7.0.

Cloudera Search

Reordered updates cause leaders and replicas to become out of sync

Bug: [SOLR-8586, SOLR-8691](#)

Solr relied on checking leader/replica document synchronization by comparing the last 100 updates on the leader and replica for significant overlap, and then applying any missing updates from the leader. In certain cases, document updates could be significantly reordered, resulting in mismatches in the index, even when the last 100 documents matched. Solr now implements hashing over the versions of all the documents to check for synchronization, eliminating a class of errors in which replicas and leaders could become out of sync.

Apache Sentry

Fixed Sentry Oracle upgrade script

Bug: [SENTRY-1066](#)

This fixes previous Sentry upgrade issues with Oracle (ORA-0955).

Tables with non-HDFS locations break Hive Metastore startup

Bug: [SENTRY-1044](#)

Tables with non-HDFS locations cause the HDFS/Sentry plugin to enter an invalid state and fail with the exception, Could not create Initial AuthzPaths or HMSHandler !!.

CDH 5 Release Notes

URI check is now case-sensitive

Bug: [SENTRY-968](#)

Sentry no longer ignores case when validating privileges for URIs.

TRUNCATE on empty partitioned table in Hive fails

Bug: [SENTRY-826](#)

PathsUpdate.parsePath(path) will throw an NPE when parsing relative paths

Bug: [SENTRY-1002](#)

Sentry now skips relative paths (that is, paths without a fully qualified scheme) rather than failing with a NPE.

The Sentry Server should be not be updated if the CREATE/DROP operations fail

Bug: [SENTRY-1008](#)

Previously, even if a CREATE TABLE operation failed, the Sentry Server would still be updated with a path to the table. This has been fixed.

SimpleDBProviderBackend should retry the authorization process

Bug: [SENTRY-902](#)

This fix includes corrections to the retry logic to remove recursive calls and include a wait time between retries when authorization fails.

Support Hive RELOAD by updating the classpath for Sentry

Bug: [SENTRY-1003](#)

When Hive issues the RELOAD command, Sentry now gets the updated auxiliary JAR path from the `hive.reloadable.aux.jars.path` property.

RealTimeGet with explicit ids can bypass document-level authorization

Bug: [SENTRY-989](#)

Users can no longer bypass security by guessing the document IDs for the RealTimeGet command.

Updated Apache Shiro dependency

Bug: [SENTRY-1054](#)

External partitions referenced by more than one table can cause some unexpected behavior with Sentry HDFS sync

Bug: [SENTRY-953](#)

INSERT INTO command no longer requires URI privilege on partition location under table

Bug: [SENTRY-1095](#)

The checks on the Hive INSERT INTO command have been relaxed. The INSERT INTO command adds location information to the partition description. Although this requires a check on URI privileges, in this case location information can be generated even if the partition is under the table directory.

Improvement to the SentryAuthFilter error message when authentication fails

Bug: [SENTRY-1060](#)

Avoid logging all DataNucleus queries when debug logging is enabled

Bug: [SENTRY-945](#)

Logging DataNucleus queries when debugging can fill up 2 GB of log file space in less than five minutes.

getGroup and getUser should always return original HDFS values for paths that are not managed by Sentry

Bug: [SENTRY-936](#)

Paths that do not correspond to Hive metastore objects should not be affected by HDFS/Sentry sync.

Exceptions in MetastoreCacheInitializer should not prevent HMS from starting up

Bug: [SENTRY-957](#)

Instead of only throwing a runtime exception, this fix ensures failed tasks are first retried.

Set maximum message size for Thrift messages

Bug: [SENTRY-904](#)

This ensures that security scans and unformatted messages do not bring down the Sentry server by going out of bounds.

Allow SentryAuthorization setter path always fall through and update HDFS

Bug: [SENTRY-988](#)

Setting HDFS rules on Sentry-managed HDFS paths should not affect original HDFS rules

Bug: [SENTRY-944](#)

Removing and modifying ACLs on Sentry-managed paths that correspond to Hive metastore objects should not affect the original HDFS access rules.

Fix inconsistency in column-level privileges

Bug: [SENTRY-847](#)

If you have column-level privileges for a table, the SHOW columns operation should not require extra table-level privileges.

Performance Improvements

Improved performance for filtering Hive SHOW commands

Bug: [SENTRY-565](#)

HiveAuthzBinding has been improved to reduce the number of RPC calls when filtering SHOW commands.

Improved Sentry column-level performance for wide tables

Bug: [SENTRY-1007](#)

Apache Spark

Certain Spark MLlib features not supported

The following Spark MLlib features are now supported:

- spark.ml
- ML pipeline APIs

Streaming incompatibility between Spark 1.2 and 1.3

Applications built as a JAR with dependencies ("uber JAR") must be built for the specific version of Spark running on the cluster.

Workaround: Rebuild the JAR with the Spark dependencies in `pom.xml` pointing to the specific version of Spark running on the target cluster.

Spark SQL cannot retrieve data from a partitioned Hive table

When reading from a partitioned Hive table, Spark SQL cannot identify the column delimiter used and reads the full record as the first column entry.

Workaround: Contact Cloudera Support for information on how to deploy a patch to resolve the issue.

Tasks that fail due to YARN preemption can cause job failure

Bug: [SPARK-8167](#)

CDH 5 Release Notes

Tasks that are running on preempted executors will count as FAILED with an `ExecutorLostFailure`.

Apache Sqoop

Oracle connector not working with lowercase columns

Bug: [SQOOP-2723](#)

The Oracle connector now works with lowercase columns.

Run only one map task attempt during export

Bug: [SQOOP-2712](#)

In most scenarios, running multiple map task attempts by default when performing an export is not required. The default is now one map task attempt during export operations.

Do not dump data on error in TextExportMapper by default

Bug: [SQOOP-2651](#)

Dumping data in the `TextExportMapper` might unintentionally leak sensitive information to logs. The `enableDataDumpOnError` key is set to `false` by default. A user can set the value to `true` to intentionally write the data to the log.

Support of glob paths during export

Bug: [SQOOP-1281](#)

Glob paths are now supported for export.

Sqoop should support importing from table with column names containing some special characters

Bug: [SQOOP-2387](#)

Sqoop supports some special characters in column names. The specific list of characters depends on those supported for a particular database.

Avro export ignores --columns option

Bug: [SQOOP-1369](#)

AvroExportMapper now supports the `--columns` option to restrict the columns to export.

JDK

Java 8 (updates 60 and higher) has problems with joda-time and S3 requests

Bug: [SPARK-11413](#)

Versions of Java 1.8, from update 60 and higher (`jdk1.8.0_60++`), cause S3 to fail because joda-time cannot format time zones.

Issues Fixed in CDH 5.6.x

The following topics describe issues fixed in CDH 5.6.x, from newest to oldest release. You can also review [What's New In CDH 5.6.x](#) on page 23 or [Known Issues in CDH 5](#) on page 111.

Issues Fixed in CDH 5.6.1

CDH 5.6.1 fixes the following issues.

Apache Hadoop

FSImage may get corrupted after deleting snapshot

Bug: [HDFS-9406](#)

When deleting a snapshot that contains the last record of a given INode, the fsimage may become corrupt because the create list of the snapshot diff in the previous snapshot and the child list of the parent `INodeDirectory` are not cleaned.

Apache HBase

The ReplicationCleaner process can abort if its connection to ZooKeeper is inconsistent

Bug: [HBASE-15234](#)

If the connection with ZooKeeper is inconsistent, the ReplicationCleaner may abort, and the following event is logged by the HMaster:

```
WARN org.apache.hadoop.hbase.replication.master.ReplicationLogCleaner: Aborting
ReplicationLogCleaner
because Failed to get list of replicators
```

Unprocessed WALs accumulate.

The seekBefore() method calculates the size of the previous data block by assuming that data blocks are contiguous, and HFile v2 and higher store Bloom blocks and leaf-level INode blocks with the data. As a result, reverse scans do not work when Bloom blocks or leaf-level INode blocks are present when HFile v2 or higher is used.

Workaround: Restart the HMaster occasionally. The ReplicationCleaner restarts if necessary and process the unprocessed WALs.

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.6.1:

- [FLUME-2632](#) - High CPU on KafkaSink
- [FLUME-2712](#) - Optional channel errors slows down the Source to Main channel event rate
- [FLUME-2781](#) - Kafka Channel with parseAsFlumeEvent=true should write data as is, not as flume events
- [FLUME-2886](#) - Optional Channels can cause OOMs
- [FLUME-2891](#) - Revert FLUME-2712 and FLUME-2886
- [FLUME-2897](#) - AsyncHBase sink NPE when Channel.getTransaction() fails
- [HADOOP-7139](#) - Allow appending to existing SequenceFiles
- [HADOOP-7817](#) - RawLocalFileSystem.append() should give FSDataOutputStream with accurate .getPos()
- [HADOOP-11171](#) - Enable using a proxy server to connect to S3a
- [HADOOP-11321](#) - copyToLocal cannot save a file to an SMB share unless the user has Full Control permissions
- [HADOOP-11687](#) - Ignore x-* and response headers when copying an Amazon S3 object
- [HADOOP-11722](#) - Some Instances of Services using ZKDelegationTokenSecretManager go down when old token cannot be deleted
- [HADOOP-12240](#) - Fix tests requiring native library to be skipped in non-native profile
- [HADOOP-12280](#) - Skip unit tests based on maven profile rather than NativeCodeLoader.isNativeCodeLoaded
- [HADOOP-12604](#) - Exception may be swallowed in KMSClientProvider
- [HADOOP-12605](#) - Fix intermittent failure of TestIPC.testIpcWithReaderQueuing
- [HADOOP-12668](#) - Support excluding weak Ciphers in HttpServer2 through ssl-server.conf
- [HADOOP-12699](#) - TestKMS#testKMSProvider intermittently fails during 'test rollover draining'
- [HADOOP-12715](#) - TestValueQueue#testgetAtMostPolicyALL fails intermittently
- [HADOOP-12718](#) - Incorrect error message by fs -put local dir without permission
- [HADOOP-12736](#) - TestTimedOutTestsListener#testThreadDumpAndDeadlocks sometimes times out
- [HADOOP-12825](#) - Log slow name resolutions
- [HADOOP-12954](#) - Add a way to change hadoop.security.token.service.use_ip
- [HADOOP-12972](#) - Lz4Compressor#getLibraryName returns the wrong version number
- [HDFS-6520](#) - hdfs fsck passes invalid length value when creating BlockReader
- [HDFS-8211](#) - DataNode UUID is always null in the JMX counter
- [HDFS-8496](#) - Calling stopWriter() with FSDataServiceImpl lock held may block other threads
- [HDFS-8576](#) - Lease recovery should return true if the lease can be released and the file can be closed
- [HDFS-8785](#) - TestDistributedFileSystem is failing in trunk
- [HDFS-8855](#) - Webhdfs client leaks active NameNode connections

CDH 5 Release Notes

- [HDFS-9264](#) - Minor cleanup of operations on FsVolumeList#volumes
- [HDFS-9289](#) - Make DataStreamer#block thread safe and verify genStamp in commitBlock
- [HDFS-9347](#) - Invariant assumption in TestQuorumJournalManager.shutdown() is wrong
- [HDFS-9350](#) - Avoid creating temporary strings in Block.toString() and getBlockName()
- [HDFS-9358](#) - TestNodeCount#testNodeCount timed out
- [HDFS-9514](#) - TestDistributedFileSystem.testDFSClientPeerWriteTimeout failing; exception being swallowed
- [HDFS-9576](#) - HTrace: collect position/length information on read operations
- [HDFS-9589](#) - Block files which have been hardlinked should be duplicated before the DataNode appends to the them
- [HDFS-9612](#) - DistCp worker threads are not terminated after jobs are done
- [HDFS-9655](#) - NN should start JVM pause monitor before loading fsimage.
- [HDFS-9688](#) - Test the effect of nested encryption zones in HDFS downgrade
- [HDFS-9701](#) - DN may deadlock when hot-swapping under load
- [HDFS-9721](#) - Allow Delimited PB OIV tool to run upon fsimage that contains INodeReference
- [HDFS-9949](#) - Add a test case to ensure that the DataNode does not regenerate its UUID when a storage directory is cleared
- [HDFS-10223](#) - peerFromSocketAndKey performs SASL exchange before setting connection timeouts
- [HDFS-10267](#) - Extra "synchronized" on FsDatasetImpl#recoverAppend and FsDatasetImpl#recoverClose
- [MAPREDUCE-4785](#) - TestMRApp occasionally fails
- [MAPREDUCE-6460](#) - TestRMContainerAllocator.testAttemptNotFoundCausesRMCommunicatorException fails
- [MAPREDUCE-6528](#) - Memory leak for HistoryFileManager.getJobSummary()
- [MAPREDUCE-6580](#) - Test failure: TestMRJobsWithProfiler
- [MAPREDUCE-6620](#) - Jobs that did not start are shown as starting in 1969 in the JHS web UI
- [YARN-2749](#) - Fix some testcases from TestLogAggregationService fails in trunk
- [YARN-2871](#) - TestRMRestart#testRMRestartGetApplicationList sometime fails in trunk
- [YARN-2902](#) - Killing a container that is localizing can orphan resources in the DOWNLOADING state
- [YARN-3104](#) - Fixed RM to not generate new AMRM tokens on every heartbeat between rolling and activation
- [YARN-3446](#) - FairScheduler headroom calculation should exclude nodes in the blacklist
- [YARN-3727](#) - For better error recovery, check if the directory exists before using it for localization
- [YARN-4155](#) - TestLogAggregationService.testLogAggregationServiceWithInterval failing
- [YARN-4168](#) - Fixed a failing test TestLogAggregationService.testLocalFileDeletionOnDiskFull
- [YARN-4354](#) - Public resource localization fails with NPE
- [YARN-4380](#) - TestResourceLocalizationService.testDownloadingResourcesOnContainerKill fails intermittently
- [YARN-4393](#) - Fix intermittent test failure for TestResourceLocalizationService#testFailedDirsResourceRelease
- [YARN-4546](#) - ResourceManager crash due to scheduling opportunity overflow
- [YARN-4573](#) - Fix test failure in TestRMAppTransitions#testAppRunningKill and testAppKilledKilled
- [YARN-4613](#) - Fix test failure in TestClientRMServer#testGetClusterNodes
- [YARN-4704](#) - TestResourceManager#testResourceAllocation() fails when using FairScheduler
- [YARN-4717](#) - TestResourceLocalizationService.testPublicResourceInitializesLocalDir fails Intermittently due to IllegalArgumentException from cleanup
- [HBASE-6617](#) - ReplicationSourceManager should be able to track multiple WAL paths
- [HBASE-14374](#) - Stuck FSHLog' issue to 1.1 Also includes HBASE-14807 TestWALockup is flakey
- [HBASE-14759](#) - Avoid using Math.abs when selecting SyncRunner in FSHLog
- [HBASE-15019](#) - Replication stuck when HDFS is restarted
- [HBASE-15052](#) - Use EnvironmentEdgeManager in ReplicationSource
- [HBASE-15152](#) - Automatically include prefix-tree module in MR jobs if present
- [HBASE-15157](#) - Add *PerformanceTest for Append, CheckAnd* Reason: Bug Author: Stack Ref:
- [HBASE-15206](#) - Fix flakey testSplitDaughtersNotInMeta
- [HBASE-15213](#) - Fix increment performance regression caused by HBASE-8763 on branch-1.0
- [HBASE-15234](#) - Don't abort ReplicationLogCleaner on ZooKeeper errors

- [HBASE-15456](#) - CreateTableProcedure/ModifyTableProcedure needs to fail when there is no family in table descriptor
- [HBASE-15479](#) - No more garbage or beware of autoboxing
- [HBASE-15582](#) - SnapshotManifestV1 too verbose when there are no regions
- [HIVE-6099](#) - Multi insert does not work properly with distinct count
- [HIVE-7653](#) - Hive AvroSerDe does not support circular references in Schema
- [HIVE-9617](#) - UDF from_utc_timestamp throws NPE if the second argument is null
- [HIVE-10115](#) - HS2 running on a Kerberized cluster should offer Kerberos(GSSAPI) and Delegation token(DIGEST) when alternate authentication is enabled
- [HIVE-10213](#) - MapReduce jobs using dynamic-partitioning fail on commit
- [HIVE-10303](#) - HIVE-9471 broke forward compatibility of ORC files
- [HIVE-11054](#) - Handle varchar/char partition columns in vectorization
- [HIVE-11097](#) - HiveInputFormat uses String.startsWith to compare splitPath and PathToAliases
- [HIVE-11135](#) - Fix the Beeline set and save command in order to avoid the NullPointerException
- [HIVE-11285](#) - ObjectInspector for partition columns in FetchOperator in SMBJoin causes exception
- [HIVE-11288](#) - Avro SerDe InstanceCache returns incorrect schema
- [HIVE-11408](#) - HiveServer2 is leaking ClassLoaders when add jar / temporary functions are used due to constructor caching in Hadoop ReflectionUtils
- [HIVE-11427](#) - Location of temporary table for CREATE TABLE SELECT broken by HIVE-7079
- [HIVE-11488](#) - Combine the following jiras for "Support sessionId and queryId logging"Add sessionId and queryId info to HS2 log
- [HIVE-12456](#): QueryId can't be stored in the configuration of the SessionState since multiple queries can run in a single session
- [HIVE-11583](#) - When PTF is used over a large partitions result could be corrupted
- [HIVE-11590](#) - AvroDeserializer is very chatty
- [HIVE-11828](#) - beeline -f fails on scripts with tabs between column type and comment
- [HIVE-11919](#) - Hive Union Type Mismatch
- [HIVE-12315](#) - Fix Vectorized double divide by zero
- [HIVE-12354](#) - MapJoin with double keys is slow on MR
- [HIVE-12431](#) - Support timeout for compile lock
- [HIVE-12469](#) - Apache Commons Collections
- [HIVE-12506](#) - SHOW CREATE TABLE command creates a table that does not work for RCFile format
- [HIVE-12706](#) - Incorrect output from from_utc_timestamp()/to_utc_timestamp when local timezone has DST
- [HIVE-12790](#) - Metastore connection leaks in HiveServer2
- [HIVE-12885](#) - LDAP Authenticator improvements
- [HIVE-12941](#) - Unexpected result when using MIN() on struct with NULL in first field
- [HIVE-12946](#) - alter table should also add default scheme and authority for the location similar to create table
- [HIVE-13039](#) - BETWEEN predicate is not functioning correctly with predicate pushdown on Parquet table
- [HIVE-13055](#) - Add unit tests for HIVE-11512
- [HIVE-13065](#) - Hive throws NPE when writing map type data to a HBase backed table
- [HIVE-13082](#) - Enable constant propagation optimization in query with left semi join
- [HIVE-13200](#) - Aggregation functions returning empty rows on partitioned columns
- [HIVE-13243](#) - Hive drop table on encryption zone fails for external tables
- [HIVE-13251](#) - Hive can't read the decimal in AVRO file generated from previous version
- [HIVE-13286](#) - Query ID is being reused across queries
- [HIVE-13295](#) - Improvement to LDAP search queries in HS2 LDAP Authenticator
- [HIVE-13401](#) - Kerberized HS2 with LDAP auth enabled fails kerberos/delegation token authentication
- [HIVE-13527](#) - Using deprecated APIs in HBase client causes zookeeper connection leaks
- [HIVE-13570](#) - Some queries with Union all fail when CBO is off
- [HUE-3106](#) - [filebrowser] Add support for full paths in zip file uploads

CDH 5 Release Notes

- [HUE-3110](#) - [oozie] Fix bundle submission when coordinator points to multiple bundles
- [HUE-3132](#) - [core] Fix Sync Ldap users and groups for anonymous binds
- [HUE-3180](#) - [useradmin] Override duplicate username validation message
- [HUE-3185](#) - [oozie] Avoid extra API calls for parent information in workflow dashboard
- [HUE-3303](#) - [core] PostgreSQL requires data update and alter table operations in separate transactions
- [HUE-3310](#) - [jobsub] Prevent browsing job designs by API
- [HUE-3334](#) - [editor] Update test, now send empty query instead of error, skip checking for multi queries if there is no semicolon
- [HUE-3398](#) - [beeswax] Filter out sessions with empty guid or secret key
- [HUE-3436](#) - [oozie] Retain old dependencies when saving a workflow
- [HUE-3437](#) - [core] PamBackend does not honor ignore_username_case
- [HUE-3523](#) - [oozie] Modify find_jobs_with_no_doc method to exclude jobs with no name
- [HUE-3528](#) - [oozie] Call correct metrics api to avoid 500 error
- [HUE-3594](#) - [fb] Smarter DOM based XSS filter on hashes
- [HUE-3637](#) - [sqoop] Avoid decode errors on attribute values
- [HUE-3650](#) - [beeswax] Notify of caught errors in the watch logs process
- [HUE-3651](#) - [core] Upgrade Moment.js
- [IMPALA-852](#), [IMPALA-2215](#) - Analyze HAVING clause before aggregation
- [IMPALA-1092](#) - Fix estimates for trivial coord-only queries
- [IMPALA-1170](#) - Fix URL parsing when path contains '@'
- [IMPALA-1934](#) - Allow shell to retrieve LDAP password from shell cmd
- [IMPALA-2093](#) - Disallow NOT IN aggregate subqueries with a constant lhs expr
- [IMPALA-2184](#) - don't inline timestamp methods with try/catch blocks in IR
- [IMPALA-2425](#) - Broadcast join hint not enforced when low memory limit is set
- [IMPALA-2503](#) - Add missing String.format() arg in error message
- [IMPALA-2539](#) - Unmark collections slots of empty union operands
- [IMPALA-2554](#) - Change default buffer size for RPC servers and clients
- [IMPALA-2565](#) - Planner tests are flaky due to file size mismatches
- [IMPALA-2592](#) - DataStreamSender::Channel::CloseInternal() does not close the channel on an error
- [IMPALA-2599](#) - Pseudo-random sleep before acquiring kerberos ticket possibly not really pseudo-random
- [IMPALA-2711](#) - Fix memory leak in Rand()
- [IMPALA-2732](#) - Timestamp formats with non-padded values
- [IMPALA-2734](#) - Correlated EXISTS subqueries with HAVING clause return wrong results
- [IMPALA-2742](#) - Avoid unbounded MemPool growth with AcquireData()
- [IMPALA-2749](#) - Fix decimal multiplication overflow
- [IMPALA-2765](#) - Preserve return type of subexpressions substituted in isTrueWithNullSlots()
- [IMPALA-2788](#) - conv(bigint num, int from_base, int to_base) returns wrong result
- [IMPALA-2798](#) - Bring in AVRO-1617 fix and add test case for it
- [IMPALA-2818](#) - Fix cancellation crashes/hangs due to BlockOnWait() race
- [IMPALA-2820](#) - Support unquoted keywords as struct-field names
- [IMPALA-2832](#) - Fix cloning of FunctionCallExpr
- [IMPALA-2844](#) - Allow count(*) on RC files with complex types
- [IMPALA-2870](#) - Fix failing metadata.test_ddl.TestDdlStatements.test_create_table test
- [IMPALA-2894](#) - Move regression test into a different .test file
- [IMPALA-2906](#) - Fix an edge case with materializing TupleIsNotNullPredicates in analytic sorts
- [IMPALA-2914](#) - Fix DCHECK Check failed: HasDateOrTime()
- [IMPALA-2926](#) - Fix off-by-one bug in SelectNode::CopyRows()
- [IMPALA-2940](#) - Fix leak of dictionaries in Parquet scanner
- [IMPALA-3000](#) - Fix BitReader::Reset()
- [IMPALA-3034](#) - Verify all consumed memory of a MemTracker is always released at destruction time

- [IMPALA-3047](#) - Separate create table test with nested types
- [IMPALA-3054](#) - Disable probe side filters when spilling
- [IMPALA-3071](#) - Fix assignment of On-clause predicates belonging to an inner join
- [IMPALA-3085](#) - Unregister data sinks' MemTrackers at their Close() functions
- [IMPALA-3093](#) - ReopenClient() could NULL out 'client_key' causing a crash
- [IMPALA-3095](#) - Add configurable whitelist of authorized internal principals
- [IMPALA-3151](#) - Impala crash for avro table when casting to char data type
- [IMPALA-3194](#) - Allow queries materializing scalar type columns in RC/sequence files
- [KITE-1114](#) - Kite CLI json-import HDFS temp file path not multiuser safe, fix missing license header
- [OOZIE-2419](#) - HBase credentials are not correctly proxied
- [OOZIE-2466](#) - Repeated failure of TestMetricsInstrumentation.testSamplers
- [OOZIE-2486](#) - TestSLAEVENTSGetForFilterJPAExecutor is flakey
- [OOZIE-2490](#) - Oozie can't set hadoop.security.token.service.use_ip
- [SENTRY-748](#) - Improve test coverage of Sentry + Hive using complex views
- [SENTRY-835](#) - Drop table leaves a connection open when using metastorelistener
- [SENTRY-922](#) - INSERT OVERWRITE DIRECTORY permission not working correctly
- [SENTRY-972](#) - Include sentry-tests-hive hadoop test script in maven project
- [SENTRY-991](#) - Roles of Sentry Permission needs to be case insensitive
- [SENTRY-994](#) - SentryAuthorizationInfoX should override isSentryManaged
- [SENTRY-1002](#) - PathsUpdate.parsePath(path) will throw an NPE when parsing relative paths
- [SENTRY-1003](#) - Support "reload" by updating the classpath of Sentry function aux jar path during runtime
- [SENTRY-1007](#) - Sentry column-level performance for wide tables
- [SENTRY-1008](#) - Path should be not be updated if the create/drop table/partition event fails
- [SENTRY-1015](#) - Improve Sentry + Hive error message when user has insufficient privileges
- [SENTRY-1044](#) - Tables with non-hdfs locations breaks HMS startup
- [SENTRY-1169](#) - MetastorePlugin#renameAuthzObject log message prints oldpathname as newpathname
- [SENTRY-1184](#) - Clean up HMSPaths.renameAuthzObject
- [SOLR-6631](#) - DistributedQueue spinning on calling zookeeper getChildren()
- [SOLR-6820](#) - The sync on the VersionInfo bucket in DistributedUpdateProcesser#addDocument appears to be a large bottleneck when using replication
- [SOLR-7281](#) - Add an overseer action to publish an entire node as 'down'
- [SOLR-7332](#) - Seed version buckets with max version from index
- [SOLR-7493](#) - Requests aren't distributed evenly if the collection isn't present locally. Merges r1683946 and r1683948 from trunk
- [SOLR-7587](#) - TestSpellCheckResponse stalled and never timed out -- possible VersionBucket bug?
- [SOLR-7625](#) - Version bucket seed not updated after new index is installed on a replica
- [SOLR-8215](#) - Only active replicas should handle incoming requests against a collection
- [SOLR-8371](#) - Try and prevent too many recovery requests from stacking up and clean up some faulty cancel recovery logic
- [SOLR-8451](#) - We should not call method.abort in HttpSolrClient or HttpSolrCall#remoteQuery and HttpSolrCall#remoteQuery should not close streams
- [SOLR-8453](#) - Solr should attempt to consume the request inputstream on errors as we cannot count on the container to do it
- [SOLR-8575](#) - Fix HDFSLogReader replay status numbers and a performance bug where we can reopen FSDataInputStream too often
- [SOLR-8578](#) - Successful or not, requests are not always fully consumed by Solrj clients and we count on HttpClient or the JVM
- [SOLR-8615](#) - Just like creating cores, we should use multiple threads when closing cores
- [SOLR-8633](#) - DistributedUpdateProcess processCommit/deleteByQuery calls finish on DUP and SolrCmdDistributor, which violates the lifecycle and can cause bugs

CDH 5 Release Notes

- [SOLR-8683](#) - Always consume the full request on the server, not just in the case of an error, tune down stream closed logging
- [SOLR-8720](#) - ZkController#publishAndWaitForDownStates should use #publishNodeAsDown
- [SOLR-8771](#) - Multi-threaded core shutdown creates executor per core
- [SOLR-8855](#) - The HDFS BlockDirectory should not clean up its cache on shutdown
- [SOLR-8856](#) - Do not cache merge or 'read once' contexts in the hdfs block cache
- [SOLR-8857](#) - HdfsUpdateLog does not use configured or new default number of version buckets and is hard coded to 256
- [SOLR-8869](#) - Optionally disable printing field cache entries in SolrFieldCacheMBean
- [SPARK-10859](#) - Predicates pushed to InmemoryColumnarTableScan are not evaluated correctly
- [SPARK-10914](#) - UnsafeRow serialization breaks when two machines have different Oops size
- [SPARK-11009](#) - RowNumber in HiveContext returns negative values in cluster mode
- [SPARK-11442](#) - Reduce numSlices for local metrics test of SparkListenerSuite
- [SPARK-12617](#) - Socket descriptor leak killing streaming app
- [SPARK-14477](#) - Allow custom mirrors for downloading artifacts in build/mvn
- [SQOOP-2847](#) - Sqoop --incremental + missing parent --target-dir reports success with no data

Issues Fixed in CDH 5.6.0

CDH 5.6.0 is a minor release that provides EMC DSSD D5 Storage Appliance Integration for Hadoop DataNodes for HDFS. It also fixes the following issues.

Apache Avro

Concurrent schema parsing can result in processes hanging because of an unsafe shared cache

Bug: [AVRO-1781](#)

Severity: High

Workaround: Use a global lock around all calls to `Schema.parse` and `Schema.Parser#parse` to ensure only one thread is parsing at a time.

Apache HBase

Values of some metrics may appear to be negative.

Bug: [HBASE-12961](#)

Some metric value are stored in integers, and cannot accommodate real-world values. This causes metric values to appear to be negative.

Workaround: None.

The HBase Shell cannot handle Scan filters which contain non-UTF8 characters.

Bug: [HBASE-15032](#)

The HBase Shell incorrectly handles filter strings which contain non-UTF8 characters.

Workaround: None.

HDFS

Checkpointing can fail due to an InvalidSignatureException in a secure cluster

Bug: [HDFS-7798](#)

Severity: High

Workaround: This problem occurs occasionally due to race condition. The error is transient, and a subsequent checkpoint may still succeed.

Issues Fixed in CDH 5.5.x

The following topics describe issues fixed in CDH 5.5.x, from newest to oldest release. You can also review [What's New In CDH 5.5.x](#) on page 23 or [Known Issues in CDH 5](#) on page 111.

Issues Fixed in CDH 5.5.5

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.5.5:

- [FLUME-2821](#) - KafkaSourceUtil Can Log Passwords at Info remove logging of security related data in older releases
- [FLUME-2913](#) - Don't strip SLF4J from imported classpaths
- [FLUME-2918](#) - Speed up TaildirSource on directories with many files
- [HADOOP-8436](#) - NPE In getLocalPathForWrite (path, conf) when the required context item is not configured
- [HADOOP-8437](#) - getLocalPathForWrite should throw IOException for invalid paths
- [HADOOP-8751](#) - NPE in Token.toString() when Token is constructed using null identifier
- [HADOOP-8934](#) - Shell command ls should include sort options
- [HADOOP-10048](#) - LocalDirAllocator should avoid holding locks while accessing the filesystem
- [HADOOP-10971](#) - Add -C flag to make `hadoop fs -ls` print filenames only
- [HADOOP-11901](#) - BytesWritable fails to support 2G chunks due to integer overflow
- [HADOOP-11984](#) - Enable parallel JUnit tests in pre-commit
- [HADOOP-12252](#) - LocalDirAllocator should not throw NPE with empty string configuration
- [HADOOP-12259](#) - Utility to Dynamic port allocation
- [HADOOP-12659](#) - Incorrect usage of config parameters in token manager of KMS
- [HADOOP-12787](#) - KMS SPNEGO sequence does not work with WEBHDFS
- [HADOOP-12841](#) - Update s3-related properties in core-default.xml.
- [HADOOP-12901](#) - Add warning log when KMSClientProvider cannot create a connection to the KMS server.
- [HADOOP-12963](#) - Allow using path style addressing for accessing the s3 endpoint.
- [HADOOP-13079](#) - Add -q option to ls to print ? instead of non-printable characters
- [HADOOP-13132](#) - Handle ClassCastException on AuthenticationException in LoadBalancingKMSClientProvider
- [HADOOP-13155](#) - Implement TokenRenewer to renew and cancel delegation tokens in KMS
- [HADOOP-13251](#) - Authenticate with Kerberos credentials when renewing KMS delegation token
- [HADOOP-13255](#) - KMSClientProvider should check and renew tgt when doing delegation token operations
- [HADOOP-13263](#) - Reload cached groups in background after expiry.
- [HADOOP-13457](#) - Remove hardcoded absolute path for shell executable.
- [HDFS-6434](#) - Default permission for creating file should be 644 for WebHdfs/HttpFS
- [HDFS-7597](#) - DelegationTokenIdentifier should cache the TokenIdentifier to UGI mapping
- [HDFS-8008](#) - Support client-side back off when the datanodes are congested
- [HDFS-8581](#) - ContentSummary on / skips further counts on yielding lock
- [HDFS-8829](#) - Make SO_RCVBUF and SO_SNDBUF size configurable for DataTransferProtocol sockets and allow configuring auto-tuning
- [HDFS-8897](#) - Balancer should handle fs.defaultFS trailing slash in HA
- [HDFS-9085](#) - Show renewer information in DelegationTokenIdentifier#toString
- [HDFS-9259](#) - Make SO_SNDBUF size configurable at DFSClient side for hdfs write scenario.
- [HDFS-9276](#) - Failed to Update HDFS Delegation Token for long running application in HA mode
- [HDFS-9365](#) - Balancer does not work with the HDFS-6376 HA setup.
- [HDFS-9405](#) - Warmup NameNode EDEK caches in background thread
- [HDFS-9466](#) - TestShortCircuitCache#testDataXceiverCleansUpSlotsOnFailure is flaky
- [HDFS-9700](#) - DFSClient and DFSOutputStream should set TCP_NODELAY on sockets for DataTransferProtocol
- [HDFS-9732](#) - Improve DelegationTokenIdentifier.toString() for better logging
- [HDFS-9805](#) - Add server-side configuration for enabling TCP_NODELAY for DataTransferProtocol and default it to true
- [HDFS-9939](#) - Increase DecompressorStream skip buffer size

- [HDFS-10360](#) - DataNode may format directory and lose blocks if current/VERSION is missing.
- [HDFS-10381](#) - DataStreamer DataNode exclusion log message should be warning.
- [HDFS-10396](#) - Using -diff option with DistCp may get "Comparison method violates its general contract" exception
- [HDFS-10481](#) - HTTPFS server should correctly impersonate as end user to open file
- [HDFS-10512](#) - VolumeScanner may terminate due to NPE in DataNode.reportBadBlocks
- [HDFS-10516](#) - Fix bug when warming up EDEK cache of more than one encryption zone
- [HDFS-10544](#) - Balancer doesn't work with IPFailoverProxyProvider.
- [HDFS-10643](#) - Namenode should use loginUser(hdfs) to generateEncryptedKey
- [MAPREDUCE-6442](#) - Stack trace is missing when error occurs in client protocol provider's constructor
- [MAPREDUCE-6473](#) - Job submission can take a long time during Cluster initialization
- [MAPREDUCE-6577](#) - MR AM unable to load native library without MR_AM_ADMIN_USER_ENV set
- [YARN-2605](#) - [RM HA] Rest api endpoints doing redirect incorrectly.
- [YARN-3055](#) - Fixed ResourceManager's DelegationTokenRenewer to not stop token renewal of applications part of a bigger workflow
- [YARN-3104](#) - Fixed RM to not generate new AMRM tokens on every heartbeat between rolling and activation
- [YARN-3832](#) - Resource Localization fails on a cluster due to existing cache directories
- [YARN-4459](#) - container-executor should only kill process groups
- [YARN-4784](#) - Fairscheduler: defaultQueueSchedulingPolicy should not accept FIFO.
- [YARN-5048](#) - DelegationTokenRenewer#skipTokenRenewal may throw NPE
- [YARN-5272](#) - Handle queue names consistently in FairScheduler.
- [HBASE-11625](#) - Verifies data before building HFileBlock.
- [HBASE-14155](#) - StackOverflowError in reverse scan
- [HBASE-14644](#) - Region in transition metric is broken
- [HBASE-14730](#) - Region server needs to log warnings when there are attributes configured for cells with hfile v2
- [HBASE-15439](#) - getMaximunAllowedTimeBetweenRuns in ScheduledChore ignores the TimeUnit
- [HBASE-15496](#) - Throw RowTooBigException only for user scan/get
- [HBASE-15707](#) - ImportTSV bulk output does not support tags with hfile.format.version=3
- [HBASE-15746](#) - Remove extra RegionCoprocessor preClose() in RSRpcServices#closeRegion
- [HBASE-15791](#) - Improve javadoc around ScheduledChore
- [HBASE-15811](#) - Batch Get after batch Put does not fetch all Cells
- [HBASE-15925](#) - Provide default values for hadoop compat module related properties that match default hadoop profile.
- [HBASE-16207](#) - Can't restore snapshot without "Admin" permission
- [HBASE-16288](#) - Revert "HFile intermediate block level indexes might recurse forever creating multi TB files"
- [HBASE-16288](#) - HFile intermediate block level indexes might recurse forever creating multi TB files
- [HIVE-7443](#) - Fix HiveConnection to communicate with Kerberized Hive JDBC server and alternative JDks
- [HIVE-9499](#) - hive.limit.query.max.table.partition makes queries fail on non-partitioned tables
- [HIVE-10685](#) - Alter table concatenate oparetor will cause duplicate data
- [HIVE-10925](#) - Non-static threadlocals in metastore code can potentially cause memory leak
- [HIVE-11031](#) - ORC concatenation of old files can fail while merging column statistics
- [HIVE-11243](#) - Changing log level in Utilities.getBaseWork
- [HIVE-11369](#) - Mapjoins in HiveServer2 fail when jmxremote is used
- [HIVE-11408](#) - HiveServer2 is leaking ClassLoaders when add jar / temporary functions are used due to constructor caching in Hadoop ReflectionUtils
- [HIVE-11427](#) - Location of temporary table for CREATE TABLE SELECT broken by HIVE-7079.
- [HIVE-11747](#) - Unnecessary error log is shown when executing a "INSERT OVERWRITE LOCAL DIRECTORY" cmd in the embedded mode
- [HIVE-11827](#) - STORED AS AVRO fails SELECT COUNT(*) when empty
- [HIVE-12481](#) - Occasionally "Request is a replay" will be thrown from HS2
- [HIVE-12635](#) - Hive should return the latest hbase cell timestamp as the row timestamp value

- [HIVE-12958](#) - Make embedded Jetty server more configurable
- [HIVE-13285](#) - Orc concatenation may drop old files from moving to final path
- [HIVE-13462](#) - HiveResultSetMetaData.getPrecision() fails for NULL columns
- [HIVE-13527](#) - Using deprecated APIs in HBase client causes zookeeper connection leaks
- [HIVE-13570](#) - Some queries with Union all fail when CBO is off
- [HIVE-13590](#) - Kerberized HS2 with LDAP auth enabled fails in multi-domain LDAP case
- [HIVE-13704](#) - Don't call DistCp.execute() instead of DistCp.run()
- [HIVE-13736](#) - View's input/output formats are TEXT by default.
- [HIVE-13932](#) - Hive SMB Map Join with small set of LIMIT failed with NPE
- [HIVE-13953](#) - Issues in HiveLockObject.equals method
- [HIVE-13991](#) - Union All on view fail with no valid permission on underneath table
- [HIVE-14006](#) - Hive query with UNION ALL fails with ArrayIndexOutOfBoundsException.
- [HIVE-14118](#) - Make the alter partition exception more meaningful
- [HUE-3520](#) - [jb] Fix backport error
- [HUE-3520](#) - [jb] Use impersonation to access JHS if security is enabled
- [HUE-3637](#) - [sqoop] Avoid decode errors on attribute values
- [HUE-3650](#) - [beeswax] Notify of caught errors in the watch logs process
- [HUE-3651](#) - [core] Upgrade Moment.js
- [HUE-3716](#) - [core] Add gen-py paths to hue.pth
- [HUE-3861](#) - [core] Upgrade Django Axes to 1.5
- [HUE-3866](#) - [core] Hue CPU reaches ~100% usage while uploading files with SSL to HTTPFS/WebHDFS
- [HUE-3880](#) - [core] Add importlib directly for Python 2.6
- [HUE-4005](#) - [oozie] Remove oozie.coord.application.path from properties when rerunning workflow
- [HUE-4006](#) - [oozie] Create new deployment directory when coordinator or bundle is copied
- [HUE-4007](#) - [oozie] Fix deployment_dir for the bundle in oozie example fixtures
- [HUE-4023](#) - [useradmin] update AuthenticationForm to allow activated users to login
- [HUE-4087](#) - [jobbrowser] Unable to kill jobs with Resource Manager HA enabled
- [HUE-4202](#) - [jb] Enable offset param for fetching jobbrowser logs
- [HUE-4215](#) - [yarn] Reset API_CACHE on logout
- [HUE-4227](#) - [yarn] Fix unittest for MR API Cache
- [HUE-4238](#) - [doc2] Ignore history docs in find_jobs_with_no_doc during sync documents
- [HUE-4252](#) - [core] Handle 307 redirect from YARN upon standby failover
- [HUE-4258](#) - [jb] Close and pool Spark History Server connections
- [HUE-4333](#) - [core] Properly reset API_CACHE on failover
- [HUE-4493](#) - [oozie] Fix sync-workflow action when Workflow includes sub-workflow
- [HUE-4515](#) - [oozie] Remove oozie.bundle.application.path from properties when rerunning workflow
- [OOZIE-2314](#) - Unable to kill old instance child job by workflow or coord rerun by Launcher
- [OOZIE-2329](#) - Make handling yarn restarts configurable
- [OOZIE-2330](#) - Spark action should take the global jobTracker and nameNode configs by default and allow file and archive elements
- [OOZIE-2345](#) - Parallel job submission for forked actions
- [OOZIE-2391](#) - spark-opts value in workflow.xml is not parsed properly
- [OOZIE-2436](#) - Fork/join workflow fails with oozie.action.yarn.tag must not be null
- [OOZIE-2481](#) - Add YARN_CONF_DIR in the Shell action
- [OOZIE-2504](#) - Create a log4j.properties under HADOOP_CONF_DIR in Shell Action
- [OOZIE-2511](#) - SubWorkflow missing variable set from option if config-default is present in parent workflow
- [OOZIE-2533](#) - Patch-1550 - workaround
- [OOZIE-2537](#) - SqoopMain does not set up log4j properly
- [SENTRY-1190](#) - IMPORT TABLE silently fails if Sentry is enabled
- [SENTRY-1201](#) - Sentry ignores database prefix for MSCK statement

CDH 5 Release Notes

- [SENTRY-1252](#) - grantServerPrivilege and revokeServerPrivilege should treat "*" and "ALL" as synonyms when action is not explicitly specified
- [SENTRY-1265](#) - Sentry service should not require a TGT as it is not talking to other kerberos services as a client
- [SENTRY-1292](#) - Reorder DBModelAction EnumSet
- [SENTRY-1293](#) - Avoid converting string permission to Privilege object
- [SOLR-6631](#) - DistributedQueue spinning on calling zookeeper getChildren()
- [SOLR-6879](#) - Have an option to disable autoAddReplicas temporarily for all collections.
- [SOLR-7178](#) - OverseerAutoReplicaFailoverThread compares Integer objects using ==
- [SOLR-8451](#) - Fix backport
- [SOLR-8497](#) - Merge indexes should mark its directories as done rather than keep them around in the directory cache.
- [SOLR-8551](#) - Make collection deletion more robust.
- [SOLR-8683](#) - Tune down stream closed logging
- [SOLR-9236](#) - AutoAddReplicas will append an extra /tlog to the update log location on replica failover.
- [SPARK-10577](#) - [PYSPARK] DataFrame hint for broadcast join
- [SPARK-11442](#) - Reduce numSlices for local metrics test of SparkListenerSuite
- [SPARK-12087](#) - [STREAMING] Create new JobConf for every batch in saveAsHadoopFiles
- [SQOOP-2846](#) - Sqoop Export with update-key failing for avro data file

Issues Fixed in CDH 5.5.4

CDH 5.5.4 fixes the following issues.

Apache Hadoop

FSImage may get corrupted after deleting snapshot

Bug: [HDFS-9406](#)

When deleting a snapshot that contains the last record of a given INode, the fsimage may become corrupt because the create list of the snapshot diff in the previous snapshot and the child list of the parent `INodeDirectory` are not cleaned.

Apache HBase

The ReplicationCleaner process can abort if its connection to ZooKeeper is inconsistent

Bug: [HBASE-15234](#)

If the connection with ZooKeeper is inconsistent, the ReplicationCleaner may abort, and the following event is logged by the HMaster:

```
WARN org.apache.hadoop.hbase.replication.master.ReplicationLogCleaner: Aborting
ReplicationLogCleaner
because Failed to get list of replicators
```

Unprocessed WALs accumulate.

The `seekBefore()` method calculates the size of the previous data block by assuming that data blocks are contiguous, and HFile v2 and higher store Bloom blocks and leaf-level INode blocks with the data. As a result, reverse scans do not work when Bloom blocks or leaf-level INode blocks are present when HFile v2 or higher is used.

Workaround: Restart the HMaster occasionally. The ReplicationCleaner restarts if necessary and process the unprocessed WALs.

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.5.4:

- [FLUME-2632](#) - High CPU on KafkaSink
- [FLUME-2712](#) - Optional channel errors slows down the Source to Main channel event rate
- [FLUME-2781](#) - Kafka Channel with `parseAsFlumeEvent=true` should write data as is, not as flume events
- [FLUME-2886](#) - Optional Channels can cause OOMs

- [FLUME-2891](#) - Revert FLUME-2712 and FLUME-2886
- [FLUME-2897](#) - AsyncHBase sink NPE when Channel.getTransaction() fails
- [HADOOP-7139](#) - Allow appending to existing SequenceFiles
- [HADOOP-7817](#) - RawLocalFileSystem.append() should give FSDataOutputStream with accurate .getPos()
- [HADOOP-11321](#) - copyToLocal cannot save a file to an SMB share unless the user has Full Control permissions
- [HADOOP-11687](#) - Ignore x-* and response headers when copying an Amazon S3 object
- [HADOOP-11722](#) - Some Instances of Services using ZKDelegationTokenSecretManager go down when old token cannot be deleted
- [HADOOP-12240](#) - Fix tests requiring native library to be skipped in non-native profile
- [HADOOP-12280](#) - Skip unit tests based on maven profile rather than NativeCodeLoader.isNativeCodeLoaded
- [HADOOP-12559](#) - KMS connection failures should trigger TGT renewal
- [HADOOP-12605](#) - Fix intermittent failure of TestIPC.testIpcWithReaderQueuing
- [HADOOP-12668](#) - Support excluding weak Ciphers in HttpServer2 through ssl-server.conf
- [HADOOP-12682](#) - Fix TestKMS#testKMSRestart* failure
- [HADOOP-12699](#) - TestKMS#testKMSProvider intermittently fails during 'test rollover draining'
- [HADOOP-12715](#) - TestValueQueue#testGetAtMostPolicyALL fails intermittently
- [HADOOP-12718](#) - Incorrect error message by fs -put local dir without permission
- [HADOOP-12736](#) - TestTimedOutTestsListener#testThreadDumpAndDeadlocks sometimes times out
- [HADOOP-12788](#) - OpensslAesCtrCryptoCodec should log which random number generator is used
- [HADOOP-12825](#) - Log slow name resolutions
- [HADOOP-12954](#) - Add a way to change hadoop.security.token.service.use_ip
- [HADOOP-12972](#) - Lz4Compressor#getLibraryName returns the wrong version number
- [HDFS-6520](#) - hdfs fsck passes invalid length value when creating BlockReader
- [HDFS-7373](#) - Clean up temporary files after fsimage transfer failures
- [HDFS-7758](#) - Retire FsDatasetSpi#getVolumes() and use FsDatasetSpi#getVolumeRefs() instead
- [HDFS-8211](#) - DataNode UUID is always null in the JMX counter
- [HDFS-8496](#) - Calling stopWriter() with FSDatasetImpl lock held may block other threads
- [HDFS-8576](#) - Lease recovery should return true if the lease can be released and the file can be closed
- [HDFS-8785](#) - TestDistributedFileSystem is failing in trunk
- [HDFS-8855](#) - Webhdfs client leaks active NameNode connections
- [HDFS-9264](#) - Minor cleanup of operations on FsVolumeList#volumes
- [HDFS-9289](#) - Make DataStreamer#block thread safe and verify genStamp in commitBlock
- [HDFS-9347](#) - Invariant assumption in TestQuorumJournalManager.shutdown() is wrong
- [HDFS-9350](#) - Avoid creating temporary strings in Block.toString() and getBlockName()
- [HDFS-9358](#) - TestNodeCount#testNodeCount timed out
- [HDFS-9406](#) - FSImage may get corrupted after deleting snapshot
- [HDFS-9514](#) - TestDistributedFileSystem.testDFSClientPeerWriteTimeout failing; exception being swallowed
- [HDFS-9576](#) - HTrace: collect position/length information on read operations
- [HDFS-9589](#) - Block files which have been hardlinked should be duplicated before the DataNode appends to them
- [HDFS-9612](#) - DistCp worker threads are not terminated after jobs are done
- [HDFS-9655](#) - NN should start JVM pause monitor before loading fsimage.
- [HDFS-9688](#) - Test the effect of nested encryption zones in HDFS downgrade
- [HDFS-9701](#) - DN may deadlock when hot-swapping under load
- [HDFS-9721](#) - Allow Delimited PB OIV tool to run upon fsimage that contains INodeReference
- [HDFS-9949](#) - Add a test case to ensure that the DataNode does not regenerate its UUID when a storage directory is cleared
- [HDFS-10223](#) - peerFromSocketAndKey performs SASL exchange before setting connection timeouts
- [HDFS-10267](#) - Extra "synchronized" on FsDatasetImpl#recoverAppend and FsDatasetImpl#recoverClose
- [MAPREDUCE-4785](#) - TestMRApp occasionally fails

CDH 5 Release Notes

- [MAPREDUCE-6460](#) - TestRMContainerAllocator.testAttemptNotFoundCausesRMCommunicatorException fails
- [MAPREDUCE-6528](#) - Memory leak for HistoryFileManager.getJobSummary()
- [MAPREDUCE-6580](#) - Test failure: TestMRJobsWithProfiler
- [MAPREDUCE-6620](#) - Jobs that did not start are shown as starting in 1969 in the JHS web UI
- [YARN-2749](#) - Fix some testcases from TestLogAggregationService fails in trunk
- [YARN-2871](#) - TestRMRestart#testRMRestartGetApplicationList sometime fails in trunk
- [YARN-2902](#) - Killing a container that is localizing can orphan resources in the DOWNLOADING state
- [YARN-3446](#) - FairScheduler headroom calculation should exclude nodes in the blacklist
- [YARN-3727](#) - For better error recovery, check if the directory exists before using it for localization
- [YARN-4155](#) - TestLogAggregationService.testLogAggregationServiceWithInterval failing
- [YARN-4168](#) - Fixed a failing test TestLogAggregationService.testLocalFileDeletionOnDiskFull
- [YARN-4354](#) - Public resource localization fails with NPE
- [YARN-4380](#) - TestResourceLocalizationService.testDownloadingResourcesOnContainerKill fails intermittently
- [YARN-4393](#) - Fix intermittent test failure for TestResourceLocalizationService#testFailedDirsResourceRelease
- [YARN-4546](#) - ResourceManager crash due to scheduling opportunity overflow
- [YARN-4573](#) - Fix test failure in TestRMAppTransitions#testAppRunningKill and testAppKilledKilled
- [YARN-4613](#) - Fix test failure in TestClientRMServer#testGetClusterNodes
- [YARN-4704](#) - TestResourceManager#testResourceAllocation() fails when using FairScheduler
- [YARN-4717](#) - TestResourceLocalizationService.testPublicResourceInitializesLocalDir fails Intermittently due to IllegalArgumentException from cleanup
- [HBASE-6617](#) - ReplicationSourceManager should be able to track multiple WAL paths (ADDENDUM)
- [HBASE-14586](#) - Use a maven profile to run Jacoco analysis
- [HBASE-14587](#) - Attach a test-sources.jar for hbase-server
- [HBASE-14588](#) - Stop accessing test resources from within src folder
- [HBASE-14759](#) - Avoid using Math.abs when selecting SyncRunner in FSHLog
- [HBASE-15019](#) - Replication stuck when HDFS is restarted
- [HBASE-15052](#) - Use EnvironmentEdgeManager in ReplicationSource
- [HBASE-15152](#) - Automatically include prefix-tree module in MR jobs if present
- [HBASE-15157](#) - Add *PerformanceTest for Append, CheckAnd*
- [HBASE-15206](#) - Fix flaky testSplitDaughtersNotInMeta
- [HBASE-15213](#) - Fix increment performance regression caused by HBASE-8763 on branch-1.0
- [HBASE-15234](#) - Don't abort ReplicationLogCleaner on ZooKeeper errors
- [HBASE-15456](#) - CreateTableProcedure/ModifyTableProcedure needs to fail when there is no family in table descriptor
- [HBASE-15479](#) - No more garbage or beware of autoboxing
- [HBASE-15582](#) - SnapshotManifestV1 too verbose when there are no regions
- [HIVE-9617](#) - UDF from _utc_timestamp throws NPE if the second argument is null
- [HIVE-9743](#) - Revert "(Tests portion only)Incorrect result set for vectorized left outer join (Matt McCline, reviewed by Vikram Dixit)"
- [HIVE-10115](#) - HS2 running on a Kerberized cluster should offer Kerberos(GSSAPI) and Delegation token(DIGEST) when alternate authentication is enabled
- [HIVE-10213](#) - MapReduce jobs using dynamic-partitioning fail on commit
- [HIVE-10303](#) - HIVE-9471 broke forward compatibility of ORC files
- [HIVE-11054](#) - Handle varchar/char partition columns in vectorization
- [HIVE-11097](#) - HiveInputFormat uses String.startsWith to compare splitPath and PathToAliases
- [HIVE-11135](#) - Fix the Beeline set and save command in order to avoid the NullPointerException
- [HIVE-11285](#) - ObjectInspector for partition columns in FetchOperator in SMBJoin causes exception
- [HIVE-11488](#) - Need to add support for sessionId and queryId logging, QueryId can't be stored in the configuration of the SessionState since multiple queries can run in a single session
- [HIVE-11583](#) - When PTF is used over a large partitions result could be corrupted

- [HIVE-11590](#) - AvroDeserializer is very chatty
- [HIVE-11828](#) - beeline -f fails on scripts with tabs between column type and comment
- [HIVE-11866](#) - Add framework to enable testing using LDAPServer using LDAP protocol
- [HIVE-11919](#) - Hive Union Type Mismatch
- [HIVE-12315](#) - Fix Vectorized double divide by zero
- [HIVE-12354](#) - MapJoin with double keys is slow on MR
- [HIVE-12431](#) - Support timeout for compile lock
- [HIVE-12506](#) - SHOW CREATE TABLE command creates a table that does not work for RCFile format
- [HIVE-12706](#) - Incorrect output from from_utc_timestamp()/to_utc_timestamp when local timezone has DST
- [HIVE-12782](#) - Update the golden files for some tests that fail
- [HIVE-12790](#) - Metastore connection leaks in HiveServer2
- [HIVE-12885](#) - LDAP Authenticator improvements
- [HIVE-12909](#) - Some encryption q-tests fail because trash is disabled in encryption_with_trash.q
- [HIVE-12941](#) - Unexpected result when using MIN() on struct with NULL in first field
- [HIVE-12946](#) - Alter table should also add default scheme and authority for the location similar to create table
- [HIVE-13039](#) - BETWEEN predicate is not functioning correctly with predicate pushdown on Parquet table
- [HIVE-13055](#) - Add unit tests for HIVE-11512
- [HIVE-13065](#) - Hive throws NPE when writing map type data to a HBase backed table
- [HIVE-13082](#) - Enable constant propagation optimization in query with left semi join
- [HIVE-13200](#) - Aggregation functions returning empty rows on partitioned columns
- [HIVE-13243](#) - Hive drop table on encryption zone fails for external tables
- [HIVE-13251](#) - Hive can't read the decimal in AVRO file generated from previous version
- [HIVE-13286](#) - Query ID is being reused across queries
- [HIVE-13295](#) - Improvement to LDAP search queries in HS2 LDAP Authenticator
- [HIVE-13401](#) - Kerberized HS2 with LDAP auth enabled fails kerberos/delegation token authentication
- [HUE-3106](#) - [filebrowser] Add support for full paths in zip file uploads
- [HUE-3110](#) - [oozie] Fix bundle submission when coordinator points to multiple bundles
- [HUE-3132](#) - [core] Fix Sync Ldap users and groups for anonymous binds
- [HUE-3180](#) - [useradmin] Override duplicate username validation message
- [HUE-3185](#) - [oozie] Avoid extra API calls for parent information in workflow dashboard
- [HUE-3303](#) - [core] PostgreSQL requires data update and alter table operations in separate transactions
- [HUE-3310](#) - [jobs] Prevent browsing job designs by API
- [HUE-3334](#) - [editor] Skip checking for multi queries if there is no semi colon, send empty query instead of error
- [HUE-3398](#) - [beeswax] Filter out sessions with empty guid or secret key
- [HUE-3436](#) - [oozie] Retain old dependencies when saving a workflow
- [HUE-3437](#) - [core] PamBackend does not honor ignore_username_case
- [HUE-3523](#) - [oozie] Modify find_jobs_with_no_doc method to exclude jobs with no name
- [HUE-3528](#) - [oozie] Call correct metrics api to avoid 500 error
- [HUE-3594](#) - [fb] Smarter DOM based XSS filter on hashes
- [IMPALA-852](#) - ,IMPALA-2215: Analyze HAVING clause before aggregation
- [IMPALA-1092](#) - Fix estimates for trivial coord-only queries
- [IMPALA-1170](#) - Fix URL parsing when path contains '@'
- [IMPALA-1934](#) - Allow shell to retrieve LDAP password from shell cmd
- [IMPALA-2093](#) - Disallow NOT IN aggregate subqueries with a constant lhs expr
- [IMPALA-2184](#) - Don't inline timestamp methods with try/catch blocks in IR
- [IMPALA-2425](#) - Broadcast join hint not enforced when low memory limit is set
- [IMPALA-2503](#) - Add missing String.format() arg in error message
- [IMPALA-2539](#) - Unmark collections slots of empty union operands
- [IMPALA-2554](#) - Change default buffer size for RPC servers and clients
- [IMPALA-2565](#) - Planner tests are flaky due to file size mismatches

CDH 5 Release Notes

- [IMPALA-2592](#) - DataStreamSender::Channel::CloseInternal() does not close the channel on an error
- [IMPALA-2599](#) - Pseudo-random sleep before acquiring kerberos ticket possibly not really pseudo-random
- [IMPALA-2711](#) - Fix memory leak in Rand()
- [IMPALA-2719](#) - test_parquet_max_page_header fails on Isilon
- [IMPALA-2732](#) - Timestamp formats with non-padded values
- [IMPALA-2734](#) - Correlated EXISTS subqueries with HAVING clause return wrong results
- [IMPALA-2742](#) - Avoid unbounded MemPool growth with AcquireData()
- [IMPALA-2749](#) - Fix decimal multiplication overflow
- [IMPALA-2765](#) - Preserve return type of subexpressions substituted in isTrueWithNullSlots()
- [IMPALA-2788](#) - conv(bigint num, int from_base, int to_base) returns wrong result
- [IMPALA-2798](#) - Bring in AVRO-1617 fix and add test case for it
- [IMPALA-2818](#) - Fix cancellation crashes/hangs due to BlockOnWait() race
- [IMPALA-2820](#) - Support unquoted keywords as struct-field names
- [IMPALA-2832](#) - Fix cloning of FunctionCallExpr
- [IMPALA-2844](#) - Allow count(*) on RC files with complex types
- [IMPALA-2870](#) - Fix failing metadata.test_ddl.TestDdlStatements.test_create_table test
- [IMPALA-2894](#) - Move regression test into a different .test file
- [IMPALA-2906](#) - Fix an edge case with materializing TupleIsNotNullPredicates in analytic sorts
- [IMPALA-2914](#) - Fix DCHECK Check failed: HasDateOrTime()
- [IMPALA-2926](#) - Fix off-by-one bug in SelectNode::CopyRows()
- [IMPALA-2940](#) - Fix leak of dictionaries in Parquet scanner
- [IMPALA-3000](#) - Fix BitReader::Reset()
- [IMPALA-3034](#) - Verify all consumed memory of a MemTracker is always released at destruction time
- [IMPALA-3047](#) - Separate create table test with nested types
- [IMPALA-3054](#) - Disable probe side filters when spilling
- [IMPALA-3071](#) - Fix assignment of On-clause predicates belonging to an inner join
- [IMPALA-3085](#) - Unregister data sinks' MemTrackers at their Close() functions
- [IMPALA-3093](#) - ReopenClient() could NULL out 'client_key' causing a crash
- [IMPALA-3095](#) - Add configurable whitelist of authorized internal principals
- [IMPALA-3151](#) - Impala crash for avro table when casting to char data type
- [IMPALA-3194](#) - Allow queries materializing scalar type columns in RC/sequence files
- [KITE-1114](#) - Kite CLI json-import HDFS temp file path not multiuser safe, fix missing license header
- [OOZIE-2419](#) - HBase credentials are not correctly proxied
- [OOZIE-2428](#) - TestSLAService, TestSLAEventGeneration flaky tests
- [OOZIE-2429](#) - TestEventGeneration test is flaky
- [OOZIE-2432](#) - TestPurgeXCommand fails
- [OOZIE-2435](#) - TestCoordChangeXCommand is flaky
- [OOZIE-2466](#) - Repeated failure of TestMetricsInstrumentation.testSamplers
- [OOZIE-2486](#) - TestSLAEventsGetForFilterJPAExecutor is flaky
- [OOZIE-2490](#) - Oozie can't set hadoop.security.token.service.use_ip
- [SENTRY-922](#) - BackportINSERT OVERWRITE DIRECTORY permission not working correctly
- [SENTRY-972](#) - backportInclude sentry-tests-hive hadoop test script in maven project
- [SENTRY-991](#) - backportRoles of Sentry Permission needs to be case insensitive
- [SENTRY-1002](#) - PathsUpdate.parsePath(path) will throw an NPE when parsing relative paths
- [SENTRY-1003](#) - Support "reload" by updating the classpath of Sentry function aux jar path during runtime
- [SENTRY-1007](#) - backportSentry column-level performance for wide tables
- [SENTRY-1008](#) - Path should be not be updated if the create/drop table/partition event fails
- [SENTRY-1015](#) - backportImprove Sentry + Hive error message when user has insufficient privileges
- [SENTRY-1044](#) - Tables with non-hdfs locations breaks HMS startup
- [SENTRY-1169](#) - MetastorePlugin#renameAuthzObject log message prints oldpathname as newpathname

- [SENTRY-1184](#) - Clean up HMSPaths.renameAuthzObject
- [SOLR-6820](#) - Make the number of version buckets used by the UpdateLog configurable as increasing beyond the default 256 has been shown to help with high volume indexing performance in SolrCloudIncrease the default number of buckets to 65536 instead of 256, fix numVersionBuckets name attribute in configsets
- [SOLR-7281](#) - Add an overseer action to publish an entire node as 'down'
- [SOLR-7332](#) - Initialize the highest value for all version buckets with the max value from the index or recent updates to avoid unnecessary lookups to the index to check for reordered updates when processing new documents
- [SOLR-7493](#) - Requests aren't distributed evenly if the collection isn't present locally. Merges r1683946 and r1683948 from trunk
- [SOLR-7587](#) - TestSpellCheckResponse stalled and never timed out -- possible VersionBucket bug?
- [SOLR-7625](#) - Version bucket seed not updated after new index is installed on a replica
- [SOLR-8215](#) - Only active replicas should handle incoming requests against a collection
- [SOLR-8371](#) - Try and prevent too many recovery requests from stacking up and clean up some faulty cancel recovery logic
- [SOLR-8451](#) - We should not call method.abort in HttpSolrClient or HttpSolrCall#remoteQuery and HttpSolrCall#remoteQuery should not close streams
- [SOLR-8453](#) - Solr should attempt to consume the request inputstream on errors as we cannot count on the container to do it
- [SOLR-8575](#) - Fix HDFSLogReader replay status numbers and a performance bug where we can reopen FSDatalInputStream too often
- [SOLR-8578](#) - Successful or not, requests are not always fully consumed by Solrj clients and we count on HttpClient or the JVM
- [SOLR-8615](#) - Just like creating cores, we should use multiple threads when closing cores
- [SOLR-8633](#) - DistributedUpdateProcess processCommit/deleteByQuery calls finish on DUP and SolrCmdDistributor, which violates the lifecycle and can cause bugs
- [SOLR-8720](#) - ZkController#publishAndWaitForDownStates should use #publishNodeAsDown
- [SOLR-8771](#) - Multi-threaded core shutdown creates executor per core
- [SOLR-8855](#) - The HDFS BlockDirectory should not clean up its cache on shutdown
- [SOLR-8856](#) - Do not cache merge or 'read once' contexts in the hdfs block cache
- [SOLR-8857](#) - HdfsUpdateLog does not use configured or new default number of version buckets and is hard coded to 256
- [SOLR-8869](#) - Optionally disable printing field cache entries in SolrFieldCacheMBean
- [SPARK-10859](#) - [SQL] Fix stats of StringType in columnar cache
- [SPARK-10914](#) - UnsafeRow serialization breaks when two machines have different Oops size
- [SPARK-11009](#) - [SQL] Fix wrong result of Window function in cluster mode
- [SPARK-11537](#) - [SQL] Fix negative hours/minutes/seconds
- [SPARK-11737](#) - [SQL] Fix serialization of UTF8String with Kyro
- [SPARK-12617](#) - [PYSPARK] Move Py4jCallbackConnectionCleaner to Streaming, clean up the leak sockets of Py4J
- [SPARK-14477](#) - [BUILD] Allow custom mirrors for downloading artifacts in build/mvn
- [SQOOP-2847](#) - Sqoop --incremental + missing parent --target-dir reports success with no data

Issues Fixed in CDH 5.5.2

Known Issues Fixed

The following topics describe known issues fixed in CDH 5.5.2.

Apache Spark

Spark SQL cannot retrieve data from a partitioned Hive table

When reading from a partitioned Hive table, Spark SQL is not able to identify the column delimiter used, and reads the full record as the first column entry.

Workaround: None.

When using Spark on YARN, the driver reports misleading error messages

CDH 5 Release Notes

The Spark driver reports misleading error messages such as:

```
ERROR ErrorMonitor: AssociationError [akka.tcp://sparkDriver@...]->
[akka.tcp://sparkExecutor@...]: Error [Association failed with [akka.tcp://sparkE
xecutor@...]]
[akka.remote.EndpointAssociationException: Association failed with
[akka.tcp://sparkExecutor@...]]
```

Workaround: Add the following property to the Spark log4j configuration file:

```
log4j.logger.org.apache.spark.rpc.akka.ErrorMonitor=FATAL..
```

Spark does not support rolling upgrades

Spark does not support rolling upgrades. Submitted Spark jobs may fail during upgrade. Jobs requiring new configuration properties will fail.

Workaround: Finish the upgrade, and then relaunch the Spark jobs.

Hue

Cannot query the `customers` table in Hue

To query the `customers` table, you must re-create the Parquet data for compatibility.

Bug: [HUE-3040](#)

Workaround: Update the parquet file of the `customers` table (`/user/hive/warehouse/customers/customers`) with the one attached to [HUE-3040](#).

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.5.2:

- [AVRO-1781](#) - Remove LogicalTypes cache
- [HADOOP-7713](#) - dfs -count -q should label output column
- [HADOOP-10406](#) - TestIPC.testIpcWithReaderQueuing may fail
- [HADOOP-10668](#) - Addendum patch to fix TestZKFailoverController
- [HADOOP-10668](#) - TestZKFailoverControllerStress#testExpireBackAndForth occasionally fails
- [HADOOP-11171](#) - Enable using a proxy server to connect to S3a
- [HADOOP-11218](#) - Add TLS 1.1, TLS 1.2 to KMS, HttpFS, SSLFactory
- [HADOOP-12269](#) - Update aws-sdk dependency to 1.10.6
- [HADOOP-12417](#) - TestWebDelegationToken failing with port in use
- [HADOOP-12418](#) - TestRPC.testRPCInterruptedSimple fails intermittently
- [HADOOP-12464](#) - Interrupted client may try to fail-over and retry
- [HADOOP-12468](#) - Partial group resolution failure should not result in user lockout
- [HADOOP-12474](#) - MiniKMS should use random ports for Jetty server by default
- [HADOOP-12568](#) - Update core-default.xml to describe posixGroups support
- [HADOOP-12573](#) - TestRPC.testClientBackOff failing
- [HADOOP-12584](#) - Disable browsing the static directory in HttpServer2
- [HADOOP-12584](#) - Revert - Disable browsing the static directory in HttpServer2
- [HADOOP-12584](#) - Disable browsing the static directory in HttpServer2
- [HADOOP-12604](#) - Exception may be swallowed in KMSClientProvider
- [HADOOP-12625](#) - Add a config to disable the /logs endpoints
- [HDFS-6101](#) - TestReplaceDatanodeOnFailure fails occasionally
- [HDFS-6533](#) - TestBPOfferService#testBasicFunctionalitytest fails intermittently
- [HDFS-6694](#) - TestPipelinesFailover.testPipelineRecoveryStress tests fail intermittently with various symptoms - debugging patch
- [HDFS-7553](#) - Fix the TestDFSUpgradeWithHA due to BindException
- [HDFS-7798](#) - Checkpointing failure caused by shared KerberosAuthenticator
- [HDFS-8647](#) - Abstract BlockManager's rack policy into BlockPlacementPolicy

- [HDFS-8722](#) - Optimize DataNode writes for small writes and flushes
- [HDFS-8772](#) - Fix TestStandbyIsHot#testDatanodeRestarts which occasionally fails
- [HDFS-8805](#) - Archival Storage: getStoragePolicy should not need superuser privilege
- [HDFS-9083](#) - Replication violates block placement policy
- [HDFS-9123](#) - Copying from the root to a subdirectory should be forbidden
- [HDFS-9160](#) - [OIV-Doc] : Missing details of 'delimited' for processor options
- [HDFS-9220](#) - Reading small file (< 512 bytes) that is open for append fails due to incorrect checksum
- [HDFS-9249](#) - NPE is thrown if an IOException is thrown in NameNode constructor
- [HDFS-9250](#) - Add precondition check to LocatedBlock#addCachedLoc
- [HDFS-9268](#) - fuse_dfs chown crashes when uid is passed as -1
- [HDFS-9273](#) - ACLs on root directory may be lost after NameNode restart
- [HDFS-9286](#) - HttpFs does not parse ACL syntax correctly for operation REMOVEAACLENTRIES
- [HDFS-9295](#) - Add a thorough test of the full KMS code path
- [HDFS-9313](#) - Possible NullPointerException in BlockManager if no excess replica can be chosen
- [HDFS-9332](#) - Fix Precondition failures from NameNodeEditLogRoller while saving namespace
- [HDFS-9339](#) - Extend full test of KMS ACLs
- [HDFS-9364](#) - Unnecessary DNS resolution attempts when creating NameNodeProxies
- [HDFS-9410](#) - Some tests should always reset sysout and syserr
- [HDFS-9429](#) - Tests in TestDFSAdminWithHA intermittently fail with EOFException
- [HDFS-9438](#) - Only collect HDFS-6694 debug data on Linux, Mac, and Solaris
- [HDFS-9445](#) - DataNode may deadlock while handling a bad volume
- [HDFS-9470](#) - Encryption zone on root not loaded from fsimage after NameNode restart
- [HDFS-9474](#) - TestPipelinesFailover should not fail when printing debug message
- [MAPREDUCE-6191](#) - Improve clearing stale state of Java serialization testcase
- [MAPREDUCE-6233](#) - org.apache.hadoop.mapreduce.TestLargeSort.testLargeSort failed in trunk
- [MAPREDUCE-6549](#) - Multibyte delimiters with LineRecordReader cause duplicate records
- [MAPREDUCE-6550](#) - archive-logs tool changes log ownership to the Yarn user when using DefaultContainerExecutor
- [YARN-3564](#) - Fix TestContainerAllocation.testAMContainerAllocationWhenDNSUnavailable fails randomly
- [YARN-3768](#) - ArrayIndexOutOfBoundsException with empty environment variables
- [YARN-4235](#) - FairScheduler PrimaryGroup does not handle empty groups returned for a user
- [YARN-4310](#) - FairScheduler: Log skipping reservation messages at DEBUG level
- [YARN-4347](#) - Resource manager fails with Null pointer exception
- [YARN-4408](#) - Fix issue that NodeManager still reports negative running containers
- [HBASE-6617](#) - ReplicationSourceManager should be able to track multiple WAL paths
- [HBASE-12961](#) - Fix negative values in read and write region server metrics
- [HBASE-13134](#) - mutateRow and checkAndMutate apis don't throw region level exceptions
- [HBASE-13703](#) - ReplicateContext should not be a member of ReplicationSource
- [HBASE-13746](#) - list_replicated_tables command is not listing table in HBase shell
- [HBASE-13833](#) - LoadIncrementalHFile.doBulkLoad(Path, HTable) does not handle unmanaged connections when using SecureBulkLoad
- [HBASE-14003](#) - Work around JDK-8044053
- [HBASE-14205](#) - RegionCoprocessorHost System.nanoTime() performance bottleneck
- [HBASE-14283](#) - Reverse scan doesn't work with HFile inline index/bloom blocks
- [HBASE-14501](#) - NPE in replication with TDE
- [HBASE-14533](#) - Connection Idle time 1 second is too short and the connection is closed too quickly by the ChoreService. Increase it to the default (10 minutes) for testAll(). The patch is not committed upstream yet.
- [HBASE-14541](#) - TestHFileOutputFormat.testMRIncrementalLoadWithSplit failed due to too many splits and few retries
- [HBASE-14547](#) - Add more debug/trace to zk-procedure
- [HBASE-14621](#) - ReplicationLogCleaner stuck on RS crash

CDH 5 Release Notes

- [HBASE-14731](#) - Add -DuseMob option to ITBLL
- [HBASE-14809](#) - Grant / revoke namespace admin permission to group
- [HBASE-14923](#) - VerifyReplication should not mask the exception during result comparison
- [HBASE-14926](#) - Hung ThriftServer; no timeout on read from client; if client crashes, worker thread gets stuck reading
- [HBASE-15031](#) - Fix merge of MVCC and SequenceID performance regression in branch-1.0
- [HBASE-15032](#) - HBase shell scan filter string assumes UTF-8 encoding
- [HBASE-15035](#) - Bulkloading HFiles with tags that require splits do not preserve tags
- [HBASE-15104](#) - Occasional failures due to NotServingRegionException in IT tests
- [HIVE-7575](#) - GetTables thrift call is very slow
- [HIVE-7653](#) - Hive AvroSerDe does not support circular references in Schema
- [HIVE-9507](#) - Make "LATERAL VIEW inline(expression) mytable" tolerant to nulls
- [HIVE-10027](#) - Use descriptions from Avro schema files in column comments
- [HIVE-10048](#) - JDBC - Support SSL encryption regardless of Authentication mechanism
- [HIVE-10083](#) - SMBJoin fails in case one table is uninitialized
- [HIVE-10265](#) - Hive CLI crashes on != inequality
- [HIVE-10514](#) - Fix MiniCliDriver tests failure
- [HIVE-10687](#) - AvroDeserializer fails to deserialize evolved union fields
- [HIVE-10697](#) - ObjectInspectorConvertors#UnionConvertor does a faulty conversion
- [HIVE-11149](#) - Fix issue with sometimes HashMap in PerfLogger.java hangs
- [HIVE-11288](#) - Backport:Avro SerDe InstanceCache returns incorrect schema
- [HIVE-11513](#) - AvroLazyObjectInspector could handle empty data better
- [HIVE-11616](#) - DelegationTokenSecretManager reuses the same objectstore, which has concurrency issues
- [HIVE-11785](#) - Revert - Support escaping carriage return and new line for LazySimpleSerDe
- [HIVE-11785](#) - Support escaping carriage return and new line for LazySimpleSerDe
- [HIVE-11826](#) - 'hadoop.proxyuser.hive.groups' configuration does not prevent unauthorized user to access metastore
- [HIVE-11977](#) - Hive should handle an external Avro table with zero length files present
- [HIVE-12008](#) - Hive queries failing when using count(*) on column in view
- [HIVE-12058](#) - Change Hive script to record errors when calling hbase fails
- [HIVE-12188](#) - DoAs does not work properly in non-Kerberos secured HiveServer2
- [HIVE-12189](#) - The list in pushdownPreds of ppd.ExprWalkerInfo should not be allowed to grow very large
- [HIVE-12218](#) - Unable to create a like table for an HBase-backed table
- [HIVE-12250](#) - Zookeeper connection leaks in Hive's HBaseHandler
- [HIVE-12265](#) - Generate lineage info only if requested
- [HIVE-12268](#) - Context leaks deleteOnExit paths
- [HIVE-12278](#) - Skip logging lineage for explain queries
- [HIVE-12287](#) - Lineage for lateral view shows wrong dependencies
- [HIVE-12330](#) - Fix precommit Spark test part2
- [HIVE-12365](#) - Added resource path is sent to cluster as an empty string when externally removed
- [HIVE-12378](#) - Exception on HBaseSerDe.serialize binary field
- [HIVE-12388](#) - GetTables cannot get external tables when TABLE type argument is given
- [HIVE-12406](#) - HIVE-9500 introduced incompatible change to LazySimpleSerDe public interface
- [HIVE-12418](#) - HiveHBaseTableInputFormat.getRecordReader() causes Zookeeper connection leak
- [HIVE-12505](#) - Backport: Insert overwrite in same encrypted zone silently fails to remove some existing files
- [HIVE-12566](#) - Incorrect result returns when using COALESCE in WHERE condition with LEFT JOIN
- [HIVE-12713](#) - Miscellaneous improvements in driver compile and execute logging
- [HIVE-12784](#) - Group by SemanticException: Invalid column reference
- [HIVE-12788](#) - Setting hive.optimize.union.remove to TRUE will break UNION ALL with aggregate functions
- [HIVE-12795](#) - Vectorized execution causes ClassCastException
- [HUE-2664](#) - Revert - [jobbrowser] Fix fetching logs from job history server

- [HUE-2997](#) - [oozie] Easier usage of email action when workflow fails
- [HUE-3035](#) - [beeswax] Optimize sample data query for partitioned tables
- [HUE-3036](#) - [beeswax] Revert get_tables to use Thrift API GetTables
- [HUE-3091](#) - [oozie] Do not remove extra new lines from email action body
- [IMPALA-1459](#) - Fix migration/assignment of On-clause predicates inside inline views.
- [IMPALA-2103](#) - Fix flaky test_impersonation test
- [IMPALA-2113](#) - Handle error when distinct and aggregates are used with a having clause
- [IMPALA-2225](#) - Handle error when star based select item and aggregate are incorrectly used
- [IMPALA-2226](#) - Throw AnalysisError if table properties are too large
- [IMPALA-2273](#) - Make MAX_PAGE_HEADER_SIZE configurable
- [IMPALA-2473](#) - Reduce scanner memory usage
- [IMPALA-2535](#) - PAGG hits mem_limit when switching to I/O buffers
- [IMPALA-2558](#) - DCHECK in Parquet scanner after block read error
- [IMPALA-2559](#) - Fix check failed: sorter_runs_.back()->is_pinned_
- [IMPALA-2591](#) - DataStreamSender::Send() does not return an error status if SendBatch() failed
- [IMPALA-2598](#) - Re-enable SSL and Kerberos on server-server
- [IMPALA-2612](#) - Free local allocations once for every row batch when building hash tables
- [IMPALA-2614](#) - Don't ignore Status returned by DataStreamRecv::CreateMerger()
- [IMPALA-2624](#) - Increase fs.trash.interval to 24 hours for test suite
- [IMPALA-2630](#) - Skip TestParquet.test_continue_on_error when using old aggs/joins
- [IMPALA-2643](#) - Prevent migrating incorrectly inferred identity predicates into inline views
- [IMPALA-2648](#) - Avoid sending large partition stats objects over thrift
- [IMPALA-2695](#) - Fix GRANTs on URIs with uppercase letters
- [IMPALA-2722](#) - Free local allocations per row batch in non-partitioned AGG and HJ
- [IMPALA-2731](#) - Refactor MemPool usage in HBase scan node
- [IMPALA-2747](#) - Thrift-client cleans openSSL state before using it in the case of the catalog
- [IMPALA-2776](#) - Remove escapechartesttable and associated tests
- [IMPALA-2812](#) - Remove additional test referencing escapecharstesttable
- [IMPALA-2829](#) - SEGV in AnalyticEvalNode touching NULL input_stream_
- [KITE-1089](#) - ReadAvroContainer morphline command should work even if the Avro writer schema of each input file is different
- [KITE-1097](#) - Add method to read the name of a Morphline command
- [OOZIE-2030](#) - Configuration properties from global section is not getting set in Hadoop job conf when using sub-workflow action in Oozie workflow.xml
- [OOZIE-2365](#) - Oozie fails to start when SMTP password not set
- [OOZIE-2380](#) - Oozie Hive action failed with wrong tmp path
- [OOZIE-2397](#) - LAST_ONLY and NONE don't properly handle READY actions
- [OOZIE-2413](#) - Kerberos credentials can expire if the KDC is slow to respond
- [OOZIE-2439](#) - FS Action no longer uses name-node from global section or default NN
- [OOZIE-2441](#) - SubWorkflow action with propagate-configuration but no global section throws NPE on submit
- [PIG-3641](#) - Split "otherwise" producing incorrect output when combined with ColumnPruning
- [SENTRY-565](#) - Improve performance of filtering Hive SHOW commands
- [SENTRY-835](#) - Drop table leaves a connection open when using metastorelistener
- [SENTRY-902](#) - SimpleDBProviderBackend should retry the authorization process properly
- [SENTRY-936](#) - getGroup and getUser should always return original HDFS values for paths in prefix which are not managed by Sentry
- [SENTRY-944](#) - Setting HDFS rules on Sentry-managed HDFS paths should not affect original HDFS rules
- [SENTRY-953](#) - External Partitions which are referenced by more than one table can cause some unexpected behavior with Sentry HDFS sync
- [SENTRY-957](#) - Exceptions in MetastoreCacheInitializer should probably not prevent HMS from starting up

CDH 5 Release Notes

- [SENTRY-960](#) - Blacklist reflect, java_method using hive.server2.builtin.udf.blacklist
- [SENTRY-988](#) - Let SentryAuthorization setter path always fall through and update HDFS
- [SENTRY-994](#) - SentryAuthorizationInfoX should override isSentryManaged
- [SOLR-6443](#) - backportDisable test that fails on Jenkins until we can determine the problem
- [SOLR-7049](#) - LIST Collections API call should be processed directly by the CollectionsHandler instead of the OverseerCollectionProcessor
- [SOLR-7989](#) - After a new leader is elected it, it should ensure it's state is ACTIVE if it has already registered with ZooKeeper
- [SOLR-8075](#) - Fix faulty implementation
- [SOLR-8152](#) - Overseer Task Processor/Queue can miss responses, leading to timeouts
- [SOLR-8223](#) - Avoid accidentally swallowing OutOfMemoryError
- [SOLR-8288](#) - DistributedUpdateProcessor#doFinish should explicitly check and ensure it does not try to put itself into LIR
- [SOLR-8353](#) - Support regex for skipping license checksums
- [SOLR-8367](#) - Fix the LeaderInitiatedRecovery 'all replicas participate' fail-safe
- [SOLR-8372](#) - backportCanceled recovery can lead to data loss
- [SOLR-8535](#) - Support forcing define-lucene-javadoc-url to be local
- [SPARK-5569](#) - [STREAMING] Fix ObjectInputStreamWithLoader for supporting load array classes
- [SPARK-8029](#) - Robust shuffle writer
- [SPARK-9735](#) - [SQL] Respect the user specified schema than the infer partition schema for HadoopFsRelation
- [SPARK-10648](#) - Oracle dialect to handle nonspecific numeric types
- [SPARK-10865](#) - [SPARK-10866] [SQL] Fix bug of ceil/floor, which should returns long instead of the Double type
- [SPARK-11105](#) - [YARN] Distribute log4j.properties to executors
- [SPARK-11126](#) - [SQL] Fix the potential flaky test
- [SPARK-11126](#) - [SQL] Fix a memory leak in SQLListener._stageIdToStageMetrics
- [SPARK-11246](#) - [SQL] Table cache for Parquet broken in 1.5
- [SPARK-11453](#) - [SQL] Append data to partitioned table will messes up the result
- [SPARK-11484](#) - [WEBUI] Using proxyBase set by Spark AM
- [SPARK-11786](#) - [CORE] Tone down messages from akka error monitor
- [SPARK-11799](#) - [CORE] Make it explicit in executor logs that uncaught exceptions are thrown during executor shutdown
- [SPARK-11929](#) - [CORE] Make the repl log4j configuration override the root logger
- [SQOOP-2745](#) - Using datetime column as a splitter for Oracle no longer works
- [SQOOP-2767](#) - Test is failing SystemImportTest
- [SQOOP-2783](#) - Query import with parquet fails on incompatible schema
- [SQOOP-2422](#) - Sqoop2: Test TestJSONIntermediateDataFormat is failing on JDK8

Issues Fixed in CDH 5.5.1

The following issues have been fixed in CDH 5.5.1:

Apache Commons Collections deserialization vulnerability

Cloudera has learned of a potential security vulnerability in a third-party library called the [Apache Commons Collections](#). This library is used in products distributed and supported by Cloudera (“Cloudera Products”), including core Apache Hadoop. The Apache Commons Collections library is also in widespread use beyond the Hadoop ecosystem. At this time, no specific attack vector for this vulnerability has been identified as present in Cloudera Products.

In an abundance of caution, we are currently in the process of incorporating a version of the Apache Commons Collections library with a fix into the Cloudera Products. In most cases, this will require coordination with the projects in the Apache community. One example of this is tracked by [HADOOP-12577](#).

The Apache Commons Collections potential security vulnerability is titled “Arbitrary remote code execution with InvokerTransformer” and is tracked by [COLLECTIONS-580](#). MITRE has not issued a CVE, but related [CVE-2015-4852](#) has been filed for the vulnerability. CERT has issued [Vulnerability Note #576313](#) for this issue.

Releases affected: CDH 5.5.0, CDH 5.4.8 and lower, CDH 5.3.8 and lower, CDH 5.2.8 and lower, CDH 5.1.7 and lower, Cloudera Manager 5.5.0, Cloudera Manager 5.4.8 and lower, Cloudera Manager 5.3.8 and lower, and Cloudera Manager 5.2.8 and lower, Cloudera Manager 5.1.6 and lower, Cloudera Navigator 2.4.0, Cloudera Navigator 2.3.8 and lower.

Users affected: All

Impact: This potential vulnerability may enable an attacker to execute arbitrary code from a remote machine without requiring authentication.

Immediate action required: Upgrade to Cloudera Manager 5.5.1 and CDH 5.5.1, Cloudera Manager 5.4.9 and CDH 5.4.9, Cloudera Manager 5.3.9 and CDH 5.3.9, and Cloudera Manager 5.2.9 and CDH 5.2.9, and Cloudera Manager 5.1.7 and CDH 5.1.7.

Apache HBase

Data may not be replicated to slave cluster if multiwal multiplicity is set to greater than 1.

Bug: [HBASE-13703](#), [HBASE-6617](#), [HBASE-14501](#).

Issues Fixed in CDH 5.5.0

The following topics describe known issues fixed in CDH 5.5.0.

Apache Flume

Fix FD leak in AsyncHBaseSink.

Bug: [FLUME-2738](#)

Fix for Kerberos configuration error when using short names.

Bug: [FLUME-2749](#)

Fix NullPointerException in KafkaSourceCounter.

Bug: [FLUME-2672](#)

Fix for Tail Directory Source FileNotFoundException.

Bug: [FLUME-2773](#)

Fix for Kafka Channel timeout property handling.

Bug: [FLUME-2734](#)

Apache Hadoop

YARN/MapReduce

Incorrect headroom leads to deadlock between mappers and reducers.

Bug: [MAPREDUCE-6302](#)

Blacklisting Support for Scheduling ApplicationMasters

When an ApplicationMaster fails, and the NodeManager on the same host has not yet been blacklisted, the framework should route the second ApplicationMaster attempt to a NodeManager on a different host.

Bug: [YARN-2005](#)

Apache HBase

HBase mlock agent is not included as part of CDH 5.x releases.

Starting in CDH 5.0, the native `mlock` daemons were not included in CDH HBase. CDH 5.5 restores the daemon in both parcels and packages.

Bug: None.

Apache Hive

Parquet Predicate pushdown for float types does not work

The Parquet predicate builder should use `PrimitiveTypeName` type to construct a predicate leaf instead of the type provided by `PredicateLeaf`.

CDH 5 Release Notes

Bug: [HIVE-11504](#)

LineageCtx should release all resources at clear()

Some maps are not released with the `clear()` method and can cause a memory leak.

Bug: [HIVE-12225](#)

LEFT JOIN query plan outputs wrong column when using subquery

Incorrect results may arise if a `LEFT OUTER JOIN` is combined with a subquery.

Bug: [HIVE-9613](#)

Map-side aggregation is extremely slow

Map-side aggregation on columns with double type is extremely slow due to [HIVE-7041](#).

Bug: [HIVE-11502](#)

Lineage does not work with dynamic partitioning query

An error message displays after running a dynamic partitioning query: `ERROR : Result schema has 2 fields, but we don't get as many dependencies.`

Bug: [HIVE-11834](#)

DROP PARTITION in encrypted zone does not remove data from HDFS

An `ALTER TABLE` query to `DROP PARTITION` removes the partition metadata from HDFS but not the data.

Bug: [HIVE-10910](#)

DROP TABLE with qualified table name ignores database name when checking partitions

`DDLTask.dropTable()` uses an older version of `Hive.getPartitionNames()`, which takes in a single string for the table name, instead of the database and table names.

Bug: [HIVE-10421](#)

INSERT INTO statement may expose data that should be encrypted

`INSERT INTO <table> VALUES()` uses a temporary table; the data in temporary tables is stored under `hive.exec.scratchdir` which is not usually encrypted.

Bug: [HIVE-10658](#)

Whitelist restrictions do not get initialized in new copy of HiveConf

Whitelist restrictions use a regex pattern in `HiveConf`, but when a new `HiveConf` object copy is created, the regex pattern is not initialized in the new `HiveConf` copy.

Bug: [HIVE-10465](#)

RuntimeException when vectorization is enabled with binary data

A `RuntimeException` is thrown when vectorization is enabled and binary data is in the `GROUP BY` clause.

Bug: [HIVE-9908](#)

Hive may return wrong results in some queries with PTF function

The select statement has an extra column with a PTF operator that is skewing results.

Bug: [HIVE-11604](#)

HiveServer2 leaks Hive Metastore Connections

HiveServer2 uses `threadlocal` to cache Hive Metastore (HMS) Thrift client in class Hive. When the thread dies, the HMS client does not close. So the connection to the HMS client leaks.

Bug: [HIVE-10956](#)

Remote Spark Client has a memory leak

In Remote Spark Client (RSC), MapWork/ReduceWork tasks build up until an `OutOfMemoryException` is thrown.

Bug: [HIVE-10006](#)

Replication factor is not properly set in `SparkHashTableSinkOperator`

The default replication factors (3) affects the Map Join performance of small files.

Bug: [HIVE-11109](#)

Hive LDAP Authenticator should allow users to set Domain without the base Distinguished Name

When the base distinguished name (DN) is not configured but only the Domain has been set in `hive-site.xml`, the LDAP authentication provider cannot locate the user in the directory. Authentication fails in such cases.

Bug: [HIVE-12007](#)

Hive should support additional LDAP authentication parameters

Currently, Hive only has the following authenticator parameters for LDAP authentication for HiveServer2:

```
<property>
  <name>hive.server2.authentication</name>
  <value>LDAP</value>
</property>
<property>
  <name>hive.server2.authentication.ldap.url</name>
  <value>ldap://our_ldap_address</value>
</property>
```

Other LDAP properties need to be included as part of Hive-LDAP authentication, for example:

```
group search base -> dc=domain,dc=com
group search filter -> member={0}
user search base -> dc=domain,dc=com
user search filter -> sAMAccountName={0}
list of valid user groups -> group1,group2,group3
```

Bug: [HIVE-7193](#)

Aggregate functions used as window functions can fail in various ways

- [HIVE-11817](#) : Window function `max()` fails with `NullPointerException`.
- [HIVE-10702](#) : `COUNT(*)` over windowing `x preceding and y preceding` returns unexpected results.
- [HIVE-10826](#) : Support `min()/max()` functions over `x preceding and y preceding` windowing.

Apache Spark

Attempts to access secure HBase from Spark executors fail when authenticating to the metastore.

An exception like the following occurs when you attempt to access kerberized HBase instance from a Spark executor.

```
GSSEException: No valid credentials provided
(Mechanism level: Failed to find any Kerberos tgt)
```

The root cause is that the HBase Kerberos authentication token is not sent to the Spark executor.

Bug: [SPARK-6918](#)

Workaround: None.

The shuffle service fails on NodeManager restarts and kills all running Spark applications

In CDH 5.4.0 through CDH 5.4.4, the shuffle service is on by default. Because it fails in NodeManager restarts, in CDH 5.4.5, and higher, the shuffle service is off by default. Dynamic allocation requires that the shuffle service be turned on.

Bug: [SPARK-9439](#)

CDH 5 Release Notes

Workaround: In CDH 5.4.5 and higher, enable the shuffle service when using dynamic allocation.

Spark not automatically picking up hive-site.xml

When you run Spark on YARN, the client `hive-site.xml` does not get picked up automatically by `spark-submit`.

Bug: [SPARK-2669](#)

Workaround: Do one of the following, depending on which deployment mode you are running in:

- Client - set `HADOOP_CONF_DIR` to `/etc/hive/conf/` (or the directory where `hive-site.xml` is located).
- Cluster - add `--files=/etc/hive/conf/hive-site.xml` (or the path for `hive-site.xml`) to the `spark-submit` script.

Apache Sentry (incubating)

Synchronize calls in SentryClient and create Sentry client once per request in SimpleDBProvider

Adds proper locking to the `SentryClient` and reduces the number of `SentryClients` created within a single request in the `SimpleDbProvider` (used by Hive). This fixes issues that may have caused transient permission failures and out of memory conditions.

Bug: [SENTRY-893](#)

Sentry-HDFS sync events should treat database and table names as case-insensitive

Sentry-HDFS Sync was treating database and table names as case-sensitive. This led to incorrect or missing ACLs being applied as part of the sync operation if the DDL operations used a different case for the catalog objects.

Bug: [SENTRY-885](#)

Hive drop database operation removes the Sentry privileges, even if drop operation fails

Even if the Hive drop database operation fails, the Sentry privileges on that database will be removed.

Bug: [SENTRY-669](#)

Nested queries in Hive on views incorrectly enforce base table privileges instead of view privileges

Nested queries in Hive on views incorrectly enforce base table privileges instead of view privileges. This leads to Hive query failures due to insufficient privileges.

Bug: [SENTRY-619](#)

Apache ZooKeeper

BinaryInputArchive readString should check length before allocating memory

This fixed a possible `OutOfMemoryError` when malformed packets were sent to the ZooKeeper server.

Bug: [ZOOKEEPER-2146](#)

Workaround: Upgrade to CDH 5.5.

Cloudera Search

The GoLive Function Does not Support Running As a Configurable User

After using `--go-live` mode with the `MapReduceIndexerTool` and `HBaseMapReduceIndexerTool`, depending on group mappings and the configured HDFS umask, Solr may not have been able to read the results of the indexing job.

With Search for CDH 5.5 and later, the `MapReduceIndexerTool` and `HBaseMapReduceIndexerTool` includes updated `--go-live` functionality. The indexers now automatically update HDFS ACLs for the specified output directory, giving Solr permission to read the indexer results.

See [MapReduceIndexerTool](#) and [HBaseMapReduceIndexerTool](#) for more information.

Bug: None.

Workaround: Do not use the `--golive` mode with `MapReduceIndexerTool` and `HBaseMapReduceIndexerTool` or use a less restrictive umask.

MapReduceIndexerTool fails to Index Documents When Sentry Is Enabled

Prior to CDH 5.5, when Sentry was enabled, the MapReduceIndexerTool was unable to index data even if the user was authorized to write to the collection according to Sentry permissions. This limitation occurred because, by default, the MapReduceIndexerTool used the underlying collection's `solrconfig.xml` from ZooKeeper to build the index using its EmbeddedSolrServers. But the embedded servers are not properly configured to use Sentry, so this process failed.

With Search for CDH 5.5, the MapReduceIndexerTool uses a default `solrconfig.xml` that is appropriate for the vast majority of collection configurations. With this configuration, the MapReduceIndexerTool is able to index data, even if Sentry is enabled. Note that this default configuration does not include any `updateRequestProcessorChains`; if your configuration requires an `updateRequestProcessorChain`, you can tell the MapReduceIndexerTool to use the configuration from ZooKeeper by specifying `--use-zk-solrconfig.xml` or from local disk by specifying `--solr-home-dir`.

Bug: None.

Workaround: To address this issue, configure the MapReduceIndexerTool to run without Sentry restrictions. This does not compromise security because this only affects the "embedded" Solr Servers in the job that are used to build the offline index; Solr's Sentry permissions are still checked when the data is merged into the cluster via `--go-live`.

Here are two ways to enable indexing:

1. If your environment uses the default configuration files, use `solrconfig.xml` for indexing jobs, rather than `solrconfig.xml.secure`. Use the `--solr-home-dir` option to specify the directory containing `solrconfig.xml`, causing the job to run with Sentry disabled.
2. Alternately, you can comment out the following line:

```
<str name="update.chain">updateIndexAuthorization</str>
```

This line must be commented out and the change saved in the `solrconfig` file used by the machine running the indexing job.

Issues Fixed in CDH 5.4.x

The following topics describe known issues fixed in CDH 5.4.x, from newest to oldest release.

Issues Fixed in CDH 5.4.11

CDH 5.4.11 fixes the following issues.

- [FLUME-2891](#) - Revert FLUME-2712 and FLUME-2886
- [FLUME-2908](#) - NetcatSource - SocketChannel not closed when session is broken
- [HADOOP-8436](#) - NPE In `getLocalPathForWrite` (`path, conf`) when the required context item is not configured
- [HADOOP-8437](#) - `getLocalPathForWrite` should throw `IOException` for invalid paths
- [HADOOP-8751](#) - NPE in `Token.toString()` when Token is constructed using null identifier
- [HADOOP-8934](#) - Shell command `ls` should include sort options
- [HADOOP-10048](#) - LocalDirAllocator should avoid holding locks while accessing the filesystem
- [HADOOP-10971](#) - Add `-C` flag to make `'hadoop fs -ls'` print filenames only
- [HADOOP-11901](#) - BytesWritable fails to support 2G chunks due to integer overflow
- [HADOOP-12252](#) - LocalDirAllocator should not throw NPE with empty string configuration
- [HADOOP-12269](#) - Update aws-sdk dependency to 1.10.6
- [HADOOP-12787](#) - KMS SPNEGO sequence does not work with WEBHDFS
- [HADOOP-12841](#) - Update s3-related properties in `core-default.xml`.
- [HADOOP-12901](#) - Add warning log when KMSClientProvider cannot create a connection to the KMS server.
- [HADOOP-12972](#) - Lz4Compressor#`getLibraryName` returns the wrong version number
- [HADOOP-13079](#) - Add `-q` option to `ls` to print `?` instead of non-printable characters
- [HADOOP-13132](#) - Handle `ClassCastException` on `AuthenticationException` in `LoadBalancingKMSClientProvider`
- [HADOOP-13155](#) - Implement `TokenRenewer` to renew and cancel delegation tokens in KMS
- [HADOOP-13251](#) - Authenticate with Kerberos credentials when renewing KMS delegation token

CDH 5 Release Notes

- [HADOOP-13255](#) - KMSClientProvider should check and renew tgt when doing delegation token operations
- [HADOOP-13263](#) - Reload cached groups in background after expiry.
- [HADOOP-13457](#) - Remove hardcoded absolute path for shell executable.
- [HDFS-4660](#) - Block corruption can happen during pipeline recovery
- [HDFS-8211](#) - DataNode UUID is always null in the JMX counter.
- [HDFS-8451](#) - DFSClient probe for encryption testing interprets empty URI property for enabled
- [HDFS-8496](#) - Calling stopWriter() with FSDataOutputStream lock held may block other threads
- [HDFS-8576](#) - Lease recovery should return true if the lease can be released and the file can be closed
- [HDFS-8722](#) - Optimize datanode writes for small writes and flushes
- [HDFS-9085](#) - Show renewer information in DelegationTokenIdentifier#toString
- [HDFS-9220](#) - Reading small file (< 512 bytes) that is open for append fails due to incorrect checksum
- [HDFS-9276](#) - Failed to Update HDFS Delegation Token for long running application in HA mode
- [HDFS-9466](#) - TestShortCircuitCache#testDataXceiverCleansUpSlotsOnFailure is flaky
- [HDFS-9589](#) - Block files which have been hardlinked should be duplicated before the DataNode appends to them
- [HDFS-9700](#) - DFSClient and DFSOutputStream should set TCP_NODELAY on sockets for DataTransferProtocol
- [HDFS-9732](#) - Improve DelegationTokenIdentifier.toString() for better logging
- [HDFS-9939](#) - Increase DecompressorStream skip buffer size
- [HDFS-9949](#) - Add a test case to ensure that the DataNode does not regenerate its UUID when a storage directory is cleared
- [HDFS-10267](#) - Extra "synchronized" on FsDatasetImpl#recoverAppend and FsDatasetImpl#recoverClose
- [HDFS-10360](#) - DataNode might format directory and lose blocks if current/VERSION is missing.
- [HDFS-10381](#) - , DataStreamer DataNode exclusion log message should be warning.
- [MAPREDUCE-4785](#) - TestMRApp occasionally fails
- [MAPREDUCE-6580](#) - Test failure: TestMRJobsWithProfiler
- [YARN-2871](#) - TestRMRestart#testRMRestartGetApplicationList sometimes fails in trunk
- [YARN-3727](#) - Check if the directory exists before using it for localization
- [YARN-4168](#) - Fixed a failing test TestLogAggregationService.testLocalFileDeletionOnDiskFull
- [YARN-4354](#) - Public resource localization fails with NPE
- [YARN-4717](#) - TestResourceLocalizationService.testPublicResourceInitializesLocalDir fails Intermittently due to IllegalArgumentException from cleanup
- [YARN-5048](#) - DelegationTokenRenewer#skipTokenRenewal might throw NPE
- [HBASE-6617](#) - ReplicationSourceManager should be able to track multiple WAL paths (ADDENDUM)
- [HBASE-11625](#) - Verifies data before building HFileBlock. - Adds HFileBlock.Header class which contains information about location of fields. Testing: Adds CorruptedFSReaderImpl to TestChecksum.
- [HBASE-11927](#) - Use Native Hadoop Library for HFile checksum.
- [HBASE-14155](#) - StackOverflowError in reverse scan
- [HBASE-14359](#) - HTable#close will hang forever if unchecked error/exception thrown in AsyncProcess#sendMultiAction
- [HBASE-14730](#) - region server needs to log warnings when there are attributes configured for cells with hfile v2
- [HBASE-14759](#) - Avoid using Math.abs when selecting SyncRunner in FSHLog
- [HBASE-15234](#) - Don't abort ReplicationLogCleaner on ZooKeeper errors
- [HBASE-15456](#) - CreateTableProcedure/ModifyTableProcedure needs to fail when there is no family in table descriptor
- [HBASE-15479](#) - No more garbage or beware of autoboxing
- [HBASE-15582](#) - SnapshotManifestV1 too verbose when there are no regions
- [HBASE-15707](#) - ImportTSV bulk output does not support tags with hfile.format.version=3
- [HBASE-15746](#) - Remove extra RegionCoprocessor preClose() in RSRpcServices#closeRegion
- [HBASE-15811](#) - Batch Get after batch Put does not fetch all Cells We were not waiting on all executors in a batch to complete. The test for no-more-executors was damaged by the 0.99/0.98.4 fix "HBASE-11403 Fix race conditions around Object#notify"

- [HBASE-15925](#) - provide default values for hadoop compat module related properties that match default hadoop profile.
- [HBASE-16207](#) - can't restore snapshot without "Admin" permission
- [HIVE-9499](#) - hive.limit.query.max.table.partition makes queries fail on non-partitioned tables
- [HIVE-10048](#) - JDBC - Support SSL encryption regardless of Authentication mechanism
- [HIVE-10303](#) - HIVE-9471 broke forward compatibility of ORC files
- [HIVE-10685](#) - Alter table concatenate oparetor will cause duplicate data
- [HIVE-10925](#) - Non-static threadlocals in metastore code can potentially cause memory leak
- [HIVE-11031](#) - ORC concatenation of old files can fail while merging column statistics
- [HIVE-11054](#) - Handle varchar/char partition columns in vectorization
- [HIVE-11243](#) - Changing log level in Utilities.getBaseWork
- [HIVE-11408](#) - HiveServer2 is leaking ClassLoaders when add jar / temporary functions are used due to constructor caching in Hadoop ReflectionUtils
- [HIVE-11427](#) - Location of temporary table for CREATE TABLE SELECT broken by HIVE-7079.
- [HIVE-11488](#) - Combine the following jiras for "Support sessionId and queryId logging"Add sessionId and queryId info to HS2 log (Aihua Xu, reviewed by Szehon Ho) HIVE-12456: QueryId can't be stored in the configuration of the SessionState since multiple queries can run in a single session
- [HIVE-11583](#) - When PTF is used over a large partitions result could be corrupted
- [HIVE-11747](#) - Unnecessary error log is shown when executing a "INSERT OVERWRITE LOCAL DIRECTORY" cmd in the embedded mode
- [HIVE-11827](#) - STORED AS AVRO fails SELECT COUNT(*) when empty
- [HIVE-11919](#) - Hive Union Type Mismatch
- [HIVE-12354](#) - MapJoin with double keys is slow on MR
- [HIVE-12431](#) - Support timeout for compile lock
- [HIVE-12481](#) - Occasionally "Request is a replay" will be thrown from HS2
- [HIVE-12635](#) - Hive should return the latest hbase cell timestamp as the row timestamp value
- [HIVE-12958](#) - Make embedded Jetty server more configurable
- [HIVE-13200](#) - Aggregation functions returning empty rows on partitioned columns
- [HIVE-13251](#) - hive can't read the decimal in AVRO file generated from previous version
- [HIVE-13285](#) - Orc concatenation may drop old files from moving to final path
- [HIVE-13286](#) - Query ID is being reused across queries
- [HIVE-13462](#) - HiveResultSetMetaData.getPrecision() fails for NULL columns
- [HIVE-13527](#) - Using deprecated APIs in HBase client causes zookeeper connection leaks
- [HIVE-13570](#) - Some queries with Union all fail when CBO is off
- [HIVE-13932](#) - Hive SMB Map Join with small set of LIMIT failed with NPE
- [HIVE-13953](#) - Issues in HiveLockObject.equals method
- [HIVE-13991](#) - Union All on view fail with no valid permission on underneath table
- [HIVE-14118](#) - Make the alter partition exception more meaningful
- [HUE-3185](#) - [oozie] Avoid extra API calls for parent information in workflow dashboard
- [HUE-3185](#) - Revert "[oozie] Avoid extra API calls for parent information in workflow dashboard"
- [HUE-3185](#) - [oozie] Avoid extra API calls for parent information in workflow dashboard
- [HUE-3437](#) - [core] PamBackend does not honor ignore_username_case
- [IMPALA-2378](#) - check proc mem limit before preparing fragment
- [IMPALA-2612](#) - Free local allocations once for every row batch when building hash tables.
- [IMPALA-2711](#) - Fix memory leak in Rand().
- [IMPALA-2722](#) - Free local allocations per row batch in non-partitioned AGG and HJ
- [OOZIE-2429](#) - TestEventGeneration test is unreliable
- [OOZIE-2466](#) - Repeated failure of TestMetricsInstrumentation.testSamplers
- [OOZIE-2486](#) - TestSLAEVENTSGetForFilterJPAExecutor is unreliable
- [SENTRY-780](#) - HDFS Plugin should not execute path callbacks for views

CDH 5 Release Notes

- [SENTRY-1184](#) - Clean up HMSPaths.renameAuthzObject
- [SENTRY-1292](#) - Reorder DBModelAction EnumSet
- [SENTRY-1293](#) - Avoid converting string permission to Privilege object
- [SOLR-6631](#) - DistributedQueue spinning on calling zookeeper getChildren()
- [SOLR-6820](#) - Make the number of version buckets used by the UpdateLog configurable as increasing beyond the default 256 has been shown to help with high volume indexing performance in SolrCloudIncrease the default number of buckets to 65536 instead of 256fix numVersionBuckets name attribute in configsets
- [SOLR-7332](#) - Initialize the highest value for all version buckets with the max value from the index or recent updates to avoid unnecessary lookups to the index to check for reordered updates when processing new documents.
- [SOLR-7587](#) - TestSpellCheckResponse stalled and never timed out
- [SOLR-7625](#) - Version bucket seed not updated after new index is installed on a replica
- [SOLR-8152](#) - Overseer Task Processor/Queue can miss responses, leading to timeouts
- [SOLR-8451](#) - Fix backport
- [SOLR-8451](#) - We should not call method.abort in HttpSolrClient or HttpSolrCall#remoteQuery and HttpSolrCall#remoteQuery should not close streams.
- [SOLR-8453](#) - Solr should attempt to consume the request inputstream on errors as we cannot count on the container to do it.
- [SOLR-8578](#) - Successful or not, requests are not always fully consumed by Solrj clients and we count on HttpClient or the JVM.
- [SOLR-8633](#) - DistributedUpdateProcess processCommit/deleteByQuery calls finish on DUP and SolrCmdDistributor, which violates the lifecycle and can cause bugs.
- [SOLR-8683](#) - Tune down stream closed logging
- [SOLR-8683](#) - Always consume the full request on the server, not just in the case of an error.
- [SOLR-8855](#) - The HDFS BlockDirectory should not clean up its cache on shutdown.
- [SOLR-8856](#) - Do not cache merge or 'read once' contexts in the hdfs block cache.
- [SOLR-8857](#) - HdfsUpdateLog does not use configured or new default number of version buckets and is hard coded to 256.
- [SOLR-8869](#) - Optionally disable printing field cache entries in SolrFieldCacheMBean
- [SPARK-12087](#) - Create new JobConf for every batch in saveAsHadoopFiles

Issues Fixed in CDH 5.4.10

CDH 5.4.10 fixes the following issues.

Apache Hadoop

FSImage may get corrupted after deleting snapshot

Bug: [HDFS-9406](#)

When deleting a snapshot that contains the last record of a given INode, the fsimage may become corrupt because the create list of the snapshot diff in the previous snapshot and the child list of the parent INodeDirectory are not cleaned.

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.4.10:

- [FLUME-2712](#) - Optional channel errors slows down the Source to Main channel event rate
- [FLUME-2886](#) - Optional Channels can cause OOMs
- [HADOOP-10406](#) - TestIPC.testIpcWithReaderQueuing may fail
- [HADOOP-10668](#) - TestZKFailoverControllerStress#testExpireBackAndForth occasionally fails
- [HADOOP-11218](#) - Add TLSv1.1,TLSv1.2 to KMS, HttpFS, SSLFactory
- [HADOOP-12200](#) - TestCryptoStreamsWithOpensslAesCtrCryptoCodec should be skipped in non-native profile
- [HADOOP-12240](#) - Fix tests requiring native library to be skipped in non-native profile
- [HADOOP-12280](#) - Skip unit tests based on Maven profile rather than NativeCodeLoader.isNativeCodeLoaded
- [HADOOP-12417](#) - TestWebDelegationToken failing with port in use

- [HADOOP-12418](#) - TestRPC.testRPCInterruptedSimple fails intermittently
- [HADOOP-12464](#) - Interrupted client may try to fail-over and retry
- [HADOOP-12468](#) - Partial group resolution failure should not result in user lockout.
- [HADOOP-12474](#) - MiniKMS should use random ports for Jetty server by default
- [HADOOP-12559](#) - KMS connection failures should trigger TGT renewal
- [HADOOP-12604](#) - Exception may be swallowed in KMSClientProvider.
- [HADOOP-12605](#) - Fix intermittent failure of TestIPC.testIpcWithReaderQueuing
- [HADOOP-12668](#) - Support excluding weak Ciphers in HttpServer2 through ssl-server.conf
- [HADOOP-12682](#) - Fix TestKMS#testKMSRestart* failure
- [HADOOP-12699](#) - TestKMS#testKMSProvider intermittently fails during 'test rollover draining'
- [HADOOP-12715](#) - TestValueQueue#testgetAtMostPolicyALL fails intermittently
- [HADOOP-12736](#) - TestTimedOutTestsListener#testThreadDumpAndDeadlocks sometimes times out
- [HADOOP-12788](#) - OpensslAesCtrCryptoCodec should log which random number generator is used
- [HDFS-6533](#) - TestBPOfferService#testBasicFunctionalitytest fails intermittently
- [HDFS-7553](#) - fix the TestDFSUpgradeWithHA due to BindException
- [HDFS-8647](#) - Abstract BlockManager's rack policy into BlockPlacementPolicy
- [HDFS-9083](#) - Replication violates block placement policy
- [HDFS-9092](#) - Nfs silently drops overlapping write requests and causes data copying to fail
- [HDFS-9289](#) - Make DataStreamer#block thread safe and verify genStamp in commitBlock
- [HDFS-9313](#) - Possible NullPointerException in BlockManager if no excess replica can be chosen
- [HDFS-9347](#) - Invariant assumption in TestQuorumJournalManager.shutdown() is wrong
- [HDFS-9358](#) - TestNodeCount#testNodeCount timed out
- [HDFS-9406](#) - FSImage may get corrupted after deleting snapshot
- [HDFS-9445](#) - Datanode may deadlock while handling a bad volume
- [HDFS-9688](#) - Test the effect of nested encryption zones in HDFS downgrade
- [HDFS-9721](#) - Allow Delimited PB OIV tool to run upon fsimage that contains INodeReference
- [MAPREDUCE-6302](#) - Incorrect headroom can lead to a deadlock between map and reduce allocations
- [MAPREDUCE-6460](#) - TestRMContainerAllocator.testAttemptNotFoundCausesRMCommunicatorException fails
- [YARN-2902](#) - Killing a container that is localizing can orphan resources in the DOWNLOADING state
- [YARN-4155](#) - TestLogAggregationService.testLogAggregationServiceWithInterval failing
- [YARN-4204](#) - ConcurrentModificationException in FairSchedulerQueueInfo
- [YARN-4347](#) - Resource manager fails with Null pointer exception
- [YARN-4380](#) - TestResourceLocalizationService.testDownloadingResourcesOnContainerKill fails intermittently
- [YARN-4393](#) - Fix intermittent test failure for TestResourceLocalizationService#testFailedDirsResourceRelease
- [YARN-4573](#) - Fix test failure in TestRMAppTransitions#testAppRunningKill and testAppKilledKilled
- [YARN-4613](#) - Fix test failure in TestClientRMSERVICE#testGetClusterNodes
- [HBASE-14205](#) - RegionCoprocessorHost System.nanoTime() performance bottleneck
- [HBASE-14621](#) - ReplicationLogCleaner stuck on RS crash
- [HBASE-14923](#) - VerifyReplication should not mask the exception during result comparison
- [HBASE-14926](#) - Hung ThriftServer; no timeout on read from client; if client crashes, worker thread gets stuck reading
- [HBASE-15019](#) - Replication stuck when HDFS is restarted
- [HBASE-15031](#) - Fix merge of MVCC and SequenceID performance regression in branch-1.0
- [HBASE-15032](#) - hbase shell scan filter string assumes UTF-8 encoding
- [HBASE-15035](#) - bulkloading hfiles with tags that require splits do not preserve tags
- [HBASE-15052](#) - Use EnvironmentEdgeManager in ReplicationSource
- [HBASE-15104](#) - Occasional failures due to NotServingRegionException in IT tests
- [HBASE-15157](#) - Add *PerformanceTest for Append, CheckAnd*
- [HBASE-15213](#) - Fix increment performance regression caused by HBASE-8763 on branch-1.0
- [HIVE-7575](#) - GetTables thrift call is very slow

- [HIVE-10213](#) - MapReduce jobs using dynamic-partitioning fail on commit
- [HIVE-10514](#) - Fix MiniCliDriver tests failure
- [HIVE-11826](#) - 'hadoop.proxyuser.hive.groups' configuration does not prevent unauthorized user to access metastore
- [HIVE-11828](#) - beeline -f fails on scripts with tabs between column type and comment
- [HIVE-11977](#) - Hive should handle an external avro table with zero length files present
- [HIVE-12008](#) - Hive queries failing when using count(*) on column in view
- [HIVE-12388](#) - GetTables cannot get external tables when TABLE type argument is given
- [HIVE-12505](#) - Insert overwrite in same encrypted zone silently fails to remove some existing files
- [HIVE-12566](#) - Incorrect result returns when using COALESCE in WHERE condition with LEFT JOIN
- [HIVE-12713](#) - Miscellaneous improvements in driver compile and execute logging
- [HIVE-12784](#) - Group by SemanticException: Invalid column reference
- [HIVE-12790](#) - Metastore connection leaks in HiveServer2
- [HIVE-12795](#) - Vectorized execution causes ClassCastException
- [HIVE-12946](#) - alter table should also add default scheme and authority for the location similar to create table
- [HIVE-13039](#) - BETWEEN predicate is not functioning correctly with predicate pushdown on Parquet table
- [HIVE-13065](#) - Hive throws NPE when writing map type data to a HBase backed table
- [HUE-3106](#) - [filebrowser] Add support for full paths in zip file uploads
- [HUE-3110](#) - [oozie] Fix bundle submission when coordinator points to multiple bundles
- [HUE-3180](#) - [useradmin] Override duplicate username validation message
- [IMPALA-1702](#) - Check for duplicate table IDs at the end of analysis (issue not entirely fixed, but now fails gracefully)
- [IMPALA-2264](#) - Implicit casts to integers from decimals with higher precision sometimes allowed
- [IMPALA-2473](#) - Excessive memory usage by scan nodes
- [IMPALA-2621](#) - Fix flaky UNIX_TIMESTAMP() test
- [IMPALA-2643](#) - Nested inline view produces incorrect result when referencing the same column implicitly
- [IMPALA-2765](#) - AnalysisException: operands of type BOOLEAN and TIMESTAMP are not comparable when OUTER JOIN with CASE statement
- [IMPALA-2798](#) - After adding a column to avro table, Impala returns weird result if codegen is enabled.
- [IMPALA-2861](#) - Fix flaky scanner test added via IMPALA-2473 backport
- [IMPALA-2914](#) - Hit DCHECK Check failed: HasDateOrTime()
- [IMPALA-3034](#) - MemTracker leak on PHJ failure to spill
- [IMPALA-3085](#) - DataSinks' MemTrackers need to unregister themselves from parent
- [IMPALA-3093](#) - ReopenClient() could NULL out 'client_key' causing a crash
- [IMPALA-3095](#) - Allow additional Kerberos users to be authorized to access internal APIs
- [KITE-1114](#) - fix test
- [KITE-1114](#) - Fix missing license header
- [KITE-1114](#) - Kite CLI json-import HDFS temp file path not multiuser safe
- [OOZIE-2413](#) - Kerberos credentials can expire if the KDC is slow to respond
- [OOZIE-2428](#) - TestSLAService, TestSLAEVENTGeneration flaky tests
- [OOZIE-2432](#) - TestPurgeXCommand fails
- [OOZIE-2435](#) - TestCoordChangeXCommand is flaky
- [SENTRY-835](#) - Drop table leaves a connection open when using MetastoreListener
- [SENTRY-885](#) - DB name should be case insensitive in HDFS sync plugin
- [SENTRY-944](#) - Setting HDFS rules on Sentry managed hdfs paths should not affect original hdfs rules
- [SENTRY-953](#) - External Partitions which are referenced by more than one table can cause some unexpected behavior with Sentry HDFS sync
- [SENTRY-957](#) - Exceptions in MetastoreCacheInitializer should probably not prevent HMS from starting up
- [SENTRY-988](#) - It's better to let SentryAuthorization setter path always fall through and update HDFS
- [SENTRY-991](#) - backportRoles of Sentry Permission needs to be case insensitive
- [SENTRY-994](#) - SentryAuthorizationInfoX should override isSentryManaged
- [SENTRY-1002](#) - PathsUpdate.parsePath(path) will throw an NPE when parsing relative paths

- [SENTRY-1003](#) - backportSupport "reload" by updating the classpath of Sentry function aux jar path during run time
- [SENTRY-1008](#) - Path should be not be updated if the create/drop table/partition event fails
- [SENTRY-1044](#) - Tables with non-HDFS locations breaks HMS startup
- [SOLR-7281](#) - Add an overseer action to publish an entire node as 'down'
- [SOLR-8367](#) - Fix the LeaderInitiatedRecovery 'all replicas participate' fail-safe
- [SOLR-8371](#) - Try and prevent too many recovery requests from stacking up and clean up some faulty cancel recovery logic
- [SOLR-8372](#) - backportCanceled recovery can lead to data loss
- [SOLR-8575](#) - Addendum to Fix HDFSLogReader replay
- [SOLR-8575](#) - Fix HDFSLogReader replay status numbers and a performance bug where we can reopen FSDataInputStream too often
- [SOLR-8615](#) - Just like creating cores, we should use multiple threads when closing cores
- [SOLR-8720](#) - ZkController#publishAndWaitForDownStates should use #publishNodeAsDown
- [SOLR-8771](#) - Multithreaded core shutdown creates executor per core
- [SQOOP-2847](#) - Sqoop --incremental + missing parent --target-dir reports success with no data
- [SQOOP-2422](#) - Sqoop2: Test TestJSONIntermediateDataFormat is failing on JDK8
- [ZOOKEEPER-442](#) - Need a way to remove watches that are no longer of interest"

Issues Fixed in CDH 5.4.9

Known Issues Fixed

The following topics describe known issues fixed in CDH 5.4.9.

Apache Commons Collections deserialization vulnerability

Cloudera has learned of a potential security vulnerability in a third-party library called the [Apache Commons Collections](#). This library is used in products distributed and supported by Cloudera ("Cloudera Products"), including core Apache Hadoop. The Apache Commons Collections library is also in widespread use beyond the Hadoop ecosystem. At this time, no specific attack vector for this vulnerability has been identified as present in Cloudera Products.

In an abundance of caution, we are currently in the process of incorporating a version of the Apache Commons Collections library with a fix into the Cloudera Products. In most cases, this will require coordination with the projects in the Apache community. One example of this is tracked by [HADOOP-12577](#).

The Apache Commons Collections potential security vulnerability is titled "Arbitrary remote code execution with InvokerTransformer" and is tracked by [COLLECTIONS-580](#). MITRE has not issued a CVE, but related [CVE-2015-4852](#) has been filed for the vulnerability. CERT has issued [Vulnerability Note #576313](#) for this issue.

Releases affected: CDH 5.5.0, CDH 5.4.8 and lower, CDH 5.3.8 and lower, CDH 5.2.8 and lower, CDH 5.1.7 and lower, Cloudera Manager 5.5.0, Cloudera Manager 5.4.8 and lower, Cloudera Manager 5.3.8 and lower, and Cloudera Manager 5.2.8 and lower, Cloudera Manager 5.1.6 and lower, Cloudera Navigator 2.4.0, Cloudera Navigator 2.3.8 and lower.

Users affected: All

Impact: This potential vulnerability may enable an attacker to execute arbitrary code from a remote machine without requiring authentication.

Immediate action required: Upgrade to Cloudera Manager 5.5.1 and CDH 5.5.1, Cloudera Manager 5.4.9 and CDH 5.4.9, Cloudera Manager 5.3.9 and CDH 5.3.9, and Cloudera Manager 5.2.9 and CDH 5.2.9, and Cloudera Manager 5.1.7 and CDH 5.1.7.

Apache HBase

Data may not be replicated to worker cluster if multiwal multiplicity is set to greater than 1

Bug: [HBASE-13703](#), [HBASE-6617](#), [HBASE-14501](#).

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.4.9:

CDH 5 Release Notes

- [FLUME-2841](#) - Upgrade commons-collections to 3.2.2
- [HADOOP-7713](#) - dfs -count -q should label output column
- [HADOOP-11171](#) - Enable using a proxy server to connect to S3a
- [HADOOP-12568](#) - Update core-default.xml to describe posixGroups support
- [HADOOP-12577](#) - Bumped up commons-collections version to 3.2.2 to address a security flaw
- [HDFS-7785](#) - Improve diagnostics information for HttpPutFailedException
- [HDFS-7798](#) - Checkpointing failure caused by shared KerberosAuthenticator
- [HDFS-7871](#) - NameNodeEditLogRoller can keep printing 'Swallowing exception' message
- [HDFS-7990](#) - IBR delete ack should not be delayed
- [HDFS-8646](#) - Prune cached replicas from DatanodeDescriptor state on replica invalidation
- [HDFS-9123](#) - Copying from the root to a subdirectory should be forbidden
- [HDFS-9250](#) - Add Precondition check to LocatedBlock#addCachedLoc
- [HDFS-9273](#) - ACLs on root directory may be lost after NN restart
- [HDFS-9332](#) - Fix Precondition failures from NameNodeEditLogRoller while saving namespace
- [HDFS-9364](#) - Unnecessary DNS resolution attempts when creating NameNodeProxies
- [HDFS-9470](#) - Encryption zone on root not loaded from fsimage after NN restart
- [MAPREDUCE-6191](#) - Improve clearing stale state of Java serialization
- [MAPREDUCE-6549](#) - Multibyte delimiters with LineRecordReader cause duplicate records
- [YARN-4235](#) - FairScheduler PrimaryGroup does not handle empty groups returned for a user
- [HBASE-6617](#) - ReplicationSourceManager should be able to track multiple WAL paths
- [HBASE-12865](#) - WALs may be deleted before they are replicated to peers
- [HBASE-13134](#) - mutateRow and checkAndMutate APIs don't throw region level exceptions
- [HBASE-13618](#) - ReplicationSource is too eager to remove sinks
- [HBASE-13703](#) - ReplicateContext should not be a member of ReplicationSource
- [HBASE-14003](#) - Work around JDK-8044053
- [HBASE-14283](#) - Reverse scan doesn't work with HFile inline index/bloom blocks
- [HBASE-14374](#) - Backport parent 'HBASE-14317 Stuck FSHLog' issue to 1.1
- [HBASE-14501](#) - NPE in replication with TDE
- [HBASE-14533](#) - Connection Idle time 1 second is too short and the connection is closed too quickly by the ChoreService
- [HBASE-14547](#) - Add more debug/trace to zk-procedure
- [HBASE-14799](#) - Commons-collections object deserialization remote command execution vulnerability
- [HBASE-14809](#) - Grant / revoke Namespace admin permission to group
- [HIVE-10265](#) - Hive CLI crashes on != inequality
- [HIVE-11149](#) - Sometimes HashMap in PerfLogger.java hangs
- [HIVE-11616](#) - DelegationTokenSecretManager reuses the same objectstore, which has concurrency issues
- [HIVE-12058](#) - Change hive script to record errors when calling hbase fails
- [HIVE-12188](#) - DoAs does not work properly in non-Kerberos secured HS2
- [HIVE-12189](#) - The list in pushdownPreds of ppd.ExprWalkerInfo should not be allowed to grow very large
- [HIVE-12250](#) - ZooKeeper connection leaks in Hive's HBaseHandler
- [HIVE-12365](#) - Added resource path is sent to cluster as an empty string when externally removed
- [HIVE-12378](#) - Exception on HBaseSerDe.serialize binary field
- [HIVE-12406](#) - HIVE-9500 introduced incompatible change to LazySimpleSerDe public interface
- [HIVE-12418](#) - HiveHBaseTableInputFormat.getRecordReader() causes ZooKeeper connection leak
- [HUE-2941](#) - [hadoop] Cache the active RM HA
- [HUE-3035](#) - [beeswax] Optimize sample data query for partitioned tables
- [IMPALA-1459](#) - Fix migration/assignment of On-clause predicates inside inline views
- [IMPALA-1675](#) - Avoid overflow when adding large intervals to TIMESTAMPs
- [IMPALA-1746](#) - QueryExecState does not check for query cancellation or errors
- [IMPALA-1917](#) - Do not register aux equivalence predicates with NULL on either side

- [IMPALA-1949](#) - Analysis exception when a binary operator contain an IN operator with
- [IMPALA-2086/IMPALA-2090](#) - Avoid boost year/month interval logic
- [IMPALA-2141](#) - UnionNode::GetNext() does not check for query errors
- [IMPALA-2252](#) - Crash (likely race) tearing down BufferedBlockMgr on query failure
- [IMPALA-2260](#) - Adding a large hour interval caused an interval overflow
- [IMPALA-2265](#) - Sorter was not checking the returned Status of PrepareRead
- [IMPALA-2273](#) - Make MAX_PAGE_HEADER_SIZE configurable
- [IMPALA-2286](#) - Fix race between ~BufferedBlockMgr() and BufferedBlockMgr::Create()
- [IMPALA-2344](#) - Work-around IMPALA-2344 Fail query with OOM in case block->Pin() fails
- [IMPALA-2357](#) - Fix spilling sorts with var-len slots that are NULL or empty
- [IMPALA-2446](#) - Fix wrong predicate assignment in outer joins
- [IMPALA-2533](#) - Impala throws IllegalStateException when inserting data into a partition
- [IMPALA-2559](#) - Fix check failed: sorter_runs_.back()->is_pinned_
- [IMPALA-2664](#) - Avoid sending large partition stats objects over thrift
- [IMPALA-2731](#) - Refactor MemPool usage in HBase scan node
- [KITE-1089](#) - readAvroContainer morphline command should work even if the Avro writer schema of each input file is different
- [PIG-3641](#) - Split "otherwise" producing incorrect output when combined with ColumnPruning
- [SENTRY-565](#) - Improve performance of filtering Hive SHOW commands
- [SENTRY-702](#) - Hive binding should support RELOAD command
- [SENTRY-936](#) - getGroup and getUser should always return orginal hdfs values for paths in prefixes which are not Sentry managed
- [SENTRY-960](#) - Blacklist reflect, java_method using hive.server2.builtin.udf.blacklist
- [SOLR-6443](#) - backportDisable test that fails on Jenkins with SolrCore.getOpenCount()==2
- [SOLR-7049](#) - LIST Collections API call should be processed directly by the CollectionsHandler instead of the OverseerCollectionProcessor
- [SOLR-7552](#) - Support using ZkCredentialsProvider/ZkACLProvider in custom filter
- [SOLR-7989](#) - After a new leader is elected, it should ensure it's state is ACTIVE if it has already registered with ZK
- [SOLR-8075](#) - Leader Initiated Recovery should not stop a leader that participated in an election with all of its replicas from becoming a valid leader
- [SOLR-8223](#) - Avoid accidentally swallowing OutOfMemoryError
- [SOLR-8288](#) - DistributedUpdateProcessor#doFinish should explicitly check and ensure it does not try to put itself into LIR
- [SPARK-11484](#) - [WEBUI] Using proxyBase set by spark AM
- [SPARK-11652](#) - [CORE] Remote code execution with InvokerTransformer

Issues Fixed in CDH 5.4.8

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.4.8:

- [FLUME-2095](#) - JMS source with TIBCO
- [HADOOP-11261](#) - Set custom endpoint for S3A
- [HADOOP-12404](#) - Disable caching for JarURLConnection to avoid sharing JarFile with other users when loading resource from URL in Configuration class
- [HADOOP-12413](#) - AccessControlList should avoid calling getGroupNames in isUserInList with empty groups
- [HDFS-7916](#) - 'reportBadBlocks' from datanodes to standby Node BPServiceActor goes for infinite loop
- [HDFS-7978](#) - Add LOG.isDebugEnabled() guard for some LOG.debug()
- [HDFS-8384](#) - Allow NN to startup if there are files having a lease but are not under construction
- [HDFS-8735](#) - Inotify: All events classes should implement toString() API
- [HDFS-8860](#) - Remove unused Replica copyOnWrite code
- [HDFS-8964](#) - When validating the edit log, do not read at or beyond the file offset that is being written

CDH 5 Release Notes

- [HDFS-8965](#) - Harden edit log reading code against out of memory errors
- [MAPREDUCE-5918](#) - LineRecordReader can return the same decompressor to CodecPool multiple times
- [MAPREDUCE-5948](#) - org.apache.hadoop.mapred.LineRecordReader does not handle multibyte record delimiters well
- [MAPREDUCE-6273](#) - HistoryFileManager should check whether summaryFile exists to avoid FileNotFoundException causing HistoryFileInfo into MOVE_FAILED state
- [MAPREDUCE-6481](#) - LineRecordReader may give incomplete record and wrong position/key information for uncompressed input sometimes
- [MAPREDUCE-6484](#) - YARN Client uses local address instead of RM address as token renewer in a secure cluster when RM HA is enabled
- [YARN-2666](#) - TestFairScheduler.testContinuousScheduling fails intermittently
- [YARN-3385](#) - Fixed a race-condition in ResourceManager's ZooKeeper based state-store to avoid crashing on duplicate deletes
- [YARN-3469](#) - ZKRMStateStore: Avoid setting watches that are not required.
- [YARN-3943](#) - Use separate threshold configurations for disk-full detection and disk-not-full detection
- [HBASE-13217](#) - Procedure fails due to ZK issue
- [HBASE-13331](#) - Exceptions from DFS client can cause CatalogJanitor to delete referenced files
- [HBASE-13388](#) - Handling NullPointer in ZKProcedureMemberRpcs while getting ZNode data
- [HBASE-13933](#) - DBE's seekBefore with tags corrupts the tag's offset information thus leading to incorrect results
- [HBASE-14196](#) - Thrift server idle connection timeout issue
- [HBASE-14302](#) - TableSnapshotInputFormat should not create back references when restoring snapshot
- [HBASE-14347](#) - Add a switch to DynamicClassLoader to disable it
- [HBASE-14385](#) - Close the sockets that is missing in connection closure
- [HBASE-14394](#) - Properly close the connection after reading records from table
- [HBASE-14471](#) - Thrift - HTTP Error 413 full HEAD if using Kerberos authentication
- [HBASE-14492](#) - Increase REST server header buffer size from 8k to 64k
- [HIVE-5545](#) - HCatRecord getInteger method returns String when used on Partition columns of type INT
- [HIVE-8529](#) - HiveSessionImpl#fetchResults should not try to fetch operation log when hive.server2.logging.operation.enabled is false
- [HIVE-9867](#) - Migrate usage of deprecated Calcite methods
- [HIVE-9984](#) - JoinReorder's getOutputSize is exponential
- [HIVE-10021](#) - "Alter index rebuild" statements submitted through HiveServer2 fail when Sentry is enabled
- [HIVE-10122](#) - Hive metastore filter-by-expression is broken for non-partition expressions
- [HIVE-10421](#) - DROP TABLE with qualified table name ignores database name when checking partitions
- [HIVE-10451](#) - PTF deserializer fails if values are not used in reducer
- [HIVE-10658](#) - Insert with values clause may expose data that should be encrypted
- [HIVE-10980](#) - Merge of dynamic partitions loads all data to default partition
- [HIVE-11077](#) - Part of Exchange partition does not properly populate fields for post/pre execute hooks.
- [HIVE-11440](#) - Create Parquet predicate push down (PPD) unit tests and q-tests
- [HIVE-11504](#) - Predicate pushing down does not work for float type for Parquet
- [HIVE-11590](#) - AvroDeserializer is very chatty
- [HIVE-11618](#) - BackportCorrect the SARG api to reunify the PredicateLeaf.Type INTEGER and LONG
- [HIVE-11657](#) - HIVE-2573 introduces some issues during metastore init (and CLI init)
- [HIVE-11695](#) - If user has no permission to create LOCAL DIRECTORY, the Hql does not throw any exception and fails silently
- [HIVE-11696](#) - Exception when table-level serde is Parquet while partition-level serde is JSON
- [HIVE-11712](#) - Duplicate groupby keys cause ClassCastException
- [HIVE-11737](#) - IndexOutOfBoundsException compiling query with duplicated groupby keys
- [HIVE-11745](#) - Alter table Exchange partition with multiple partition_spec is not working
- [HIVE-11816](#) - Upgrade groovy to 2.4.4
- [HIVE-11824](#) - Insert to local directory causes staging directory to be copied

- [HIVE-11843](#) - Add 'sort by c' to Parquet PPD q-tests to avoid different output issues with hadoop-1
- [HIVE-11891](#) - Add basic performance logging to metastore calls
- [HIVE-11926](#) - Backport:Stats annotation might not extract stats for varchar/decimal columns
- [HIVE-11982](#) - Some test cases for union all fail with recent changes
- [HIVE-11995](#) - Remove repetitively setting permissions in insert/load overwrite partition
- [HUE-2881](#) - [oozie] A fork can point to a deleted node
- [IMPALA-1136](#) - Support loading Avro tables without an explicit Avro schema
- [IMPALA-1899](#) - Cleanup handling of Hive's field schema3e0fee5 IMPALA-2369, IMPALA-2435: Impala crashes when the sorter hits an OOM error
- [IMPALA-2130](#) - Wrong verification of Parquet file version
- [IMPALA-2161](#) - Skip \u0000 characters when dealing Avro schemas
- [IMPALA-2165](#) - Avoid cardinality 0 in scan nodes of small tables and low selectivity
- [IMPALA-2168](#) - Do not try to access streams of repartitioned spilled partition in right-joins
- [IMPALA-2213](#) - Make Parquet scanner fail query if the file size metadata is stale
- [IMPALA-2249](#) - Avoid allocating StringBuffer > 1GB in ScannerContext::Stream::GetBytesInternal()
- [IMPALA-2256](#) - Handle joins with right side of high cardinality and zero materialized slots
- [IMPALA-2270](#) - Avoid FnvHash64to32 with empty inputs
- [IMPALA-2284](#) - Disallow long (1<<30) strings in group_concat()
- [IMPALA-2292](#) - Change the type of timestamp_col to string in the table no_avro_schema.
- [IMPALA-2314](#) - LargestSpilledPartition was not checking if partition is closed
- [IMPALA-2348](#) - The catalog does not close the connection to HMS during table invalidation
- [IMPALA-2364](#) - Wrong DCHECK in PHJ::ProcessProbeBatch
- [IMPALA-2366](#) - Check fread return code correctly
- [IMPALA-2440](#) - Fix old HJ full outer join with no rows
- [IMPALA-2477](#) - Parquet metadata randomly 'appears stale'
- [IMPALA-2514](#) - DCHECK on destroying an ExprContext
- [KITE-1069](#) - Make zkClientSessionTimeout and zkClientConnectTimeout configurable in SolrLocator
- [KITE-1074](#) - Partial updates aka Atomic updates with loadSolr aren't recognized with SolrCloud
- [MAHOUT-1771](#) - Cluster dumper omits indices and 0 elements for dense vector or sparse containing 0s
- [OOZIE-2376](#) - Default action configs not honored if no <configuration> section in workflow
- [SENTRY-878](#) - collect_list missing from HIVE_UDF_WHITE_LIST
- [SENTRY-884](#) - Give execute permission by default to paths managed by sentry
- [SENTRY-893](#) - Synchronize calls in SentryClient and create sentry client once per request in SimpleDBProvider
- [SOLR-5776](#) - Use less TLS/SSL in a test run
- [SOLR-7109](#) - Indexing threads stuck during network partition can put leader into down state
- [SOLR-7844](#) - Zookeeper session expiry during shard leader election can cause multiple leaders
- [SOLR-7956](#) - There are interrupts on shutdown in places that can cause ChannelAlreadyClosed exceptions which prevents proper closing of transaction logs and can poison the IndexWriter and interfere with the HDFS client
- [SOLR-8046](#) - HdfsCollectionsAPIDistributedZkTest checks that no transaction logs failed to be opened during the test but does not isolate this to the test and could fail due to other tests
- [SOLR-8069](#) - Ensure that only the valid ZooKeeper registered leader can put a replica into Leader Initiated Recovery
- [SOLR-8075](#) - Leader Initiated Recovery should not stop a leader that participated in an election with all of it's replicas from becoming a valid leader
- [SOLR-8077](#) - Replication can still cause index corruption
- [SOLR-8085](#) - Fix a variety of issues that can result in replicas getting out of sync
- [SOLR-8094](#) - HdfsUpdateLog should not replay buffered documents as a replacement to dropping them
- [SOLR-8095](#) - Add enable prop for HDFS Locality Metrics
- [SOLR-8121](#) - It looks like ChaosMonkeySafeLeader test can fail with replica inconsistency because waitForThingsToLevelOut can pass while state is still changing.
- [SPARK-6880](#) - Spark Shutdowns with NoSuchElementException when running parallel collect on cachedRDD

CDH 5 Release Notes

- [SQOOP-2597](#) - Missing method AvroSchemaGenerator.generate()

Published Known Issues Fixed

As a result of the above fixes, the following issues, previously published as [Known Issues in CDH 5](#) on page 111, are also fixed.

Spurious warning in MRv1 jobs

The `mapreduce.client.genericoptionsparser.used` property is not correctly checked by `JobClient` and this leads to a spurious warning.

Bug: None

Workaround: MapReduce jobs using `GenericOptionsParser` or implementing `Tool` can remove the warning by setting this property to true.

Spark Sink requires spark-assembly.jar in Flume classpath

In CDH 5.4.0, Flume requires `spark-assembly.jar` in the Flume classpath to use the Spark Sink. Without this, the sink fails with a dependency issue.

Bug: [SPARK-7038](#)

Workaround: Use the Spark Sink from CDH 5.3 with Spark from CDH 5.4, or add `spark-assembly.jar` to the `FLUME_CLASSPATH`.

Streaming incompatibility between Spark 1.2 and 1.3

Applications built as a JAR with dependencies ("uber JAR") must be built for the specific version of Spark running on the cluster.

Workaround: Rebuild the JAR with the Spark dependencies in `pom.xml` pointing to the specific version of Spark running on the target cluster.

Configuring more than one NT domain does not work in CDH 5.4.0

Trying to add users and groups using the multi-NT domain feature (<http://gethue.com/hadoop-tutorial-make-hadoop-more-accessible-by-integrating-multiple-ldap-servers/>) produces an error.

Bug: [HUE-2665](#)

Workaround: None.

If Sentry is enabled, the RELOAD command cannot be executed in the Hive CLI or Beeline.

Bug: [SENTRY-702](#)

Workaround: None.

Issues Fixed in CDH 5.4.7

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.4.7:

- [CRUNCH-525](#) - Correct (more) accurate default scale factors for built-in MapFn implementations
- [CRUNCH-527](#) - Use hash smearing for partitioning
- [CRUNCH-528](#) - Improve Pair comparison
- [CRUNCH-531](#) - Fix split graph rendering typo.
- [CRUNCH-535](#) - call initCredentials on the job
- [CRUNCH-536](#) - Refactor CrunchControlledJob.Hook interface and make it client-accessible
- [CRUNCH-539](#) - Fix reading WritableComparables bimap
- [CRUNCH-540](#) - Make AvroReflectDeepCopier serializable
- [CRUNCH-543](#) - Have AvroPathPerKeyTarget handle child directories properly
- [CRUNCH-544](#) - Improve performance/serializability of materialized toMap.

- [CRUNCH-546](#) - Remove calls to CellUtil.cloneXXX
- [CRUNCH-547](#) - Properly handle nullability for Avro union types
- [CRUNCH-548](#) - Have the AvroReflectDeepCopier use the class of the source object when constructing new instances instead of the target class
- [CRUNCH-551](#) - Make the use of Configuration objects consistent in CrunchInputSplit and CrunchRecordReader
- [CRUNCH-553](#) - Fix record drop issue that can occur w/From.formattedFile TableSources
- [FLUME-1934](#) - Spooling Directory Source dies on encountering zero-byte files.
- [FLUME-2753](#) - Error when specifying empty replace string in Search and Replace Interceptor
- [HADOOP-12317](#) - Applications fail on NM restart on some linux distro because NM container recovery declares AM container as LOST
- [HDFS-8806](#) - Inconsistent metrics: number of missing blocks with replication factor 1 not properly cleared
- [HDFS-8850](#) - VolumeScanner thread exits with exception if there is no block pool to be scanned but there are suspicious blocks.
- [MAPREDUCE-5817](#) - Mappers get rescheduled on node transition even after all reducers are completed.
- [MAPREDUCE-6277](#) - Job can post multiple history files if attempt loses connection to the RM
- [MAPREDUCE-6439](#) - AM may fail instead of retrying if RM shuts down during the allocate call.
- [YARN-2921](#) - Fix MockRM/MockAM#waitForState sleep too long.
- [YARN-3823](#) - Fix mismatch in default values for yarn.scheduler.maximum-allocation-vcores property
- [YARN-3990](#) - AsyncDispatcher may overloaded with RMAppNodeUpdateEvent when Node is connected/disconnected
- [HBASE-13329](#) - ArrayIndexOutOfBoundsException in CellComparator#getMinimumMidpointArray.
- [HBASE-13437](#) - ThriftServer leaks ZooKeeper connections
- [HBASE-13471](#) - Fix a possible infinite loop in doMiniBatchMutation
- [HBASE-13684](#) - Allow mlockagent to be used when not starting as root
- [HBASE-14162](#) - Fixing maven target for regenerating thrift classes fails against 0.9.2
- [HBASE-14354](#) - Minor improvements for usage of the mlock agent
- [HIVE-7476](#) - CTAS does not work properly for s3
- [HIVE-9327](#) - CBO (Calcite Return Path): Removing Row Resolvers from ParseContext
- [HIVE-9512](#) - HIVE-9327 causing regression in stats annotation
- [HIVE-9580](#) - Server returns incorrect result from JOIN ON VARCHAR columns
- [HIVE-9613](#) - Left join query plan outputs wrong column when using subquery
- [HIVE-10085](#) - Lateral view on top of a view throws RuntimeException
- [HIVE-10140](#) - Window boundary is not compared correctly
- [HIVE-10288](#) - Cannot call permanent UDFs
- [HIVE-10319](#) - Hive CLI startup takes a long time with a large number of databases
- [HIVE-10719](#) - Hive metastore failure when alter table rename is attempted.
- [HIVE-10875](#) - Select query with view in subquery adds underlying table as direct input
- [HIVE-10906](#) - Value based UDAF function without orderby expression throws NPE
- [HIVE-10911](#) - Add support for date datatype in the value based windowing function
- [HIVE-10972](#) - DummyTxnManager always locks the current database in shared mode, which is incorrect
- [HIVE-10985](#) - Value based windowing on timestamp and double can't handle NULL value
- [HIVE-10996](#) - Aggregation / Projection over Multi-Join Inner Query producing incorrect results
- [HIVE-11139](#) - PROPOSEDQTest combine2_hadoop20.q fails when using -Phadoop-1 profile due to
- [HIVE-11172](#) - Vectorization wrong results for aggregate query with where clause without group by
- [HIVE-11203](#) - Beeline force option does not force execution when errors occurred in a script.
- [HIVE-11250](#) - Change in spark.executor.instances (and others) does not take effect after RSC is launched for HS2
- [HIVE-11255](#) - get_table_objects_by_name() in HiveMetaStore.java needs to retrieve table objects in multiple batches
- [HIVE-11258](#) - The function drop_database_core() of HiveMetaStore.java may not drop all the tables
- [HIVE-11271](#) - java.lang.IndexOutOfBoundsException when union all with if function
- [HIVE-11288](#) - Avro SerDe InstanceCache returns incorrect schema

CDH 5 Release Notes

- [HIVE-1133](#) - ColumnPruner prunes columns of UnionOperator that should be kept
- [HIVE-11502](#) - Map side aggregation is extremely slow
- [HIVE-11604](#) - HIVE return wrong results in some queries with PTF function
- [HIVE-11620](#) - Fix several qtest output order
- [HUE-2873](#) - [oozie] Handle TransactionManagementError on workflow dashboard
- [HUE-2877](#) - [desktop] Add pyasn1 and ndg_httpsclient to support SSL Server Name Indication
- [HUE-2880](#) - [hadoop] Fix uploading large files to a kerberized HTTPFS
- [HUE-2882](#) - [oozie] Fix parsing error when workflow job uses Australian timezone
- [HUE-2883](#) - [impala] Canceling a query shows an error message
- [HUE-2885](#) - [oozie] Java options java-opts not generated correctly in XML
- [HUE-2893](#) - [desktop] Backport CherryPy SSL file upload fix
- [HUE-2903](#) - [oozie] Fix error with Workflow parameter on rerun
- [IMPALA-1737](#) - Substitute an InsertStmt's partition key exprs with the root node's smap.
- [IMPALA-1756](#) - Constant filter expressions are not checked for errors and state cleanup is not done before throwing exception.
- [IMPALA-1898](#) - Explicit aliases + ordinals analysis bug
- [IMPALA-1983](#) - Warn if table stats are potentially corrupt.
- [IMPALA-1987](#) - Fix TupleIsNotNullPredicate to return false if no tuples are nullable.
- [IMPALA-2088](#) - Fix planning of empty union operands with analytics.
- [IMPALA-2089](#) - Retain eq predicates bound by grouping slots with complex grouping exprs.
- [IMPALA-2178](#) - fix Expr::ComputeResultsLayout() logic.
- [IMPALA-2199](#) - Row count not set for empty partition when spec is used with compute incremental stats
- [IMPALA-2201](#) - Unconditionally update the partition stats and row count.
- [IMPALA-2203](#) - Set an InsertStmt's result exprs from the source statement's result exprs.
- [IMPALA-2216](#) - Set the output smap of an EmptySetNode produced from an empty inline view.
- [IMPALA-2239](#) - update misc.test to match the new .test file format.
- [IMPALA-2266](#) - Pass correct child node in 2nd phase merge aggregation.
- [KITE-1053](#) - Fix int overflow bug in FS writer.
- [SENTRY-810](#) - CTAS without location is not verified properly
- [SOLR-7135](#) - Allow the server build.xml 'sync-hack' target to be skipped by specifying a system property.
- [SOLR-7999](#) - SolrRequestParserTest#testStreamURL started failing.
- [SPARK-8606](#) - Prevent exceptions in RDD.getPreferredLocations() from crashing DAGScheduler
- [ZOOKEEPER-442](#) - need a way to remove watches that are no longer of interest

Issues Fixed in CDH 5.4.5

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.4.5:

- [CRUNCH-508](#) - Improve performance of Scala Enumeration counters in Scrunch
- [CRUNCH-511](#) - Scrunch product type support should use derived() instead of derivedImmutable()
- [CRUNCH-514](#) - AvroDerivedDeepCopier should initialize delegate MapFns
- [CRUNCH-516](#) - Scrunch needs some additional null checks
- [CRUNCH-530](#) - Fix object reuse bug in GenericRecordToTuple
- [CRUNCH-542](#) - Wider tolerance for flaky scrunch PCollectionTest
- [FLUME-2215](#) - ResettableFileInputStream can't support ucs-4 character
- [FLUME-2732](#) - Make maximum tolerated failures before shutting down and recreating client in AsyncHbaseSink configurable
- [FLUME-2738](#) - Async HBase sink FD leak on client shutdown
- [FLUME-2749](#) - Kerberos configuration error when using short names in multiple HDFS Sinks
- [HADOOP-12017](#) - Hadoop archives command should use configurable replication factor when closing
- [HADOOP-12103](#) - Small refactoring of DelegationTokenAuthenticationFilter to allow code sharing

- [HADOOP-8151](#) - Error handling in snappy decompressor throws invalid exceptions
- [HDFS-7501](#) - TransactionsSinceLastCheckpoint can be negative on SBNs
- [HDFS-7546](#) - Document, and set an accepting default for dfs.namenode.kerberos.principal.pattern
- [HDFS-7890](#) - Improve information on Top users for metrics in RollingWindowsManager and lower log level
- [HDFS-7894](#) - Rolling upgrade readiness is not updated in jmx until query command is issued.
- [HDFS-8072](#) - Reserved RBW space is not released if client terminates while writing block
- [HDFS-8337](#) - Accessing https via webhdfs doesn't work from a jar with kerberos
- [HDFS-8656](#) - Preserve compatibility of ClientProtocol#rollingUpgrade after finalization
- [HDFS-8681](#) - BlockScanner is incorrectly disabled by default
- [MAPREDUCE-5965](#) - Hadoop streaming throws error if list of input files is high.
- [YARN-3143](#) - RM Apps REST API can return NPE or entries missing id and other fields
- [YARN-3453](#) - Fair Scheduler: Parts of preemption logic uses DefaultResourceCalculator even in DRF mode causing thrashing
- [YARN-3535](#) - Scheduler must re-request container resources when RMContainer transitions from ALLOCATED to KILLED
- [YARN-3793](#) - Several NPEs when deleting local files on NM recovery
- [YARN-3842](#) - NMProxy should retry on NMNotYetReadyException
- [HBASE-13342](#) - Fix incorrect interface annotations
- [HBASE-13419](#) - Thrift gateway should propagate text from exception causes.
- [HBASE-13491](#) - Fix bug in FuzzyRowFilter#getNextForFuzzyRule
- [HBASE-13851](#) - RpcClientImpl.close() can hang with cancelled replica RPCs
- [HBASE-13885](#) - ZK watches leaks during snapshots
- [HBASE-13958](#) - RESTApiClusterManager calls kill() instead of suspend() and resume()
- [HBASE-13995](#) - ServerName is not fully case insensitive
- [HBASE-14027](#) - Clean up netty dependencies
- [HBASE-14045](#) - Bumping thrift version to 0.9.2.
- [HBASE-14076](#) - ResultSerialization and MutationSerialization can throw InvalidProtocolBufferException when serializing a cell larger than 64MB
- [HIVE-10252](#) - Make PPD work for Parquet in row group level
- [HIVE-10270](#) - Cannot use Decimal constants less than 0.1BD
- [HIVE-10553](#) - Remove hardcoded Parquet references from SearchArgumentImpl SearchArgumentImpl
- [HIVE-10706](#) - Make vectorized_timestamp_funcs test more stable
- [HIVE-10801](#) - 'drop view' fails throwing java.lang.NullPointerException
- [HIVE-10808](#) - Inner join on Null throwing Cast Exception
- [HIVE-11150](#) - Remove wrong warning message related to chgrp
- [HIVE-11174](#) - Hive does not treat floating point signed zeros as equal
- [HIVE-11216](#) - UDF GenericUDFMapKeys throws NPE when a null map value is passed in
- [HIVE-11401](#) - Predicate push down does not work with Parquet when partitions are in the expression expression
- [HIVE-6099](#) - Multi insert does not work properly with distinct count
- [HIVE-9500](#) - Support nested structs over 24 levels
- [HIVE-9665](#) - Parallel move task optimization causes race condition
- [HIVE-10427](#) - collect_list() and collect_set() should accept struct types as argument
- [HIVE-10437](#) - NullPointerException on queries where map/reduce is not involved on tables with partitions
- [HIVE-10895](#) - ObjectStore does not close Query objects in some calls, causing a potential leak in some metastore db resources
- [HIVE-10976](#) - Redundant HiveMetaStore connect check in HS2 CLIService start
- [HIVE-10977](#) - No need to instantiate MetaStoreDirectSql when HMS DirectSql is disabled
- [HIVE-11095](#) - Fix SerDeUtils bug when Text is reused
- [HIVE-11100](#) - Beeline should escape semi-colon in queries
- [HIVE-11112](#) - ISO-8859-1 text output has fragments of previous longer rows appended

CDH 5 Release Notes

- [HIVE-11157](#) - Hive.get(HiveConf) returns same Hive object to different user sessions
- [HIVE-11194](#) - Exchange partition on external tables should fail with error message when target folder already exists
- [HIVE-11433](#) - NPE for a multiple inner join query
- [HIVE-9767](#) - Fixes in Hive UDF to be usable in Pig
- [HIVE-10629](#) - Dropping table in an encrypted zone does not drop warehouse directory
- [HIVE-10630](#) - Renaming tables across encryption zones renames table even though the operation throws error
- [HIVE-10659](#) - Beeline command which contains semi-colon as a non-command terminator will fail
- [HIVE-10788](#) - Change sort_array to support non-primitive types
- [HIVE-10895](#) - ObjectStore does not close Query objects in some calls causing potential leak.
- [HIVE-11109](#) - Replication factor is not properly set in SparkHashTableSinkOperator [Spark Branch]
- [HIVE-10594](#) - Remote Spark client doesn't use Kerberos keytab to authenticate [Spark Branch]
- [HUE-2618](#) - [hive] Recent query results show character encoding in view
- [HUE-2767](#) - [impala] Issue showing sample data for a table
- [HUE-2796](#) - sync_groups_on_login doesn't work with posixGroups
- [HUE-2807](#) - [useradmin] Support deleting numeric groups
- [HUE-2808](#) - [dbquery] Add row numbers to support default order by
- [HUE-2813](#) - [hive] Report when Hue server is down when trying to execute a query
- [HUE-2814](#) - Revert pyopenssl 0.13.1
- [HUE-2835](#) - Fixed issue with DN's that have weird comma location
- [HUE-2840](#) - [useradmin] Fix create home directories for Add/Sync LDAP group
- [HUE-2849](#) - [useradmin] Fix exception in Add/Sync LDAP group for undefined group name
- [IMPALA-1929](#) - Avoiding a DCHECK of NULL hash table in spilled right joins
- [IMPALA-2136](#) - Bug in PrintTColumnValue caused wrong stats for TINYINT partition cols
- [IMPALA-2133](#) - Properly unescape string value for HBase filters
- [IMPALA-2018](#) - Where clause does not propagate to joins inside nested views
- [IMPALA-2064](#) - Add effective_user() builtin
- [IMPALA-2125](#) - Make UTC to local TimestampValue conversion faster.
- [IMPALA-2048](#) - Impala DML/DDL operations corrupt table metadata leading to Hive query failures
- [KITE-1014](#) - Fix support for Hive datasets on Kerberos enabled clusters.
- [KITE-1015](#) - Add "replaceValues" morphline command that replaces all matching record field values with a given replacement string
- [KITE-462](#) - Oozie jobs do not pass credentials
- [KITE-976](#) - DatasetKeyInputFormat/DatasetKeyOutputFormat not setting job configuration before loading dataset
- [KITE-1030](#) - readCSV WARN log msg on overly long lines where quoteChar is non-empty should print the whole record seen so far
- [OOZIE-2268](#) - Update ActiveMQ version for security and other fixes
- [OOZIE-2286](#) - Update Log4j and Log4j-extras to latest 1.2.x release
- [PIG-4053](#) - PIG-4053: TestMRCompiler succeeded with sun jdk 1.6 while failed with sun jdk 1.7
- [PIG-4338](#) - PIG-4338: Fix test failures with JDK8
- [PIG-4326](#) - PIG-4326: AvroStorageSchemaConversionUtilities does not properly convert schema for maps of arrays of records
- [SENTRY-695](#) - Sentry service should read the hadoop group mapping properties from core-site
- [SENTRY-721](#) - HDFS Cascading permissions not applied to child file ACLs if a direct grant exists
- [SENTRY-752](#) - Sentry service audit log file name format should be consistent
- [SOLR-7457](#) - Make DirectoryFactory publishing MBeanInfo extensible
- [SOLR-7458](#) - Expose HDFS Block Locality Metrics
- [SPARK-6480](#) - histogram() bucket function is wrong in some simple edge cases
- [SPARK-6954](#) - ExecutorAllocationManager can end up requesting a negative number of executors
- [SPARK-7503](#) - Resources in .sparkStaging directory can't be cleaned up on error

- [SPARK-7705](#) - Cleanup of .sparkStaging directory fails if application is killed
- [SQOOP-2103](#) - Not able define Decimal(n,p) data type in map-column-hive option
- [SQOOP-2149](#) - Update Kite dependency to 1.0.0
- [SQOOP-2252](#) - Add default to Avro Schema
- [SQOOP-2294](#) - Change to Avro schema name breaks some use cases
- [SQOOP-2295](#) - Hive import with Parquet should append automatically
- [SQOOP-2327](#) - Sqoop2: Change package name from Authorization to authorization
- [SQOOP-2339](#) - Move sub-directory might fail in append mode
- [SQOOP-2362](#) - Add oracle direct mode in list of supported databases
- [SQOOP-2400](#) - hive.metastore.sasl.enabled should be set to true for Oozie integration
- [SQOOP-2406](#) - Add support for secure mode when importing Parquet files into Hive
- [SQOOP-2437](#) - Use hive configuration to connect to secure metastore

Issues Fixed in CDH 5.4.4

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.4.4:

- [HIVE-10572](#) - Improve Hive service test to check empty string
- [HIVE-9934](#) - Vulnerability in LdapAuthenticationProviderImpl enables HiveServer2 client to degrade the authentication mechanism to "none", allowing authentication without password
- [HIVE-10006](#) - RSC has memory leak while execute multi queries.
- [HUE-2814](#) - Revert pyopenssl 0.13.1

Published Known Issues Fixed

As a result of the above fixes, the following issues, previously published as [Known Issues in CDH 5](#) on page 111, are also fixed.

Hue with TLS/SSL Fails to Start in CDH 5.4.3

In CDH 5.4.3, Hue with TLS/SSL fails to start because a pyOpenSSL package is missing in the parcel. This applies to both new installs and upgrades and is not operating-system specific.

Bug: None

Release affected: CDH 5.4.3

Release containing the fix: CDH 5.4.4

Workaround:

1. Download the package:

```
a. cd /tmp
b. curl -O https://pypi.python.org/packages/source/p/pyOpenSSL/pyOpenSSL-0.13.tar.gz
```

2. Determine the Hue installation directory:

- Parcels:

```
export HUE_DIR=/opt/cloudera/parcels/CDH-5.4.3-1.cdh5.4.3.p0.6/lib/hue
```

- Packages:

```
export HUE_DIR=/usr/lib/hue
```

3. Change to the Hue installation directory:

```
cd $HUE_DIR
```

4. Do the following, depending on your OS:

- On CentOS/RedHat 6.x:
 1. `sudo yum install gcc python-devel openssl-devel`
 2. `sudo ./build/env/bin/python ./build/env/bin/pip -v install /tmp/pyOpenSSL-0.13.tar.gz`
- On Ubuntu 14.04:
 1. `sudo apt-get install gcc python-dev python-pip libssl-dev`
 2. `sudo pip install --target=`pwd`/`ls -d build/env/lib/python*/site-packages` /tmp/pyOpenSSL-0.13.tar.gz`
- On other platforms, [contact Support](#) for assistance.

Issues Fixed in CDH 5.4.3

Upgrades to CDH 5.4.1 from Releases Earlier than 5.4.0 May Fail

Problem: Because of a change in the implementation of the NameNode metadata upgrade mechanism, upgrading to CDH 5.4.1 from a version lower than 5.4.0 can take an inordinately long time. In a cluster with NameNode high availability (HA) configured and a large number of edit logs, the upgrade can fail, with errors indicating a timeout in the pre-upgrade step on JournalNodes.

What to do:

To avoid the problem: Do not upgrade to CDH 5.4.1; upgrade to CDH 5.4.2 instead.

If you experience the problem: If you have already started an upgrade and seen it fail, contact Cloudera Support. This problem involves no risk of data loss, and manual recovery is possible.

If you have already completed an upgrade to CDH 5.4.1, or are installing a new cluster: In this case you are not affected and can continue to run CDH 5.4.1.

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.4.3:

- [HADOOP-12043](#) - Display warning if defaultFs is not set when running fs commands.
- [HADOOP-11969](#) - ThreadLocal initialization in several classes is not thread safe
- [HADOOP-11402](#) - Negative user-to-group cache entries are never cleared for never-again-accessed users
- [HADOOP-11238](#) - Update the NameNode's Group Cache in the background when possible
- [HDFS-8535](#) - Clarify that dfs usage in dfsadmin -report output includes all block replicas.
- [HDFS-8486](#) - DN startup may cause severe data loss
- [HDFS-7917](#) - Use file to replace data dirs in test to simulate a disk failure.
- [HDFS-7833](#) - DataNode reconfiguration does not recalculate valid volumes required, based on configured failed volumes tolerated.
- [HDFS-7604](#) - Track and display failed DataNode storage locations in NameNode.
- [HDFS-8380](#) - Always call addStoredBlock on blocks which have been shifted from one storage to another
- [HDFS-7980](#) - Incremental BlockReport will dramatically slow down the startup of a namenode
- [HDFS-8305](#) - HDFS INotify: the destination field of RenameOp should always end with the file name
- [YARN-3842](#) - NMProxy should retry on NMNotYetReadyException
- [YARN-3467](#) - Expose allocatedMB, allocatedVCores, and runningContainers metrics on running Applications in RM Web UI
- [YARN-3762](#) - FairScheduler: CME on FSParentQueue#getQueueUserAclInfo
- [YARN-3675](#) - FairScheduler: RM quits when node removal races with continuous scheduling on the same node
- [YARN-3491](#) - PublicLocalizer#addResource is too slow.
- [MAPREDUCE-6387](#) - Serialize the recently added Task#encryptedSpillKey field at the end
- [HBASE-13481](#) - Master should respect master (old) DNS/bind related configurations

- [HBASE-13826](#) - Unable to create table when group acls are appropriately set.
- [HBASE-13729](#) - Old hbase.regionserver.global.memstore.upperLimit and hbase.regionserver.global.memstore.lowerLimit properties are ignored if present
- [HBASE-13789](#) - ForeignException should not be sent to the client
- [HBASE-13779](#) - Calling table.exists() before table.get() end up with an empty Result
- [HBASE-13780](#) - Default to 700 for HDFS root dir permissions for secure deployments
- [HBASE-13768](#) - ZooKeeper znodes are bootstrapped with insecure ACLs in a secure configuration
- [HBASE-13767](#) - Allow ZKAclReset to set and not just clear ZK ACLs
- [HBASE-13086](#) - Show ZK root node on Master WebUI
- [HBASE-13413](#) - Create an integration test for Replication
- [HBASE-13611](#) - update clover to work for current versions
- [HIVE-10841](#) - [WHERE col is not null] does not work sometimes for queries with many JOIN statements
- [HIVE-10956](#) - HS2 leaks HMS connections
- [HIVE-10571](#) - HiveMetaStoreClient should close existing thrift connection before its reconnect
- [HIVE-10835](#) - Concurrency issues in JDBC driver
- [HIVE-10802](#) - Table join query with some constant field in select fails
- [HIVE-10538](#) - Fix NPE in FileSinkOperator from hashcode mismatch
- [HIVE-10771](#) - "separatorChar" has no effect in "CREATE TABLE AS SELECT" statement
- [HIVE-10732](#) - Hive JDBC driver does not close operation for metadata queries
- [HIVE-10151](#) - insert into A select from B is broken when both A and B are Acid tables and bucketed the same way
- [HIVE-10483](#) - insert overwrite partition deadlocks on itself with DbTxnManager
- [HIVE-10050](#) - Support overriding memory configuration for AM launched for TempletonControllerJob
- [HIVE-10242](#) - ACID: insert overwrite prevents create table command
- [HIVE-10481](#) - ACID table update finishes but values not really updated if column names are not all lower case
- [HIVE-10150](#) - delete from acidTbl where a in(select a from nonAcidOrcTbl) fails
- [HIVE-10721](#) - SparkSessionManagerImpl leaks SparkSessions [Spark Branch]
- [HIVE-10671](#) - yarn-cluster mode offers a degraded performance from yarn-client [Spark Branch]
- [HIVE-10453](#) - HS2 leaking open file descriptors when using UDFs
- [HIVE-2573](#) - Create per-session function registry
- [HIVE-9520](#) - Create NEXT_DAY UDF
- [HIVE-9143](#) - select user(), current_user()
- [HIVE-5472](#) - support a simple scalar which returns the current timestamp
- [HIVE-10646](#) - ColumnValue does not handle NULL_TYPE
- [HUE-2784](#) - [oozie] Coordinator editor generate wrong Monday cron expression
- [HUE-2793](#) - [JB] Fix Mapper & Reducer counts in job page
- [HUE-2776](#) - [jb] Fix "View All Tasks" pagination
- [HUE-2778](#) - [jobbrowser] Fix "Text Filter" search box in JB "View All Tasks" page
- [HUE-2767](#) - [impala] Issue showing sample data for a table
- [HUE-2754](#) - [oozie] Sqoop action with variable adds an empty argument
- [HUE-2743](#) - [search] Error HTML style leaks in the UI
- [HUE-2587](#) - [jb] Kill jobs in accepted state
- [HUE-2731](#) - [core] Validate that Hue is running in collect data script
- [HUE-2687](#) - [core] Create script to gather hue process info for troubleshooting
- [HUE-2656](#) - [tools] Add cron scripts for restart when mem usage is high
- [HUE-2701](#) - [oozie] Java action relative jar path results in error on submit
- [HUE-2703](#) - [sentry] Make more obvious why a user is not a Sentry admin
- [HUE-2741](#) - [home] Hide the document move dialog
- [HUE-2739](#) - [metastore] Autocomplete with databases/tables with built in names fails
- [HUE-2732](#) - Hue isn't correctly doing add_column migrations with non-blank defaults
- [IMPALA-1963](#): Impala Timestamp ISO-8601 Support.

CDH 5 Release Notes

- [IMPALA-2043](#): skip metadata/testddl.py#test_create_alter_bulk_partition on S3
- [IMPALA-1968](#): Part 1: Improve planner numNodes estimate for remote scans
- [IMPALA-1730](#): reduce scanner thread spinning windows
- [IMPALA-2002](#): Provide way to cache ext data source classes
- [IMPALA-2008](#): Fix wrong warning when insert overwrite to empty table
- [IMPALA-1381](#): Expand set of supported timezones.
- [IMPALA-1952](#): Expand parsing of decimals to include scientific notation
- [SENTRY-227](#) - Fix for "Unsupported entity type DUMMYPARTITION"
- [SOLR-7503](#) - Recovery after ZK session expiration happens in a single thread for all cores in a node
- [SPARK-6299](#) - ClassNotFoundException in standalone mode when running groupByKey with class defined in REPL.
- [SPARK-5522](#) - Accelerate the History Server start

Published Known Issues Fixed

Migrations to MySQL fail if multiple Hue users have the same name but different upper/lower case letters

Bug: None

Workaround: None.

Issues Fixed in CDH 5.4.2

Upgrades to CDH 5.4.1 from Releases Earlier than 5.4.0 May Fail

Problem: Because of a change in the implementation of the NameNode metadata upgrade mechanism, upgrading to CDH 5.4.1 from a version lower than 5.4.0 can take an inordinately long time. In a cluster with NameNode high availability (HA) configured and a large number of edit logs, the upgrade can fail, with errors indicating a timeout in the pre-upgrade step on JournalNodes.

What to do:

To avoid the problem: Do not upgrade to CDH 5.4.1; upgrade to CDH 5.4.2 instead.

If you experience the problem: If you have already started an upgrade and seen it fail, contact Cloudera Support. This problem involves no risk of data loss, and manual recovery is possible.

If you have already completed an upgrade to CDH 5.4.1, or are installing a new cluster: In this case you are not affected and can continue to run CDH 5.4.1.

Issues Fixed in CDH 5.4.1

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.4.1:

- [HADOOP-11891](#) - OsSecureRandom should lazily fill its reservoir to avoid open too many file descriptors.
- [HADOOP-11802](#) - DomainSocketWatcher thread terminates sometimes after there is an I/O error during requestShortCircuitShm
- [HADOOP-11724](#) - DistCp throws NPE when the target directory is root.
- [HDFS-7645](#) - Rolling upgrade is restoring blocks from trash multiple times which could cause significant and unnecessary block churn.
- [HDFS-7869](#) - Inconsistency in the return information while performing rolling upgrade
- [HDFS-8127](#) - NameNode Failover during HA upgrade can cause DataNode to finalize upgrade
- [HDFS-3443](#) - Fix NPE when NameNode transition to active during startup.
- [HDFS-7312](#) - Update DistCp v1 to optionally not use tmp location
- [HDFS-8292](#) - Move conditional in fmt_time from dfs-dust.js to status.html
- [HDFS-6673](#) - Add delimited format support to PB OIV tool
- [HDFS-8214](#) - Secondary NN Web UI shows wrong date for Last Checkpoint
- [HDFS-7884](#) - Fix NullPointerException in BlockSender when the generation stamp provided by the client is larger than the one stored in the DataNode
- [HDFS-4448](#) - Allow HA NN to start in secure mode with wildcard address configured
- [HDFS-8070](#) - Fix issue that Pre-HDFS-7915 DFSClient cannot use short circuit on post-HDFS-7915 DataNode

- [HDFS-7915](#) - The DataNode can sometimes allocate a ShortCircuitShm slot and fail to tell the DFSClient about it because of a network error
- [HDFS-7931](#) - DistributedFileSystem should not look for keyProvider in cache if Encryption is disabled
- [HDFS-7916](#) - 'reportBadBlocks' from DataNodes to standby Node BPServiceActor goes for infinite loop
- [HDFS-8099](#) - Change "DFSIInputStream has been closed already" message to debug log level
- [HDFS-7996](#) - After swapping a volume, BlockReceiver reports ReplicaNotFoundException
- [HDFS-7587](#) - Edit log corruption can happen if append fails with a quota violation
- [HDFS-7881](#) - TestHftpFileSystem#testSeek fails
- [HDFS-7929](#) - inotify is unable to fetch pre-upgrade edit log segments once upgrade starts
- [YARN-3363](#) - Add localization and container launch time to ContainerMetrics at NM to show these timing information for each active container.
- [YARN-3485](#) - FairScheduler headroom calculation doesn't consider maxResources for Fifo and FairShare policies
- [YARN-3464](#) - Race condition in LocalizerRunner kills localizer before localizing all resources
- [YARN-3516](#) - Killing ContainerLocalizer action doesn't take effect when private localizer receives FETCH_FAILURE status.
- [YARN-3021](#) - YARN's delegation-token handling disallows certain trust setups to operate properly over DistCp
- [YARN-3241](#) - FairScheduler handles invalid queue names inconsistently.
- [YARN-2868](#) - FairScheduler: Add a metric for measuring latency of allocating first container for an application
- [YARN-3428](#) - Add debug logs to capture the resources being localized for a container.
- [MAPREDUCE-6339](#) - Job history file is not flushed correctly because isTimerActive flag is not set true when flushTimerTask is scheduled.
- [MAPREDUCE-5710](#) - Running distcp with -delete incurs avoidable penalties
- [MAPREDUCE-6343](#) - JobConf.parseMaximumHeapSizeMB() fails to parse value greater than 2GB expressed in bytes
- [MAPREDUCE-6238](#) - MR2 can't run local jobs with -libjars command options which is a regression from MR1
- [MAPREDUCE-6076](#) - Zero map split input length combined with none zero map split input length may cause MR1 job hung sometimes.
- [HBASE-13374](#) - Small scanners (with particular configurations) do not return all rows
- [HBASE-13269](#) - Limit result array pre-allocation to avoid OOME with large scan caching values
- [HBASE-13335](#) - Update ClientSmallScanner and ClientSmallReversedScanner to use serverHasMoreResults context
- [HBASE-13534](#) - Change HBase master WebUI to explicitly mention if it is a backup master
- [HBASE-13111](#) - truncate_preserve command is failing with undefined method error
- [HBASE-13430](#) - HFiles that are in use by a table cloned from a snapshot may be deleted when that snapshot is deleted
- [HBASE-13546](#) - NPE on RegionServer status page if all masters are down
- [HBASE-13350](#) - Add a debug-warning if we fail HTD checks even if table.sanity.checks is disabled
- [HBASE-13262](#) - ResultScanner doesn't return all rows in Scan
- [HIVE-10452](#) - Avoid sending Beeline prompt+query to the standard output/error only when in script mode.
- [HIVE-10541](#) - Beeline requires newline at the end of each query in a file
- [HIVE-9625](#) - Delegation tokens for HMS are not renewed
- [HIVE-10499](#) - Ensure Session/ZooKeeperClient instances are closed
- [HIVE-10312](#) - SASL.QOP in JDBC URL is ignored for Delegation token Authentication
- [HIVE-10324](#) - Hive metatool should take table_param_key to allow for changes to avro serde's schema url key
- [HIVE-10202](#) - Beeline outputs prompt+query on standard output when used in non-interactive mode
- [HIVE-10087](#) - Beeline's --silent option should suppress query from being echoed when running with -f option
- [HIVE-10098](#) - HS2 local task for map join fails in KMS encrypted cluster
- [HIVE-10146](#) - Add option to not count session as idle if query is running
- [HIVE-10108](#) - Index#getIndexTableName() should return db.index_table_name instead of qualified table name
- [HIVE-10093](#) - Unnecessary HMSHandler initialization for default MemoryTokenStore on HS2
- [HIVE-10085](#) - Lateral view on top of a view throws RuntimeException
- [HIVE-10086](#) - Parquet file using column index access throws error in Hive

- [HIVE-9839](#) - HiveServer2 leaks OperationHandle on async queries which fail at compile phase
- [HIVE-9920](#) - DROP DATABASE IF EXISTS throws exception if database does not exist
- [HIVE-10476](#) - Hive query should fail when it fails to initialize a session in SetSparkReducerParallelism
- [HIVE-10434](#) - Cancel connection when remote Spark driver process has failed
- [HIVE-10473](#) - Spark client is recreated even spark configuration is not changed
- [HIVE-10291](#) - Hive on Spark job configuration needs to be logged
- [HIVE-10143](#) - HS2 fails to clean up Spark client state on timeout
- [HIVE-10073](#) - Runtime exception when querying HBase with Spark
- [HUE-2723](#) - [hive] Listing table information in non default DB fails
- [HUE-2722](#) - [hive] Query returns wrong number of rows when HiveServer2 returns data not encoded properly
- [HUE-2713](#) - [oozie] Deleting a Fork of Fork can break the workflow
- [HUE-2717](#) - [oozie] Coordinator editor does not save non-default schedules
- [HUE-2716](#) - [pig] Scripts fail on hcat auth with org.apache.hive.hcatalog.pig.HCatLoader()
- [HUE-2707](#) - [hive] Allow sample of data on partitioned tables in strict mode
- [HUE-2720](#) - [oozie] Intermittent 500s when trying to view oozie workflow history v1
- [HUE-2712](#) - [oozie] Creating a fork can error
- [HUE-2710](#) - [search] Heatmap select on yelp example errors
- [HUE-2686](#) - [impala] Explain button is erroring
- [HUE-2671](#) - [core] sync_groups_on_login doesn't work with NT Domain
- [IMPALA-1519/IMPALA-1946](#) - Fix wrapping of exprs via a TupleIsNotNullPredicate with analytics.
- [IMPALA-1900](#) - Assign predicates below analytic functions with a compatible partition by clause for partition pruning.
- [IMPALA-1919](#) - When out_batch->AtCapacity(), avoid calling ProcessBatch in right joins.
- [IMPALA-1960](#) - Illegal reference to non-materialized tuple when query has an empty select-project-join block.
- [IMPALA-1969](#) - OpenSSL init must not be called concurrently.
- [IMPALA-1973](#) - Fixing crash when uninitialized, empty row is added in HdfsTextScanner due to missing newline at the end of file.
- [OOZIE-2218](#) - META-INF directories in the war file have 777 permissions
- [OOZIE-2170](#) - Oozie should automatically set configs to make Spark jobs show up in the Spark History Server
- [SENTRY-699](#) - Memory leak when running Sentry with HiveServer2
- [SENTRY-703](#) - Calls to add_partition fail when passed a Partition object with a null location
- [SENTRY-696](#) - Improve Metastoreplugin Cache Initialization time
- [SENTRY-683](#) - HDFS service client should ensure the kerberos ticket is valid before new service connection
- [SOLR-7478](#) - UpdateLog#close shuts down its executor with interrupts before running close, possibly preventing a clean close.
- [SOLR-7437](#) - Make HDFS transaction log replication factor configurable.
- [SOLR-7338/SOLR-6583](#) - A reloaded core will never register itself as active after a ZK session expiration.
- [SPARK-7281](#) - No option for AM native library path in yarn-client mode.
- [SPARK-6087](#) - Provide actionable exception if Kryo buffer is not large enough
- [SPARK-6868](#) - Container link broken on Spark UI Executors page when YARN is set to HTTPS_ONLY
- [SPARK-6506](#) - python support in yarn cluster mode requires SPARK_HOME to be set
- [SPARK-6650](#) - ExecutorAllocationManager never stops
- [SPARK-6578](#) - Outbound channel in network library is not thread-safe, can lead to fetch failures
- [SQOOP-2343](#) - AsyncSqlRecordWriter stuck if any exception is thrown out in its close method
- [SQOOP-2286](#) - Ensure Sqoop generates valid avro column names
- [SQOOP-2283](#) - Support usage of --exec and --password-alias
- [SQOOP-2281](#) - Set overwrite on kite dataset
- [SQOOP-2282](#) - Add validation check for --hive-import and --append
- [SQOOP-2257](#) - Import Parquet data into a hive table with --hive-overwrite option does not work
- [ZOOKEEPER-2146](#) - BinaryInputArchive readString should check length before allocating memory

- [ZOOKEEPER-2149](#) - Log client address when socket connection established

Published Known Issues Fixed

As a result of the above fixes, the following issues, previously published as [Known Issues in CDH 5](#) on page 111, are also fixed.

Apache Hadoop

NameNode cannot use wildcard address in a secure cluster

In a secure cluster, you cannot use a wildcard for the NameNode's RPC or HTTP bind address. For example, `dfs.namenode.http-address` must be a real, routable address and port, not `0.0.0.<port>`. This should affect you only if you are running a secure cluster *and* your NameNode needs to bind to multiple local addresses.

Bug: [HDFS-4448](#)

Workaround: None

Offline Image Viewer (OIV) tool regression: missing Delimited outputs.

Bugs: [HDFS-6673](#), [HDFS-5952](#)

Severity: Medium

Workaround: Set up `dfs.namenode.legacy-oiv-image.dir` to an appropriate directory on the secondary NameNode (or standby NameNode in an HA configuration), and use `hdfs oiv_legacy` to process the legacy format of the OIV fsimage.

Apache HBase

Setting `maxResultSize` Incorrectly On a Scan May Cause Client Data Loss

Scanners may not return all the results from a region if a scan is configured with a `maxResultSize` limit that could be reached before the caching limit. Results are missed because the scanner jumps to the next region preemptively.

The default value for `maxResultSize` is `Long.MAX_VALUE` and the default value of caching is 100, so with the default configuration, the caching limit will always be reached before the `maxResultSize` and the issue will not appear. If the `maxResultSize` is configured to any limit that may be reached before the caching limit, the issue may occur.

Bug: [HBASE-13262](#)

Severity: Low

Workaround: Never configure a scan with a `maxResultSize` other than `Long.MAX_VALUE` (never change it from its default value) because that will ensure that the `maxResultSize` limit is never reached before the caching limit.

Apache Hive

Hive metatool does not fix Avro schema URL setting in an HDFS HA upgrade

When you upgrade Hive in an HDFS HA configuration, and the `avro.schema.url` is set in an Avro table's properties instead of the SerDe properties, the metatool will not correct the problem.

Bug: [HIVE-10324](#)

Workaround: Use `alter table.. set tblproperties` to fix the `avro.schema.url`.

Hive metastore `getIndexTableName` returns qualified table name

In CDH 5.4.0, `getIndexTableName` returns a qualified table name such as

```
database_name
:
index_table_name
```

whereas in previous releases it returns an unqualified table name, such as

```
index_table_name
```

CDH 5 Release Notes

Bug: [HIVE-10108](#)

Workaround: None

HiveServer2 has an unexpected Derby metastore directory in secure clusters

Bug: [HIVE-10093](#)

Workaround: None; ignore the Derby database.

Apache Oozie

Spark jobs run from the Spark action don't show up in the Spark History Server or properly link to it from the Spark AM

Bug: [OOZIE-2170](#)

Severity: Low

Workaround: Specify these configuration properties in the `spark-opts` element of your Spark action in the `workflow.xml` file:

```
--conf spark.yarn.historyServer.address=http://SPH:18088 --conf  
spark.eventLog.dir=hdfs://NN:8020/user/spark/applicationHistory --conf  
spark.eventLog.enabled=true
```

where `SPH` is the hostname of the Spark History Server and `NN` is the hostname of the NameNode. You can also find these values in `/etc/spark/conf/spark-defaults.conf` on the gateway host when Spark is installed from Cloudera Manager.

Apache Sentry

Hive binding should support enforcing URI privilege for transforms

Bug: [SENTRY-598](#)

Severity: Medium

Workaround: None.

Issues Fixed in CDH 5.4.0

The following topics describe known issues fixed in CDH 5.4.0.

For the latest Impala fixed issues, see [Issues Fixed in the 2.2.0 Release / CDH 5.4.0](#) on page 325.

Apache Hadoop

HDFS

After upgrade from a release earlier than CDH 5.2.0, storage IDs may no longer be unique

As of CDH 5.2, each storage volume on a DataNode should have its own unique `storageID`, but in clusters upgraded from CDH 4, or CDH 5 releases earlier than CDH 5.2.0, each volume on a given DataNode shares the same `storageID`, because the HDFS upgrade does not properly update the IDs to reflect the new naming scheme. This causes problems with load balancing. The problem affects only clusters upgraded from CDH 5.1.x and earlier to CDH 5.2 or later. Clusters that are new as of CDH 5.2.0 or later do not have the problem.

Bug: [HDFS-7575](#)

Severity: Medium

Workaround: Upgrade to a later or patched version of CDH.

Apache Hive

UDF infile() does not accept arguments of type CHAR or VARCHAR

Bug: [HIVE-6637](#)

Severity: Low

Workaround: Cast the argument to type String.

Hive's Decimal type cannot be stored in Parquet and Avro

Tables containing decimal columns cannot use Parquet or the Avro storage engine.

Bug: [HIVE-6367](#) and [HIVE-5823](#)

Severity: Low

Workaround: Use a different file format.

Apache Oozie

Executing oozie job -config properties file -dryrun fails because of a code defect in argument parsing

Bug: [OOZIE-1878](#)

Severity: Low

Workaround: None.

When you use Hive Server 2 from Oozie, Oozie won't collect or print out the Hadoop Job IDs of any jobs launched by Hive Server 2

Bug: None

Severity: Low

Workaround: You can get the Hadoop IDs from the Resource Manager or JobTracker.

Cloudera Search

Spark indexer failed if configured to use security.

Spark indexing jobs failed when Kerberos authentication was enabled.

With Search for CDH 5.4 and later, Spark indexing jobs succeed, even when Kerberos authentication is required.

Bug: None.

Severity: Medium.

Workaround: Disable Kerberos authentication or use another indexer.

Mapper-only HBase batch indexer failed if configured to use security.

Attempts to complete an HBase batch indexing job failed when Kerberos authentication was enabled and reducers were set to 0.

With Search for CDH 5.4 and later, mapper-only HBase batch indexer succeeds, even when Kerberos authentication is required.

Bug: None.

Severity: Medium.

Workaround: Either disable Kerberos authentication or use one or more reducers.

Shard splitting support is experimental.

Cloudera anticipated shard splitting to function as expected with Cloudera Search, but this interaction had not been thoroughly tested.

As of the release of Search for CDH 5.4, additional testing of shard splitting has been completed, so this functionality can be safely used.

Severity: Low

Workaround: Use shard splitting for test and development purposes, but be aware of the risks of using shard splitting in production environments. To avoid using shard splitting, use the source data to create a new index with a new sharding count by re-indexing the data to a new collection. You can enable this using the MapReduceIndexerTool.

CDH 5 Release Notes

TrieDateField defaulted OMIT_NORMS to True.

All primitive field types were intended to omit norms by default with schema version 1.5 or higher. This change was not applied to TrieDateField.

With Search for CDH 5.4, TrieDateField is set to omit norms by default.

Bug: [SOLR-6211](#)

Severity: Low

Fields or Types outside <field> or <types> tags are silently ignored.

In previous releases, Solr silently ignored definitions such as <fieldType>, <field>, and <copyField> if those definitions were not contained in <fields> or <types> tags.

With Search 5.4 for CDH, these tags are no longer required for definitions to be included. These tags are supported so either style may be implemented.

Bug: [SOLR-5228](#)

Apache Sentry (incubating)

INSERT OVERWRITE LOCAL fails if you use only the Linux pathname

Bug: None

Severity: Low

Workaround: Prefix the path of the local file with `file://` when using `INSERT OVERWRITE LOCAL`.

INSERT OVERWRITE and CREATE EXTERNAL commands fail because of HDFS URI permissions

When you use Sentry to secure Hive, and use HDFS URIs in a HiveQL statement, the query will fail with an HDFS permissions error unless you specify the NameNode and port.

Bug: None

Severity: Low

Workaround: Specify the NameNode and port, where applicable, in the URI; for example specify `hdfs://nn-uri:port/user/warehouse/hive/tab` rather than simply `/user/warehouse/hive/tab`. In a high-availability deployment, specify the value of `FS.defaultFS`.

Issues Fixed in CDH 5.3.x

The following topics describe issues fixed in CDH 5.3.x, from newest to oldest release. You can also review [What's New In CDH 5.3.x](#) on page 34 or [Known Issues in CDH 5](#) on page 111.

Issues Fixed in CDH 5.3.10

CDH 5.3.10 fixes the following issues.

Apache Hadoop

FSImage may get corrupted after deleting snapshot

Bug: [HDFS-9406](#)

When deleting a snapshot that contains the last record of a given INode, the fsimage may become corrupt because the create list of the snapshot diff in the previous snapshot and the child list of the parent INodeDirectory are not cleaned.

Apache HBase

The ReplicationCleaner process can abort if its connection to ZooKeeper is inconsistent

Bug: [HBASE-15234](#)

If the connection with ZooKeeper is inconsistent, the ReplicationCleaner may abort, and the following event is logged by the HMaster:

```
WARN org.apache.hadoop.hbase.replication.master.ReplicationLogCleaner: Aborting
ReplicationLogCleaner
because Failed to get list of replicators
```

Unprocessed WALs accumulate.

The seekBefore() method calculates the size of the previous data block by assuming that data blocks are contiguous, and HFile v2 and higher store Bloom blocks and leaf-level INode blocks with the data. As a result, reverse scans do not work when Bloom blocks or leaf-level INode blocks are present when HFile v2 or higher is used.

Workaround: Restart the HMaster occasionally. The ReplicationCleaner restarts if necessary and process the unprocessed WALs.

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.3.10:

- [HADOOP-7713](#) - dfs -count -q should label output column
- [HADOOP-8944](#) - Shell command fs -count should include human readable option
- [HADOOP-10406](#) - TestIPC.testIpcWithReaderQueuing may fail
- [HADOOP-12200](#) - TestCryptoStreamsWithOpensslAesCtrCryptoCodec should be skipped in non-native profile
- [HADOOP-12240](#) - Fix tests requiring native library to be skipped in non-native profile
- [HADOOP-12280](#) - Skip unit tests based on maven profile rather than NativeCodeLoader.isNativeCodeLoaded
- [HADOOP-12418](#) - TestRPC.testRPCInterruptedSimple fails intermittently
- [HADOOP-12464](#) - Interrupted client may try to fail over and retry
- [HADOOP-12468](#) - Partial group resolution failure should not result in user lockout
- [HADOOP-12559](#) - KMS connection failures should trigger TGT renewal
- [HADOOP-12604](#) - Exception may be swallowed in KMSClientProvider
- [HADOOP-12605](#) - Fix intermittent failure of TestIPC.testIpcWithReaderQueuing
- [HADOOP-12682](#) - Fix TestKMS#testKMSRestart* failure
- [HADOOP-12699](#) - TestKMS#testKMSProvider intermittently fails during 'test rollover draining'
- [HADOOP-12715](#) - TestValueQueue#testGetAtMostPolicyALL fails intermittently
- [HADOOP-12736](#) - TestTimedOutTestsListener#testThreadDumpAndDeadlocks sometimes times out
- [HADOOP-12788](#) - OpensslAesCtrCryptoCodec should log which random number generator is used
- [HDFS-6533](#) - TestBPOfferService#testBasicFunctionalitytest fails intermittently
- [HDFS-6673](#) - Add delimited format support to PB OIV tool
- [HDFS-6799](#) - The invalidate method in SimulatedFSDataset failed to remove (invalidate) blocks from the file system
- [HDFS-7423](#) - Various typos and message formatting fixes in nfs daemon and doc
- [HDFS-7553](#) - Fix the TestDFSUpgradeWithHA due to BindException
- [HDFS-7990](#) - IBR delete ack should not be delayed
- [HDFS-8211](#) - DataNode UUID is always null in the JMX counter
- [HDFS-8646](#) - Prune cached replicas from DatanodeDescriptor state on replica invalidation
- [HDFS-9092](#) - NFS silently drops overlapping write requests and causes data copying to fail
- [HDFS-9250](#) - Add Precondition check to LocatedBlock#addCachedLoc
- [HDFS-9347](#) - Invariant assumption in TestQuorumJournalManager.shutdown() is wrong
- [HDFS-9358](#) - TestNodeCount#testNodeCount timed out
- [HDFS-9364](#) - Unnecessary DNS resolution attempts when creating NameNodeProxies
- [HDFS-9406](#) - FSImage may get corrupted after deleting snapshot
- [HDFS-9949](#) - Add a test case to ensure that the DataNode does not regenerate its UUID when a storage directory is cleared
- [MAPREDUCE-6302](#) - Incorrect headroom can lead to a deadlock between map and reduce allocations
- [MAPREDUCE-6387](#) - Serialize the recently added Task#encryptedSpillKey field at the end

- [MAPREDUCE-6460](#) - TestRMContainerAllocator.testAttemptNotFoundCausesRMCommunicatorException fails
- [YARN-2377](#) - Localization exception stack traces are not passed as diagnostic info
- [YARN-2785](#) - Fixed intermittent TestContainerResourceUsage failure
- [YARN-3024](#) - LocalizerRunner should give DIE action when all resources are localized
- [YARN-3074](#) - Nodemanager dies when localizer runner tries to write to a full disk
- [YARN-3464](#) - Race condition in LocalizerRunner kills localizer before localizing all resources.
- [YARN-3516](#) - Killing ContainerLocalizer action does not take effect when private localizer receives FETCH_FAILURE status
- [YARN-3727](#) - For better error recovery, check if the directory exists before using it for localization
- [YARN-3762](#) - FairScheduler: CME on FSParentQueue#getQueueUserAclInfo
- [YARN-4204](#) - ConcurrentModificationException in FairSchedulerQueueInfo
- [YARN-4235](#) - FairScheduler PrimaryGroup does not handle empty groups returned for a user
- [YARN-4354](#) - Public resource localization fails with NPE
- [YARN-4380](#) - TestResourceLocalizationService.testDownloadingResourcesOnContainerKill fails intermittently
- [YARN-4393](#) - Fix intermittent test failure for TestResourceLocalizationService#testFailedDirsResourceRelease
- [YARN-4613](#) - Fix test failure in TestClientRMServices#testGetClusterNodes
- [YARN-4717](#) - TestResourceLocalizationService.testPublicResourceInitializesLocalDir fails Intermittently due to IllegalArgumentException from cleanup
- [HBASE-10153](#) - Improve VerifyReplication to compute BADROWS more accurately
- [HBASE-11394](#) - AmendReplication can have data loss if peer id contains hyphen
- [HBASE-11394](#) - Replication can have data loss if peer id contains hyphen "-"
- [HBASE-11992](#) - Backport HBASE-11367 (Pluggable replication endpoint) to 0.98
- [HBASE-12136](#) - Race condition between client adding tableCF replication znode and server triggering TableCFsTracker
- [HBASE-12150](#) - Backport replication changes from HBASE-12145
- [HBASE-12336](#) - RegionServer failed to shutdown for NodeFailoverWorker thread
- [HBASE-12631](#) - Backport HBASE-12576 (Add metrics for rolling the HLog if there are too few DNs in the write pipeline) to 0.98
- [HBASE-12658](#) - Backport HBASE-12574 (Update replication metrics to not do so many map look ups) to 0.98
- [HBASE-12865](#) - WALS may be deleted before they are replicated to peers
- [HBASE-13035](#) - Backport HBASE-12867 Shell does not support custom replication endpoint specification
- [HBASE-13084](#) - Add labels to VisibilityLabelsCache asynchronously causes TestShell flaky
- [HBASE-13437](#) - ThriftServer leaks ZooKeeper connections
- [HBASE-13703](#) - ReplicateContext should not be a member of ReplicationSource
- [HBASE-13746](#) - list_replicated_tables command is not listing table in hbase shell
- [HBASE-14146](#) - Fix Once replication sees an error it slows down forever
- [HBASE-14501](#) - NPE in replication with TDE
- [HBASE-14621](#) - ReplicationLogCleaner gets stuck when a regionserver crashes
- [HBASE-14923](#) - VerifyReplication should not mask the exception during result comparison
- [HBASE-15019](#) - Replication stuck when HDFS is restarted
- [HBASE-15032](#) - hbase shell scan filter string assumes UTF-8 encoding
- [HBASE-15035](#) - bulkloading hfiles with tags that require splits do not preserve tags
- [HBASE-15052](#) - Use EnvironmentEdgeManager in ReplicationSource
- [HIVE-7524](#) - Enable auto conversion of SMBjoin in presence of constant propagate optimization
- [HIVE-7575](#) - GetTables thrift call is very slow
- [HIVE-8115](#) - Fixing text failures caused in CDH
- [HIVE-8115](#) - Hive select query hang when fields contain map
- [HIVE-8184](#) - Inconsistency between colList and columnExprMap when ConstantPropagate is applied to subquery
- [HIVE-9112](#) - Query may generate different results depending on the number of reducers
- [HIVE-9500](#) - Support nested structs over 24 levels
- [HIVE-9860](#) - MapredLocalTask/SecureCmdDoAs leaks local files

- [HIVE-10956](#) - Fallout fix from backport to CDH 5.3.x
- [HIVE-11977](#) - Hive should handle an external avro table with zero length files present
- [HIVE-12388](#) - GetTables cannot get external tables when TABLE type argument is given
- [HIVE-12406](#) - HIVE-9500 introduced incompatible change to LazySimpleSerDe public interface
- [HIVE-12713](#) - Miscellaneous improvements in driver compile and execute logging
- [HIVE-12790](#) - Metastore connection leaks in HiveServer2
- [HIVE-12946](#) - alter table should also add default schema and authority for the location similar to create table
- [HUE-2767](#) - [impala] Issue showing sample data for a table
- [HUE-2941](#) - [hadoop] Cache the active RM HA
- [IMPALA-1702](#) - "invalidate metadata" can cause duplicate TableIds (issue not entirely fixed, but now fails gracefully)
- [IMPALA-2125](#) - Improve perf when reading timestamps from parquet files written by hive
- [IMPALA-2565](#) - Planner tests are flaky due to file size mismatches
- [IMPALA-3095](#) - Allow additional Kerberos users to be authorized to access internal APIs
- [OOZIE-2432](#) - TestPurgeXCommand fails
- [SENTRY-565](#) - Improve performance of filtering Hive SHOW commands
- [SENTRY-780](#) - HDFS Plugin should not execute path callbacks for views
- [SENTRY-835](#) - Drop table leaves a connection open when using metastorelistener
- [SENTRY-885](#) - DB name should be case insensitive in HDFS sync plugin.
- [SENTRY-936](#) - getGroup and getUser should always return orginal hdfs values for paths in prefix which are not sentry managed
- [SENTRY-944](#) - Setting HDFS rules on Sentry managed hdfs paths should not affect original hdfs rules
- [SENTRY-957](#) - Exceptions in MetastoreCacheInitializer should probably not prevent HMS from starting up
- [SENTRY-988](#) - It is better to let SentryAuthorization setter path always fall through and update HDFS
- [SENTRY-994](#) - SentryAuthorizationInfoX should override isSentryManaged
- [SENTRY-1002](#) - PathsUpdate.parsePath(path) will throw an NPE when parsing relative paths
- [SENTRY-1044](#) - Tables with non-hdfs locations break HMS startup
- [SPARK-12617](#) - [PYSPARK] Move Py4jCallbackConnectionCleaner to Streaming

Issues Fixed in CDH 5.3.9

Apache Commons Collections deserialization vulnerability

Cloudera has learned of a potential security vulnerability in a third-party library called the [Apache Commons Collections](#). This library is used in products distributed and supported by Cloudera (“Cloudera Products”), including core Apache Hadoop. The Apache Commons Collections library is also in widespread use beyond the Hadoop ecosystem. At this time, no specific attack vector for this vulnerability has been identified as present in Cloudera Products.

In an abundance of caution, we are currently in the process of incorporating a version of the Apache Commons Collections library with a fix into the Cloudera Products. In most cases, this will require coordination with the projects in the Apache community. One example of this is tracked by [HADOOP-12577](#).

The Apache Commons Collections potential security vulnerability is titled “Arbitrary remote code execution with InvokerTransformer” and is tracked by [COLLECTIONS-580](#). MITRE has not issued a CVE, but related [CVE-2015-4852](#) has been filed for the vulnerability. CERT has issued [Vulnerability Note #576313](#) for this issue.

Releases affected: CDH 5.5.0, CDH 5.4.8 and lower, CDH 5.3.8 and lower, CDH 5.2.8 and lower, CDH 5.1.7 and lower, Cloudera Manager 5.5.0, Cloudera Manager 5.4.8 and lower, Cloudera Manager 5.3.8 and lower, and Cloudera Manager 5.2.8 and lower, Cloudera Manager 5.1.6 and lower, Cloudera Navigator 2.4.0, Cloudera Navigator 2.3.8 and lower.

Users affected: All

Impact: This potential vulnerability may enable an attacker to execute arbitrary code from a remote machine without requiring authentication.

Immediate action required: Upgrade to Cloudera Manager 5.5.1 and CDH 5.5.1, Cloudera Manager 5.4.9 and CDH 5.4.9, Cloudera Manager 5.3.9 and CDH 5.3.9, and Cloudera Manager 5.2.9 and CDH 5.2.9, and Cloudera Manager 5.1.7 and CDH 5.1.7.

CDH 5 Release Notes

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.3.9:

- [FLUME-2841](#) - Upgrade commons-collections to 3.2.2
- [HADOOP-12577](#) - Bumped up commons-collections version to 3.2.2 to address a security flaw
- [HDFS-7785](#) - Improve diagnostics information for HttpPutFailedException
- [HDFS-7798](#) - Checkpointing failure caused by shared KerberosAuthenticator
- [HDFS-7871](#) - NameNodeEditLogRoller can keep printing 'Swallowing exception' message
- [HDFS-9123](#) - Copying from the root to a subdirectory should be forbidden
- [HDFS-9273](#) - ACLs on root directory may be lost after NN restart
- [HDFS-9332](#) - Fix Precondition failures from NameNodeEditLogRoller while saving namespace
- [HDFS-9470](#) - Encryption zone on root not loaded from fsimage after NN restart
- [MAPREDUCE-6191](#) - Improve clearing stale state of Java serialization testcase
- [MAPREDUCE-6233](#) - org.apache.hadoop.mapreduce.TestLargeSort.testLargeSort failed in trunk
- [MAPREDUCE-6549](#) - Multibyte delimiters with LineRecordReader cause duplicate records
- [YARN-3564](#) - Fix TestContainerAllocation.testAMContainerAllocationWhenDNSUnavailable fails randomly
- [YARN-3602](#) - TestResourceLocalizationService.testPublicResourceInitializesLocalDir fails Intermittently due to IOException from cleanup
- [YARN-3675](#) - FairScheduler: RM quits when node removal races with continuous-scheduling on the same node
- [HBASE-13134](#) - mutateRow and checkAndMutate APIs do not throw region level exceptions
- [HBASE-14196](#) - Thrift server idle connection timeout issue
- [HBASE-14283](#) - Reverse scan does not work with HFile inline index/bloom blocks
- [HBASE-14533](#) - Thrift client gets "AsyncProcess: Failed to get region location closed"
- [HBASE-14799](#) - Commons-collections object deserialization remote command execution vulnerability
- [HIVE-6099](#) - Multi insert does not work properly with distinct count
- [HIVE-7146](#) - poseplode() UDTF fails with a NullPointerException on NULL columns
- [HIVE-8612](#) - Support metadata result filter hooks
- [HIVE-9475](#) - HiveMetastoreClient.tableExists does not work
- [HIVE-10895](#) - ObjectStore does not close Query objects in some calls, causing a potential leak in some metastore db resources
- [HIVE-11255](#) - get_table_objects_by_name() in HiveMetaStore.java needs to retrieve table objects in multiple batches
- [HIVE-12378](#) - Exception on HBaseSerDe.serialize binary field
- [HUE-3035](#) - [beeswax] Optimize sample data query for partitioned tables
- [IMPALA-1746](#) - QueryExecState does not check for query cancellation or errors
- [IMPALA-1756](#) - Constant filter expressions are not checked for errors and state cleanup not done before throwing exception
- [IMPALA-1917](#) - DCHECK on destroying an ExprContext
- [IMPALA-2141](#) - UnionNode::GetNext() does not check for query errors
- [IMPALA-2264](#) - Fix edge cases for decimal/integer cast
- [IMPALA-2514](#) - DCHECK on destroying an ExprContext
- [OOZIE-2413](#) - Kerberos credentials can expire if the KDC is slow to respond
- [PIG-3641](#) - Split "otherwise" producing incorrect output when combined with ColumnPruning
- [SPARK-11484](#) - [WEBUI] Using proxyBase set by spark instead of env
- [SPARK-11652](#) - [CORE] Remote code execution with InvokerTransformer

Issues Fixed in CDH 5.3.8

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.3.8:

- [CRUNCH-525](#) - Correct (more) accurate default scale factors for built-in MapFn implementations

- [CRUNCH-527](#) - Use hash smearing for partitioning
- [CRUNCH-528](#) - Improve Pair comparison
- [CRUNCH-535](#) - call initCredentials on the job
- [CRUNCH-536](#) - Refactor CrunchControlledJob.Hook interface and make it client-accessible
- [CRUNCH-539](#) - Fix reading WritableComparables bimap
- [CRUNCH-540](#) - Make AvroReflectDeepCopier serializable
- [CRUNCH-542](#) - Eliminate flaky Scrunch sampling test.
- [CRUNCH-543](#) - Have AvroPathPerKeyTarget handle child directories properly
- [CRUNCH-544](#) - Improve performance/serializability of materialized toMap.
- [CRUNCH-547](#) - Properly handle nullability for Avro union types
- [CRUNCH-548](#) - Have the AvroReflectDeepCopier use the class of the source object when constructing new instances instead of the target class
- [CRUNCH-551](#) - Make the use of Configuration objects consistent in CrunchInputSplit and CrunchRecordReader
- [CRUNCH-553](#) - Fix record drop issue that can occur w/From.formattedFile TableSources
- [FLUME-1934](#) - Spooling Directory Source dies on encountering zero-byte files.
- [FLUME-2095](#) - JMS source with TIBCO
- [FLUME-2385](#) - Remove incorrect log message at INFO level in Spool Directory Source.
- [FLUME-2753](#) - Error when specifying empty replace string in Search and Replace Interceptor
- [HADOOP-11105](#) - MetricsSystemImpl could leak memory in registered callbacks
- [HADOOP-11446](#) - S3AOutputStream should use shared thread pool to avoid OutOfMemoryError
- [HADOOP-11463](#) - Replace method-local TransferManager object with S3AFileSystem#transfers.
- [HADOOP-11584](#) - s3a file block size set to 0 in getFileStatus.
- [HADOOP-11607](#) - Reduce log spew in S3AFileSystem.
- [HADOOP-12317](#) - Applications fail on NM restart on some linux distro because NM container recovery declares AM container as LOST
- [HADOOP-12404](#) - Disable caching for JarURLConnection to avoid sharing JarFile with other users when loading resource from URL in Configuration class
- [HADOOP-12413](#) - AccessControlList should avoid calling getGroupNames in isUserInList with empty groups
- [HDFS-7978](#) - Add LOG.isDebugEnabled() guard for some LOG.debug(..)
- [HDFS-8384](#) - Allow NN to startup if there are files having a lease but are not under construction
- [HDFS-8964](#) - When validating the edit log, do not read at or beyond the file offset that is being written
- [HDFS-8965](#) - Harden edit log reading code against out of memory errors
- [MAPREDUCE-5918](#) - LineRecordReader can return the same decompressor to CodecPool multiple times
- [MAPREDUCE-5948](#) - org.apache.hadoop.mapred.LineRecordReader does not handle multibyte record delimiters well
- [MAPREDUCE-6277](#) - Job can post multiple history files if attempt loses connection to the RM
- [MAPREDUCE-6439](#) - AM may fail instead of retrying if RM shuts down during the allocate call.
- [MAPREDUCE-6481](#) - LineRecordReader may give incomplete record and wrong position/key information for uncompressed input sometimes
- [MAPREDUCE-6484](#) - Yarn Client uses local address instead of RM address as token renewer in a secure cluster when RM HA is enabled
- [YARN-3385](#) - Fixed a race-condition in ResourceManager's ZooKeeper based state-store to avoid crashing on duplicate deletes
- [YARN-3469](#) - ZKRMStateStore: Avoid setting watches that are not required.
- [YARN-3990](#) - AsyncDispatcher may overloaded with RMAppNodeUpdateEvent when Node is connected/disconnected
- [HBASE-12639](#) - Backport HBASE-12565 Race condition in HRegion.batchMutate() causes partial data to be written when region closes
- [HBASE-13217](#) - Procedure fails due to ZK issue
- [HBASE-13388](#) - Handling NullPointer in ZKProcedureMemberRpcs while getting ZNode data
- [HBASE-13437](#) - ThriftServer leaks ZooKeeper connections
- [HBASE-13471](#) - Fix a possible infinite loop in doMiniBatchMutation

- [HBASE-13684](#) - Allow mlockagent to be used when not starting as root
- [HBASE-13885](#) - ZK watches leaks during snapshots.
- [HBASE-14045](#) - Bumping thrift version to 0.9.2.
- [HBASE-14302](#) - TableSnapshotInputFormat should not create back references when restoring snapshot
- [HBASE-14354](#) - Minor improvements for usage of the mlock agent
- [HIVE-4867](#) - Deduplicate columns appearing in both the key list and value list of ReduceSinkOperator
- [HIVE-7012](#) - Wrong RS de-duplication in the ReduceSinkDeDuplication Optimizer
- [HIVE-8162](#) - Dynamic sort optimization propagates additional columns even in the absence of order by
- [HIVE-8398](#) - ExprNodeColumnDesc cannot be cast to ExprNodeConstantDesc
- [HIVE-8404](#) - ColumnPruner doesn't prune columns from limit operator
- [HIVE-8560](#) - SerDes that do not inherit AbstractSerDe do not get table properties during initialize()
- [HIVE-9195](#) - CBO changes constant to column type
- [HIVE-9450](#) - Merge[Parquet] Check all data types work for Parquet in Group
- [HIVE-9613](#) - Left join query plan outputs wrong column when using subquery
- [HIVE-9984](#) - JoinReorder's getOutputSize is exponential
- [HIVE-10319](#) - Hive CLI startup takes a long time with a large number of databases
- [HIVE-10572](#) - Improve Hive service test to check empty string
- [HIVE-11077](#) - part ofExchange partition does not properly populate fields for post/pre execute hooks.
- [HIVE-11172](#) - Retrofit Q-Test + Vectorization wrong results for aggregate query with where clause without group by
- [HIVE-11172](#) - Vectorization wrong results for aggregate query with where clause without group by
- [HIVE-11174](#) - Hive does not treat floating point signed zeros as equal (-0.0 should equal 0.0 according to IEEE floating point spec)
- [HIVE-11203](#) - Beeline force option does not force execution when errors occurred in a script.
- [HIVE-11216](#) - UDF GenericUDFMapKeys throws NPE when a null map value is passed in
- [HIVE-11271](#) - java.lang.IndexOutOfBoundsException when union all with if function
- [HIVE-11288](#) - Avro SerDe InstanceCache returns incorrect schema
- [HIVE-11333](#) - ColumnPruner prunes columns of UnionOperator that should be kept
- [HIVE-11590](#) - AvroDeserializer is very chatty
- [HIVE-11657](#) - HIVE-2573 introduces some issues during metastore init (and CLI init)
- [HIVE-11695](#) - If user have no permission to create LOCAL DIRECTORY `Hive` does not throw any exception and fail silently.
- [HIVE-11696](#) - Exception when table-level serde is Parquet while partition-level serde is JSON
- [HIVE-11816](#) - Upgrade groovy to 2.4.4
- [HIVE-11824](#) - Insert to local directory causes staging directory to be copied
- [HIVE-11995](#) - Remove repetitively setting permissions in insert/load overwrite partition
- [HUE-2880](#) - [hadoop] Fix uploading large files to a kerberized HTTPFS
- [HUE-2893](#) - [desktop] Backport CherryPy SSL file upload fix
- [IMPALA-1929](#) - Avoiding a DCHECK of NULL hash table in spilled right joins
- [IMPALA-2133](#) - Properly unescape string value for HBase filters
- [IMPALA-2165](#) - Avoid cardinality 0 in scan nodes of small tables and low selectivity
- [IMPALA-2178](#) - fix Expr::ComputeResultsLayout() logic
- [IMPALA-2314](#) - LargestSpilledPartition was not checking if partition is closed
- [IMPALA-2364](#) - Wrong DCHECK in PHJ::ProcessProbeBatch
- [KITE-1053](#) - Fix int overflow bug in FS writer.
- [KITE-1074](#) - Partial updates aka Atomic updates with loadSolr aren't recognized with Solrcloud
- [MAHOUT-1771](#) - Cluster dumper omits indices and 0 elements for dense vector or sparse containing 0s, this closes apache/mahout#158
- [MAHOUT-1771](#) - Cluster dumper omits indices and 0 elements for dense vector or sparse containing 0s closes apache/mahout #158

- [PIG-4024](#) - TestPigStreamingUDF and TestPigStreaming fail on IBM JDK
- [PIG-4326](#) - AvroStorageSchemaConversionUtilities does not properly convert schema for maps of arrays of records
- [PIG-4338](#) - Fix test failures with JDK8
- [SENTRY-799](#) - unit test forFix TestDbEndToEnd flaky test - drop table/dbs before creating
- [SENTRY-878](#) - collect_list missing from HIVE_UDF_WHITE_LIST
- [SENTRY-893](#) - Synchronize calls in SentryClient and create sentry client once per request in SimpleDBProvider
- [SOLR-5496](#) - Ensure all http CMs get shutdown.
- [SOLR-7956](#) - There are interrupts on shutdown in places that can cause ChannelAlreadyClose
- [SOLR-7999](#) - SolrRequestParserTest#testStreamURL started failing.
- [SPARK-6480](#) - [CORE] histogram() bucket function is wrong in some simple edge cases
- [SPARK-6880](#) - [CORE]Fixed null check when all the dependent stages are cancelled due to previous stage failure
- [SPARK-8606](#) - Prevent exceptions in RDD.getPreferredLocations() from crashing DAGScheduler

Issues Fixed in CDH 5.3.6

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.3.6:

- [CRUNCH-516](#) - Scrunch needs some additional null checks
- [CRUNCH-508](#) - Improve performance of Scala Enumeration counters in Scrunch
- [CRUNCH-514](#) - AvroDerivedDeepCopier should initialize delegate MapFns
- [CRUNCH-530](#) - Fix object reuse bug in GenericRecordToTuple
- [HADOOP-12158](#) - Improve error message in TestCryptoStreamsWithOpensslAesCtrCryptoCodec when OpenSSL is not installed
- [HADOOP-11711](#) - Provide a default value for AES/CTR/NoPadding CryptoCodec classes
- [HADOOP-12103](#) - Small refactoring of DelegationTokenAuthenticationFilter to allow code sharing
- [HADOOP-8151](#) - Error handling in snappy decompressor throws invalid exceptions
- [HADOOP-11969](#) - ThreadLocal initialization in several classes is not thread safe
- [HDFS-7443](#) - Datanode upgrade to BLOCKID_BASED_LAYOUT fails if duplicate block files are present in the same volume
- [HDFS-8337](#) - Accessing https via webhdfs doesn't work from a jar with kerberos
- [HDFS-7546](#) - Document, and set an accepting default for dfs.namenode.kerberos.principal.pattern
- [HDFS-8656](#) - Preserve compatibility of ClientProtocol#rollingUpgrade after finalization
- [HDFS-7894](#) - Rolling upgrade readiness is not updated in jmx until query command is issued.
- [HDFS-8127](#) - NameNode Failover during HA upgrade can cause DataNode to finalize upgrade
- [HDFS-3443](#) - Fix NPE when namenode transition to active during startup by adding checkNNStartup() in NameNodeRpcServer
- [YARN-3143](#) - RM Apps REST API can return NPE or entries missing id and other fields
- [HBASE-13995](#) - ServerName is not fully case insensitive
- [HBASE-13430](#) - HFiles that are in use by a table cloned from a snapshot may be deleted when that snapshot is deleted
- [HBASE-12539](#) - HFileLinkCleaner logs are uselessly noisy
- [HBASE-11898](#) - CoprocessorHost.Environment should cache class loader instance
- [HBASE-13826](#) - Unable to create table when group acls are appropriately set.
- [HBASE-13241](#) - Add tests for group level grants
- [HBASE-13239](#) - HBase grant at specific column level does not work for Groups
- [HBASE-13789](#) - ForeignException should not be sent to the client
- [HBASE-13779](#) - Calling table.exists() before table.get() end up with an empty Result
- [HBASE-13780](#) - Default to 700 for HDFS root dir permissions for secure deployments
- [HBASE-13768](#) - ZooKeeper znodes are bootstrapped with insecure ACLs in a secure configuration
- [HBASE-13767](#) - Allow ZKAcReset to set and not just clear ZK ACLs
- [HBASE-13086](#) - Show ZK root node on Master WebUI

CDH 5 Release Notes

- [HBASE-13342](#) - Fix incorrect interface annotations
- [HBASE-13162](#) - Add capability for cleaning hbase acls to hbase cleanup script.
- [HBASE-12641](#) - Grant all permissions of hbase zookeeper node to hbase superuser in a secure cluster
- [HBASE-12414](#) - Move HFileLink.exists() to base class
- [HIVE-11150](#) - Remove wrong warning message related to chgrp
- [HIVE-8318](#) - Null Scan optimizer throws exception when no partitions are selected
- [HIVE-7385](#) - Optimize for empty relation scans
- [HIVE-7299](#) - Enable metadata only optimization on Tez
- [HIVE-10808](#) - Inner join on Null throwing Cast Exception
- [HIVE-9087](#) - The move task does not handle properly in the case of loading data from the local file system path.
- [HIVE-9325](#) - Handle the case of insert overwrite statement with a qualified path that the destination path does not have a schema.
- [HIVE-9349](#) - Remove the schema in the getQualifiedPathWithoutSchemeAndAuthority method
- [HIVE-9328](#) - Tests cannot move files due to change on HIVE-9325
- [HIVE-6024](#) - Load data local inpath unnecessarily creates a copy task
- [HIVE-10841](#) - [WHERE col is not null] does not work sometimes for queries with many JOIN statements
- [HIVE-9620](#) - Cannot retrieve column statistics using HMS API if column name contains uppercase characters
- [HIVE-8863](#) - Cannot drop table with uppercase name after "compute statistics for columns"
- [HIVE-10629](#) - Dropping table in an encrypted zone does not drop warehouse directory
- [HIVE-10630](#) - Renaming tables across encryption zones renames table even though the operation throws error
- [HIVE-10956](#) - HS2 leaks HMS connections
- [HIVE-8298](#) - Incorrect results for n-way join when join expressions are not in same order across joins
- [HIVE-8895](#) - bugs in mergejoin
- [HIVE-10771](#) - "separatorChar" has no effect in "CREATE TABLE AS SELECT" statement
- [HIVE-6679](#) - HiveServer2 should support configurable the server side socket timeout and keepalive for various transports types where applicable
- [HIVE-10732](#) - Hive JDBC driver does not close operation for metadata queries
- [HIVE-7027](#) - Hive job fails when referencing a view that explodes an array
- [IMPALA-1774](#) - Allow querying Parquet tables with complex-typed columns as long as those columns are not selected
- [IMPALA-1919](#) - Avoid calling ProcessBatch with out_batch->AtCapacity in right joins
- [IMPALA-2002](#) - Provide way to cache ext data source classes
- [IMPALA-1726](#) - Move JNI / Thrift utilities to separate header
- [HUE-2813](#) - [hive] Report when Hue server is down when trying to execute a query
- [HUE-2243](#) - [metastore] Listing tables can be very slow
- [OOZIE-1944](#) - Recursive variable resolution broken when same parameter name in config-default and action conf
- [PIG-4053](#) - TestMRCompiler succeeded with sun jdk 1.6 while failed with sun jdk 1.7
- [SENTRY-721](#) - HDFS Cascading permissions not applied to child file ACLs if a direct grant exists
- [SENTRY-699](#) - Memory leak when running Sentry w/ HiveServer2
- [SOLR-6146](#) - Leak in CloudSolrServer causing "Too many open files"
- [SOLR-7503](#) - Recovery after ZK session expiration happens in a single thread for all cores in a node

Issues Fixed in CDH 5.3.5

Potential job failures during YARN rolling upgrades to CDH 5.3.4

Problem: A MapReduce security fix introduced a compatibility issue that results in job failures during YARN rolling upgrades from CDH 5.3.3 to CDH 5.3.4.

Release affected: CDH 5.3.4

Release containing the fix: CDH 5.3.5

Workarounds: You can use any one of the following workarounds for this issue:

- Upgrade to CDH 5.3.5.
- Restart any jobs that might have failed during the upgrade.
- Explicitly set the version of MapReduce to be used so it is picked on a per-job basis.
 1. Update the YARN property, **MR Application Classpath** (`mapreduce.application.classpath`), either in Cloudera Manager or in the `mapred-site.xml` file. Remove all existing values and add a new entry: `<parcel-path>/lib/hadoop-mapreduce/*`, where `<parcel-path>` is the absolute path to the parcel installation. For example, the default installation path for the CDH 5.3.3 parcel would be: `/opt/cloudera/parcels/CDH-5.3.3-1.cdh5.3.3.p0.5/lib/hadoop-mapreduce/*`.
 2. Wait until jobs submitted with the above client configuration change have run to completion.
 3. Upgrade to CDH 5.3.4.
 4. Update the **MR Application Classpath** (`mapreduce.application.classpath`) property to point to the new CDH 5.3.4 parcel.

Do not delete the old parcel until after all jobs submitted prior to the upgrade have finished running.

Upstream Issues Fixed

The following upstream issue has been fixed in CDH 5.3.5:

- [YARN-3811](#) - NodeManager restarts could lead to application failures

Issues Fixed in CDH 5.3.4

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.3.4:

- [HDFS-7980](#) - Incremental BlockReport will dramatically slow down the startup of a namenode
- [HDFS-8380](#) - Always call addStoredBlock on blocks which have been shifted from one storage to another
- [HDFS-7645](#) - Rolling upgrade is restoring blocks from trash multiple times
- [HDFS-7869](#) - Inconsistency in the return information while performing rolling upgrade
- [HDFS-7340](#) - make rollingUpgrade start/finalize idempotent
- [HDFS-7312](#) - Update DistCp v1 to optionally not use tmp location (branch-1 only)
- [HDFS-7530](#) - Allow renaming of encryption zone roots
- [HDFS-7587](#) - Edit log corruption can happen if append fails with a quota violation
- [YARN-3485](#) - FairScheduler headroom calculation doesn't consider maxResources for Fifo and FairShare policies
- [YARN-3491](#) - PublicLocalizer#addResource is too slow.
- [YARN-3021](#) - YARN's delegation-token handling disallows certain trust setups to operate properly over DistCp
- [YARN-3241](#) - FairScheduler handles "invalid" queue names inconsistently
- [YARN-3022](#) - Expose Container resource information from NodeManager for monitoring
- [YARN-2984](#) - Metrics for container's actual memory usage
- [YARN-3465](#) - Use LinkedHashMap to preserve order of resource requests
- [MAPREDUCE-6339](#) - Job history file is not flushed correctly because isTimerActive flag is not set true when flushTimerTask is scheduled.
- [MAPREDUCE-5710](#) - Backport MAPREDUCE-1305 to branch-1
- [MAPREDUCE-6238](#) - MR2 can't run local jobs with -libjars command options which is a regression from MR1
- [MAPREDUCE-6076](#) - Zero map split input length combine with none zero map split input length may cause MR1 job hung sometimes.
- [HBASE-13374](#) - Small scanners (with particular configurations) do not return all rows
- [HBASE-13269](#) - Limit result array preallocation to avoid OOME with large scan caching values
- [HBASE-13422](#) - remove use of StandardCharsets in 0.98
- [HBASE-13335](#) - Update ClientSmallScanner and ClientSmallReversedScanner
- [HBASE-13262](#) - ResultScanner doesn't return all rows in Scan
- [HIVE-10646](#) - ColumnValue does not handle NULL_TYPE
- [HIVE-10453](#) - HS2 leaking open file descriptors when using UDFs

CDH 5 Release Notes

- [HIVE-9655](#) - Dynamic partition table insertion error
- [HIVE-10452](#) - Followup fix for HIVE-10202 to restrict it for script mode.
- [HIVE-10312](#) - SASL.QOP in JDBC URL is ignored for Delegation token Authentication
- [HIVE-10202](#) - Beeline outputs prompt+query on standard output when used in non-interactive mode
- [HIVE-10087](#) - Beeline's --silent option should suppress query from being echoed when running with -f option
- [HIVE-10085](#) - Lateral view on top of a view throws RuntimeException
- [HIVE-2828](#) - make timestamp accessible in the hbase KeyValue
- [HUE-2741](#) - [home] Hide the document move dialog
- [HUE-2732](#) - Hue isn't correctly doing add_column migrations with non-blank defaults
- [HUE-2513](#) - [fb] File list column sorting is broken
- [IMPALA-1519](#) - Fix wrapping of exprs via a TupleIsNotNullPredicate with analytics
- [IMPALA-1952](#) - Expand parsing of decimals to include scientific notation
- [IMPALA-1860](#) - INSERT/CTAS evaluates and applies constant predicates.
- [IMPALA-1900](#) - Assign predicates below analytic functions with a compatible partition by clause
- [IMPALA-1376](#) - Split up Planner into multiple classes.
- [IMPALA-1888](#) - FIRST_VALUE may produce incorrect results with preceding windows
- [IMPALA-1559](#) - FIRST_VALUE rewrite fn type might not match slot type
- [IMPALA-1808](#) - AnalyticEvalNode cannot handle partition/order by exprs with NaN
- [IMPALA-1562](#) - AnalyticEvalNode not properly handling nullable tuples
- [OOZIE-2063](#) - Cron syntax creates duplicate actions
- [OOZIE-2218](#) - META-INF directories in the war file have 777 permissions
- [OOZIE-1878](#) - Can't execute dryrun on the CLI
- [SENTRY-696](#) - Improve Metastoreplugin Cache Initialization time
- [SENTRY-703](#) - Calls to add_partition fail when passed a Partition object with a null location
- [SENTRY-408](#) - The URI permission should support more filesystem prefixes
- [SOLR-7478](#) - UpdateLog#close shutdown it's executor with interrupts before running close, preventing a clean close.
- [SOLR-7437](#) - Make HDFS transaction log replication factor configurable.
- [SOLR-7338](#) - A reloaded core will never register itself as active after a ZK session expiration
- [SOLR-7370](#) - FSHDFSUtils#recoverFileLease tries to recover the lease every one second after the first four second wait.
- [SPARK-6578](#) - Outbound channel in network library is not thread-safe, can lead to fetch failures
- [SQOOP-2343](#) - AsyncSqlRecordWriter stuck if any exception is thrown out in its close method
- [SQOOP-2286](#) - Ensure Sqoop generates valid avro column names
- [SQOOP-2283](#) - Support usage of --exec and --password-alias
- [SQOOP-2281](#) - Set overwrite on kite dataset
- [SQOOP-2282](#) - Add validation check for --hive-import and --append
- [SQOOP-2257](#) - Parquet target for imports with Hive overwrite option does not work
- [ZOOKEEPER-2146](#) - BinaryInputArchive readString should check length before allocating memory
- [ZOOKEEPER-2149](#) - Logging of client address when socket connection established

Published Known Issues Fixed

As a result of the above fixes, the following issues, previously published as [Known Issues in CDH 5](#) on page 111, are also fixed.

Executing oozie job -config properties file -dryrun fails because of a code defect in argument parsing

Bug: [OOZIE-1878](#)

Severity: Low

Workaround: None.

Issues Fixed in CDH 5.3.3

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.3.3:

- [HADOOP-11722](#) - Some Instances of Services using ZKDelegationTokenSecretManager go down when old token cannot be deleted
- [HADOOP-11469](#) - KMS should skip default.key.acl and whitelist.key.acl when loading key acl
- [HADOOP-11710](#) - Make CryptoOutputStream behave like DFSOutputStream wrt synchronization
- [HADOOP-11674](#) - oneByteBuf in CryptoInputStream and CryptoOutputStream should be non static
- [HADOOP-11445](#) - Bzip2Codec: Data block is skipped when position of newly created stream is equal to start of split
- [HADOOP-11620](#) - Add support for load balancing across a group of KMS for HA
- [HDFS-6830](#) - BlockInfo.addStorage fails when DN changes the storage for a block replica
- [HDFS-7961](#) - Trigger full block report after hot swapping disk
- [HDFS-7960](#) - The full block report should prune zombie storages even if they're not empty
- [HDFS-7575](#) - Upgrade should generate a unique storage ID for each volume
- [HDFS-7596](#) - NameNode should prune dead storages from storageMap
- [HDFS-7579](#) - Improve log reporting during block report rpc failure
- [HDFS-7208](#) - NN doesn't schedule replication when a DN storage fails
- [HDFS-6899](#) - Allow changing MiniDFSCluster volumes per DN and capacity per volume
- [HDFS-6878](#) - Change MiniDFSCluster to support StorageType configuration for individual directories
- [HDFS-6678](#) - MiniDFSCluster may still be partially running after initialization fails.
- [YARN-3351](#) - AppMaster tracking URL is broken in HA
- [YARN-3242](#) - Asynchrony in ZK-close can lead to ZKRMStateStore watcher receiving events for old client
- [YARN-2865](#) - Application recovery continuously fails with "Application with id already present. Cannot duplicate"
- [MAPREDUCE-6275](#) - Race condition in FileOutputCommitter v2 for user-specified task output subdirs
- [MAPREDUCE-4815](#) - Speed up FileOutputCommitter#commitJob for many output files
- [HBASE-13131](#) - ReplicationAdmin leaks connections if there's an error in the constructor
- [HIVE-10086](#) - Hive throws error when accessing Parquet file schema using field name match
- [HIVE-10098](#) - HS2 local task for map join fails in KMS encrypted cluster
- [HIVE-7426](#) - ClassCastException: ...IntWritable cannot be cast to ...Text involving ql.udf.generic.GenericUDFBasePad.evaluate
- [HIVE-7737](#) - Hive logs full exception for table not found
- [HIVE-9749](#) - ObjectStore schema verification logic is incorrect
- [HIVE-9788](#) - Make double quote optional in tsv/csv/dsv output
- [HIVE-9755](#) - Hive built-in "ngram" UDAF fails when a mapper has no matches.
- [HIVE-9770](#) - Beeline ignores --showHeader for non-tabular output formats i.e csv,tsv,dsv
- [HIVE-8688](#) - serialized plan OutputStream is not being closed
- [HIVE-9716](#) - Map job fails when table's LOCATION does not have scheme
- [HIVE-5857](#) - Reduce tasks do not work in uber mode in YARN
- [HIVE-8938](#) - Compiler should save the transform URI as input entity
- [HUE-2569](#) - [home] Delete project is broken
- [HUE-2529](#) - Increase the character limit of 'Name' Textfield in Useradmin Ldap Sync Groups
- [HUE-2506](#) - [search] Marker map does not display with HTML widget
- [HUE-1663](#) - [core] Option to either follow or not LDAP referrals for auth
- [HUE-2198](#) - [core] Reduce noise such as "handle_other(): Mutual authentication unavailable on 200 response"
- [SENTRY-683](#) - HDFS service client should ensure the kerberos ticket validity before new service connection
- [SENTRY-654](#) - Calls to append_partition fail when Sentry is enabled
- [SENTRY-664](#) - After Namenode is restarted, Path updates remain unsynced
- [SENTRY-665](#) - PathsUpdate.parsePath needs to handle special characters
- [SENTRY-652](#) - Sentry fails to parse spaces when HDFS ACL sync enabled

CDH 5 Release Notes

- [SOLR-7092](#) - Stop the HDFS lease recovery retries on HdfsTransactionLog on close and try to avoid lease recovery on closed files.
- [SOLR-7141](#) - RecoveryStrategy: Raise time that we wait for any updates from the leader before they saw the recovery state to have finished.
- [SOLR-7113](#) - Multiple calls to UpdateLog#init is not thread safe with respect to the HDFS FileSystem client object usage.
- [SOLR-7134](#) - Replication can still cause index corruption.
- [SQOOP-1764](#) - Numeric Overflow when getting extent map
- [IMPALA-1658](#) - Add compatibility flag for Hive-Parquet-Timestamps
- [IMPALA-1820](#) - Start with small pages for hash tables during repartitioning
- [IMPALA-1897](#) - Fixes for old hash join and agg
- [IMPALA-1894](#) - Fix old aggregation node hash table cleanup
- [IMPALA-1863](#) - Avoid deadlock across fragment instances
- [IMPALA-1915](#) - Fix query hang in BufferedBlockMgr:FindBlock()
- [IMPALA-1890](#) - Fixing a race between ~BufferedBlockMgr() and the WriteComplete() call
- [IMPALA-1738](#) - Use snprintf() instead of lexical_cast() in float-to-string casts
- [IMPALA-1865](#) - Fix partition spilling cleanup when new stream OOMs
- [IMPALA-1835](#) - Keep the fragment alive for TransmitData()
- [IMPALA-1805](#) - Impala's ACLs check do not consider all group ACLs, only checked first one.
- [IMPALA-1794](#) - Fix infinite loop opening or closing file with invalid metadata
- [IMPALA-1801](#) - external-data-source-executor leaking global jni refs
- [IMPALA-1712](#) - Unexpected remote bytes read counter was not being reset properly
- [IMPALA-1636](#) - Generalize index-based partition pruning to allow constant expressions

Published Known Issues Fixed

As a result of the above fixes, the following issues, previously published as [Known Issues in CDH 5](#) on page 111, are also fixed.

After upgrade from a release earlier than CDH 5.2.0, storage IDs may no longer be unique

As of CDH 5.2, each storage volume on a DataNode should have its own unique `storageID`, but in clusters upgraded from CDH 4, or CDH 5 releases earlier than CDH 5.2.0, each volume on a given DataNode shares the same `storageID`, because the HDFS upgrade does not properly update the IDs to reflect the new naming scheme. This causes problems with load balancing. The problem affects only clusters upgraded from CDH 5.1.x and earlier to CDH 5.2 or later. Clusters that are new as of CDH 5.2.0 or later do not have the problem.

Bug: [HDFS-7575](#)

Severity: Medium

Workaround: Upgrade to a later or patched version of CDH.

Issues Fixed in CDH 5.3.2

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.3.2:

- [AVRO-1630](#) - Creating Builder from instance loses data
- [AVRO-1628](#) - Add Schema.createUnion(Schema... type)
- [AVRO-1539](#) - Add FileSystem-based FslInput Constructor
- [AVRO-1623](#) - GenericData#validate() of enum: IndexOutOfBoundsException
- [AVRO-1614](#) - Always getting a value...
- [AVRO-1592](#) - Java keyword as an enum constant in Avro schema file causes deserialization to fail.
- [AVRO-1619](#) - Generate better JavaDoc
- [AVRO-1622](#) - Add missing license headers
- [AVRO-1604](#) - ReflectData.AllowNull fails to generate schemas when @Nullable is present.

- [AVRO-1407](#) - NettyTransceiver can cause a infinite loop when slow to connect
- [AVRO-834](#) - Data File corruption recovery tool
- [AVRO-1596](#) - Cannot read past corrupted block in Avro data file
- [HADOOP-11350](#) - The size of header buffer of HttpServer is too small when HTTPS is enabled
- [HDFS-7707](#) - Edit log corruption due to delayed block removal again
- [HDFS-7718](#) - Store KeyProvider in ClientContext to avoid leaking key provider threads when using FileContext
- [HDFS-6425](#) - Large postponedMisreplicatedBlocks has impact on blockReport latency
- [HDFS-7560](#) - ACLs removed by removeDefaultAcl() will be back after NameNode restart/failover
- [HDFS-7513](#) - HDFS inotify: add defaultBlockSize to CreateEvent
- [HDFS-7158](#) - Reduce the memory usage of WebImageViewer
- [HDFS-7497](#) - Inconsistent report of decommissioning DataNodes between dfsadmin and NameNode webui
- [HDFS-6917](#) - Add an hdfs debug command to validate blocks, call recoverlease, etc.
- [HDFS-6779](#) - Add missing version subcommand for hdfs
- [YARN-2697](#) - RMAuthenticationHandler is no longer useful
- [YARN-2656](#) - RM web services authentication filter should add support for proxy user
- [YARN-3082](#) - Non thread safe access to systemCredentials in NodeHeartbeatResponse processing
- [YARN-3079](#) - Scheduler should also update maximumAllocation when updateNodeResource.
- [YARN-2992](#) - ZKRMStateStore crashes due to session expiry
- [YARN-2675](#) - containersKilled metrics is not updated when the container is killed during localization
- [YARN-2715](#) - Proxy user is problem for RPC interface if yarn.resourcemanager.webapp.proxyuser is not set
- [MAPREDUCE-6198](#) - NPE from JobTracker#resolveAndAddToTopology in MR1 cause initJob and heartbeat failure.
- [MAPREDUCE-6196](#) - Fix BigDecimal ArithmeticException in PiEstimator
- [HBASE-12540](#) - TestRegionServerMetrics#testMobMetrics test failure
- [HBASE-12533](#) - staging directories are not deleted after secure bulk load
- [HBASE-12077](#) - FilterLists create many ArrayList\$Iter objects per row.
- [HBASE-12386](#) - Replication gets stuck following a transient zookeeper error to remote peer cluster
- [HBASE-11979](#) - Compaction progress reporting is wrong
- [HBASE-12445](#) - hbase is removing all remaining cells immediately after the cell marked with marker = KeyValue.Type.DeleteColumn via PUT
- [HIVE-7647](#) - Beeline does not honor --headerInterval and --color when executing with "-e"
- [HIVE-7733](#) - Ambiguous column reference error on query
- [HIVE-9303](#) - Parquet files are written with incorrect definition levels
- [HIVE-8444](#) - update pom to junit 4.11
- [HIVE-9474](#) - truncate table changes permissions on the target
- [HIVE-9462](#) - HIVE-8577 - breaks type evolution
- [HIVE-9482](#) - Hive parquet timestamp compatibility
- [HIVE-6308](#) - COLUMNS_V2 Metastore table not populated for tables created without an explicit column list.
- [HIVE-9502](#) - Parquet cannot read Map types from files written with Hive 0.12 or earlier
- [HIVE-9445](#) - Revert HIVE-5700 - enforce single date format for partition column storage
- [HIVE-9393](#) - reduce noisy log level of ColumnarSerDe.java:116 from INFO to DEBUG
- [HIVE-7800](#) - Parquet Column Index Access Schema Size Checking
- [HIVE-9330](#) - DummyTxnManager will throw NPE if WriteEntity writeType has not been set
- [HIVE-9265](#) - Hive with encryption throws NPE to fs path without schema
- [HIVE-9199](#) - Excessive exclusive lock used in some DDLs with DummyTxnManager
- [HIVE-6978](#) - beeline always exits with 0 status, should exit with non-zero status on error
- [HUE-2556](#) - [core] Cannot update project tags of a document
- [HUE-2528](#) - Partitions limit gets capped to 1000 despite configuration
- [HUE-2548](#) - [metastore] Create table then load data does redirect to the table page
- [HUE-2525](#) - [core] Fix manual install of samples
- [HUE-2501](#) - [metastore] Creating a table with header files bigger than 64MB truncates it

CDH 5 Release Notes

- [HUE-2484](#) - [beeswax] Configure support for Hive Server2 LDAP authentication
- [HUE-2532](#) - [search] Fix share URL on Internet Explorer
- [HUE-2531](#) - [impala] Autogrow missing result list
- [HUE-2524](#) - [impala] Sort numerically recent queries tab
- [HUE-2495](#) - [oozie] Improve dashboards sorting mechanism
- [HUE-2511](#) - [impala] Infinite scroll keeps fetching results even if finished
- [HUE-2102](#) - [oozie] Workflow with credentials can't be used with Coordinator
- [HUE-2152](#) - [pig] Credentials support in editor
- [OOZIE-2131](#) - Add flag to sqoop action to skip hbase delegation token generation
- [OOZIE-2047](#) - Oozie does not support Hive tables that use datatypes introduced since Hive 0.8
- [OOZIE-2102](#) - Streaming actions are broken cause of incorrect method signature
- [PARQUET-173](#) - StatisticsFilter doesn't handle And properly
- [PARQUET-157](#) - Divide by zero in logging code
- [PARQUET-142](#) - parquet-tools doesn't filter _SUCCESS file
- [PARQUET-124](#) - parquet.hadoop.ParquetOutputCommitter.commitJob() throws parquet.io.ParquetEncodingException
- [PARQUET-136](#) - NPE thrown in StatisticsFilter when all values in a string/binary column trunk are null
- [PARQUET-168](#) - Wrong command line option description in parquet-tools
- [PARQUET-145](#) - InternalParquetRecordReader.close() should not throw an exception if initialization has failed
- [PARQUET-140](#) - Allow clients to control the GenericData object that is used to read Avro records
- [SOLR-7033](#) - [RecoveryStrategy should not publish any state when closed / cancelled.
- [SOLR-5961](#) - Solr gets crazy on /overseer/queue state change
- [SOLR-6640](#) - Replication can cause index corruption
- [SOLR-5875](#) - QueryComponent.mergeIds() unmarshals all docs' sort field values once per doc instead of once per shard
- [SOLR-6919](#) - Log REST info before executing
- [SOLR-6969](#) - When opening an HDFSTransactionLog for append we must first attempt to recover it's lease to prevent data loss.
- [SOLR-5515](#) - NPE when getting stats on date field with empty result on solrcloud
- [SPARK-3778](#) - newAPIHadoopRDD doesn't properly pass credentials for secure hdfs on yarn
- [SPARK-4835](#) - Streaming saveAs*HadoopFiles() methods may throw FileAlreadyExistsException during checkpoint recovery
- [SQOOP-2057](#) - Skip delegation token generation flag during hbase import
- [SQOOP-1779](#) - Add support for --hive-database when importing Parquet files into Hive
- [IMPALA-1622](#) - Fix overflow in StringParser::StringToFloatInternal()
- [IMPALA-1614](#) - Compute stats fails if table name starts with number
- [IMPALA-1623](#) - unix_timestamp() does not return correct time
- [IMPALA-1535](#) - Partition pruning with NULL
- [IMPALA-1606](#) - Impala does not always give short name to Llama
- [IMPALA-1120](#) - Fetch column statistics using Hive 0.13 bulk API

In addition, CDH 5.3.2 reverts [YARN-2713](#), which has caused problems since its inclusion in CDH 5.3.0.

Published Known Issues Fixed

As a result of the above fixes, the following issues, previously published as [Known Issues in CDH 5](#) on page 111, are also fixed.

Hive does not support Parquet schema evolution

Adding a new column to a Parquet table causes queries on that table to fail with a `column not found` error.

Bug: [HIVE-7800](#)

Severity: Medium

Workaround: Use Impala instead; Impala handles Parquet schema evolution correctly.

Issues Fixed in CDH 5.3.1

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.3.1:

- [YARN-2975](#) - FSLeafQueue app lists are accessed without required locks
- [YARN-2010](#) - Handle app-recovery failures gracefully
- [YARN-3027](#) - Scheduler should use totalAvailable resource from node instead of availableResource for maxAllocation
- [HIVE-9445](#) - Revert HIVE-5700 - enforce single date format for partition column storage
- [IMPALA-1668](#) - TSaslServerTransport::Factory::getTransport() leaks transport map entries
- [IMPALA-1674](#) - IMPALA-1556 causes memory leak with secure connections

Published Known Issues Fixed

As a result of the above fixes, the following issues, previously published as [Known Issues in CDH 5](#) on page 111, are also fixed.

Upgrading a PostgreSQL Hive Metastore from Hive 0.12 to Hive 0.13 may result in a corrupt metastore

[HIVE-5700](#) introduced a serious bug into the Hive Metastore upgrade scripts. This bug affects users who have a PostgreSQL Hive Metastore and have *at least one table* which is partitioned by date and the value is stored as a date type (not string).

Bug: [HIVE-5700](#)

Severity: High

Workaround: None. Do not upgrade your PostgreSQL metastore to version 0.13 if you satisfy the condition stated above.

Issues Fixed in CDH 5.3.0

The following topics describe known issues fixed in CDH 5.3.0.

[Apache Hadoop](#)

[HDFS](#)

Kerberos re-login attempts fail when using JDK 1.7.0_80

On clusters using JDK 1.7.0_80, long running HDFS clients are unable to re-authenticate using Kerberos once their ticket expires. Due to this authentication failure, any jobs triggered from these clients will fail.

Releases Affected: CDH 5.1, 5.2

Bug: [HADOOP-10786](#)

Workaround: Upgrade to CDH 5.3.2 (or higher).

NameNode - KMS communication fails after long periods of inactivity

Encrypted files and encryption zones cannot be created if a long period of time (by default, 20 hours) has passed since the last time the KMS and NameNode communicated.

Bug: [HADOOP-11187](#)

Workaround: There are two possible workarounds to this issue:

- You can increase the KMS authentication token validity period to a very high number. Since the default value is 10 hours, this bug will only be encountered after 20 hours of no communication between the NameNode and the KMS. Add the following property to the `kms-site.xml` Safety Valve:

```
<property>
<name>hadoop.kms.authentication.token.validity</name>
<value>SOME VERY HIGH NUMBER</value>
</property>
```

CDH 5 Release Notes

- You can switch the KMS signature secret provider to the string secret provider by adding the following property to the `kms-site.xml` Safety Valve:

```
<property>
<name>hadoop.kms.authentication.signature.secret</name>
<value>SOME VERY SECRET STRING</value>
</property>
```

DataNodes may become unresponsive to block creation requests

In releases earlier than CDH 5.2.3, DataNodes may become unresponsive to block creation requests from clients when the directory scanner is running.

Bug: [HDFS-7489](#)

Workaround: Upgrade to CDH 5.2.3 or later.

Apache Hive

UDF translate() does not accept arguments of type CHAR or VARCHAR

Bug: [HIVE-6622](#)

Workaround: Cast the argument to type String.

Hive's Timestamp type cannot be stored in Parquet

Tables containing timestamp columns cannot use Parquet as the storage engine.

Bug: [HIVE-6394](#)

Workaround: Use a different file format.

Apache Spark

Spark sort-based shuffle is affected by a kernel bug

Spark sort-based shuffle is affected by a kernel bug

(<http://git.kernel.org/cgit/linux/kernel/git/torvalds/linux.git/commit/?id=2cb4b05e7647891b46b91c07c9a60304803d1688>). The kernel bug was fixed in RHEL/CentOS 6.2.



Note: Previously CDH defaulted to hash-based shuffle. It now defaults to sort-based shuffle.

Bug: [SPARK-3948](#)

Issues Fixed in CDH 5.2.x

The following topics describe known issues fixed in CDH 5.2.x, from newest to oldest release.

CDH 5.2.6

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.2.6:

- [CRUNCH-516](#) - Scrunch needs some additional null checks
- [CRUNCH-508](#) - Improve performance of Scala Enumeration counters in Scrunch
- [CRUNCH-514](#) - AvroDerivedDeepCopier should initialize delegate MapFns
- [CRUNCH-530](#) - Fix object reuse bug in GenericRecordToTuple
- [HADOOP-12103](#) - Small refactoring of DelegationTokenAuthenticationFilter to allow code sharing
- [HADOOP-10839](#) - Add unregisterSource() to MetricsSystem API
- [HDFS-8337](#) - Accessing https via webhdfs doesn't work from a jar with kerberos
- [HDFS-7546](#) - Document, and set an accepting default for dfs.namenode.kerberos.principal.pattern
- [HDFS-6997](#) - Archival Storage: add more tests for data migration and replicaion

- [HDFS-7980](#) - Incremental BlockReport will dramatically slow down the startup of a namenode
- [HDFS-8380](#) - Always call addStoredBlock on blocks which have been shifted from one storage to another
- [HDFS-7312](#) - Update DistCp v1 to optionally not use tmp location (branch-1 only)
- [YARN-3485](#) - FairScheduler headroom calculation doesn't consider maxResources for Fifo and FairShare policies
- [YARN-3241](#) - FairScheduler handles "invalid" queue names inconsistently
- [YARN-2669](#) - FairScheduler: queue names shouldn't allow periods
- [YARN-3022](#) - Expose Container resource information from NodeManager for monitoring
- [YARN-2984](#) - Metrics for container's actual memory usage
- [YARN-3465](#) - Use LinkedHashMap to preserve order of resource requests
- [MAPREDUCE-6387](#) - Serialize the recently added Task#encryptedSpillKey field at the end
- [MAPREDUCE-6339](#) - Job history file is not flushed correctly because isTimerActive flag is not set true when flushTimerTask is scheduled.
- [MAPREDUCE-5710](#) - Backport MAPREDUCE-1305 to branch-1
- [MAPREDUCE-6238](#) - MR2 can't run local jobs with -libjars command options which is a regression from MR1
- [HBASE-13826](#) - Unable to create table when group acls are appropriately set.
- [HBASE-13241](#) - Add tests for group level grants
- [HBASE-13239](#) - HBase grant at specific column level does not work for Groups
- [HBASE-13768](#) - ZooKeeper znodes are bootstrapped with insecure ACLs in a secure configuration
- [HBASE-13789](#) - ForeignException should not be sent to the client
- [HBASE-13779](#) - Calling table.exists() before table.get() end up with an empty Result
- [HBASE-13780](#) - Default to 700 for HDFS root dir permissions for secure deployments
- [HBASE-13768](#) - ZooKeeper znodes are bootstrapped with insecure ACLs in a secure configuration
- [HBASE-13767](#) - Allow ZKAcLReset to set and not just clear ZK ACLs
- [HBASE-13086](#) - Show ZK root node on Master WebUI
- [HBASE-13342](#) - Fix incorrect interface annotations
- [HBASE-13162](#) - Add capability for cleaning hbase acls to hbase cleanup script.
- [HBASE-12641](#) - Grant all permissions of hbase zookeeper node to hbase superuser in a secure cluster
- [HBASE-13374](#) - Small scanners (with particular configurations) do not return all rows
- [HBASE-13269](#) - Limit result array preallocation to avoid OOME with large scan caching values
- [HBASE-13422](#) - remove use of StandardCharsets in 0.98
- [HBASE-13335](#) - Update ClientSmallScanner and ClientSmallReversedScanner
- [HBASE-13262](#) - ResultScanner doesn't return all rows in Scan
- [HIVE-10841](#) - [WHERE col is not null] does not work sometimes for queries with many JOIN statements
- [HIVE-9620](#) - Cannot retrieve column statistics using HMS API if column name contains uppercase characters
- [HIVE-8863](#) - Cannot drop table with uppercase name after "compute statistics for columns"
- [HIVE-6679](#) - HiveServer2 should support configurable the server side socket timeout and keepalive for various transports types where applicable
- [OOZIE-1944](#) - Recursive variable resolution broken when same parameter name in config-default and action conf
- [OOZIE-2218](#) - META-INF directories in the war file have 777 permissions
- [SENTRY-540](#) - Fix Sentry test validating special chars in username due to HIVE-8916
- [SENTRY-227](#) - Fix for "Unsupported entity type DUMMYPARTITION"
- [SOLR-7478](#) - UpdateLog#close shutdown it's executor with interrupts before running close, preventing a clean close.
- [SOLR-7338](#) - A reloaded core will never register itself as active after a ZK session expiration
- [SOLR-7370](#) - FSHDFSUtils#recoverFileLease tries to recover the lease every one second after the first four second wait.
- [ZOOKEEPER-2146](#) - BinaryInputArchive readString should check length before allocating memory
- [ZOOKEEPER-2149](#) - Logging of client address when socket connection established
- [IMPALA-1726](#): Move JNI / Thrift utilities to separate header
- [IMPALA-2002](#): Provide way to cache ext data source classes

CDH 5.2.5

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.2.5:

- [HADOOP-11350](#) - The size of header buffer of HttpServer is too small when HTTPS is enabled
- [HADOOP-11710](#) - Make CryptoOutputStream behave like DFSOutputStream wrt synchronization
- [HADOOP-11674](#) - oneByteBuf in CryptoInputStream and CryptoOutputStream should be non static
- [HDFS-6830](#) - BlockInfo.addStorage fails when DN changes the storage for a block replica
- [HDFS-7960](#) - The full block report should prune zombie storages even if they're not empty
- [HDFS-6425](#) - Large postponedMisreplicatedBlocks has impact on blockReport latency
- [HDFS-7575](#) - Upgrade should generate a unique storage ID for each volume
- [HDFS-7596](#) - NameNode should prune dead storages from storageMap
- [HDFS-7579](#) - Improve log reporting during block report rpc failure
- [HDFS-7208](#) - NN doesn't schedule replication when a DN storage fails
- [HDFS-6899](#) - Allow changing MiniDFSCluster volumes per DN and capacity per volume
- [HDFS-6878](#) - Change MiniDFSCluster to support StorageType configuration for individual directories
- [HDFS-6678](#) - MiniDFSCluster may still be partially running after initialization fails.
- [HDFS-7575](#) - Upgrade should generate a unique storage ID for each volume
- [HDFS-7960](#) - The full block report should prune zombie storages even if they're not empty
- [YARN-570](#) - Time strings are formatted in different timezone
- [YARN-2251](#) - Avoid negative elapsed time in JHS/MRAM web UI and services
- [YARN-3242](#) - Asynchrony in ZK-close can lead to ZKRMStateStore watcher receiving events for old client
- [MAPREDUCE-5957](#) - AM throws ClassNotFoundException with job classloader enabled if custom output format/committer is used
- [MAPREDUCE-6076](#) - Zero map split input length combine with none zero map split input length may cause MR1 job hung sometimes.
- [MAPREDUCE-6275](#) - Race condition in FileOutputCommitter v2 for user-specified task output subdirs
- [MAPREDUCE-4815](#) - Speed up FileOutputCommitter#commitJob for many output files
- [HIVE-2828](#) - make timestamp accessible in the hbase KeyValue
- [HIVE-2828](#) - make timestamp accessible in the hbase KeyValue
- [HIVE-7433](#) - ColumnMappins.ColumnMapping should expose public accessors for its fields
- [HIVE-6148](#) - Support arbitrary structs stored in HBase
- [HIVE-6147](#) - Support avro data stored in HBase columns
- [HIVE-6584](#) - Add HiveHBaseTableSnapshotInputFormat
- [HIVE-6411](#) - Support more generic way of using composite key for HBaseHandler
- [HIVE-6677](#) - HBaseSerDe needs to be refactored
- [HIVE-9934](#) - Vulnerability in LdapAuthenticationProviderImpl enables HiveServer2 client to degrade the authentication mechanism to "none", allowing authentication without password
- [HIVE-7737](#) - Hive logs full exception for table not found
- [HIVE-9716](#) - Map job fails when table's LOCATION does not have scheme
- [HIVE-8688](#) - serialized plan OutputStream is not being closed
- [HIVE-5857](#) - Reduce tasks do not work in uber mode in YARN
- [HUE-2446](#) - Migrating from CDH 4.7 to CDH 5.0.1+/Hue 3.5+ will fail
- [HUE-2371](#) - [sentry] Sentry URI should be created only with a ALL permission
- [HUE-1663](#) - [core] Option to either follow or not LDAP referrals for auth
- [SENTRY-654](#) - Calls to append_partition fail when Sentry is enabled
- [SOLR-7092](#) - Stop the HDFS lease recovery retries on HdfsTransactionLog on close and try to avoid lease recovery on closed files.
- [SOLR-7134](#) - Replication can still cause index corruption.
- [SOLR-7113](#) - Multiple calls to UpdateLog#init is not thread safe with respect to the HDFS FileSystem client object usage.

- [SOLR-7141](#) - RecoveryStrategy: Raise time that we wait for any updates from the leader before they saw the recovery state to have finished.
- [SQOOP-1764](#) - Numeric Overflow when getting extent map
- [IMPALA-1658](#): Add compatibility flag for Hive-Parquet-Timestamps
- [IMPALA-1794](#): Fix infinite loop opening/closing file w/ invalid metadata
- [IMPALA-1801](#): external-data-source-executor leaking global jni refs

CDH 5.2.4

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.2.4:

- [HDFS-7707](#) - Edit log corruption due to delayed block removal again
- [YARN-2846](#) - Incorrect persist exit code for running containers in reacquireContainer() that interrupted by NodeManager restart.
- [HIVE-7733](#) - Ambiguous column reference error on query
- [HIVE-8444](#) - update pom to junit 4.11
- [HIVE-9474](#) - truncate table changes permissions on the target
- [HIVE-6308](#) - COLUMNS_V2 Metastore table not populated for tables created without an explicit column list.
- [HIVE-9445](#) - Revert HIVE-5700 - enforce single date format for partition column storage
- [HIVE-7800](#) - Parquet Column Index Access Schema Size Checking Checking
- [HIVE-9393](#) - reduce noisy log level of ColumnarSerDe.java:116 from INFO to DEBUG
- [HUE-2501](#) - [metastore] Creating a table with header files bigger than 64MB truncates it
- [SOLR-7033](#) - [RecoveryStrategy should not publish any state when closed / cancelled.
- [SOLR-5961](#) - Solr gets crazy on /overseer/queue state change
- [SOLR-6640](#) - Replication can cause index corruption
- [SOLR-6920](#) - During replication use checksums to verify if files are the same
- [SOLR-5875](#) - QueryComponent.mergeIds() unmarshals all docs' sort field values once per doc instead of once per shard
- [SOLR-6919](#) - Log REST info before executing
- [SOLR-6969](#) - When opening an HDFSTransactionLog for append we must first attempt to recover its lease to prevent data loss
- [IMPALA-1471](#): Bug in spilling of PHJ that was affecting left anti and outer joins.
- [IMPALA-1451](#): Empty Row in HBase triggers NPE in Planner
- [IMPALA-1535](#): Partition pruning with NULL
- [IMPALA-1483](#): Substitute TupleIsNotNullPredicates to refer to physical analytic output.
- [IMPALA-1674](#): Fix serious memory leak in TSaslTransport
- [IMPALA-1668](#): Fix leak of transport objects in TSaslServerTransport::Factory
- [IMPALA-1565](#): Python sasl client transport perf issue
- [IMPALA-1556](#): Kerberos fetches 3x slower
- [IMPALA-1120](#): Fetch column statistics using Hive 0.13 bulk API

Published Known Issues Fixed

As a result of the above fixes, the following issues, previously published as [Known Issues in CDH 5](#) on page 111, are also fixed:

Hive does not support Parquet schema evolution

Adding a new column to a Parquet table causes queries on that table to fail with a `column not found` error.

Bug: [HIVE-7800](#)

Workaround: Use Impala instead; Impala handles Parquet schema evolution correctly.

CDH 5 Release Notes

CDH 5.2.3

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.2.3:

- [AVRO-1623](#) - GenericData#validate() of enum: IndexOutOfBoundsException
- [AVRO-1622](#) - Add missing license headers
- [AVRO-1604](#) - ReflectData.AllowNull fails to generate schemas when @Nullable is present.
- [AVRO-1407](#) - NettyTransceiver can cause a infinite loop when slow to connect
- [AVRO-834](#) - Data File corruption recovery tool
- [AVRO-1596](#) - Cannot read past corrupted block in Avro data file
- [CRUNCH-480](#) - AvroParquetFileSource does not properly configure user-supplied read schema
- [CRUNCH-479](#) - Writing to target with WriteMode.APPEND merges values into PCollection
- [CRUNCH-477](#) - Fix HFileTargetIT failures on hadoop1 under Java 1.7/1.8
- [CRUNCH-473](#) - Use specific class type for case class serialization
- [CRUNCH-473](#) - Use specific class type for case class serialization
- [CRUNCH-472](#) - Add Scrunch serialization support for Java Enums
- [HADOOP-11068](#) - Match hadoop.auth cookie format to jetty output
- [HADOOP-11343](#) - Overflow is not properly handled in calculating final iv for AES CTR
- [HADOOP-11301](#) - [optionally] update jmx cache to drop old metrics
- [HADOOP-11085](#) - Excessive logging by org.apache.hadoop.util.Progress when value is NaN
- [HADOOP-11247](#) - Fix a couple javac warnings in NFS
- [HADOOP-11195](#) - Move Id-Name mapping in NFS to the hadoop-common area for better maintenance
- [HADOOP-11130](#) - NFS updateMaps OS check is reversed
- [HADOOP-10990](#) - Add missed NFSv3 request and response classes
- [HADOOP-11323](#) - WritableComparator#compare keeps reference to byte array
- [HDFS-7560](#) - ACLs removed by removeDefaultAcl() will be back after NameNode restart/failover
- [HDFS-7367](#) - HDFS short-circuit read cannot negotiate shared memory slot and file descriptors when SASL is enabled on DataTransferProtocol.
- [HDFS-7489](#) - Incorrect locking in FsVolumeList#checkDirs can hang datanodes
- [HDFS-7158](#) - Reduce the memory usage of WebImageViewer
- [HDFS-7497](#) - Inconsistent report of decommissioning DataNodes between dfsadmin and NameNode webui
- [HDFS-7146](#) - NFS ID/Group lookup requires SSSD enumeration on the server
- [HDFS-7387](#) - NFS may only do partial commit due to a race between COMMIT and write
- [HDFS-7356](#) - Use DirectoryListing.hasMore() directly in nfs
- [HDFS-7180](#) - NFSv3 gateway frequently gets stuck due to GC
- [HDFS-7259](#) - Unresponsive NFS mount point due to deferred COMMIT response
- [HDFS-6894](#) - Add XDR parser method for each NFS response
- [HDFS-6850](#) - Move NFS out of order write unit tests into TestWrites class
- [HDFS-7385](#) - ThreadLocal used in FSEditLog class causes FSImage permission mess up
- [HDFS-7409](#) - Allow dead nodes to finish decommissioning if all files are fully replicated
- [HDFS-7373](#) - Clean up temporary files after fsimage transfer failures
- [HDFS-7225](#) - Remove stale block invalidation work when DN re-registers with different UUID
- [YARN-2721](#) - Race condition: ZKRMStateStore retry logic may throw NodeExist exception
- [YARN-2975](#) - FSLeafQueue app lists are accessed without required locks
- [YARN-2992](#) - ZKRMStateStore crashes due to session expiry
- [YARN-2910](#) - FSLeafQueue can throw ConcurrentModificationException
- [YARN-2816](#) - NM fail to start with NPE during container recovery
- [MAPREDUCE-6198](#) - NPE from JobTracker#resolveAndAddToTopology in MR1 cause initJob and heartbeat failure.
- [MAPREDUCE-6169](#) - MergeQueue should release reference to the current item from key and value at the end of the iteration to save memory.
- [HBASE-11794](#) - StripeStoreFlusher causes NullPointerException

- [HBASE-12077](#) - FilterLists create many ArrayList\$Iter objects per row.
- [HBASE-12386](#) - Replication gets stuck following a transient zookeeper error to remote peer cluster
- [HBASE-11979](#) - Compaction progress reporting is wrong
- [HBASE-12529](#) - Use ThreadLocalRandom for RandomQueueBalancer
- [HBASE-12445](#) - hbase is removing all remaining cells immediately after the cell marked with marker = KeyValue.Type.DeleteColumn via PUT
- [HBASE-12460](#) - Moving Chore to hbase-common module.
- [HBASE-12366](#) - Add login code to HBase Canary tool.
- [HBASE-12447](#) - Add support for setTimeRange for RowCounter and CellCounter
- [HIVE-9330](#) - DummyTxnManager will throw NPE if WriteEntity writeType has not been set
- [HIVE-9199](#) - Excessive exclusive lock used in some DDLs with DummyTxnManager
- [HIVE-6835](#) - Reading of partitioned Avro data fails if partition schema does not match table schema
- [HIVE-6978](#) - beeline always exits with 0 status, should exit with non-zero status on error
- [HIVE-8891](#) - Another possible cause to NucleusObjectNotFoundException from drops/rollback
- [HIVE-8874](#) - Error Accessing HBase from Hive via Oozie on Kerberos 5.0.1 cluster
- [HIVE-8916](#) - Handle user@domain username under LDAP authentication
- [HIVE-8889](#) - JDBC Driver ResultSet.getXXXXXX(String columnLabel) methods Broken
- [HIVE-9445](#) - Revert HIVE-5700 - enforce single date format for partition column storage
- [HIVE-5454](#) - HCatalog runs a partition listing with an empty filter
- [HIVE-8784](#) - Querying partition does not work with JDO enabled against PostgreSQL
- [HUE-2484](#) - [beeswax] Configure support for Hive Server2 LDAP authentication
- [HUE-2102](#) - [oozie] Workflow with credentials can't be used with Coordinator
- [HUE-2152](#) - [pig] Credentials support in editor
- [HUE-2472](#) - [impala] Stabilize result retrieval
- [HUE-2406](#) - [search] New dashboard page has a margin problem
- [HUE-2373](#) - [search] Heatmap can break
- [HUE-2395](#) - [search] Broken widget in Solr Apache logs example
- [HUE-2414](#) - [search] Timeline chart breaks when there's no extraSeries defined
- [HUE-2342](#) - [impala] TLS/SSL encryption
- [HUE-2426](#) - [pig] Dashboard gives a 500 error
- [HUE-2430](#) - [pig] Progress bars of running scripts not updated on Dashboard
- [HUE-2411](#) - [useradmin] Lazy load user and group list in permission sharing popup
- [HUE-2398](#) - [fb] Drag and Drop hover message should not appear when elements originating in DOM are dragged
- [HUE-2401](#) - [search] Visually report selected and excluded values for ranges too
- [HUE-2389](#) - [impala] Expand results table after the results are added to datatables
- [HUE-2360](#) - [sentry] Sometimes Groups are not loaded we see the input box instead
- [IMPALA-1453](#) - Fix many bugs with HS2 FETCH_FIRST
- [IMPALA-1623](#) - unix_timestamp() does not return correct time
- [IMPALA-1606](#) - Impala does not always give short name to Llama
- [IMPALA-1475](#) - accept unmangled native UDF symbols
- [OOZIE-2102](#) - Streaming actions are broken cause of incorrect method signature
- [PARQUET-145](#) - InternalParquetRecordReader.close() should not throw an exception if initialization has failed
- [PARQUET-140](#) - Allow clients to control the GenericData object that is used to read Avro records
- [PIG-4330](#) - Regression test for PIG-3584 - AvroStorage does not correctly translate arrays of strings
- [PIG-3584](#) - AvroStorage does not correctly translate arrays of strings
- [SOLR-5515](#) - NPE when getting stats on date field with empty result on solrcloud

Published Known Issues Fixed

As a result of the above fixes, the following issues, previously published as [Known Issues in CDH 5](#) on page 111, are also fixed.

CDH 5 Release Notes

Upgrading a PostgreSQL Hive Metastore from Hive 0.12 to Hive 0.13 may result in a corrupt metastore

[HIVE-5700](#) introduced a serious bug into the Hive Metastore upgrade scripts. This bug affects users who have a PostgreSQL Hive Metastore and have *at least one table* which is partitioned by date and the value is stored as a date type (not string).

Bug: [HIVE-5700](#)

Workaround: None. Do not upgrade your PostgreSQL metastore to version 0.13 if you satisfy the condition stated above.

DataNodes may become unresponsive to block creation requests

DataNodes may become unresponsive to block creation requests from clients when the directory scanner is running.

Bug: [HDFS-7489](#)

Workaround: Disable the directory scanner by setting `dfs.datanode.directoryscan.interval` to -1.

CDH 5.2.2

There is no CDH 5.2.2 release.

CDH 5.2.1

Apache Hadoop

Files inside encryption zones cannot be read in Hue

Hue uses either WebHDFS or HttpFS to access files. Both are proxy user clients of KMS and the KMS client library does not currently handle proxy users correctly.

Bug: [HADOOP-11176](#)

Workaround: None

Both ResourceManagers can end up in Standby mode

After a restart, if an application fails to recover, both Resource Managers could end up in Standby mode.

Bug: [YARN-2588](#), [YARN-2010](#)

Workarounds:

- Stop the RM. Format the state store using `yarn resourcemanager -format-state-store`. Applications that were running before the ResourceManager went down will not be recovered.
- You can limit the number of completed applications the RM state-store stores (`yarn.resourcemanager.state-store.max-completed-applications`) to reduce the chances of running into this problem.

Apache Oozie

Using cron-like syntax for Coordinator frequencies can result in duplicate actions

Every `throttle` number of actions will be a duplicate. For example, if the throttle is set to 5, every fifth action will be a duplicate.

Bug: [OOZIE-2063](#)

Workaround: If possible, use the older syntax to specify an equivalent frequency.

Upstream Fixes

CDH 5.2.1 also fixes the following issues:

- [HADOOP-11243](#) - SSLFactory shouldn't allow SSLv3
- [HADOOP-11217](#) - Disable SSLv3 in KMS
- [HADOOP-11156](#) - DelegateToFileSystem should implement `getFsStatus(final Path f)`.
- [HADOOP-11176](#) - KMSClientProvider authentication fails when both `currentUgi` and `loginUgi` are a proxied user
- [HDFS-7235](#) - DataNode#transferBlock should report blocks that don't exist using `reportBadBlock`

- [HDFS-7274](#) - Disable SSLv3 in HttpFS
- [HDFS-7391](#) - Reenable SSLv2Hello in HttpFS
- [HDFS-6781](#) - Separate HDFS commands from CommandsManual.apt.vm
- [HDFS-6831](#) - Inconsistency between 'hdfs dfsadmin' and 'hdfs dfsadmin -help'
- [HDFS-7278](#) - Add a command that allows sysadmins to manually trigger full block reports from a DN
- [YARN-2010](#) - Handle app-recovery failures gracefully
- [YARN-2588](#) - Standby RM does not transitionToActive if previous transitionToActive is failed with ZK exception.
- [YARN-2566](#) - DefaultContainerExecutor should pick a working directory randomly
- [YARN-2641](#) - Decommission nodes on -refreshNodes instead of next NM-RM heartbeat
- [MAPREDUCE-6147](#) - Support mapreduce.input.fileinputformat.split.maxsize
- [HBASE-12376](#) - HBaseAdmin leaks ZK connections if failure starting watchers (ConnectionLossException)
- [HBASE-12201](#) - Close the writers in the MOB sweep tool
- [HBASE-12220](#) - Add hedgedReads and hedgedReadWins metrics
- [HIVE-8693](#) - Separate out fair scheduler dependency from hadoop 0.23 shim
- [HIVE-8634](#) - HiveServer2 fair scheduler queue mapping does not handle the secondary groups rules correctly
- [HIVE-8675](#) - Increase thrift server protocol test coverage
- [HIVE-8827](#) - Remove SSLv2Hello from list of disabled protocols protocols
- [HIVE-8615](#) - beeline csv,tsv outputformat needs backward compatibility mode
- [HIVE-8627](#) - Compute stats on a table from impala caused the table to be corrupted
- [HIVE-7764](#) - Support all JDBC-HiveServer2 authentication modes on a secure cluster cluster
- [HIVE-8182](#) - beeline fails when executing multiple-line queries with trailing spaces
- [HUE-2438](#) - [core] Disable SSLv3 for Poodle vulnerability
- [IMPALA-1361](#): FE Exceptions with BETWEEN predicates
- [IMPALA-1397](#): free local expr allocations in scanner threads
- [IMPALA-1400](#): Window function insert issue (LAG() + OVER)
- [IMPALA-1401](#): raise MAX_PAGE_HEADER_SIZE and use scanner context to stitch together header buffer
- [IMPALA-1410](#): accept "single character" character classes in regex functions
- [IMPALA-1411](#): Create table as select produces incorrect results
- [IMPALA-1416](#) - Queries fail with metastore exception after upgrade and compute stats
- [OOZIE-2034](#) - Disable SSLv3 (POODLEbleed vulnerability)
- [OOZIE-2063](#) - Cron syntax creates duplicate actions
- [PARQUET-107](#) - Add option to disable summary metadata aggregation after MR jobs
- [SPARK-3788](#) - Yarn dist cache code is not friendly to HDFS HA, Federation
- [SPARK-3661](#) - spark.*.memory is ignored in cluster mode
- [SPARK-3979](#) - Yarn backend's default file replication should match HDFS' default one
- [SPARK-1720](#) - use LD_LIBRARY_PATH instead of -Djava.library.path

CDH 5.2.0

Apache HBase

`hbase.zookeeper.useMulti set to false by default`

The default value of `hbase.zookeeper.useMulti` was changed from `true` to `false` in CDH 5. In CDH 5.2, the default is changed back to `true`. This affects environments with HBase replication enabled and large replication queues.

Bug: None

Workaround: Enable `hbase.zookeeper.useMulti` by setting the value to `true` in `hbase-site.xml`.

Apache Hadoop

Hadoop shell commands which reference the root directory ("") do not work

Bug: [HDFS-5888](#)

Workaround: None

CDH 5 Release Notes

ResourceManager High Availability with manual failover does not work on secure clusters

Bug: [YARN-1640](#)

Workaround: Enable automatic failover; this requires ZooKeeper.

Apache Hive

*Select * fails on Parquet tables with the map data type*

Bug: [HIVE-6575](#)

Severity: Low

Workaround: Use the map's column name in the SELECT statement.

Cloudera Search

Malicious users could update information by circumventing Sentry checks

A sophisticated malicious user could update restricted content by setting the `update.distrib` parameter to bypass Sentry's index-level checks.

With Search for CDH 5.2 and later, Sentry always checks for index-level access control settings, preventing unauthorized updates.

Bug: None.

Severity: Medium.

Workaround: None.

Kerberos name rules were not followed

`SOLR_AUTHENTICATION_KERBEROS_NAME_RULES`, which is specified in `/etc/default/solr` or `/opt/cloudera/parcels/CDH-*/*/etc/default/solr` would sometimes not be respected, even if those rules were also specified using the `hadoop.security.auth_to_local` property in `SOLR_HDFS_CONFIG/core-site.xml`.

With Search for CDH 5.2 and later, Kerberos name rules are followed.

Bug: None.

Workaround: None.

Issues Fixed in CDH 5.1.x

The following topics describe issues fixed in CDH 5.1.x, from newest to oldest release. You can also review [What's New in CDH 5.1.x](#) on page 43 or [Known Issues in CDH 5](#) on page 111.

Issues Fixed in CDH 5.1.7

Upstream Issues Fixed

CDH 5.1.7 includes the following issues fixed upstream:

- [HDFS-7960](#) - The full block report should prune zombie storages even if they're not empty
- [HDFS-7278](#) - Add a command that allows sysadmins to manually trigger full block reports from a DN
- [HDFS-6831](#) - Inconsistency between 'hdfs dfsadmin' and 'hdfs dfsadmin -help'
- [HDFS-7596](#) - NameNode should prune dead storages from storageMap
- [HDFS-7208](#) - NN doesn't schedule replication when a DN storage fails
- [HDFS-7575](#) - Upgrade should generate a unique storage ID for each volume
- [HDFS-6529](#) - Trace logging for RemoteBlockReader2 to identify remote datanode and file being read
- [YARN-570](#) - Time strings are formatted in different timezone
- [YARN-2251](#) - Avoid negative elapsed time in JHS/MRAM web UI and services
- [YARN-2588](#) - Standby RM does not transitionToActive if previous transitionToActive is failed with ZK exception.
- [HIVE-8634](#) - HiveServer2 fair scheduler queue mapping doesn't handle the secondary groups rules correctly
- [HIVE-8634](#) - HiveServer2 fair scheduler queue mapping doesn't handle the secondary groups rules correctly
- [HIVE-6403](#) - uncorrelated subquery is failing with auto.convert.join=true

- [HIVE-5945](#) - ql.plan.ConditionalResolverCommonJoin.resolveMapJoinTask also sums those tables which are not used in the child of this conditional task.
- [HIVE-8916](#) - Handle user@domain username under LDAP authentication
- [HIVE-8874](#) - Error Accessing HBase from Hive via Oozie on Kerberos 5.0.1 cluster
- [HIVE-9716](#) - Map job fails when table's LOCATION does not have schema
- [HIVE-8784](#) - Querying partition does not work with JDO enabled against PostgreSQL
- [HUE-2484](#) - [beeswax] Configure support for Hive Server2 LDAP authentication
- [HUE-2446](#) - Migrating from CDH 4.7 to CDH 5.0.1+/Hue 3.5+ will fail
- [PARQUET-107](#) - Add option to disable summary metadata aggregation after MR jobs
- [SOLR-6268](#) - HdfsUpdateLog has a race condition that can expose a closed HDFS FileSystem instance and should close it's FileSystem instance if either inherited close method is called.
- [SOLR-6393](#) - Improve transaction log replay speed on HDFS.
- [SOLR-6403](#) - TransactionLog replay status logging.
- [IMPALA-1801](#) - external-data-source-executor leaking global jni refs
- [IMPALA-1794](#) - Fix infinite loop opening/closing file w/ invalid metadata
- [IMPALA-1674](#) - Fix serious memory leak in TSaslTransport
- [IMPALA-1668](#) - Fix leak of transport objects in TSaslServerTransport::Factory
- [IMPALA-1556](#) - TSaslTransport.read() should return available data before next frame
- [IMPALA-1565](#) - Python sasl client transport perf issue
- [IMPALA-1556](#) - Sasl transport should be wrapped with buffered transport
- [IMPALA-1442](#) - Better fix for non-buffered SASL transports The Thrift SASL implementation relies on the

Apache Commons Collections deserialization vulnerability

Cloudera has learned of a potential security vulnerability in a third-party library called the [Apache Commons Collections](#). This library is used in products distributed and supported by Cloudera (“Cloudera Products”), including core Apache Hadoop. The Apache Commons Collections library is also in widespread use beyond the Hadoop ecosystem. At this time, no specific attack vector for this vulnerability has been identified as present in Cloudera Products.

In an abundance of caution, we are currently in the process of incorporating a version of the Apache Commons Collections library with a fix into the Cloudera Products. In most cases, this will require coordination with the projects in the Apache community. One example of this is tracked by [HADOOP-12577](#).

The Apache Commons Collections potential security vulnerability is titled “Arbitrary remote code execution with InvokerTransformer” and is tracked by [COLLECTIONS-580](#). MITRE has not issued a CVE, but related [CVE-2015-4852](#) has been filed for the vulnerability. CERT has issued [Vulnerability Note #576313](#) for this issue.

Releases affected: CDH 5.5.0, CDH 5.4.8 and lower, CDH 5.3.8 and lower, CDH 5.2.8 and lower, CDH 5.1.7 and lower, Cloudera Manager 5.5.0, Cloudera Manager 5.4.8 and lower, Cloudera Manager 5.3.8 and lower, and Cloudera Manager 5.2.8 and lower, Cloudera Manager 5.1.6 and lower, Cloudera Navigator 2.4.0, Cloudera Navigator 2.3.8 and lower.

Users affected: All

Impact: This potential vulnerability may enable an attacker to execute arbitrary code from a remote machine without requiring authentication.

Immediate action required: Upgrade to Cloudera Manager 5.5.1 and CDH 5.5.1, Cloudera Manager 5.4.9 and CDH 5.4.9, Cloudera Manager 5.3.9 and CDH 5.3.9, and Cloudera Manager 5.2.9 and CDH 5.2.9, and Cloudera Manager 5.1.7 and CDH 5.1.7.

Issues Fixed in CDH 5.1.5

This is a maintenance release that fixes the following issues:

- [HDFS-7960](#) - The full block report should prune zombie storages even if they're not empty
- [HDFS-7278](#) - Add a command that allows sysadmins to manually trigger full block reports from a DN
- [HDFS-6831](#) - Inconsistency between 'hdfs dfsadmin' and 'hdfs dfsadmin -help'
- [HDFS-7596](#) - NameNode should prune dead storages from storageMap

CDH 5 Release Notes

- [HDFS-7208](#) - NN doesn't schedule replication when a DN storage fails
- [HDFS-7575](#) - Upgrade should generate a unique storage ID for each volume
- [HDFS-6529](#) - Trace logging for RemoteBlockReader2 to identify remote datanode and file being read
- [YARN-570](#) - Time strings are formatted in different timezone
- [YARN-2251](#) - Avoid negative elapsed time in JHS/MRAM web UI and services
- [YARN-2588](#) - Standby RM does not transitionToActive if previous transitionToActive is failed with ZK exception.
- [HIVE-8634](#) - HiveServer2 fair scheduler queue mapping doesn't handle the secondary groups rules correctly
- [HIVE-8634](#) - HiveServer2 fair scheduler queue mapping doesn't handle the secondary groups rules correctly
- [HIVE-6403](#) - uncorrelated subquery is failing with auto.convert.join=true
- [HIVE-5945](#) - ql.plan.ConditionalResolverCommonJoin.resolveMapJoinTask also sums those tables which are not used in the child of this conditional task.
- [HIVE-8916](#) - Handle user@domain username under LDAP authentication
- [HIVE-8874](#) - Error Accessing HBase from Hive via Oozie on Kerberos 5.0.1 cluster
- [HIVE-9716](#) - Map job fails when table's LOCATION does not have scheme
- [HIVE-8784](#) - Querying partition does not work with JDO enabled against PostgreSQL
- [HUE-2484](#) - [beeswax] Configure support for Hive Server2 LDAP authentication
- [HUE-2446](#) - Migrating from CDH 4.7 to CDH 5.0.1+/Hue 3.5+ will fail
- [PARQUET-107](#) - Add option to disable summary metadata aggregation after MR jobs
- [SOLR-6268](#) - HdfsUpdateLog has a race condition that can expose a closed HDFS FileSystem instance and should close its FileSystem instance if either inherited close method is called.
- [SOLR-6393](#) - Improve transaction log replay speed on HDFS.
- [SOLR-6403](#) - TransactionLog replay status logging.
- [IMPALA-1801](#) - external-data-source-executor leaking global jni refs
- [IMPALA-1794](#) - Fix infinite loop opening/closing file w/ invalid metadata
- [IMPALA-1674](#) - Fix serious memory leak in TSaslTransport
- [IMPALA-1668](#) - Fix leak of transport objects in TSaslServerTransport::Factory
- [IMPALA-1556](#) - TSaslTransport.read() should return available data before next frame
- [IMPALA-1565](#) - Python sasl client transport perf issue
- [IMPALA-1556](#) - Sasl transport should be wrapped with buffered transport
- [IMPALA-1442](#) - Better fix for non-buffered SASL transports The Thrift SASL implementation relies on the

Issues Fixed in CDH 5.1.4

CDH 5.1.4 fixes the following issues, organized by component. See [What's New in CDH 5.1.x](#) on page 43 for a list of the most important upstream problems fixed in this release.

HTTPS does not work on the HTTPS configured port

If you enable HTTPS (TLS/SSL) for YARN services, these services (including ResourceManager, NodeManager, and Job History Server) will not continue to use non-secure HTTP, but HTTPS does not work on the HTTPS configured port.

Bug: [YARN-1553](#)

Workaround: None.

Upstream Issues Fixed

In addition to the above, CDH 5.1.4 includes the following issues fixed upstream:

- [DATAFU-68](#) - SampleByKey can throw NullPointerException
- [HADOOP-11243](#) - SSLFactory shouldn't allow SSLv3
- [HADOOP-11156](#) - DelegateToFileSystem should implement getFsStatus(final Path f).
- [HDFS-7391](#) - Reenable SSLv2Hello in HttpFS
- [HDFS-7235](#) - DataNode#transferBlock should report blocks that don't exist using reportBadBlock
- [HDFS-7274](#) - Disable SSLv3 in HttpFS
- [HDFS-7005](#) - DFS input streams do not timeout

- [HDFS-6376](#) - Distcp data between two HA clusters requires another configuration
- [HDFS-6621](#) - Hadoop Balancer prematurely exits iterations
- [YARN-2273](#) - NPE in ContinuousScheduling thread when we lose a node
- [YARN-2566](#) - DefaultContainerExecutor should pick a working directory randomly
- [YARN-2588](#) - Standby RM does not transitionToActive if previous transitionToActive is failed with ZK exception.
- [YARN-2641](#) - Decommission nodes on -refreshNodes instead of next NM-RM heartbeat
- [YARN-2608](#) - FairScheduler: Potential deadlocks in loading alloc files and clock access
- [HBASE-12376](#) - HBaseAdmin leaks ZK connections if failure starting watchers (ConnectionLossException)
- [HBASE-12366](#) - Add login code to HBase Canary tool
- [HBASE-12098](#) - User granted namespace table create permissions can't create a table
- [HBASE-12087](#) - [0.98] Changing the default setting of hbase.security.access.early_out to true
- [HBASE-11896](#) - LoadIncrementalHFiles fails in secure mode if the namespace is specified
- [HBASE-12054](#) - bad state after NamespaceUpgrade with reserved table names
- [HBASE-12460](#) - Moving Chore to hbase-common module
- [HIVE-5643](#) - ZooKeeperHiveLockManager.getQuorumServers incorrectly appends the custom zk port to quorum hosts
- [HIVE-8675](#) - Increase thrift server protocol test coverage
- [HIVE-8827](#) - Remove SSLv2Hello from list of disabled protocols
- [HIVE-8182](#) - beeline fails when executing multiple-line queries with trailing spaces
- [HIVE-8330](#) - HiveResultSet.findColumn() parameters are case sensitive
- [HIVE-5994](#) - ORC RLEv2 encodes wrongly for large negative BIGINTs (64 bits)
- [HIVE-7629](#) - Problem in SMB Joins between two Parquet tables
- [HIVE-6670](#) - ClassNotFound with Serde
- [HIVE-6409](#) - FileOutputCommitterContainer::commitJob() cancels delegation tokens too early.
- [HIVE-7647](#) - Beeline does not honor --headerInterval and --color when executing with \
- [HIVE-7441](#) - Custom partition scheme gets rewritten with hive scheme upon concatenate
- [HIVE-5871](#) - Use multiple-characters as field delimiter
- [HIVE-1363](#) - SHOW TABLE EXTENDED LIKE command does not strip single/double quotes
- [HIVE-5989](#) - Hive metastore authorization check is not threadsafe
- [HUE-2438](#) - [core] Disable SSLv3 for Poodle vulnerability
- [HUE-2291](#) - [oozie] Faster dashboard display
- [IMPALA-1334](#) - Impala does not map principals to lowercase, affecting Sentry authorisation
- [IMPALA-1251](#) - High-offset queries hang
- [IMPALA-1338](#) - HDFS does not return all ACLs in getAclStatus()
- [IMPALA-1279](#) - Impala does not employ ACLs when checking path permissions for LOAD and INSERT
- [OOZIE-2034](#) - Disable SSLv3 (POODLEbleed vulnerability)
- [OOZIE-2063](#) - Cron syntax creates duplicate actions
- [SENTRY-428](#) - Sentry service should periodically renew the server kerberos ticket
- [SENTRY-431](#) - Sentry db provider client should attempt to refresh kerberos ticket before connection
- [SPARK-3606](#) - Spark-on-Yarn AmlpFilter does not work with Yarn HA

Issues Fixed in CDH 5.1.3

CDH 5.1.3 fixes the following issues, organized by component. See [New Features and Changes in CDH 5](#) on page 14 for a list of the most important upstream problems fixed in this release.

Apache Hadoop

The default setting of dfs.client.block.write.replace-datanode-on-failure.policy can cause an unrecoverable error in small clusters

The default setting of `dfs.client.block.write.replace-datanode-on-failure.policy` (DEFAULT) can cause an unrecoverable error in a small cluster during HBase rolling restart.

CDH 5 Release Notes

Bug: [HDFS-4257](#)

Workaround: Set `dfs.client.block.write.replace-datanode-on-failure.policy` to `NEVER` for 1-2- or 3-node clusters, and leave it as `DEFAULT` for all other clusters. Leave `dfs.client.block.write.replace-datanode-on-failure.enable` set to `true`.

Upstream Issues Fixed

In addition to the above, CDH 5.1.3 includes the following issues fixed upstream.

- [HADOOP-11035](#) - distcp on mr1(branch-1) fails with NPE using a short relative source path.
- [HBASE-10188](#) - Hide ServerName constructor
- [HBASE-10012](#) - Hide ServerName constructor
- [HBASE-11349](#) - [Thrift] support authentication/impersonation
- [HBASE-11446](#) - Reduce the frequency of RNG calls in SecureWALCellCodec#EncryptedKvEncoder
- [HBASE-11457](#) - Increment HFile block encoding IVs accounting for cipher's internal use
- [HBASE-11474](#) - [Thrift2] support authentication/impersonation
- [HBASE-11565](#) - Stale connection could stay for a while
- [HBASE-11627](#) - RegionSplitter's rollingSplit terminated with "/ by zero", and the _balancedSplit file was not deleted properly
- [HBASE-11788](#) - hbase is not deleting the cell when a Put with a KeyValue, KeyValue.Type.Delete is submitted
- [HBASE-11828](#) - Callers of ServerName.valueOf should use equals and not ==
- [HDFS-4257](#) - The ReplaceDatanodeOnFailure policies could have a forgiving option
- [HDFS-6776](#) - Using distcp to copy data between insecure and secure cluster via webhdfs does not work
- [HDFS-6908](#) - Incorrect snapshot directory diff generated by snapshot deletion
- [HUE-2247](#) - [Impala] Support pass-through LDAP authentication
- [HUE-2295](#) - [librdbms] External oracle DB connection is broken due to a typo
- [HUE-2273](#) - [desktop] Blacklisting apps with existing document will break home page
- [HUE-2318](#) - [desktop] Documents shared with write group permissions are not editable
- [HIVE-5087](#) - Rename npath UDF to matchpath
- [HIVE-6820](#) - HiveServer(2) ignores HIVE_OPTS
- [HIVE-7635](#) - Query having same aggregate functions but different case throws IndexOutOfBoundsException
- [IMPALA-958](#) - Excessively long query plan serialization time in FE when querying huge tables
- [IMPALA-1091](#) - Improve TScanRangeLocation struct and associated code
- [OOZIE-1989](#) - NPE during a rerun with forks
- [YARN-1458](#) - FairScheduler: Zero weight can lead to livelock

Issues Fixed in CDH 5.1.2

CDH 5.1.2 fixes the following issues, organized by component. See [What's New in CDH 5.1.x](#) on page 43 for a list of the most important upstream problems fixed in this release.

Apache Hadoop

Jobs can hang on NodeManager decommission owing to a race condition when continuous scheduling is enabled.

Bug: [YARN-2273](#)

Workaround: Disable continuous scheduling by setting `yarn.scheduler.fair.continuous-scheduling-enabled` to false

Apache HBase

Sending a large amount of invalid data to the Thrift service can cause it to crash

Bug: [HBASE-11052](#)

Workaround: None. This is a longstanding problem, not a new issue in CDH 5.1.

The metric ageOfLastShippedOp never decreases

This can cause it to appear as though the cluster is in an inconsistent state even when there is no problem.

Bug: [HBASE-11143](#).

Workaround: None.

Upstream Issues Fixed

In addition to the above, CDH 5.1.2 includes the following issues fixed upstream.

- [FLUME-2438](#)
- [HBASE-11052](#)
- [HBASE-11143](#)
- [HBASE-11609](#)
- [HDFS-6114](#)
- [HDFS-6640](#)
- [HDFS-6703](#)
- [HDFS-6788](#)
- [HDFS-6825](#)
- [HUE-2211](#)
- [HUE-2223](#)
- [HUE-2232](#)
- [HIVE-5515](#)
- [HIVE-6495](#)
- [HIVE-7450](#)
- [IMPALA-1093](#)
- [IMPALA-1107](#)
- [IMPALA-1131](#)
- [IMPALA-1142](#)
- [IMPALA-1149](#)
- [MAPREDUCE-5966](#)
- [MAPREDUCE-5979](#)
- [MAPREDUCE-6012](#)
- [OOZIE-1920](#)
- [PARQUET-19](#)
- [SENTRY-363](#)
- [YARN-2273](#)
- [YARN-2274](#)
- [YARN-2313](#)
- [YARN-2352](#)
- [YARN-2359](#)

Issues Fixed in CDH 5.1.0

Apache Hadoop

HDFS

The same DataNodes may appear in the NameNode web UI in both the live and dead node lists

Bug: [HDFS-6180](#)

Workaround: None

MapReduce

YARN Fair Scheduler's Cluster Utilization Threshold check is broken

Bug: [YARN-1640](#)

Workaround: Set the `yarn.scheduler.fair.preemption.cluster-utilization-threshold` property in `yarn-site.xml` to -1.

CDH 5 Release Notes

ResourceManager High Availability with manual failover does not work on secure clusters

Bug: [YARN-2155](#)

Workaround: Enable automatic failover; this requires ZooKeeper.

Apache HBase

MapReduce over HBase Snapshot bypasses HBase-level security

The MapReduce over HBase Snapshot bypasses HBase-level security completely since the files are read from the HDFS directly. The user who is running the scan/job has to have read permissions to the data and snapshot files.

Bug: [HBASE-8369](#)

Workaround: MapReduce users must be trusted to process/view all data in HBase.

HBase snapshots now saved to the /<hbase>/.hbase-snapshot directory

HBase snapshots are now saved to the /<hbase>/.hbase-snapshot directory instead of the /.snapshot directory. This was a conflict introduced by the HDFS snapshot feature in Hadoop 2.2/CDH 5 HDFS.

Bug: [HBASE-8352](#)

Workaround: This should be handled in the upgrade process.

Hue

Oozie jobs don't support ResourceManager HA in YARN

If the ResourceManager fails, the workflow will fail.

Bug: None

Severity: Medium

Workaround: None

Apache Oozie

Oozie HA does not work properly with HCatalog integration or SLA notifications

This issue appears when you are using HCatalog as a data dependency in a coordinator; using HCatalog from an action (for example, Pig) works correctly.

Bug: [OOZIE-1492](#)

Workaround: None

Issues Fixed in CDH 5.0.x

The following topics describe issues fixed in CDH 5.0.x, from newest to oldest release. You can also review [What's New in CDH 5.0.x](#) on page 47 or [Known Issues in CDH 5](#) on page 111.

Issues Fixed in CDH 5.0.6

Upstream Issues Fixed

The following upstream issues are fixed in CDH 5.0.6:

- [HDFS-7960](#) - The full block report should prune zombie storages even if they're not empty
- [HDFS-7278](#) - Add a command that allows sysadmins to manually trigger full block reports from a DN
- [HDFS-6831](#) - Inconsistency between hdfs dfsadmin and hdfs dfsadmin -help
- [HDFS-7596](#) - NameNode should prune dead storages from storageMap
- [HDFS-7208](#) - NN does not schedule replication when a DN storage fails
- [HDFS-7575](#) - Upgrade should generate a unique storage ID for each volume
- [YARN-570](#) - Time strings are formatted in different timezone
- [YARN-2251](#) - Avoid negative elapsed time in JHS/MRAM web UI and services
- [HIVE-8874](#) - Error Accessing HBase from Hive via Oozie on Kerberos 5.0.1 cluster

- [SOLR-6268](#) - HdfsUpdateLog has a race condition that can expose a closed HDFS FileSystem instance and should close its FileSystem instance if either inherited close method is called.
- [SOLR-6393](#) - Improve transaction log replay speed on HDFS.
- [SOLR-6403](#) - TransactionLog replay status logging.

Issues Fixed in CDH 5.0.5

“POODLE” Vulnerability on TLS/SSL enabled ports

The POODLE (Padding Oracle On Downgraded Legacy Encryption) attack takes advantage of a cryptographic flaw in the obsolete SSLv3 protocol, after first forcing the use of that protocol. The only solution is to disable SSLv3 entirely. This requires changes across a wide variety of components of CDH and Cloudera Manager in all current versions. CDH 5.0.5 provides these changes for CDH 5.0.x deployments.

For more information, see the [Cloudera Security Bulletin](#).

Apache Hadoop Distributed Cache Vulnerability

The Distributed Cache Vulnerability allows a malicious cluster user to expose private files owned by the user running the YARN NodeManager process. For more information, see the [Cloudera Security Bulletin](#).

Upstream Issues Fixed

CDH 5.0.4 includes the following issues fixed upstream.

- [HADOOP-11243](#) - SSLFactory shouldn't allow SSLv3
- [HDFS-7274](#) - Disable SSLv3 in HttpFS
- [HDFS-7391](#) - Reenable SSLv2Hello in HttpFS
- [HBASE-12376](#) - HBaseAdmin leaks ZK connections if failure starting watchers (ConnectionLossException)
- [HIVE-8675](#) - Increase thrift server protocol test coverage
- [HIVE-8827](#) - Remove SSLv2Hello from list of disabled protocols
- [HUE-2438](#) - [core] Disable SSLv3 for Poodle vulnerability
- [OOZIE-2034](#) - Disable SSLv3 (POODLEbleed vulnerability)
- [OOZIE-2063](#) - Cron syntax creates duplicate actions

Issues Fixed in CDH 5.0.4

Upstream Issues Fixed

CDH 5.0.4 includes the following issues fixed upstream.

- [FLUME-2438](#)
- [HBASE-11609](#)
- [HDFS-6044](#)
- [HDFS-6529](#)
- [HDFS-6618](#)
- [HDFS-6622](#)
- [HDFS-6640](#)
- [HDFS-6647](#)
- [HDFS-6703](#)
- [HDFS-6788](#)
- [HIVE-5515](#)
- [HIVE-7459](#)
- [HUE-2166](#)
- [HUE-2249](#)
- [IMPALA-1019](#)
- [MAPREDUCE-5966](#)
- [MAPREDUCE-5979](#)
- [OOZIE-1920](#)

CDH 5 Release Notes

- [PARQUET-19](#)
- [SPARK-1930](#)
- [YARN-1550](#)
- [YARN-2061](#)
- [YARN-2132](#)

Issues Fixed in CDH 5.0.3

The following topics describe known issues fixed in CDH 5.0.3. See [What's New in CDH 5.0.x](#) on page 47 for a list of the most important upstream problems fixed in this release.

Apache Hadoop *MapReduce*

YARN Fair Scheduler's Cluster Utilization Threshold check is broken

Bug: [YARN-2155](#)

Workaround: Set the `yarn.scheduler.fair.preemption.cluster-utilization-threshold` property in `yarn-site.xml` to -1.

Apache Oozie

When Oozie is configured to use MRv1 and TLS/SSL, YARN / MRv2 libraries are erroneously included in the classpath instead

This problem causes much of the configured Oozie functionality to be unusable.

Bug: None

Workaround: Use a different configuration (non-TLS/SSL or YARN), if possible.

Upstream Issues Fixed

In addition to the above, CDH 5.0.2 includes the following issues fixed upstream.

- [FLUME-2245](#)
- [FLUME-2416](#)
- [HBASE-10871](#)
- [HDFS-5891](#)
- [HDFS-6021](#)
- [HDFS-6077](#)
- [HDFS-6340](#)
- [HDFS-6475](#)
- [HDFS-6510](#)
- [HDFS-6527](#)
- [HDFS-6563](#)
- [HUE-1928](#)
- [HUE-2184](#)
- [HUE-2085](#)
- [HUE-2192](#)
- [HUE-2193](#)
- [OOZIE-1621](#)
- [OOZIE-1890](#)
- [OOZIE-1907](#)
- [SOLR-5593](#)
- [SOLR-5915](#)
- [SOLR-6161](#)
- [YARN-1550](#)
- [YARN-2155](#)

Issues Fixed in CDH 5.0.2

The following topics describe known issues fixed in CDH 5.0.2. See [What's New in CDH 5.0.x](#) on page 47 for a list of the most important upstream problems fixed in this release.

Apache Hadoop

CDH 5 clients running releases 5.0.1 and earlier cannot use WebHDFS to connect to a CDH 4 cluster

For example, a `hadoop fs -ls webhdfs` command run from the CDH 5 client to the CDH 4 cluster produces an error such as the following:

```
Found 21 items
ls: Invalid value for webhdfs parameter "op": No enum const class
org.apache.hadoop.hdfs.web.resources.GetOpParam.Op.GETACLSTATUS
```

Bug: [HDFS-6326](#)

Workaround: None; note that this is fixed as of CDH 5.0.2.

Apache HBase

Endless Compaction Loop

If an empty HFile whose max timestamp is past its TTL (time-to-live) is selected for compaction, it is compacted into another empty HFile, which is selected for compaction, creating an endless compaction loop.

Bug: [HBASE-10371](#)

Workaround: None

Upstream Issues Fixed

In addition to the above, CDH 5.0.2 includes the following issues fixed upstream.

- [HADOOP-10556](#)
- [HADOOP-10638](#)
- [HADOOP-10639](#)
- [HADOOP-10658](#)
- [HBASE-6690](#)
- [HBASE-10312](#)
- [HBASE-10371](#)
- [HDFS-6326](#)
- [HIVE-5380](#)
- [HIVE-6913](#)
- [PIG-3677](#)
- [YARN-2073](#)

Issues Fixed in CDH 5.0.1

Apache Hadoop

HDFS

NameNode LeaseManager may crash

Bug: [HDFS-6148/HDFS-6094](#)

Workaround: Restart the NameNode.

Some group mapping providers can cause the NameNode to crash

In certain environments, some group mapping providers can cause the NameNode to segfault and crash.

Bug: [HADOOP-10442](#)

Workaround: Configure either `ShellBasedUnixGroupsMapping` in Hadoop or configure SSSD in the operating system on the NameNode.

Apache Hive

CREATE TABLE AS SELECT (CTAS) does not work with Parquet files

Since CTAS does not work with Parquet files, the following example will return null values.

```
CREATE TABLE test_data(column1 string);
LOAD DATA LOCAL INPATH './data.txt' OVERWRITE INTO TABLE test_data;

CREATE TABLE parquet_test
ROW FORMAT SERDE 'parquet.hive.serde.ParquetHiveSerDe'
STORED AS
  INPUTFORMAT 'parquet.hive.DeprecatedParquetInputFormat'
  OUTPUTFORMAT 'parquet.hive.DeprecatedParquetOutputFormat'
AS
  SELECT column1 FROM test_data;

SELECT * FROM parquet_test;
SELECT column1 FROM parquet_test;
```

Bug: [HIVE-6375](#)

Workaround: A workaround for this is to follow up a CREATE TABLE query with an INSERT OVERWRITE TABLE SELECT * as in the example below.

```
CREATE TABLE parquet_test (column1 string)
ROW FORMAT SERDE 'parquet.hive.serde.ParquetHiveSerDe'
STORED AS
  INPUTFORMAT 'parquet.hive.DeprecatedParquetInputFormat'
  OUTPUTFORMAT 'parquet.hive.DeprecatedParquetOutputFormat';
INSERT OVERWRITE TABLE parquet_test SELECT * from test_data;
```

Apache Oozie

The oozie-workflow-0.4.5 schema has been removed

Workflows using schema 0.4.5 will no longer be accepted by Oozie because this schema definition version has been removed.

Bug: [OOZIE-1768](#)

Workaround: Use schema 0.5. It's backwards compatible with 0.4.5, so updating the workflow is as simple as changing the schema version number.

Upstream Issues Fixed

In addition to the above, CDH 5.0.1 includes the following issues fixed upstream.

- [HADOOP-10442](#) - Group look-up can cause segmentation fault when a certain JNI-based mapping module is used.
- [HADOOP-10456](#) - Bug in Configuration.java exposed by Spark (ConcurrentModificationException)
- [HDFS-5064](#) - Standby checkpoints should not block concurrent readers
- [HDFS-6039](#) - Uploading a File under a Dir with default ACLs throws "Duplicated ACLFeature"
- [HDFS-6094](#) - The same block can be counted twice towards safe mode threshold
- [HDFS-6231](#) - DFSClient hangs infinitely if using hedged reads and all eligible DataNodes die
- [HIVE-6495](#) - TableDesc.getDeserializer() should use correct classloader when calling Class.forName()
- [HIVE-6575](#) - select * fails on parquet table with map data type
- [HIVE-6648](#) - Fixed permission inheritance for multi-partitioned tables
- [HIVE-6740](#) - Fixed addition of Avro JARs to classpath
- [HUE-2061](#) - Task logs are not retrieved if containers not on the same host
- [OOZIE-1794](#) - java-opts and java-opt in the Java action don't always work properly in YARN
- [SOLR-5608](#) - Frequently reproducible failures in CollectionsAPIDistributedZkTest#testDistribSearch
- [YARN-1924](#) - STATE_STORE_OP_FAILED happens when ZKRMStateStore tries to update app(attempt) before storing it

Issues Fixed in CDH 5.0.0**Apache Flume***AsyncHBaseSink does not work in CDH 5 Beta 1 and CDH 5 Beta 2***Bug:** None**Workaround:** Use the HBASE sink (`org.apache.flume.sink.hbase.HBaseSink`) to write to HBase in CDH 5 Beta releases.**Apache Hadoop****HDFS**

DataNode can consume 100 percent of one CPU

A narrow race condition can cause one of the threads in the DataNode process to get stuck in a tight loop and consume 100 percent of one CPU.

Bug: [HDFS-5922](#)**Workaround:** Restart the DataNode process.

HDFS NFS gateway does not work with Kerberos-enabled clusters

Bug: [HDFS-5898](#)**Workaround:** None.

Cannot browse filesystem via NameNode Web UI if any directory has the sticky bit set

When listing any directory which contains an entry that has the sticky bit permission set, for example `/tmp` is often set this way, nothing will appear where the list of files or directories should be.**Bug:** [HDFS-5921](#)**Workaround:** Use the Hue File Browser.

Appending to a file that has been snapshotted previously will append to the snapshotted file as well

If you append content to a file that exists in snapshot, the file in snapshot will have the same content appended to it, invalidating the original snapshot.

Bug: See also [HDFS-5343](#)**Workaround:** None**MapReduce**

In MRv2 (YARN), the JobHistory Server has no information about a job if the ApplicationMasters fails while the job is running

Bug: None**Workaround:** None.**Apache HBase***An empty rowkey is treated as the first row of a table*

An empty rowkey is allowed in HBase, but it was treated as the first row of the table, even if it was not in fact the first row. Also, multiple rows with empty rowkeys caused issues.

Bug: [HBASE-3170](#)**Workaround:** Do not use empty rowkeys.**Apache Hive***Hive queries that combine multiple splits and query large tables fail on YARN*

Hive queries that scan large tables, or perform map side joins may fail with the following exception when the query is run using YARN:

```
java.io.IOException: Max block location exceeded for split:  
InputFormatClass: org.apache.hadoop.mapred.TextInputFormat
```

```
splitsize: 21 maxsize: 10
at
org.apache.hadoop.mapreduce.split.JobSplitWriter.writeOldSplits(JobSplitWriter.java:162)
at
org.apache.hadoop.mapreduce.split.JobSplitWriter.createSplitFiles(JobSplitWriter.java:87)
at org.apache.hadoop.mapreduce.JobSubmitter.writeOldSplits(JobSubmitter.java:540)
at org.apache.hadoop.mapreduce.JobSubmitter.writeSplits(JobSubmitter.java:510)
at org.apache.hadoop.mapreduce.JobSubmitter.submitJobInternal(JobSubmitter.java:392)
at org.apache.hadoop.mapreduce.Job$10.run(Job.java:1268)
at org.apache.hadoop.mapreduce.Job$10.run(Job.java:1265)
at java.security.AccessController.doPrivileged(Native Method)
at javax.security.auth.Subject.doAs(Subject.java:415)
at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1491)
at org.apache.hadoop.mapreduce.Job.submit(Job.java:1265)
at org.apache.hadoop.mapred.JobClient$1.run(JobClient.java:562)
at org.apache.hadoop.mapred.JobClient$1.run(JobClient.java:557)
at java.security.AccessController.doPrivileged(Native Method)
at javax.security.auth.Subject.doAs(Subject.java:415)
at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1491)
at org.apache.hadoop.mapred.JobClient.submitJobInternal(JobClient.java:557)
at org.apache.hadoop.mapred.JobClient.submitJob(JobClient.java:548)
at org.apache.hadoop.hive.ql.exec.mr.ExecDriver.execute(ExecDriver.java:425)
at org.apache.hadoop.hive.ql.exec.mr.MapRedTask.execute(MapRedTask.java:136)
at org.apache.hadoop.hive.ql.exec.Task.executeTask(Task.java:151)
at org.apache.hadoop.hive.ql.exec.TaskRunner.runSequential(TaskRunner.java:65)
```

Bug: [MAPREDUCE-5186](#)

Workaround: Set `mapreduce.job.max.split.locations` to a high value such as 100.

Files in Avro tables no longer have .avro extension

As of CDH 4.3.0 Hive no longer creates files in Avro tables with the `.avro` extension by default. This does not cause any problems in Hive, but could affect downstream components such as Pig, MapReduce, or Sqoop 1 that expect files with the `.avro` extension.

Bug: None

Workaround: Manually set the extension to `.avro` before using a query that inserts data into your Avro table. Use the following `set` statement:

```
set hive.output.file.extension=".avro";
```

Apache Oozie

Oozie does not work seamlessly with ResourceManager HA

Oozie workflows are not recovered on ResourceManager failover when ResourceManager HA is enabled. Further, users cannot specify the `clusterId` for JobTracker to work against either ResourceManager.

Bug: None

Workaround: On non-secure clusters, users are required to specify either of the ResourceManagers' `host:port`. For secure clusters, users are required to specify the Active ResourceManager's `host:port`.

When using Oozie HA with security enabled, some znodes have world ACLs

Oozie High Availability with security enabled will still work, but a malicious user or program can alter znodes used by Oozie for locking, possibly causing Oozie to be unable to finish processing certain jobs.

Bug: [OOZIE-1608](#)

Workaround: None

Oozie and Sqoop 2 may need additional configuration to work with YARN

In CDH 5, MRv2 (YARN) MapReduce 2.0 is recommended over the Hadoop 0.20-based MRv1. The default configuration may not reflect this in Oozie and Sqoop 2 in CDH 5 Beta 2, however, unless you are using Cloudera Manager.

Bug: None

Workaround: Check the value of CATALINA_BASE in /etc/oozie/conf/oozie-env.sh (if you are running an Oozie server) and /etc/default/sqoop2-server (if you are using a Sqoop 2 server). You should also ensure that CATALINA_BASE is correctly set in your environment if you are invoking /usr/bin/sqoop2-server directly instead of using the service init scripts. For Oozie, CATALINA_BASE should be set to /usr/lib/oozie/oozie-server for YARN, or /usr/lib/oozie/oozie-server-0.20 for MRv1. For Sqoop 2, CATALINA_BASE should be set to /usr/lib/sqoop2/sqoop-server for YARN, or /usr/lib/sqoop2/sqoop-server-0.20 on MRv1.

Cloudera Search

Creating cores using the web UI with default values causes the system to become unresponsive

You can use the Solr Server web UI to create new cores. If you click **Create Core** without making any changes to the default attributes, the server may become unresponsive. Checking the log for the server shows a repeated error that begins:

```
ERROR org.apache.solr.cloud.Overseer: Exception in Overseer main queue loop
java.lang.IllegalArgumentException: Path must not end with / character
```

Bug: Solr-5813

Workaround: To avoid this issue, do not create cores without first updating values for the new core in the web UI. For example, you might enter a new name for the core to be created.

If you created a core with default settings and are seeing this error, you can address the problem by finding which node is having problems and removing that node. Find the problematic node by using a tool that can inspect ZooKeeper, such as the Solr Admin UI. Using such a tool, examine items in the ZooKeeper queue, reviewing the properties for the item. The problematic node will have an item in its queue with the property collection="".

Remove the node with the item with the collection="" property using a ZooKeeper management tool. For example, you can remove nodes using the ZooKeeper command line tool or recent versions of HUE.

Issues Fixed in CDH 5 Beta Releases

The following topics describe issues fixed in CDH 5 Beta 2, after being discovered in CDH 5 Beta 1. You can also review [What's New In CDH 5 Beta Releases](#) on page 50 or [Known Issues in CDH 5](#) on page 111.

Issues Fixed in CDH 5 Beta 2

Apache Hadoop

MapReduce

ResourceManager High Availability does not work on secure clusters

If JobTrackers in an High Availability configuration are shut down, migrated to new hosts, then restarted, no JobTracker becomes active. The logs show a Mismatched address exception.

Bug: None

Workaround: None.

Default port conflicts

By default, the Shuffle Handler (which runs inside the YARN NodeManager), the REST server, and many third-party applications, all use port 8080. This will result in conflicts if you deploy more than one of them without reconfiguring the default port.

Bug: None

Workaround: Make sure at most one service uses port 8080. To reconfigure the REST server, follow [these instructions](#). To change the default port for the Shuffle Handler, set the value of mapreduce.shuffle.port in mapred-site.xml to an unused port.

JobTracker memory leak

The JobTracker has a memory leak caused by subtleties in the way UserGroupInformation interacts with the file-system cache. The number of cached file system objects can grow without bound.

Bug: [MAPREDUCE-5508](#)

CDH 5 Release Notes

Workaround: Set `keep.failed.task.files` to true, which will sidestep the memory leak but require job staging directories to be cleaned out manually.

Hue

Running a Hive Beeswax metastore on the same host as the Hue server will result in Simple Authentication and Security Layer (SASL) authentication failures on a Kerberos-enabled cluster

Bug: None

Workaround: The simple workaround is to run the metastore server remotely on a different host and make sure all Hive and Hue configurations properly refer to it. A more complex workaround is to adjust network configurations to ensure that reverse DNS properly resolves the host's address to its fully qualified-domain name (FQDN) rather than localhost.

The Pig shell does not work when NameNode uses a wildcard address

The Pig shell does not work from Hue if you use a wildcard for the NameNode's RPC or HTTP bind address. For example, `dfs.namenode.http-address` must be a real, routable address and port, not `0.0.0.<port>`.

Bug: [HUE-1060](#)

Workaround: Use a real, routable address and port, not `0.0.0.0.<port>`, for the NameNode; or use the Pig application directly, rather than from Hue.

Apache Sqoop

Oozie and Sqoop 2 may need additional configuration to work with YARN

In CDH 5, MRv2 (YARN) MapReduce 2.0 is recommended over the Hadoop 0.20-based MRv1. The default configuration may not reflect this in Oozie and Sqoop 2 in CDH 5 Beta 2, however, unless you are using Cloudera Manager.

Bug: None

Workaround: Check the value of `CATALINA_BASE` in `/etc/oozie/conf/oozie-env.sh` (if you are running an Oozie server) and `/etc/default/sqoop2-server` (if you are using a Sqoop 2 server). You should also ensure that `CATALINA_BASE` is correctly set in your environment if you are invoking `/usr/bin/sqoop2-server` directly instead of using the service init scripts. For Oozie, `CATALINA_BASE` should be set to `/usr/lib/oozie/oozie-server` for YARN, or `/usr/lib/oozie/oozie-server-0.20` for MRv1. For Sqoop 2, `CATALINA_BASE` should be set to `/usr/lib/sqoop2/sqoop-server` for YARN, or `/usr/lib/sqoop2/sqoop-server-0.20` on MRv1.

Apache Sentry (incubating)

Sentry allows unauthorized access to a directory whose name includes the scratch directory name as a prefix

As an example, if the scratch directory path is `/tmp/hive`, and you create a directory `/tmp/hive-data`, Sentry allows unauthorized read/write access to `/tmp/hive-data`.

Bug: None

Workaround: For external tables or data export location, do not use a pathname that includes the scratch directory name as a prefix. For example, if the scratch directory is `/tmp/hive`, do not locate external tables or exported data in `/tmp/hive-data` or any directory whose path uses `"/tmp/hive-"` as a prefix.

Apache Oozie

Oozie Hive action against HiveServer2 fails on a secure cluster

Workaround: None

Fixed Issues in Apache Impala (incubating)

The following sections describe the major issues fixed in each Impala release.

For known issues that are currently unresolved, see [Apache Impala \(incubating\) Known Issues](#) on page 135.

Issues Fixed in Impala for CDH 5.9.0

For the full list of Impala fixed issues in Impala 2.7.0, see [this report in the Impala JIRA tracker](#).

For the full list of fixed issues for all CDH components in CDH 5.9.0, see [Issues Fixed in CDH 5.9.x](#) on page 159.

- [IMPALA-1112](#) - Remove some unnecessary code from cross-compilation
- [IMPALA-1240](#) - add back spilling sort now that sorter is not flaky
- [IMPALA-1440](#) - test for insert mem limit
- [IMPALA-3018](#) - Address various small memory allocation related bugs
- [IMPALA-1619](#) - Support 64-bit allocations
- [IMPALA-1633](#) - GetOperationStatus should set errorMessage and sqlState
- [IMPALA-1671](#) - Print time and link to coordinator web UI once query is submitted in shell
- [IMPALA-1683](#) - Allow REFRESH on a single partition
- [IMPALA-2347](#) - Reuse metastore client connections in Catalog
- [IMPALA-2459](#) - Implement next_day date/time UDF
- [IMPALA-2700](#) - ASCII NUL characters are doubled on insert into text tables
- [IMPALA-2767](#) - Web UI call to force expire sessions
- [IMPALA-2878](#) - Fix Base64Decode error and remove duplicate codes
- [IMPALA-2885](#) - ScannerContext::Stream objects should be owned by ScannerContext
- [IMPALA-2979](#) - Fix scheduling on remote hosts
- [IMPALA-3018](#) - Don't return NULL on zero length allocations
- [IMPALA-3063](#) - Separate join inversion from join ordering
- [IMPALA-3084](#) - Cache the sequence of table ref and materialized tuple ids during analysis
- [IMPALA-3181](#) - Add noexcept to some functions
- [IMPALA-3201](#) - buffer pool header only
- [IMPALA-3206](#) - Enable codegen for AVRO_DECIMAL
- [IMPALA-3210](#) - last/first_value() support for IGNORE NULLS
- [IMPALA-3223](#) - Remove boost multiprecision in thirdparty
- [IMPALA-3225](#) - Add script to push from gerrit to ASF
- [IMPALA-3227](#) - generate test TPC data sets during data load
- [IMPALA-3253](#) - Modify gen_build_version.sh to always output the right version
- [IMPALA-3336](#) - qgen: do not randomly generate query options
- [IMPALA-3376](#) - Extra definition level when writing Parquet files
- [IMPALA-3418](#) - The Impala FE project relies on Z-tools snapshot builds
- [IMPALA-3442](#) - Replace '> >' with '>>' in template decls
- [IMPALA-3449](#) - Kudu deploy.py should find clusters by displayName
- [IMPALA-3454](#) - Kudu deletes may fail if subqueries are used
- [IMPALA-3470](#) - DecompressorTest is flaky
- [IMPALA-3491](#) - Merge test_hbase_metadata.py into compute_stats.py. Use unique db fixture
- [IMPALA-3501](#) - ee tests: detect build type and support different timeouts based on the same
- [IMPALA-3507](#) - update binutils version to fix slow linking
- [IMPALA-3521](#) - Impalad should communicate with the statestore after binding to the hs2 and besswax ports
- [IMPALA-3530](#) - Clean up test_ddl.py. Part 1
- [IMPALA-3567](#) - Part 1: groundwork to make Join build sides DataSinks
- [IMPALA-3575](#) - Add retry to backend connection request and rpc timeout
- [IMPALA-3587](#) - Get rid of not_default_fs skip marker
- [IMPALA-3600](#) - Add missing admission control tests
- [IMPALA-3606](#) - Fix Java NPE when trying to add an existing partition
- [IMPALA-3611](#) - track unused Disk IO buffer memory
- [IMPALA-3627](#) - Clean up RPC structures in ImpalaInternalService
- [IMPALA-3632](#) - Add script for running cppclean over the BE code
- [IMPALA-3647](#) - track runtime filter memory in separate tracker
- [IMPALA-3656](#) - Hitting DCHECK/CHECK does not write minidumps
- [IMPALA-3664](#) - S3A test_keys_do_not_work fails
- [IMPALA-3674](#) - Lazy materialization of LLVM module bitcode

CDH 5 Release Notes

- [IMPALA-3677](#) - Write minidump on SIGUSR1
- [IMPALA-3682](#) - Don't retry unrecoverable socket creation errors
- [IMPALA-3687](#) - Prefer Avro field name during schema reconciliation
- [IMPALA-3715](#) - Include total usage of JVM memory
- [IMPALA-3715](#) - Include more info by default in Impala debug memz webpage
- [IMPALA-3716](#) - Add Memory Tab in query's Details page
- [IMPALA-3727](#) - Change microbenchmarks to use percentile-based reporting
- [IMPALA-3729](#) - batch_size=1 coverage for avro scanner
- [IMPALA-3734](#) - C++11 - Replace boost::shared_ptr with std:: equivalent
- [IMPALA-3736](#) - Move Impala HTTP handlers to a separate class
- [IMPALA-3737](#) - Local filesystem build failed loading custom schemas
- [IMPALA-3751](#) - fix clang build errors and warnings
- [IMPALA-3753](#) - Disable create table test for old aggs and joins
- [IMPALA-3756](#) - Fix wrong argument type in HiveStringsTest
- [IMPALA-3757](#) - Add missing lock in RuntimeProfile::ComputeTimelineProfile
- [IMPALA-3762](#) - Download Python requirements before they are needed
- [IMPALA-3763](#) - download_requirements fixes
- [IMPALA-3764](#) - fuzz test HDFS scanners and fix parquet bugs found
- [IMPALA-3767](#) - bootstrap_virtualenv fails to find cython distribution
- [IMPALA-3774](#) - fix download_requirements for older Python versions
- [IMPALA-3778](#) - Fix ASF packaging build
- [IMPALA-3779](#) - Disable cache pool reader thread when HDFS isn't running
- [IMPALA-3780](#) - avoid many small reads past end of block
- [IMPALA-3786](#) - Remove "Cloudera" from impalad webpage title
- [IMPALA-3790](#) - AC tests timeout in codecoverage builds
- [IMPALA-3799](#) - Make MAX_SCAN_RANGE_LENGTH accept formatted quantities
- [IMPALA-3806](#) - remove a few modern shell idioms to improve RHEL5 support
- [IMPALA-3817](#) - Ensure filter hash function is the same on all hardware
- [IMPALA-3839](#) - Fix race condition in impala_cluster.py
- [IMPALA-3843](#) - Update warning for non-SSE3 CPUs
- [IMPALA-3845](#) - Split up hdfs-parquet-scanner.cc into more files/components
- [IMPALA-3852](#) - Remove Derby and Shiro FE dependencies
- [IMPALA-3854](#) - Fix use-after-free in HdfsTextScanner::Close()
- [IMPALA-3856](#) - Fix BinaryPredicate normalization for Kudu
- [IMPALA-3857](#) - KuduScanNode race on returning "optional" threads
- [IMPALA-3864](#) - qgen: reduce likelihood of create_query() exceptions
- [IMPALA-3866](#) - consistent user-facing terminology for scratch dirs
- [IMPALA-3881](#) - Add DataTables 1.10.12 to www/
- [IMPALA-3886](#) - Improve log of pip_download.py
- [IMPALA-3892](#) - qgen: always run Impala with -convert_legacy_hive_parquet_utc_timestamps=true
- [IMPALA-3905](#) - Add HdfsScanner::GetNext() interface and implementation for Parquet
- [IMPALA-3906](#) - Materialize implicitly referenced IR functions
- [IMPALA-3914](#) - SKIP_TOOLCHAIN_BOOTSTRAP skips Python package downloads
- [IMPALA-3918](#) - Remove Cloudera copyrights and add ASF license header
- [IMPALA-3923](#) - fix overflow in BufferedTupleStream::GetRows()
- [IMPALA-3924](#) - Ubuntu16 support
- [IMPALA-3936](#) - BufferedBlockMgr fixes for Pin() while write in flight
- [IMPALA-3939](#) - Data loading may fail on tpch kudu
- [IMPALA-3943](#) - Do not throw scan errors for empty Parquet files
- [IMPALA-3946](#) - fix MemPool integrity issues with empty chunks

- [IMPALA-3952](#) - Clear scratch batch mem pool if Open() failed
- [IMPALA-3953](#) - Fixes for KuduScanNode BE test failure
- [IMPALA-3954](#) - Add unique_database to scanner test
- [IMPALA-3957](#) - Test failure in S3 build: TestLoadData.test_load
- [IMPALA-3964](#) - Fix crash when a count(*) is performed on a nested collection
- [IMPALA-3969](#) - stress test: add option to set common query options
- [IMPALA-3972](#) - Improve display of /varz page
- [IMPALA-3992](#) - bad shell error message when running nonexistent file

Issues Fixed in Impala for CDH 5.8.3

For the full list of fixed issues for all CDH components, see [Upstream Issues Fixed](#) on page 184.

- [IMPALA-1619](#) - Support 64-bit allocations
- [IMPALA-3687](#) - Prefer Avro field name during schema reconciliation
- [IMPALA-3751](#) - Fix clang build errors and warnings
- [IMPALA-4135](#) - Thrift threaded server times-out connections during high load
- [IMPALA-4170](#) - Fix identifier quoting in COMPUTE INCREMENTAL STATS
- [IMPALA-4180](#) - Synchronize accesses to RuntimeState::reader_contexts_
- [IMPALA-4196](#) - Cross compile bit-byte-functions
- [IMPALA-4237](#) - Fix materialization of 4-byte decimals in data source scan node

Issues Fixed in Impala for CDH 5.8.2

For the full list of fixed issues for all CDH components, see [Upstream Issues Fixed](#) on page 186.

- [IMPALA-1346](#) - /1590/2344: fix sorter buffer mgmt when spilling
- [IMPALA-3159](#) - impala-shell does not accept wildcard or SAN certificates
- [IMPALA-3344](#) - Simplify sorter and document/enforce invariants.
- [IMPALA-3441](#) - , IMPALA-3659: check for malformed Avro data
- [IMPALA-3499](#) - Split catalog update.
- [IMPALA-3628](#) - Fix cancellation from shell when security is enabled
- [IMPALA-3633](#) - cancel fragment if coordinator is gone
- [IMPALA-3646](#) - Handle corrupt RLE literal or repeat counts of 0.
- [IMPALA-3670](#) - fix sorter buffer mgmt bugs
- [IMPALA-3678](#) - Fix migration of predicates into union operands with an order by + limit.
- [IMPALA-3680](#) - Cleanup the scan range state after failed hdfs cache reads
- [IMPALA-3711](#) - Remove unnecessary privilege checks in getDbsMetadata().
- [IMPALA-3732](#) - handle string length overflow in avro files
- [IMPALA-3745](#) - parquet invalid data handling
- [IMPALA-3754](#) - fix TestParquet.test_corrupt_rle_counts flakiness
- [IMPALA-3772](#) - Fix admission control stress test.
- [IMPALA-3776](#) - fix 'describe formatted' for Avro tables
- [IMPALA-3820](#) - Handle linkage errors while loading Java UDFs in Catalog
- [IMPALA-3861](#) - Replace BetweenPredicates with their equivalent CompoundPredicate.
- [IMPALA-3915](#) - Register privilege and audit requests when analyzing resolved table refs.
- [IMPALA-3930](#) - Fix shuffle insert hint with constant partition exprs.
- [IMPALA-3940](#) - Fix getting column stats through views.
- [IMPALA-3965](#) - TSSLocketWithWildcardSAN.py not exported as part of impala-shell build lib
- [IMPALA-4020](#) - Handle external conflicting changes to HMS gracefully
- [IMPALA-4049](#) - fix empty batch handling NLJ build side

CDH 5 Release Notes

Issues Fixed in Impala for CDH 5.8.0

The following list contains the most critical fixed issues (`priority='Blocker'`) from the JIRA system. For the full list of fixed issues in CDH 5.8.0 / Impala 2.6.0, see [this report in the Impala JIRA tracker](#).

RuntimeState::error_log_crashes

A crash could occur, with stack trace pointing to `impala::RuntimeState::ErrorLog`.

Bug: [IMPALA-3385](#)

Severity: High

HiveUdfCall::Open() produces unsynchronized access to JniUtil::global_refs_vector

A crash could occur because of contention between multiple calls to Java UDFs.

Bug: [IMPALA-3378](#)

Severity: High

HBaseTableWriter::CreatePutList() produces unsynchronized access to JniUtil::global_refs_vector

A crash could occur because of contention between multiple concurrent statements writing to HBase.

Bug: [IMPALA-3379](#)

Severity: High

Stress test failure: `sorter.cc:745] Check failed: i == 0 (1 vs. 0)`

A crash or wrong results could occur if the spill-to-disk mechanism encountered a zero-length string at the very end of a data block.

Bug: [IMPALA-3317](#)

Severity: High

String data coming out of agg can be corrupted by blocking operators

If a query plan contains an aggregation node producing string values anywhere within a subplan (that is, if in the SQL statement, the aggregate function appears within an inline view over a collection column), the results of the aggregation may be incorrect.

Bug: [IMPALA-3311](#)

Severity: High

CTAS with subquery throws AuthzException

A `CREATE TABLE AS SELECT` operation could fail with an authorization error, due to a slight difference in the privilege checking for the CTAS operation.

Bug: [IMPALA-3269](#)

Severity: High

Crash on inserting into table with binary and parquet

Impala incorrectly allowed `BINARY` to be specified as a column type, resulting in a crash during a write to a Parquet table with a column of that type.

Bug: [IMPALA-3237](#)

Severity: High

RowBatch::MaxTupleBufferSize() calculation incorrect, may lead to memory corruption

A crash could occur while querying tables with very large rows, for example wide tables with many columns or very large string values. This problem was identified in Impala 2.3, but had low reproducibility in subsequent releases. The fix ensures the memory allocation size is correct.

Bug: [IMPALA-3105](#)

Severity: High

[Thrift buffer overflows when serialize more than 3355443200 bytes in impala](#)

A very large memory allocation within the catalogd daemon could exceed an internal Thrift limit, causing a crash.

Bug: [IMPALA-3494](#)

Severity: High

[Altering table partition's storage format is not working and crashing the daemon](#)

If a partitioned table used a file format other than Avro, and the file format of an individual partition was changed to Avro, subsequent queries could encounter a crash.

Bug: [IMPALA-3314](#)

Severity: High

[Race condition may cause scanners to spin with runtime filters on Avro or Sequence files](#)

A timing problem during runtime filter processing could cause queries against Avro or SequenceFile tables to hang.

Bug: [IMPALA-3798](#)

Severity: High

Issues Fixed in Impala for CDH 5.7.5

For the full list of fixed issues for all CDH components, see [Upstream Issues Fixed](#) on page 196.

- [IMPALA-1619](#) - Support 64-bit allocations
- [IMPALA-1740](#) - Add support for skip.header.line.count
- [IMPALA-3458](#) - Fix table creation to test insert with header lines
- [IMPALA-3949](#) - Log the error message in FileSystemUtil.copyToLocal()
- [IMPALA-4037](#) - Fx locking during query cancellation
- [IMPALA-4076](#) - Fix runtime filter sort compare method
- [IMPALA-4099](#) - Fix the error message while loading UDFs with no JARs
- [IMPALA-4120](#) - Incorrect results with LEAD() analytic function
- [IMPALA-4135](#) - Thrift threaded server times-out connections during high load
- [IMPALA-4170](#) - Fix identifier quoting in COMPUTE INCREMENTAL STATS
- [IMPALA-4196](#) - Cross compile bit-byte functions
- [IMPALA-4237](#) - Fix materialization of 4 byte decimals in data source scan node
- [IMPALA-4246](#) - SleepForMs() utility function has undefined behavior for > 1s

Issues Fixed in Impala for CDH 5.7.4

For the full list of fixed issues for all CDH components, see [Upstream Issues Fixed](#) on page 198.

- [IMPALA-3081](#) - Increase memory limit for TestWideRow
- [IMPALA-3311](#) - Fix string data coming out of aggs in subplans
- [IMPALA-3575](#) - Add retry to back end connection request and rpc timeout
- [IMPALA-3678](#) - Fix migration of predicates into union operands with an order by + limit.
- [IMPALA-3682](#) - Do not retry unrecoverable socket creation errors
- [IMPALA-3687](#) - Fix test failure introduced by backporting
- [IMPALA-3687](#) - Prefer Avro field name during schema reconciliation
- [IMPALA-3820](#) - Handle linkage errors while loading Java UDFs in Catalog
- [IMPALA-3930](#) - Fix shuffle insert hint with constant partition exprs
- [IMPALA-3940](#) - Fix getting column stats through views
- [IMPALA-4020](#) - Handle external conflicting changes to HMS gracefully

CDH 5 Release Notes

- [IMPALA-4049](#) - Fix empty batch handling NLJ build side

Issues Fixed in Impala for CDH 5.7.2

For the full list of fixed issues for all CDH components, see [Upstream Issues Fixed](#) on page 202.

- [IMPALA-1928](#) - Fix Thrift client transport wrapping order
- [IMPALA-2660](#) - Respect auth_to_local configs from hdfs configs
- [IMPALA-3276](#) - Consistently handle pin failure in BTS::PrepareForRead()
- [IMPALA-3369](#) - Add ALTER TABLE SET COLUMN STATS statement.
- [IMPALA-3441](#) - Impala should not crash for invalid avro serialized data
- [IMPALA-3499](#) - Split catalog update
- [IMPALA-3502](#) - Fix race in the coordinator while updating filter routing table
- [IMPALA-3633](#) - Cancel fragment if coordinator is gone
- [IMPALA-3732](#) - Handle string length overflow in Avro files
- [IMPALA-3745](#) - Corrupt encoded values in parquet files can cause crashes
- [IMPALA-3751](#) - Fix clang build errors and warnings
- [IMPALA-3754](#) - Fix TestParquet.test_corrupt_rle_counts flakiness

Issues Fixed in Impala for CDH 5.7.1

For the full list of fixed issues for all CDH components, see [Upstream Issues Fixed](#) on page 205.

- [IMPALA-2076](#) - Correct execution time tracking for DataStreamSender
- [IMPALA-2502](#) - Don't redundantly repartition grouping aggregations
- [IMPALA-2892](#) - Buffered-tuple-stream-ir.cc is not cross-compiled
- [IMPALA-3133](#) - Wrong privileges after a REVOKE ALL ON SERVER statement
- [IMPALA-3139](#) - Fix drop table statement to not drop views and vice versa
- [IMPALA-3141](#) - Send dummy filters when filter production is disabled
- [IMPALA-3194](#) - Allow queries materializing scalar type columns in RC/sequence files
- [IMPALA-3220](#) - Skip logging empty ScannerContext's stream in parse error
- [IMPALA-3236](#) - Increase timeout for runtime filter tests
- [IMPALA-3238](#) - Avoid log spam for very large hash tables
- [IMPALA-3245](#), [IMPALA-3305](#): Fix crash with global filters when NUM_NODES=1
- [IMPALA-3269](#) - Remove authz checks on default table location in CTAS queries
- [IMPALA-3285](#) - Fix ASAN failure in webserver-test
- [IMPALA-3317](#) - Fix crash in sorter when spilling zero-length strings
- [IMPALA-3334](#) - Fix some bugs in query options' parsing.
- [IMPALA-3367](#) - Ensure runtime filters tests run on 3 nodes
- [IMPALA-3378](#), [IMPALA-3379](#): fix various JNI issues
- [IMPALA-3385](#) - Fix crashes on accessing error_log
- [IMPALA-3395](#) - Old HT filter code uses wrong expr type
- [IMPALA-3396](#) - Fix ConcurrentTimerCounter unit test "TimerCounterTest" failure.
- [IMPALA-3412](#) - Fix CHAR codegen crash in tuple comparator
- [IMPALA-3420](#) - Set IMPALA_THRIFT_VERSION patch level to +4

Issues Fixed in Impala for CDH 5.7.0

The following list contains the most critical issues (`priority='Blocker'`) from the JIRA system. For the full list of fixed issues in CDH 5.7.0 / Impala 2.5.0, see [this report in the Impala JIRA tracker](#).

Stress test hit assert in LLVM: external function could not be resolved

Bug: [IMPALA-2683](#)

The stress test was running a build with the TPC-H, TPC-DS, and TPC-H nested queries with scale factor 3.

Impalad is crashing if udf jar is not available in hdfs location for first time

Bug: [IMPALA-2365](#)

If a UDF JAR was not available in the HDFS location specified in the `CREATE FUNCTION` statement, the `impalad` daemon could crash.

PAGG hits `mem_limit` when switching to I/O buffers

Bug: [IMPALA-2535](#)

A join query could fail with an out-of-memory error despite the apparent presence of sufficient memory. The cause was the internal ordering of operations that could cause a later phase of the query to allocate memory required by an earlier phase of the query. The workaround was to either increase or decrease the `MEM_LIMIT` query option, because the issue would only occur for a specific combination of memory limit and data volume.

Prevent migrating incorrectly inferred identity predicates into inline views

Bug: [IMPALA-2643](#)

Referring to the same column twice in a view definition could cause the view to omit rows where that column contained a `NULL` value. This could cause incorrect results due to an inaccurate `COUNT(*)` value or rows missing from the result set.

Fix migration/assignment of `On`-clause predicates inside inline views

Bug: [IMPALA-1459](#)

Some combinations of `ON` clauses in join queries could result in comparisons being applied at the wrong stage of query processing, leading to incorrect results. Wrong predicate assignment could happen under the following conditions:

- The query includes an inline view that contains an outer join.
- That inline view is joined with another table in the enclosing query block.
- That join has an `ON` clause containing a predicate that only references columns originating from the outer-joined tables inside the inline view.

Wrong plan of `NOT IN` aggregate subquery when a constant is used in subquery predicate

Bug: [IMPALA-2093](#)

`IN` subqueries might return wrong results if the left-hand side of the `IN` is a constant. For example:

```
select * from alltypesintiny t1
  where 10 not in (select sum(int_col) from alltypesintiny);
```

Parquet DictDecoders accumulate throughout query

Bug: [IMPALA-2940](#)

Parquet dictionary decoders can accumulate throughout query execution, leading to excessive memory usage. One decoder is created per-column per-split.

Planner doesn't set the `has_local_target` field correctly

Bug: [IMPALA-3056](#)

MemPool allocation growth behavior

Bug: [IMPALA-2742](#)

Currently, the MemPool would always double the size of the last allocation. This can lead to bad behavior if the MemPool transferred the ownership of all its data except the last chunk. In the next allocation, the next allocated chunk would double the size of this large chunk, which can be undesirable.

Drop partition operations don't follow the catalog's locking protocol

Bug: [IMPALA-3035](#)

CDH 5 Release Notes

The CatalogOpExecutor.alterTableDropPartition() function violates the locking protocol used in the catalog that requires catalogLock_ to be acquired before any table-level lock. That may cause deadlocks when ALTER TABLE DROP PARTITION is executed concurrently with other DDL operations.

HAVING clause without aggregation not applied properly

Bug: [IMPALA-2215](#)

A query with a HAVING clause but no GROUP BY clause was not being rejected, despite being invalid syntax. For example:

```
select case when 1=1 then 'diddit' end as c1 from (select 1 as one) a having 1!=1;
```

Hit DCHECK Check failed: HasDateOrTime()

Bug: [IMPALA-2914](#)

TimestampValue::ToTimestampVal() requires a valid TimestampValue as input. This requirement was not enforced in some places, leading to serious errors.

Aggregation spill loop gives up too early leading to mem limit exceeded errors

Bug: [IMPALA-2986](#)

An aggregation query could fail with an out-of-memory error, despite sufficient memory being reported as available.

DataStreamSender::Channel::CloseInternal() does not close the channel on an error.

Bug: [IMPALA-2592](#)

Some queries do not close an internal communication channel on an error. This will cause the node on the other side of the channel to wait indefinitely, causing the query to hang. For example, this issue could happen on a Kerberos-enabled system if the credential cache was outdated. Although the affected query hangs, the impalad daemons continue processing other queries.

Codegen does not catch exceptions in FROM_UNIXTIME()

Bug: [IMPALA-2184](#)

Querying for the min or max value of a timestamp cast from a bigint via from_unixtime() fails silently and crashes instances of impalad when the input includes a value outside of the valid range.

Workaround: Disable native code generation with:

```
SET disable_codegen=true;
```

Impala returns wrong result for function 'conv(bigint, from_base, to_base)'

Bug: [IMPALA-2788](#)

Impala returns wrong result for function conv(). Function conv(bigint, from_base, to_base) returns an correct result, while conv(string, from_base, to_base) returns the correct value. For example:

```
select 2061013007, conv(2061013007, 16, 10), conv('2061013007', 16, 10);
+-----+-----+-----+
| 2061013007 | conv(2061013007, 16, 10) | conv('2061013007', 16, 10) |
+-----+-----+-----+
| 2061013007 | 1627467783           | 139066421255          |
+-----+-----+-----+
Fetched 1 row(s) in 0.65s

select 2061013007, conv(cast(2061013007 as bigint), 16, 10), conv('2061013007', 16, 10);
+-----+-----+-----+
| 2061013007 | conv(cast(2061013007 as bigint), 16, 10) | conv('2061013007', 16, 10) |
+-----+-----+-----+
```

```
+-----+-----+-----+
| 2061013007 | 1627467783 | 139066421255 |
+-----+-----+-----+
select 2061013007, conv(cast(2061013007 as string), 16, 10), conv('2061013007', 16, 10);
+-----+-----+-----+
| 2061013007 | conv(cast(2061013007 as string), 16, 10) | conv('2061013007', 16, 10) |
+-----+-----+-----+
| 2061013007 | 139066421255 | 139066421255 |
+-----+-----+-----+
select 2061013007, conv(cast(cast(2061013007 as decimal(20,0)) as bigint), 16, 10),
conv('2061013007', 16, 10);
+-----+-----+-----+
| 2061013007 | conv(cast(cast(2061013007 as decimal(20,0)) as bigint), 16, 10) |
conv('2061013007', 16, 10) |
+-----+-----+-----+
| 2061013007 | 1627467783 |
139066421255 |
+-----+-----+
```

Workaround: Cast the value to string and use `conv(string, from_base, to_base)` for conversion.

Issues Fixed in Impala for CDH 5.6.1

For the full list of fixed issues for all CDH components, see [Upstream Issues Fixed](#) on page 219.

- [IMPALA-852](#), [IMPALA-2215](#) - Analyze HAVING clause before aggregation
- [IMPALA-1092](#) - Fix estimates for trivial coord-only queries
- [IMPALA-1170](#) - Fix URL parsing when path contains '@'
- [IMPALA-1934](#) - Allow shell to retrieve LDAP password from shell cmd
- [IMPALA-2093](#) - Disallow NOT IN aggregate subqueries with a constant lhs expr
- [IMPALA-2184](#) - don't inline timestamp methods with try/catch blocks in IR
- [IMPALA-2425](#) - Broadcast join hint not enforced when low memory limit is set
- [IMPALA-2503](#) - Add missing String.format() arg in error message
- [IMPALA-2539](#) - Unmark collections slots of empty union operands
- [IMPALA-2554](#) - Change default buffer size for RPC servers and clients
- [IMPALA-2565](#) - Planner tests are flaky due to file size mismatches
- [IMPALA-2592](#) - DataStreamSender::Channel::CloseInternal() does not close the channel on an error
- [IMPALA-2599](#) - Pseudo-random sleep before acquiring kerberos ticket possibly not really pseudo-random
- [IMPALA-2711](#) - Fix memory leak in Rand()
- [IMPALA-2732](#) - Timestamp formats with non-padded values
- [IMPALA-2734](#) - Correlated EXISTS subqueries with HAVING clause return wrong results
- [IMPALA-2742](#) - Avoid unbounded MemPool growth with AcquireData()
- [IMPALA-2749](#) - Fix decimal multiplication overflow
- [IMPALA-2765](#) - Preserve return type of subexpressions substituted in isTrueWithNullSlots()
- [IMPALA-2788](#) - conv(bigint num, int from_base, int to_base) returns wrong result
- [IMPALA-2798](#) - Bring in AVRO-1617 fix and add test case for it
- [IMPALA-2818](#) - Fix cancellation crashes/hangs due to BlockOnWait() race
- [IMPALA-2820](#) - Support unquoted keywords as struct-field names
- [IMPALA-2832](#) - Fix cloning of FunctionCallExpr
- [IMPALA-2844](#) - Allow count(*) on RC files with complex types
- [IMPALA-2870](#) - Fix failing metadata.test_ddl.TestDdlStatements.test_create_table test
- [IMPALA-2894](#) - Move regression test into a different .test file
- [IMPALA-2906](#) - Fix an edge case with materializing TupleIsNotNullPredicates in analytic sorts
- [IMPALA-2914](#) - Fix DCHECK Check failed: HasDateOrTime()
- [IMPALA-2926](#) - Fix off-by-one bug in SelectNode::CopyRows()

CDH 5 Release Notes

- [IMPALA-2940](#) - Fix leak of dictionaries in Parquet scanner
- [IMPALA-3000](#) - Fix BitReader::Reset()
- [IMPALA-3034](#) - Verify all consumed memory of a MemTracker is always released at destruction time
- [IMPALA-3047](#) - Separate create table test with nested types
- [IMPALA-3054](#) - Disable probe side filters when spilling
- [IMPALA-3071](#) - Fix assignment of On-clause predicates belonging to an inner join
- [IMPALA-3085](#) - Unregister data sinks' MemTrackers at their Close() functions
- [IMPALA-3093](#) - ReopenClient() could NULL out 'client_key' causing a crash
- [IMPALA-3095](#) - Add configurable whitelist of authorized internal principals
- [IMPALA-3151](#) - Impala crash for avro table when casting to char data type
- [IMPALA-3194](#) - Allow queries materializing scalar type columns in RC/sequence files

Issues Fixed in Impala for CDH 5.6.0

The set of fixes for Impala in CDH 5.6.0 is the same as in CDH 5.5.2. See [Issues Fixed in Impala for CDH 5.5.2](#) on page 317 for details.

Issues Fixed in Impala for CDH 5.5.4

For the full list of fixed issues for all CDH components, see [Issues Fixed in CDH 5.5.4](#) on page 228.



Note: Impala 2.3.x is available as part of CDH 5.5.x and is not available for CDH 4. Cloudera does not intend to release future versions of Impala for CDH 4 outside patch and maintenance releases if required. Given the end-of-maintenance status for CDH 4, Cloudera recommends all customers to migrate to a recent CDH 5 release.

- [IMPALA-852](#) - ,IMPALA-2215: Analyze HAVING clause before aggregation
- [IMPALA-1092](#) - Fix estimates for trivial coord-only queries
- [IMPALA-1170](#) - Fix URL parsing when path contains '@'
- [IMPALA-1934](#) - Allow shell to retrieve LDAP password from shell cmd
- [IMPALA-2093](#) - Disallow NOT IN aggregate subqueries with a constant lhs expr
- [IMPALA-2184](#) - Don't inline timestamp methods with try/catch blocks in IR
- [IMPALA-2425](#) - Broadcast join hint not enforced when low memory limit is set
- [IMPALA-2503](#) - Add missing String.format() arg in error message
- [IMPALA-2539](#) - Unmark collections slots of empty union operands
- [IMPALA-2554](#) - Change default buffer size for RPC servers and clients
- [IMPALA-2565](#) - Planner tests are flaky due to file size mismatches
- [IMPALA-2592](#) - DataStreamSender::Channel::CloseInternal() does not close the channel on an error
- [IMPALA-2599](#) - Pseudo-random sleep before acquiring kerberos ticket possibly not really pseudo-random
- [IMPALA-2711](#) - Fix memory leak in Rand()
- [IMPALA-2719](#) - test_parquet_max_page_header fails on Isilon
- [IMPALA-2732](#) - Timestamp formats with non-padded values
- [IMPALA-2734](#) - Correlated EXISTS subqueries with HAVING clause return wrong results
- [IMPALA-2742](#) - Avoid unbounded MemPool growth with AcquireData()
- [IMPALA-2749](#) - Fix decimal multiplication overflow
- [IMPALA-2765](#) - Preserve return type of subexpressions substituted in isTrueWithNullSlots()
- [IMPALA-2788](#) - conv(bignum num, int from_base, int to_base) returns wrong result
- [IMPALA-2798](#) - Bring in AVRO-1617 fix and add test case for it
- [IMPALA-2818](#) - Fix cancellation crashes/hangs due to BlockOnWait() race
- [IMPALA-2820](#) - Support unquoted keywords as struct-field names
- [IMPALA-2832](#) - Fix cloning of FunctionCallExpr
- [IMPALA-2844](#) - Allow count(*) on RC files with complex types
- [IMPALA-2870](#) - Fix failing metadata.test_ddl.TestDdlStatements.test_create_table test

- [IMPALA-2894](#) - Move regression test into a different .test file
- [IMPALA-2906](#) - Fix an edge case with materializing TupleIsNotNullPredicates in analytic sorts
- [IMPALA-2914](#) - Fix DCHECK Check failed: HasDateOrTime()
- [IMPALA-2926](#) - Fix off-by-one bug in SelectNode::CopyRows()
- [IMPALA-2940](#) - Fix leak of dictionaries in Parquet scanner
- [IMPALA-3000](#) - Fix BitReader::Reset()
- [IMPALA-3034](#) - Verify all consumed memory of a MemTracker is always released at destruction time
- [IMPALA-3047](#) - Separate create table test with nested types
- [IMPALA-3054](#) - Disable probe side filters when spilling
- [IMPALA-3071](#) - Fix assignment of On-clause predicates belonging to an inner join
- [IMPALA-3085](#) - Unregister data sinks' MemTrackers at their Close() functions
- [IMPALA-3093](#) - ReopenClient() could NULL out 'client_key' causing a crash
- [IMPALA-3095](#) - Add configurable whitelist of authorized internal principals
- [IMPALA-3151](#) - Impala crash for avro table when casting to char data type
- [IMPALA-3194](#) - Allow queries materializing scalar type columns in RC/sequence files

Issues Fixed in Impala for CDH 5.5.2

This section lists the most serious or frequently encountered customer issues fixed in CDH 5.5.2 / Impala 2.3.2. For the full list of fixed Impala issues, see [Issues Fixed in CDH 5.5.2](#) on page 233.

[SEGV in AnalyticEvalNode touching NULL input_stream_](#)

A query involving an analytic function could encounter a serious error. This issue was encountered infrequently, depending upon specific combinations of queries and data.

Bug: [IMPALA-2829](#)

[Free local allocations per row batch in non-partitioned AGG and HJ](#)

An outer join query could fail unexpectedly with an out-of-memory error when the “spill to disk” mechanism was turned off.

Bug: [IMPALA-2722](#)

[Free local allocations once for every row batch when building hash tables](#)

A join query could encounter a serious error due to an internal failure to allocate memory, which resulted in dereferencing a NULL pointer.

Bug: [IMPALA-2612](#)

[Prevent migrating incorrectly inferred identity predicates into inline views](#)

Referring to the same column twice in a view definition could cause the view to omit rows where that column contained a NULL value. This could cause incorrect results due to an inaccurate COUNT(*) value or rows missing from the result set.

Bug: [IMPALA-2643](#)

[Fix GRANTs on URIs with uppercase letters](#)

A GRANT statement for a URI could be ineffective if the URI contained uppercase letters, for example in an uppercase directory name. Subsequent statements, such as CREATE EXTERNAL TABLE with a LOCATION clause, could fail with an authorization exception.

Bug: [IMPALA-2695](#)

[Avoid sending large partition stats objects over thrift](#)

The catalogd daemon could encounter a serious error when loading the incremental statistics metadata for tables with large numbers of partitions and columns. The problem occurred when the internal representation of metadata

CDH 5 Release Notes

for the table exceeded 2 GB, for example in a table with 20K partitions and 77 columns. The fix causes a COMPUTE INCREMENTAL STATS operation to fail if it would produce metadata that exceeded the maximum size.

Bug: [IMPALA-2664](#), [IMPALA-2648](#)

Throw AnalysisError if table properties are too large (for the Hive metastore)

CREATE TABLE or ALTER TABLE statements could fail with metastore database errors due to length limits on the SERDEPROPERTIES and TBLPROPERTIES clauses. (The limit on key size is 256, while the limit on value size is 4000.) The fix makes Impala handle these error conditions more cleanly, by detecting too-long values rather than passing them to the metastore database.

Bug: [IMPALA-2226](#)

Make MAX_PAGE_HEADER_SIZE configurable

Impala could fail to access Parquet data files with page headers larger than 8 MB, which could occur, for example, if the minimum or maximum values for a column were long strings. The fix adds a configuration setting --max_page_header_size, which you can use to increase the Impala size limit to a value higher than 8 MB.

Bug: [IMPALA-2273](#)

reduce scanner memory usage

Queries on Parquet tables could consume excessive memory (potentially multiple gigabytes) due to producing large intermediate data values while evaluating groups of rows. The workaround was to reduce the size of the NUM_SCANNER_THREADS query option, the BATCH_SIZE query option, or both.

Bug: [IMPALA-2473](#)

Handle error when distinct and aggregates are used with a having clause

A query that included a DISTINCT operator and a HAVING clause, but no aggregate functions or GROUP BY, would fail with an uninformative error message.

Bug: [IMPALA-2113](#)

Handle error when star based select item and aggregate are incorrectly used

A query that included * in the SELECT list, in addition to an aggregate function call, would fail with an uninformative message if the query had no GROUP BY clause.

Bug: [IMPALA-2225](#)

Refactor MemPool usage in HBase scan node

Queries involving HBase tables used substantially more memory than in earlier Impala versions. The problem occurred starting in Impala 2.2.8, as a result of the changes for IMPALA-2284. The fix for this issue involves removing a separate memory work area for HBase queries and reusing other memory that was already allocated.

Bug: [IMPALA-2731](#)

Fix migration/assignment of On-clause predicates inside inline views

Some combinations of ON clauses in join queries could result in comparisons being applied at the wrong stage of query processing, leading to incorrect results. Wrong predicate assignment could happen under the following conditions:

- The query includes an inline view that contains an outer join.
- That inline view is joined with another table in the enclosing query block.
- That join has an ON clause containing a predicate that only references columns originating from the outer-joined tables inside the inline view.

Bug: [IMPALA-1459](#)

DCHECK in parquet scanner after block read error

A debug build of Impala could encounter a serious error after encountering some kinds of I/O errors for Parquet files. This issue only occurred in debug builds, not release builds.

Bug: [IMPALA-2558](#)

PAGG hits mem_limit when switching to I/O buffers

A join query could fail with an out-of-memory error despite the apparent presence of sufficient memory. The cause was the internal ordering of operations that could cause a later phase of the query to allocate memory required by an earlier phase of the query. The workaround was to either increase or decrease the `MEM_LIMIT` query option, because the issue would only occur for a specific combination of memory limit and data volume.

Bug: [IMPALA-2535](#)

Fix check failed: sorter_runs_back()->is_pinned_

A query could fail with an internal error while calculating the memory limit. This was an infrequent condition uncovered during stress testing.

Bug: [IMPALA-2559](#)

Don't ignore Status returned by `DataStreamRecv::CreateMerger()`

A query could fail with an internal error while calculating the memory limit. This was an infrequent condition uncovered during stress testing.

Bug: [IMPALA-2614](#), [IMPALA-2559](#)

`DataStreamSender::Send()` does not return an error status if `SendBatch()` failed

Bug: [IMPALA-2591](#)

Re-enable SSL and Kerberos on server-server

These fixes lift the restriction on using SSL encryption and Kerberos authentication together for internal communication between Impala components.

Bug: [IMPALA-2598](#), [IMPALA-2747](#)

Issues Fixed in Impala for CDH 5.5.1

The version of Impala that is included with CDH 5.5.1 is identical to the Impala for CDH 5.5.0. There are no new bug fixes, new features, or incompatible changes.

Issues Fixed in Impala for CDH 5.5.0

This section lists the most serious or frequently encountered customer issues fixed in CDH 5.5.0 / Impala 2.3.0. Any issues already fixed in CDH 5.4 maintenance releases (up through CDH 5.4.8) are also included. Those issues are listed under the respective CDH 5.4 sections and are not repeated here. For the full list of fixed Impala issues, see [Issues Fixed in CDH 5.5.0](#) on page 239.

Fixes for Serious Errors

A number of issues were resolved that could result in serious errors when encountered. The most critical or commonly encountered are listed here.

Bugs: [IMPALA-2168](#), [IMPALA-2378](#), [IMPALA-2369](#), [IMPALA-2357](#), [IMPALA-2319](#), [IMPALA-2314](#), [IMPALA-2016](#)

Fixes for Correctness Errors

A number of issues were resolved that could result in wrong results when encountered. The most critical or commonly encountered are listed here.

Bugs: [IMPALA-2192](#), [IMPALA-2440](#), [IMPALA-2090](#), [IMPALA-2086](#), [IMPALA-1947](#), [IMPALA-1917](#)

Issues Fixed in Impala for CDH 5.4.10

For the full list of fixed issues for all CDH components, see [Issues Fixed in CDH 5.4.10](#) on page 246.



Note: The Impala 2.2.x maintenance releases now use the CDH 5.4.x numbering system rather than increasing the Impala version numbers. Impala 2.2 and higher are not available under CDH 4.

- [IMPALA-1702](#) - Check for duplicate table IDs at the end of analysis (issue not entirely fixed, but now fails gracefully)
- [IMPALA-2264](#) - Implicit casts to integers from decimals with higher precision sometimes allowed
- [IMPALA-2473](#) - Excessive memory usage by scan nodes
- [IMPALA-2621](#) - Fix flaky UNIX_TIMESTAMP() test
- [IMPALA-2643](#) - Nested inline view produces incorrect result when referencing the same column implicitly
- [IMPALA-2765](#) - AnalysisException: operands of type BOOLEAN and TIMESTAMP are not comparable when OUTER JOIN with CASE statement
- [IMPALA-2798](#) - After adding a column to avro table, Impala returns weird result if codegen is enabled.
- [IMPALA-2861](#) - Fix flaky scanner test added via IMPALA-2473 backport
- [IMPALA-2914](#) - Hit DCHECK Check failed: HasDateOrTime()
- [IMPALA-3034](#) - MemTracker leak on PHJ failure to spill
- [IMPALA-3085](#) - DataSinks' MemTrackers need to unregister themselves from parent
- [IMPALA-3093](#) - ReopenClient() could NULL out 'client_key' causing a crash
- [IMPALA-3095](#) - Allow additional Kerberos users to be authorized to access internal APIs

Issues Fixed in Impala for CDH 5.4.9

This section lists the most frequently encountered customer issues fixed in Impala for CDH 5.4.9.



Note: The Impala 2.2.x maintenance releases now use the CDH 5.4.x numbering system rather than increasing the Impala version numbers. Impala 2.2 and higher are not available under CDH 4.

For the full list of fixed issues, see [Issues Fixed in CDH 5.4.9](#) on page 249.

Query return empty result if it contains NullLiteral in inlineview

If an inline view in a FROM clause contained a NULL literal, the result set was empty.

Bug: [IMPALA-1917](#)

HBase scan node uses 2-4x memory after upgrade to Impala 2.2.8

Queries involving HBase tables used substantially more memory than in earlier Impala versions. The problem occurred starting in Impala 2.2.8, as a result of the changes for IMPALA-2284. The fix for this issue involves removing a separate memory work area for HBase queries and reusing other memory that was already allocated.

Bug: [IMPALA-2731](#)

Fix migration/assignment of On-clause predicates inside inline views

Some combinations of ON clauses in join queries could result in comparisons being applied at the wrong stage of query processing, leading to incorrect results. Wrong predicate assignment could happen under the following conditions:

- The query includes an inline view that contains an outer join.
- That inline view is joined with another table in the enclosing query block.
- That join has an ON clause containing a predicate that only references columns originating from the outer-joined tables inside the inline view.

Bug: [IMPALA-1459](#)

Fix wrong predicate assignment in outer joins

The join predicate for an OUTER JOIN clause could be applied at the wrong stage of query processing, leading to incorrect results.

Bug: [IMPALA-2446](#)

Avoid sending large partition stats objects over thrift

The `catalogd` daemon could encounter a serious error when loading the incremental statistics metadata for tables with large numbers of partitions and columns. The problem occurred when the internal representation of metadata for the table exceeded 2 GB, for example in a table with 20K partitions and 77 columns. The fix causes a `COMPUTE INCREMENTAL STATS` operation to fail if it would produce metadata that exceeded the maximum size.

Bug: [IMPALA-2648](#), [IMPALA-2664](#)

Avoid overflow when adding large intervals to TIMESTAMPs

Adding or subtracting a large `INTERVAL` value to a `TIMESTAMP` value could produce an incorrect result, with the value wrapping instead of returning an out-of-range error.

Bug: [IMPALA-1675](#)

Analysis exception when a binary operator contains an IN operator with values

An `IN` operator with literal values could cause a statement to fail if used as the argument to a binary operator, such as an equality test for a `BOOLEAN` value.

Bug: [IMPALA-1949](#)

Make MAX_PAGE_HEADER_SIZE configurable

Impala could fail to access Parquet data files with page headers larger than 8 MB, which could occur, for example, if the minimum or maximum values for a column were long strings. The fix adds a configuration setting `--max_page_header_size`, which you can use to increase the Impala size limit to a value higher than 8 MB.

Bug: [IMPALA-2273](#)

Fix spilling sorts with var-len slots that are NULL or empty.

A query that activated the spill-to-disk mechanism could fail if it contained a sort expression involving certain combinations of fixed-length or variable-length types.

Bug: [IMPALA-2357](#)

Work-around IMPALA-2344: Fail query with OOM in case block->Pin() fails

Some queries that activated the spill-to-disk mechanism could produce a serious error if there was insufficient memory to set up internal work areas. Now those queries produce normal out-of-memory errors instead.

Bug: [IMPALA-2344](#)

Crash (likely race) tearing down BufferedBlockMgr on query failure

A serious error could occur under rare circumstances, due to a race condition while freeing memory during heavily concurrent workloads.

Bug: [IMPALA-2252](#)

QueryExecState doesn't check for query cancellation or errors

A call to `setError()` in a user-defined function (UDF) would not cause the query to fail as expected.

Bug: [IMPALA-1746](#)

Impala throws IllegalStateException when inserting data into a partition while select subquery group by partition columns

An `INSERT ... SELECT` operation into a partitioned table could fail if the `SELECT` query included a `GROUP BY` clause referring to the partition key columns.

Bug: [IMPALA-2533](#)

Issues Fixed in Impala for CDH 5.4.8

This section lists the most frequently encountered customer issues fixed in Impala for CDH 5.4.8.



Note: The Impala 2.2.x maintenance releases now use the CDH 5.4.x numbering system rather than increasing the Impala version numbers. Impala 2.2 and higher are not available under CDH 4.

For the full list of fixed issues, see [Issues Fixed in CDH 5.4.8](#) on page 251.

Impala is unable to read hive tables created with the "STORED AS AVRO" clause

Impala could not read Avro tables created in Hive with the STORED AS AVRO clause.

Bug: [IMPALA-1136](#), [IMPALA-2161](#)

make Parquet scanner fail query if the file size metadata is stale

If a Parquet file in HDFS was overwritten by a smaller file, Impala could encounter a serious error. Issuing a INVALIDATE METADATA statement before a subsequent query would avoid the error. The fix allows Impala to handle such inconsistencies in Parquet file length cleanly regardless of whether the table metadata is up-to-date.

Bug: [IMPALA-2213](#)

Avoid allocating StringBuffer > 1GB in ScannerContext::Stream::GetBytesInternal()

Impala could encounter a serious error when reading compressed text files larger than 1 GB. The fix causes Impala to issue an error message instead in this case.

Bug: [IMPALA-2249](#)

Disallow long (1<<30) strings in group_concat()

A query using the group_concat() function could encounter a serious error if the returned string value was larger than 1 GB. Now the query fails with an error message in this case.

Bug: [IMPALA-2284](#)

avoid FnvHash64to32 with empty inputs

An edge case in the algorithm used to distribute data among nodes could result in uneven distribution of work for some queries, with all data sent to the same node.

Bug: [IMPALA-2270](#)

The catalog does not close the connection to HMS during table invalidation

A communication error could occur between Impala and the Hive metastore database, causing Impala operations that update table metadata to fail.

Bug: [IMPALA-2348](#)

Wrong DCHECK in PHJ::ProcessProbeBatch

Certain queries could encounter a serious error if the spill-to-disk mechanism was activated.

Bug: [IMPALA-2364](#)

Avoid cardinality 0 in scan nodes of small tables and low selectivity

Impala could generate a suboptimal query plan for some queries involving small tables.

Bug: [IMPALA-2165](#)

Issues Fixed in Impala for CDH 5.4.7

This section lists the most frequently encountered customer issues fixed in Impala for CDH 5.4.7.



Note: The Impala 2.2.x maintenance releases now use the CDH 5.4.x numbering system rather than increasing the Impala version numbers. Impala 2.2 and higher are not available under CDH 4.

For the full list of fixed issues, see [Issues Fixed in CDH 5.4.7](#) on page 254.

Warn if table stats are potentially corrupt.

Impala warns if it detects a discrepancy in table statistics: a table considered to have zero rows even though there are data files present. In this case, Impala also skips query optimizations that are normally applied to very small tables.

Bug: [IMPALA-1983](#):

Pass correct child node in 2nd phase merge aggregation.

A query could encounter a serious error if it included a particular combination of aggregate functions and inline views.

Bug: [IMPALA-2266](#)

Set the output smap of an EmptySetNode produced from an empty inline view.

A query could encounter a serious error if it included an inline view whose subquery had no `FROM` clause.

Bug: [IMPALA-2216](#)

Set an InsertStmt's result exprs from the source statement's result exprs.

A `CREATE TABLE AS SELECT` or `INSERT ... SELECT` statement could produce different results than a `SELECT` statement, for queries including a `FULL JOIN` clause and including literal values in the select list.

Bug: [IMPALA-2203](#)

Fix planning of empty union operands with analytics.

A query could return incorrect results if it contained a `UNION` clause, calls to analytic functions, and a constant expression that evaluated to `FALSE`.

Bug: [IMPALA-2088](#)

Retain eq predicates bound by grouping slots with complex grouping exprs.

A query containing an `INNER JOIN` clause could return undesired rows. Some predicate specified in the `ON` clause could be omitted from the filtering operation.

Bug: [IMPALA-2089](#)

Row count not set for empty partition when spec is used with compute incremental stats

A `COMPUTE INCREMENTAL STATS` statement could leave the row count for an empty partition as `-1`, rather than initializing the row count to `0`. The missing statistic value could result in reduced query performance.

Bug: [IMPALA-2199](#)

Explicit aliases + ordinals analysis bug

A query could encounter a serious error if it included column aliases with the same names as table columns, and used ordinal numbers in an `ORDER BY` or `GROUP BY` clause.

Bug: [IMPALA-1898](#)

Fix `TupleIsNullPredicate` to return false if no tuples are nullable.

A query could return incorrect results if it included an outer join clause, inline views, and calls to functions such as `coalesce()` that can generate `NULL` values.

Bug: [IMPALA-1987](#)

fix `Expr::ComputeResultsLayout()` logic

A query could return incorrect results if the table contained multiple `CHAR` columns with length of 2 or less, and the query included a `GROUP BY` clause that referred to multiple such columns.

Bug: [IMPALA-2178](#)

Substitute an InsertStmt's partition key exprs with the root node's smap.

An `INSERT` statement could encounter a serious error if the `SELECT` portion called an analytic function.

Bug: [IMPALA-1737](#)

Issues Fixed in Impala for CDH 5.4.5

This section lists the most frequently encountered customer issues fixed in Impala for CDH 5.4.5.



Note: The Impala 2.2.x maintenance releases now use the CDH 5.4.x numbering system rather than increasing the Impala version numbers. Impala 2.2 and higher are not available under CDH 4.

For the full list of fixed issues, see [Issues Fixed in CDH 5.4.5](#) on page 256.

Impala DML/DDL operations corrupt table metadata leading to Hive query failures

When the Impala `COMPUTE STATS` statement was run on a partitioned Parquet table that was created in Hive, the table subsequently became inaccessible in Hive. The table was still accessible to Impala. Regaining access in Hive required a workaround of creating a new table. The error displayed in Hive was:

```
Error: Error while compiling statement: FAILED: SemanticException Class not found:  
com.cloudera.impala.hive.serde.ParquetInputFormat (state=42000,code=40000)
```

Bug: [IMPALA-2048](#)

Avoiding a DCHECK of NULL hash table in spilled right joins

A query could encounter a serious error if it contained a `RIGHT OUTER`, `RIGHT ANTI`, or `FULL OUTER` join clause and approached the memory limit on a host so that the “spill to disk” mechanism was activated.

Bug: [IMPALA-1929](#)

Bug in PrintTColumnValue caused wrong stats for TINYINT partition cols

Declaring a partition key column as a `TINYINT` caused problems with the `COMPUTE STATS` statement. The associated partitions would always have zero estimated rows, leading to potential inefficient query plans.

Bug: [IMPALA-2136](#)

Where clause does not propagate to joins inside nested views

A query that referred to a view whose query referred to another view containing a join, could return incorrect results. `WHERE` clauses for the outermost query were not always applied, causing the result set to include additional rows that should have been filtered out.

Bug: [IMPALA-2018](#)

Add effective_user() builtin

The `user()` function returned the name of the logged-in user, which might not be the same as the user name being checked for authorization if, for example, delegation was enabled.

Bug: [IMPALA-2064](#)

Resolution: Rather than change the behavior of the `user()` function, the fix introduces an additional function `effective_user()` that returns the user name that is checked during authorization.

Make UTC to local TimestampValue conversion faster.

Query performance was improved substantially for Parquet files containing `TIMESTAMP` data written by Hive, when the `-convert_legacy_hive_parquet_utc_timestamps=true` setting is in effect.

Bug: [IMPALA-2125](#)

Workaround IMPALA-1619 in BufferedBlockMgr::ConsumeMemory()

A join query could encounter a serious error if the query approached the memory limit on a host so that the “spill to disk” mechanism was activated, and data volume in the join was large enough that an internal memory buffer exceeded 1 GB in size on a particular host. (Exceeding this limit would only happen for huge join queries, because Impala could

split this intermediate data into 16 parts during the join query, and the buffer only contains compact bookkeeping data rather than the actual join column data.)

Bug: [IMPALA-2065](#)

Issues Fixed in Impala for CDH 5.4.3

This section lists the most frequently encountered customer issues fixed in Impala for CDH 5.4.3.



Note: The Impala 2.2.x maintenance releases now use the CDH 5.4.x numbering system rather than increasing the Impala version numbers. Impala 2.2 and higher are not available under CDH 4.

For the full list of fixed issues, see [Issues Fixed in CDH 5.4.3](#) on page 260.

Enable using Isilon as the underlying filesystem.

The certification of CDH and Impala with the Isilon filesystem involves a number of fixes to performance and flexibility for dealing with I/O using remote reads. See [Using Impala with Isilon Storage](#) for details on using Impala and Isilon together.

Bug: [IMPALA-1968](#), [IMPALA-1730](#)

Expand set of supported timezones.

The set of timezones recognized by Impala was expanded. You can always find the latest list of supported timezones in the Impala source code, in the file [timezone_db.cc](#).

Bug: [IMPALA-1381](#)

Impala Timestamp ISO-8601 Support.

Impala can now process `TIMESTAMP` literals including a trailing `z`, signifying “Zulu” time, a synonym for UTC.

Bug: [IMPALA-1963](#)

Fix wrong warning when insert overwrite to empty table

An `INSERT OVERWRITE` operation would encounter an error if the `SELECT` portion of the statement returned zero rows, such as with a `LIMIT 0` clause.

Bug: [IMPALA-2008](#)

Expand parsing of decimals to include scientific notation

`DECIMAL` literals can now include e scientific notation. For example, now `CAST(1e3 AS DECIMAL(5,3))` is a valid expression. Formerly it returned `NULL`. Some scientific expressions might have worked before in `DECIMAL` context, but only when the scale was 0.

Bug: <https://issues.cloudera.org/browse/>

Issues Fixed in Impala for CDH 5.4.1

This section lists the most frequently encountered customer issues fixed in Impala for CDH 5.4.1.



Note: The Impala 2.2.x maintenance releases now use the CDH 5.4.x numbering system rather than increasing the Impala version numbers. Impala 2.2 and higher are not available under CDH 4.

For the full list of fixed issues, see [Issues Fixed in CDH 5.4.1](#) on page 262.

Issues Fixed in the 2.2.0 Release / CDH 5.4.0

This section lists the most frequently encountered customer issues fixed in Impala 2.2.0.

For the full list of fixed issues in Impala 2.2.0, including over 40 critical issues, see [this report in the JIRA system](#).



Note: Impala 2.2.0 is available as part of CDH 5.4.0 and is not available for CDH 4. Cloudera does not intend to release future versions of Impala for CDH 4 outside patch and maintenance releases if required. Given the end-of-maintenance status for CDH 4, Cloudera recommends all customers to migrate to a recent CDH 5 release.

Altering a column's type causes column stats to stop sticking for that column

When the type of a column was changed in either Hive or Impala through `ALTER TABLE CHANGE COLUMN`, the metastore database did not correctly propagate that change to the table that contains the column statistics. The statistics (particularly the `NDV`) for that column were permanently reset and could not be changed by Impala's `COMPUTE STATS` command. The underlying cause is a Hive bug (HIVE-9866).

Bug: [IMPALA-1607](#)

Resolution: Resolved by incorporating the fix for [HIVE-9866](#).

Workaround: On systems without the corresponding Hive fix, change the column back to its original type. The stats reappear and you can recompute or drop them.

Impala may leak or use too many file descriptors

If a file was truncated in HDFS without a corresponding `REFRESH` in Impala, Impala could allocate memory for file descriptors and not free that memory.

Bug: [IMPALA-1854](#)

Spurious stale block locality messages

Impala could issue messages stating the block locality metadata was stale, when the metadata was actually fine. The internal “remote bytes read” counter was not being reset properly. This issue did not cause an actual slowdown in query execution, but the spurious error could result in unnecessary debugging work and unnecessary use of the `INVALIDATE METADATA` statement.

Bug: [IMPALA-1712](#)

DROP TABLE fails after COMPUTE STATS and ALTER TABLE RENAME to a different database.

When a table was moved from one database to another, the column statistics were not pointed to the new database. This could result in lower performance for queries due to unavailable statistics, and also an inability to drop the table.

Bug: [IMPALA-1711](#)

IMPALA-1556 causes memory leak with secure connections

impalad daemons could experience a memory leak on clusters using Kerberos authentication, with memory usage growing as more data is transferred across the secure channel, either to the client program or between Impala nodes. The same issue affected LDAP-secured clusters to a lesser degree, because the LDAP security only covers data transferred back to client programs.

Bug: [IMPALA-1674](#)

unix_timestamp() does not return correct time

The `unix_timestamp()` function could return an incorrect value (a constant value of 1).

Bug: [IMPALA-1623](#)

Impala incorrectly handles text data missing a newline on the last line

Some queries did not recognize the final line of a text data file if the line did not end with a newline character. This could lead to inconsistent results, such as a different number of rows for `SELECT COUNT(*)` as opposed to `SELECT *.`

Bug: [IMPALA-1476](#)

Impala's ACLs check do not consider all group ACLs, only checked first one.

If the HDFS user ID associated with the `impalad` process had read or write access in HDFS based on group membership, Impala statements could still fail with HDFS permission errors if that group was not the first listed group for that user ID.

Bug: [IMPALA-1805](#)

Fix infinite loop opening or closing file with invalid metadata

Truncating a file in HDFS, after Impala had cached the file metadata, could produce a hang when Impala queried a table containing that file.

Bug: [IMPALA-1794](#)

Cannot write Parquet files when values are larger than 64KB

Impala could sometimes fail to `INSERT` into a Parquet table if a column value such as a `STRING` was larger than 64 KB.

Bug: [IMPALA-1705](#)

Impala Will Not Run on Certain Intel CPUs

This fix relaxes the CPU requirement for Impala. Now only the SSSE3 instruction set is required. Formerly, SSE4.1 instructions were generated, making Impala refuse to start on some older CPUs.

Bug: [IMPALA-1646](#)

Issues Fixed in Impala for CDH 5.3.10

For the full list of fixed issues for all CDH components, see [Issues Fixed in CDH 5.3.10](#) on page 268.



Note: This Impala maintenance release is only available as part of CDH 5, not under CDH 4.

- [IMPALA-1702](#) - "invalidate metadata" can cause duplicate TableIds (issue not entirely fixed, but now fails gracefully)
- [IMPALA-2125](#) - Improve perf when reading timestamps from parquet files written by hive
- [IMPALA-2565](#) - Planner tests are flaky due to file size mismatches
- [IMPALA-3095](#) - Allow additional Kerberos users to be authorized to access internal APIs

Issues Fixed in the 2.1.7 Release / CDH 5.3.9

This section lists the most significant Impala issues fixed in Impala 2.1.7 for CDH 5.3.9.

For the full list of Impala fixed issues in this release, see [Issues Fixed in CDH 5.3.9](#) on page 271.



Note: This Impala maintenance release is only available as part of CDH 5, not under CDH 4.

Query return empty result if it contains NullLiteral in inlineview

If an inline view in a `FROM` clause contained a `NULL` literal, the result set was empty.

Bug: [IMPALA-1917](#)

Fix edge cases for decimal/integer cast

A value of type `DECIMAL(3,0)` could be incorrectly cast to `TINYINT`. The resulting out-of-range value could be incorrect. After the fix, the smallest type that is allowed for this cast is `INT`, and attempting to use `DECIMAL(3,0)` in a `TINYINT` context produces a "loss of precision" error.

Bug: [IMPALA-2264](#)

CDH 5 Release Notes

Constant filter expressions are not checked for errors and state cleanup on exception / DCHECK on destroying an ExprContext

An invalid constant expression in a WHERE clause (for example, an invalid regular expression pattern) could fail to clean up internal state after raising a query error. Therefore, certain combinations of invalid expressions in a query could cause a crash, or cause a query to continue when it should halt with an error.

Bug: [IMPALA-1756](#), [IMPALA-2514](#)

QueryExecState does not check for query cancellation or errors

A call to SetError() in a user-defined function (UDF) would not cause the query to fail as expected.

Bug: [IMPALA-1746](#), [IMPALA-2141](#)

Issues Fixed in the 2.1.6 Release / CDH 5.3.8

This section lists the most significant Impala issues fixed in Impala 2.1.6 for CDH 5.3.8.

For the full list of Impala fixed issues in this release, see [Issues Fixed in CDH 5.3.8](#) on page 272.



Note: This Impala maintenance release is only available as part of CDH 5, not under CDH 4.

Wrong DCHECK in PHJ::ProcessProbeBatch

Certain queries could encounter a serious error if the spill-to-disk mechanism was activated.

Bug: [IMPALA-2364](#)

LargestSpilledPartition was not checking if partition is closed

Certain queries could encounter a serious error if the spill-to-disk mechanism was activated.

Bug: [IMPALA-2314](#)

Avoid cardinality 0 in scan nodes of small tables and low selectivity

Impala could generate a suboptimal query plan for some queries involving small tables.

Bug: [IMPALA-2165](#)

fix Expr::ComputeResultsLayout() logic

Queries using the GROUP BY operator on multiple CHAR columns with length less than or equal to 2 characters could return incorrect results for some columns.

Bug: [IMPALA-2178](#)

Properly unescape string value for HBase filters

Queries against HBase tables could return incomplete results if the WHERE clause included string comparisons using literals containing escaped quotation marks.

Bug: [IMPALA-2133](#)

Avoiding a DCHECK of NULL hash table in spilled right joins

A query could encounter a serious error if it contained a RIGHT OUTER, RIGHT ANTI, or FULL OUTER join clause and approached the memory limit on a host so that the “spill to disk” mechanism was activated.

Bug: [IMPALA-1929](#)

Issues Fixed in the 2.1.5 Release / CDH 5.3.6

This section lists the most significant Impala issues fixed in Impala 2.1.5 for CDH 5.3.6.

For the full list of Impala fixed issues in this release, see [Issues Fixed in CDH 5.3.6](#) on page 275.



Note: This Impala maintenance release is only available as part of CDH 5, not under CDH 4.

Avoid calling ProcessBatch with out_batch->AtCapacity in right joins

Queries including RIGHT OUTER JOIN, RIGHT ANTI JOIN, or FULL OUTER JOIN clauses and involving a high volume of data could encounter a serious error.

Bug: [IMPALA-1919](#)

Issues Fixed in the 2.1.4 Release / CDH 5.3.4

This section lists the most significant Impala issues fixed in Impala 2.1.4 for CDH 5.3.4. Because CDH 5.3.5 does not include any code changes for Impala, Impala 2.1.4 is included with both CDH 5.3.4 and 5.3.5.

For the full list of Impala fixed issues in Impala 2.1.4 for CDH 5.3.4, see [Issues Fixed in CDH 5.3.4](#) on page 277.



Note: This Impala maintenance release is only available as part of CDH 5, not under CDH 4.

Crash: impala::TupleIsNotNullPredicate::Prepare

When expressions that tested for NULL were used in combination with analytic functions, an error could occur. (The original crash issue was fixed by an earlier patch.)

Bug: [IMPALA-1519](#)

Expand parsing of decimals to include scientific notation

DECIMAL literals could include e scientific notation. For example, now CAST(1e3 AS DECIMAL(5,3)) is a valid expression. Formerly it returned NULL. Some scientific expressions might have worked before in DECIMAL context, but only when the scale was 0.

Bug: [IMPALA-1952](#)

INSERT/CTAS evaluates and applies constant predicates.

An INSERT OVERWRITE statement would write new data, even if a constant clause such as WHERE 1 = 0 should have prevented it from writing any rows.

Bug: [IMPALA-1860](#)

Assign predicates below analytic functions with a compatible partition by clause

If the PARTITION BY clause in an analytic function refers to partition key columns in a partitioned table, now Impala can perform partition pruning during the analytic query.

Bug: [IMPALA-1900](#)

FIRST_VALUE may produce incorrect results with preceding windows

A query using the FIRST_VALUE analytic function and a window defined with the PRECEDING keyword could produce wrong results.

Bug: [IMPALA-1888](#)

FIRST_VALUE rewrite fn type might not match slot type

A query referencing a DECIMAL column with the FIRST_VALUE analytic function could encounter an error.

Bug: [IMPALA-1559](#)

CDH 5 Release Notes

AnalyticEvalNode cannot handle partition/order by exprs with NaN

A query using an analytic function could encounter an error if the evaluation of an analytic ORDER BY or PARTITION expression resulted in a NaN value, for example if the ORDER BY or PARTITION contained a division operation where both operands were zero.

Bug: [IMPALA-1808](#)

AnalyticEvalNode not properly handling nullable tuples

An analytic function containing only an OVER clause could encounter an error if another part of the query (specifically an outer join) produced all-NULL tuples.

Bug: [IMPALA-1562](#)

Issues Fixed in the 2.1.3 Release / CDH 5.3.3

This section lists the most significant issues fixed in Impala 2.1.3.

For the full list of fixed issues in Impala 2.1.3, see [Issues Fixed in CDH 5.3.3](#) on page 279.



Note: Impala 2.1.3 is available as part of CDH 5.3.3, not under CDH 4.

Add compatibility flag for Hive-Parquet-Timestamps

When Hive writes TIMESTAMP values, it represents them in the local time zone of the server. Impala expects TIMESTAMP values to always be in the UTC time zone, possibly leading to inconsistent results depending on which component created the data files. This patch introduces a new startup flag, -convert_legacy_hive_parquet_utc_timestamps for the impalad daemon. Specify -convert_legacy_hive_parquet_utc_timestamps=true to make Impala recognize Parquet data files written by Hive and automatically adjust TIMESTAMP values read from those files into the UTC time zone for compatibility with other Impala TIMESTAMP processing. Although this setting is currently turned off by default, consider enabling it if practical in your environment, for maximum interoperability with Hive-created Parquet files.

Bug: [IMPALA-1658](#)

Use sprintf() instead of lexical_cast() in float-to-string casts

Converting a floating-point value to a STRING could be slower than necessary.

Bug: [IMPALA-1738](#)

Fix partition spilling cleanup when new stream OOMs

Certain calls to aggregate functions with STRING arguments could encounter a serious error when the system ran low on memory and attempted to activate the spill-to-disk mechanism. The error message referenced the function impala::AggregateFunctions::StringValGetValue.

Bug: [IMPALA-1865](#)

Impala's ACLs check do not consider all group ACLs, only checked first one.

If the HDFS user ID associated with the impalad process had read or write access in HDFS based on group membership, Impala statements could still fail with HDFS permission errors if that group was not the first listed group for that user ID.

Bug: [IMPALA-1805](#)

Fix infinite loop opening or closing file with invalid metadata

Truncating a file in HDFS, after Impala had cached the file metadata, could produce a hang when Impala queried a table containing that file.

Bug: [IMPALA-1794](#)

external-data-source-executor leaking global jni refs

Successive calls to the data source API could result in excessive memory consumption, with memory allocated but never freed.

Bug: [IMPALA-1801](#)

Spurious stale block locality messages

Impala could issue messages stating the block locality metadata was stale, when the metadata was actually fine. The internal “remote bytes read” counter was not being reset properly. This issue did not cause an actual slowdown in query execution, but the spurious error could result in unnecessary debugging work and unnecessary use of the INVALIDATE_METADATA statement.

Bug: [IMPALA-1712](#)

Issues Fixed in the 2.1.2 Release / CDH 5.3.2

This section lists the most significant issues fixed in Impala 2.1.2.

For the full list of fixed issues in Impala 2.1.2, see [this report in the JIRA system](#).



Note: Impala 2.1.2 is available as part of CDH 5.3.2, not under CDH 4.

Impala incorrectly handles double numbers with more than 19 significant decimal digits

When a floating-point value was read from a text file and interpreted as a FLOAT or DOUBLE value, it could be incorrectly interpreted if it included more than 19 significant digits.

Bug: [IMPALA-1622](#)

unix_timestamp() does not return correct time

The unix_timestamp() function could return an incorrect value (a constant value of 1).

Bug: [IMPALA-1623](#)

Row Count Mismatch: Partition pruning with NULL

A query against a partitioned table could return incorrect results if the WHERE clause compared the partition key to NULL using operators such as = or !=.

Bug: [IMPALA-1535](#)

Fetch column stats in bulk using new (Hive .13) HMS APIs

The performance of the COMPUTE_STATS statement and queries was improved, particularly for wide tables.

Bug: [IMPALA-1120](#)

Issues Fixed in the 2.1.1 Release / CDH 5.3.1

This section lists the most significant issues fixed in Impala 2.1.1.

For the full list of fixed issues in Impala 2.1.1, see [this report in the JIRA system](#).

IMPALA-1556 causes memory leak with secure connections

impalad daemons could experience a memory leak on clusters using Kerberos authentication, with memory usage growing as more data is transferred across the secure channel, either to the client program or between Impala nodes. The same issue affected LDAP-secured clusters to a lesser degree, because the LDAP security only covers data transferred back to client programs.

Bug: <https://issues.cloudera.org/browse/IMPALA-1674> IMPALA-1674

CDH 5 Release Notes

TSaslServerTransport::Factory::getTransport() leaks transport map entries

impalad daemons in clusters secured by Kerberos or LDAP could experience a slight memory leak on each connection. The accumulation of unreleased memory could cause problems on long-running clusters.

Bug: [IMPALA-1668](#)

Issues Fixed in the 2.1.0 Release / CDH 5.3.0

This section lists the most significant issues fixed in Impala 2.1.0.

For the full list of fixed issues in Impala 2.1.0, see [this report in the JIRA system](#).

Kerberos fetches 3x slower

Transferring large result sets back to the client application on Kerberos

Bug: [IMPALA-1455](#)

Compressed file needs to be held on entirely in Memory

Queries on gzipped text files required holding the entire data file and its uncompressed representation in memory at the same time. SELECT and COMPUTE STATS statements could fail or perform inefficiently as a result. The fix enables streaming reads for gzipped text, so that the data is uncompressed as it is read.

Bug: [IMPALA-1556](#)

Cannot read hbase metadata with NullPointerException: null

Impala might not be able to access HBase tables, depending on the associated levels of Impala and HBase on the system.

Bug: [IMPALA-1611](#)

Serious errors / crashes

Improved code coverage in Impala testing uncovered a number of potentially serious errors that could occur with specific query syntax. These errors are resolved in Impala 2.1.

Bug: [IMPALA-1553](#), [IMPALA-1528](#), [IMPALA-1526](#), [IMPALA-1524](#), [IMPALA-1508](#), [IMPALA-1493](#), [IMPALA-1501](#), [IMPALA-1483](#)

Issues Fixed in the 2.0.5 Release / CDH 5.2.6

For the full list of fixed issues in Impala 2.0.5, see [this report in the JIRA system](#).



Note: Impala 2.0.5 is available as part of CDH 5.2.6, not under CDH 4.

Issues Fixed in the 2.0.4 Release / CDH 5.2.5

This section lists the most significant issues fixed in Impala 2.0.4.

For the full list of fixed issues in Impala 2.0.4, see [this report in the JIRA system](#).



Note: Impala 2.0.4 is available as part of CDH 5.2.5, not under CDH 4.

Add compatibility flag for Hive-Parquet-Timestamps

When Hive writes TIMESTAMP values, it represents them in the local time zone of the server. Impala expects TIMESTAMP values to always be in the UTC time zone, possibly leading to inconsistent results depending on which component created the data files. This patch introduces a new startup flag, `-convert_legacy_hive_parquet_utc_timestamps` for the impalad daemon. Specify `-convert_legacy_hive_parquet_utc_timestamps=true` to make Impala recognize Parquet data files written by Hive and automatically adjust TIMESTAMP values read from those files into the

UTC time zone for compatibility with other Impala `TIMESTAMP` processing. Although this setting is currently turned off by default, consider enabling it if practical in your environment, for maximum interoperability with Hive-created Parquet files.

Bug: [IMPALA-1658](#)

IoMgr infinite loop opening/closing file when shorter than cached metadata size

If a table data file was replaced by a shorter file outside of Impala, such as with `INSERT OVERWRITE` in Hive producing an empty output file, subsequent Impala queries could hang.

Bug: [IMPALA-1794](#)

Issues Fixed in the 2.0.3 Release / CDH 5.2.4

This section lists the most significant issues fixed in Impala 2.0.3.

For the full list of fixed issues in Impala 2.0.3, see [this report in the JIRA system](#).



Note: Impala 2.0.3 is available as part of CDH 5.2.4, not under CDH 4.

Anti join could produce incorrect results when spilling

An anti-join query (or a `NOT EXISTS` operation that was rewritten internally into an anti-join) could produce incorrect results if Impala reached its memory limit, causing the query to write temporary results to disk.

Bug: [IMPALA-1471](#)

Row Count Mismatch: Partition pruning with NULL

A query against a partitioned table could return incorrect results if the `WHERE` clause compared the partition key to `NULL` using operators such as `=` or `!=`.

Bug: [IMPALA-1535](#)

Fetch column stats in bulk using new (Hive .13) HMS APIs

The performance of the `COMPUTE STATS` statement and queries was improved, particularly for wide tables.

Bug: [IMPALA-1120](#)

Issues Fixed in the 2.0.2 Release / CDH 5.2.3

This section lists the most significant issues fixed in Impala 2.0.2.

For the full list of fixed issues in Impala 2.0.2, see [this report in the JIRA system](#).



Note: Impala 2.0.2 is available as part of CDH 5.2.3, not under CDH 4.

GROUP BY on STRING column produces inconsistent results

Some operations in queries submitted through Hue or other HiveServer2 clients could produce inconsistent results.

Bug: [IMPALA-1453](#)

Fix leaked file descriptor and excessive file descriptor use

Impala could encounter an error from running out of file descriptors. The fix reduces the amount of time file descriptors are kept open, and avoids leaking file descriptors when read operations encounter errors.

unix_timestamp() does not return correct time

The `unix_timestamp()` function could return a constant value 1 instead of a representation of the time.

Bug: [IMPALA-1623](#)

CDH 5 Release Notes

[Impala should randomly select cached replica](#)

To avoid putting too heavy a load on any one node, Impala now randomizes which scan node processes each HDFS data block rather than choosing the first cached block replica.

Bug: [IMPALA-1586](#)

[Impala does not always give short name to Llama.](#)

In clusters secured by Kerberos or LDAP, a discrepancy in internal transmission of user names could cause a communication error with Llama.

Bug: [IMPALA-1606](#)

[accept unmangled native UDF symbols](#)

The `CREATE FUNCTION` statement could report that it could not find a function entry point within the `.so` file for a UDF written in C++, even if the corresponding function was present.

Bug: [IMPALA-1475](#)

[Issues Fixed in the 2.0.1 Release / CDH 5.2.1](#)

This section lists the most significant issues fixed in Impala 2.0.1.

For the full list of fixed issues in Impala 2.0.1, see [this report in the JIRA system](#).

[Queries fail with metastore exception after upgrade and compute stats](#)

After running the `COMPUTE STATS` statement on an Impala table, subsequent queries on that table could fail with the exception message `Failed to load metadata for table: default.stats_test`.

Bug: <https://issues.cloudera.org/browse/IMPALA-1416> IMPALA-1416

Workaround: Upgrading to CDH 5.2.1, or another level of CDH that includes the fix for HIVE-8627, prevents the problem from affecting future `COMPUTE STATS` statements. On affected levels of CDH, or for Impala tables that have become inaccessible, the workaround is to disable the `hive.metastore.try.direct.sql` setting in the Hive metastore `hive-site.xml` file and issue the `INVALIDATE METADATA` statement for the affected table. You do not need to rerun the `COMPUTE STATS` statement for the table.

[Issues Fixed in the 2.0.0 Release / CDH 5.2.0](#)

This section lists the most significant issues fixed in Impala 2.0.0.

For the full list of fixed issues in Impala 2.0.0, see [this report in the JIRA system](#).

[Join Hint is dropped when used inside a view](#)

Hints specified within a view query did not take effect when the view was queried, leading to slow performance. As part of this fix, Impala now supports hints embedded within comments.

Bug: [IMPALA-995](#)

[WHERE condition ignored in simple query with RIGHT JOIN](#)

Potential wrong results for some types of queries.

Bug: [IMPALA-1101](#)

[Query with self joined table may produce incorrect results](#)

Potential wrong results for some types of queries.

Bug: [IMPALA-1102](#)

[Incorrect plan after reordering predicates \(inner join following outer join\)](#)

Potential wrong results for some types of queries.

Bug: [IMPALA-1118](#)

Combining fragments with compatible data partitions can lead to incorrect results due to type incompatibilities (missing casts).

Potential wrong results for some types of queries.

Bug: [IMPALA-1123](#)"

Predicate dropped: Inline view + DISTINCT aggregate in outer query

Potential wrong results for some types of queries.

Bug: [IMPALA-1165](#)"

Reuse of a column in JOIN predicate may lead to incorrect results

Potential wrong results for some types of queries.

Bug: [IMPALA-1353](#)"

Usage of TRUNC with string timestamp reliably crashes node

Serious error for certain combinations of function calls and data types.

Bug: [IMPALA-1105](#)"

Timestamp Cast Returns invalid TIMESTAMP

Serious error for certain combinations of function calls and data types.

Bug: [IMPALA-1109](#)"

IllegalStateException upon JOIN of DECIMAL columns with different precision

DECIMAL columns with different precision could not be compared in join predicates.

Bug: [IMPALA-1121](#)"

Allow creating Avro tables without column definitions. Allow COMPUTE STATS to always work on Impala-created Avro tables.

Hive-created Avro tables with columns specified by a JSON file or literal could produce errors when queried in Impala, and could not be used with the COMPUTE STATS statement. Now you can create such tables in Impala to avoid such errors.

Bug: [IMPALA-1104](#)"

Ensure all webserver output is escaped

The Impala debug web UI did not properly encode all output.

Bug: [IMPALA-1133](#)"

Queries with union in inline view have empty resource requests

Certain queries could run without obeying the limits imposed by resource management.

Bug: [IMPALA-1236](#)"

Impala does not employ ACLs when checking path permissions for LOAD and INSERT

Certain INSERT and LOAD DATA statements could fail unnecessarily, if the target directories in HDFS had restrictive HDFS permissions, but those permissions were overridden by HDFS extended ACLs.

Bug: [IMPALA-1279](#)"

Impala does not map principals to lowercase, affecting Sentry authorisation

In a Kerberos environment, the principal name was not mapped to lowercase, causing issues when a user logged in with an uppercase principal name and Sentry authorization was enabled.

Bug: [IMPALA-1334](#)"

CDH 5 Release Notes

Issues Fixed in the 1.4.4 Release / CDH 5.1.5

For the list of fixed issues, see [Issues Fixed in CDH 5.1.5](#) in the *CDH 5 Release Notes*.



Note: Impala 1.4.4 is available as part of CDH 5.1.5, not under CDH 4.

Issues Fixed in the 1.4.3 Release / CDH 5.1.4

Impala 1.4.3 includes fixes to address what is known as the POODLE vulnerability in SSLv3. SSLv3 access is disabled in the Impala debug web UI.



Note: Impala 1.4.3 is available as part of CDH 5.1.4, and under CDH 4.

Issues Fixed in the 1.4.2 Release / CDH 5.1.3

This section lists the most significant issues fixed in Impala 1.4.2.

For the full list of fixed issues in Impala 1.4.2, see [this report in the JIRA system](#).



Note: Impala 1.4.3 is available as part of CDH 5.1.4, and under CDH 4.

Issues Fixed in the 1.4.1 Release / CDH 5.1.2

This section lists the most significant issues fixed in Impala 1.4.1.

For the full list of fixed issues in Impala 1.4.1, see [this report in the JIRA system](#).



Note: Impala 1.4.1 is only available as part of CDH 5.1.2, not under CDH 4.

impalad terminating with Boost exception

Occasionally, a non-trivial query run through Llama could encounter a serious error. The detailed error in the log was:

```
boost::exception_detail::clone_impl
<boost::exception_detail::error_info_injector<boost::lock_error> >
```

Severity: High

Impalad uses wrong string format when writing logs

Impala log files could contain internal error messages due to a problem formatting certain strings. The messages consisted of a Java call stack starting with:

```
jni-util.cc:177] java.util.MissingFormatArgumentException: Format specifier 's'
```

Update HS2 client API.

A downlevel version of the HiveServer2 API could cause difficulty retrieving the precision and scale of a DECIMAL value.

Bug: [IMPALA-1107](#)

Impalad catalog updates can fail with error: "IllegalArgumentException: fromKey out of range" at com.cloudera.impala.catalog.CatalogDeltaLog

The error in the title could occur following a DDL statement. This issue was discovered during internal testing and has not been reported in customer environments.

Bug: [IMPALA-1093](#)

"Total" time counter does not capture all the network transmit time

The time for some network operations was not counted in the report of total time for a query, making it difficult to diagnose network-related performance issues.

Bug: [IMPALA-1131](#)

Impala will crash when reading certain Avro files containing bytes data

Certain Avro fields for byte data could cause Impala to be unable to read an Avro data file, even if the field was not part of the Impala table definition. With this fix, Impala can now read these Avro data files, although Impala queries cannot refer to the "bytes" fields.

Bug: [IMPALA-1149](#)

Support specifying a custom AuthorizationProvider in Impala

The `--authorization_policy_provider_class` option for `impalad` was added back. This option specifies a custom `AuthorizationProvider` class rather than the default `HadoopGroupAuthorizationProvider`. It had been used for internal testing, then removed in Impala 1.4.0, but it was considered useful by some customers.

Bug: [IMPALA-1142](#)

Issues Fixed in the 1.4.0 Release / CDH 5.1.0

This section lists the most significant issues fixed in Impala 1.4.0.

For the full list of fixed issues in Impala 1.4.0, see [this report in the JIRA system](#).

Failed DCHECK in disk-io-mgr-reader-context.cc:174

The serious error in the title could occur, with the supplemental message:

```
num_used_buffers_ < 0: #used=-1 during cancellation HDFS cached data
```

The issue was due to the use of HDFS caching with data files accessed by Impala. Support for HDFS caching in Impala was introduced in Impala 1.4.0 for CDH 5.1.0. The fix for this issue was backported to Impala 1.3.x, and is the only change in Impala 1.3.2 for CDH 5.0.4.

Bug: [IMPALA-1019](#)

Workaround: On CDH 5.0.x, upgrade to CDH 5.0.4 with Impala 1.3.2, where this issue is fixed. In Impala 1.3.0 or 1.3.1 on CDH 5.0.x, do not use HDFS caching for Impala data files in Impala internal or external tables. If some of these data files are cached (for example because they are used by other components that take advantage of HDFS caching), set the query option `DISABLE_CACHED_READS=true`. To set that option for all Impala queries across all sessions, start `impalad` with the `-default_query_options` option and include this setting in the option argument, or on a cluster managed by Cloudera Manager, fill in this option setting on the **Impala Daemon** options page.

Resolution: This issue is fixed in Impala 1.3.2 for CDH 5.0.4. The addition of HDFS caching support in Impala 1.4 means that this issue does not apply to any new level of Impala on CDH 5.

impala-shell only works with ASCII characters

The `impala-shell` interpreter could encounter errors processing SQL statements containing non-ASCII characters.

Bug: [IMPALA-489](#)

The extended view definition SQL text in Views created by Impala should always have fully-qualified table names

When a view was accessed while inside a different database, references to tables were not resolved unless the names were fully qualified when the view was created.

Bug: [IMPALA-962](#)

CDH 5 Release Notes

Impala forgets about partitions with non-existent locations

If an `ALTER TABLE` specified a non-existent HDFS location for a partition, afterwards Impala would not be able to access the partition at all.

Bug: [IMPALA-741](#)

CREATE TABLE LIKE fails if source is a view

The `CREATE TABLE LIKE` clause was enhanced to be able to create a table with the same column definitions as a view. The resulting table is a text table unless the `STORED AS` clause is specified, because a view does not have an associated file format to inherit.

Bug: [IMPALA-834](#)

Improve partition pruning time

Operations on tables with many partitions could be slow due to the time to evaluate which partitions were affected. The partition pruning code was speeded up substantially.

Bug: [IMPALA-887](#)

Improve compute stats performance

The performance of the `COMPUTE STATS` statement was improved substantially. The efficiency of its internal operations was improved, and some statistics are no longer gathered because they are not currently used for planning Impala queries.

Bug: [IMPALA-1003](#)

When I run `CREATE TABLE new_table LIKE avro_table`, the schema does not get mapped properly from an avro schema to a hive schema

After a `CREATE TABLE LIKE` statement using an Avro table as the source, the new table could have incorrect metadata and be inaccessible, depending on how the original Avro table was created.

Bug: [IMPALA-185](#)

Race condition in IoMgr. Blocked ranges enqueued after cancel.

Impala could encounter a serious error after a query was cancelled.

Bug: [IMPALA-1046](#)

Deadlock in scan node

A deadlock condition could make all `impalad` daemons hang, making the cluster unresponsive for Impala queries.

Bug: [IMPALA-1083](#)

Issues Fixed in the 1.3.3 Release / CDH 5.0.5

Impala 1.3.3 includes fixes to address what is known as the POODLE vulnerability in SSLv3. SSLv3 access is disabled in the Impala debug web UI.



Note: Impala 1.3.3 is only available as part of CDH 5.0.5, not under CDH 4.

Issues Fixed in the 1.3.2 Release / CDH 5.0.4

This backported bug fix is the only change between Impala 1.3.1 and Impala 1.3.2.



Note: Impala 1.3.3 is only available as part of CDH 5.0.5, not under CDH 4.

Failed DCHECK in disk-io-mgr-reader-context.cc:174

The serious error in the title could occur, with the supplemental message:

```
num_used_buffers_ < 0: #used=-1 during cancellation HDFS cached data
```

The issue was due to the use of HDFS caching with data files accessed by Impala. Support for HDFS caching in Impala was introduced in Impala 1.4.0 for CDH 5.1.0. The fix for this issue was backported to Impala 1.3.x, and is the only change in Impala 1.3.2 for CDH 5.0.4.

Bug: [IMPALA-1019](#)

Workaround: On CDH 5.0.x, upgrade to CDH 5.0.4 with Impala 1.3.2, where this issue is fixed. In Impala 1.3.0 or 1.3.1 on CDH 5.0.x, do not use HDFS caching for Impala data files in Impala internal or external tables. If some of these data files are cached (for example because they are used by other components that take advantage of HDFS caching), set the query option `DISABLE_CACHED_READS=true`. To set that option for all Impala queries across all sessions, start `impalad` with the `-default_query_options` option and include this setting in the option argument, or on a cluster managed by Cloudera Manager, fill in this option setting on the **Impala Daemon** options page.

Resolution: This issue is fixed in Impala 1.3.2 for CDH 5.0.4. The addition of HDFS caching support in Impala 1.4 means that this issue does not apply to any new level of Impala on CDH 5.

Issues Fixed in the 1.3.1 Release / CDH 5.0.3

This section lists the most significant issues fixed in Impala 1.3.1.

For the full list of fixed issues in Impala 1.3.1, see [this report in the JIRA system](#). Because 1.3.1 is the first 1.3.x release for CDH 4, if you are on CDH 4, also consult [Issues Fixed in the 1.3.0 Release / CDH 5.0.0](#) on page 340.

Impalad crashes when left joining inline view that has aggregate using distinct

Impala could encounter a severe error in a query combining a left outer join with an inline view containing a `COUNT(DISTINCT)` operation.

Bug: [IMPALA-904](#)

Incorrect result with group by query with null value in group by data

If the result of a `GROUP BY` operation is `NULL`, the resulting row might be omitted from the result set. This issue depends on the data values and data types in the table.

Bug: [IMPALA-901](#)

Drop Function does not clear local library cache

When a UDF is dropped through the `DROP FUNCTION` statement, and then the UDF is re-created with a new .so library or JAR file, the original version of the UDF is still used when the UDF is called from queries.

Bug: [IMPALA-786](#)

Workaround: Restart the `impalad` daemon on all nodes.

Compute stats doesn't propagate underlying error correctly

If a `COMPUTE STATS` statement encountered an error, the error message is “Query aborted” with no further detail. Common reasons why a `COMPUTE STATS` statement might fail include network errors causing the coordinator node to lose contact with other `impalad` instances, and column names that match Impala [reserved words](#). (Currently, if a column name is an Impala reserved word, `COMPUTE STATS` always returns an error.)

Bug: [IMPALA-762](#)

Inserts should respect changes in partition location

After an `ALTER TABLE` statement that changes the `LOCATION` property of a partition, a subsequent `INSERT` statement would always use a path derived from the base data directory for the table.

Bug: [IMPALA-624](#)

CDH 5 Release Notes

Text data with carriage returns generates wrong results for count(*)

A COUNT(*) operation could return the wrong result for text tables using nul characters (ASCII value 0) as delimiters.

Bug: [IMPALA-13](#)

Workaround: Impala adds support for ASCII 0 characters as delimiters through the clause FIELDS TERMINATED BY '\0'.

IO Mgr should take instance memory limit into account when creating io buffers

Impala could allocate more memory than necessary during certain operations.

Bug: [IMPALA-488](#)

Workaround: Before issuing a COMPUTE STATS statement for a Parquet table, reduce the number of threads used in that operation by issuing SET NUM_SCANNER_THREADS=2 in impala-shell. Then issue UNSET NUM_SCANNER_THREADS before continuing with queries.

Impala should provide an option for new sub directories to automatically inherit the permissions of the parent directory

When new subdirectories are created underneath a partitioned table by an INSERT statement, previously the new subdirectories always used the default HDFS permissions for the impala user, which might not be suitable for directories intended to be read and written by other components also.

Bug: [IMPALA-827](#)

Resolution: In Impala 1.3.1 and higher, you can specify the --insert_inherit_permissions configuration when starting the impalad daemon.

Illegal state exception (or crash) in query with UNION in inline view

Impala could encounter a severe error in a query where the FROM list contains an inline view that includes a UNION. The exact type of the error varies.

Bug: [IMPALA-888](#)

INSERT column reordering doesn't work with SELECT clause

The ability to specify a subset of columns in an INSERT statement, with order different than in the target table, was not working as intended.

Bug: [IMPALA-945](#)

Issues Fixed in the 1.3.0 Release / CDH 5.0.0

This section lists the most significant issues fixed in Impala 1.3.0, primarily issues that could cause wrong results, or cause problems running the COMPUTE STATS statement, which is very important for performance and scalability.

For the full list of fixed issues, see [this report in the JIRA system](#).

Inner join after right join may produce wrong results

The automatic join reordering optimization could incorrectly reorder queries with an outer join or semi join followed by an inner join, producing incorrect results.

Bug: [IMPALA-860](#)

Workaround: Including the STRAIGHT_JOIN keyword in the query prevented the issue from occurring.

Incorrect results with codegen on multi-column group by with NULLs.

A query with a GROUP BY clause referencing multiple columns could introduce incorrect NULL values in some columns of the result set. The incorrect NULL values could appear in rows where a different GROUP BY column actually did return NULL.

Bug: [IMPALA-850](#)

Using distinct inside aggregate function may cause incorrect result when using having clause

A query could return incorrect results if it combined an aggregate function call, a `DISTINCT` operator, and a `HAVING` clause, without a `GROUP BY` clause.

Bug: [IMPALA-845](#)

Aggregation on union inside (inline) view not distributed properly.

An aggregation query or a query with `ORDER BY` and `LIMIT` could be executed on a single node in some cases, rather than distributed across the cluster. This issue affected queries whose `FROM` clause referenced an inline view containing a `UNION`.

Bug: [IMPALA-831](#)

Wrong expression may be used in aggregate query if there are multiple similar expressions

If a `GROUP BY` query referenced the same columns multiple times using different operators, result rows could contain multiple copies of the same expression.

Bug: [IMPALA-817](#)

Incorrect results when changing the order of aggregates in the select list with codegen enabled

Referencing the same columns in both a `COUNT()` and a `SUM()` call in the same query, or some other combinations of aggregate function calls, could incorrectly return a result of 0 from one of the aggregate functions. This issue affected references to `TINYINT` and `SMALLINT` columns, but not `INT` or `BIGINT` columns.

Bug: [IMPALA-765](#)

Workaround: Setting the query option `DISABLE_CODEGEN=TRUE` prevented the incorrect results. Switching the order of the function calls could also prevent the issue from occurring.

Union queries give Wrong result in a UNION followed by SIGSEGV in another union

A `UNION` query could produce a wrong result, followed by a serious error for a subsequent `UNION` query.

Bug: [IMPALA-723](#)

String data in MR-produced parquet files may be read incorrectly

Impala could return incorrect string results when reading uncompressed Parquet data files containing multiple row groups. This issue only affected Parquet data files produced by MapReduce jobs.

Bug: [IMPALA-729](#)

Compute stats need to use quotes with identifiers that are Impala keywords

Using a column or table name that conflicted with Impala keywords could prevent running the `COMPUTE STATS` statement for the table.

Bug: [IMPALA-777](#)

COMPUTE STATS child queries do not inherit parent query options.

The `COMPUTE STATS` statement did not use the setting of the `MEM_LIMIT` query option in `impala-shell`, potentially causing problems gathering statistics for wide Parquet tables.

Bug: [IMPALA-903](#)

COMPUTE STATS should update partitions in batches

The `COMPUTE STATS` statement could be slow or encounter a timeout while analyzing a table with many partitions.

Bug: [IMPALA-880](#)

Fail early (in analysis) when COMPUTE STATS is run against Avro table with no columns

If the columns for an Avro table were all defined in the `TBLPROPERTIES` or `SERDEPROPERTIES` clauses, the `COMPUTE STATS` statement would fail after completely analyzing the table, potentially causing a long delay. Although the `COMPUTE STATS` statement still does not work for such tables, now the problem is detected and reported immediately.

Bug: [IMPALA-867](#)

Workaround: Re-create the Avro table with columns defined in SQL style, using the output of `SHOW CREATE TABLE`. (See the JIRA page for detailed steps.)

Issues Fixed in the 1.2.4 Release

This section lists the most significant issues fixed in Impala 1.2.4. For the full list of fixed issues, see [this report in the JIRA system](#).

The Catalog Server exits with an OOM error after a certain number of CREATE statements

A large number of concurrent `CREATE TABLE` statements can cause the `catalogd` process to consume excessive memory, and potentially be killed due to an out-of-memory condition.

Bug: [IMPALA-818](#)

Workaround: Restart the `catalogd` service and re-try the DDL operations that failed.

Catalog Server consumes excessive cpu cycle

A large number of tables and partitions could result in unnecessary CPU overhead during Impala idle time and background operations.

Bug: [IMPALA-821](#)

Resolution: Catalog server processing was optimized in several ways.

Query against Avro table crashes Impala with codegen enabled

A query against a `TIMESTAMP` column in an Avro table could encounter a serious issue.

Bug: [IMPALA-828](#)

Workaround: Set the query option `DISABLE_CODEGEN=TRUE`

Statestore seems to send concurrent heartbeats to the same subscriber leading to repeated "Subscriber '*hostname*' is registering with statestore, ignoring update" messages

Impala nodes could produce repeated error messages after recovering from a communication error with the statestore service.

Bug: [IMPALA-809](#)

Join predicate incorrectly ignored

A join query could produce wrong results if multiple equality comparisons between the same tables referred to the same column.

Bug: [IMPALA-805](#)

Query result differing between Impala and Hive

Certain outer join queries could return wrong results. If one of the tables involved in the join was an inline view, some tests from the `WHERE` clauses could be applied to the wrong phase of the query.

`ArrayIndexOutOfBoundsException / Invalid query handle when reading large HBase cell`

An HBase cell could contain a value larger than 32 KB, leading to a serious error when Impala queries that table. The error could occur even if the applicable row is not part of the result set.

Bug: [IMPALA-715](#)

Workaround: Use smaller values in the HBase table, or exclude the column containing the large value from the result set.

`select with distinct and full outer join, impalad coredump`

A query involving a `DISTINCT` operator combined with a `FULL OUTER JOIN` could encounter a serious error.

Bug: [IMPALA-735](#)

Workaround: Set the query option `DISABLE_CODEGEN=TRUE`

Impala cannot load tables with more than Short.MAX_VALUE number of partitions

If a table had more than 32,767 partitions, Impala would not recognize the partitions above the 32K limit and query results could be incomplete.

Bug: [IMPALA-749](#)

Various issues with HBase row key specification

Queries against HBase tables could fail with an error if the row key was compared to a function return value rather than a string constant. Also, queries against HBase tables could fail if the `WHERE` clause contained combinations of comparisons that could not possibly match any row key.

Resolution: Queries now return appropriate results when function calls are used in the row key comparison. For queries involving non-existent row keys, such as `WHERE row_key IS NULL` or where the lower bound is greater than the upper bound, the query succeeds and returns an empty result set.

Issues Fixed in the 1.2.3 Release

This release is a fix release that supercedes Impala 1.2.2, with the same features and fixes as 1.2.2 plus one additional fix for compatibility with Parquet files generated outside of Impala by components such as Hive, Pig, or MapReduce.

Impala cannot read Parquet files with multiple row groups

The `parquet-mr` library included with CDH4.5 writes files that are not readable by Impala, due to the presence of multiple row groups. Queries involving these data files might result in a crash or a failure with an error such as “Column chunk should not contain two dictionary pages”.

This issue does not occur for Parquet files produced by Impala `INSERT` statements, because Impala only produces files with a single row group.

Bug: [IMPALA-720](#)

Issues Fixed in the 1.2.2 Release

This section lists the most significant issues fixed in Impala 1.2.2. For the full list of fixed issues, see [this report in the JIRA system](#).

Order of table references in FROM clause is critical for optimal performance

Impala does not currently optimize the join order of queries; instead, it joins tables in the order in which they are listed in the `FROM` clause. Queries that contain one or more large tables on the right hand side of joins (either an explicit join expressed as a `JOIN` statement or a join implicit in the list of table references in the `FROM` clause) may run slowly or crash Impala due to out-of-memory errors. For example:

```
SELECT ... FROM small_table JOIN large_table
```

Anticipated Resolution: Fixed in Impala 1.2.2.

Workaround: In Impala 1.2.2 and higher, use the `COMPUTE STATS` statement to gather statistics for each table involved in the join query, after data is loaded. Prior to Impala 1.2.2, modify the query, if possible, to join the largest table first. For example:

```
SELECT ... FROM small_table JOIN large_table
```

should be modified to:

```
SELECT ... FROM large_table JOIN small_table
```

Parquet in CDH4.5 writes data files that are sometimes unreadable by Impala

Some Parquet files could be generated by other components that Impala could not read.

CDH 5 Release Notes

Bug: [IMPALA-694](#)

Resolution: The underlying issue is being addressed by a fix in the CDH Parquet libraries. Impala 1.2.2 works around the problem and reads the existing data files.

Deadlock in statestore when unregistering a subscriber and building a topic update

The statestore service could experience an internal error leading to a hang.

Bug: [IMPALA-699](#)

IllegalStateException when doing a union involving a group by

A UNION query where one side involved a GROUP BY operation could cause a serious error.

Bug: [IMPALA-687](#)

Impala Parquet Writer hit DCHECK in RleEncoder

A serious error could occur when doing an INSERT into a Parquet table.

Bug: [IMPALA-689](#)

Hive UDF jars cannot be loaded by the FE

If the JAR file for a Java-based Hive UDF was not in the CLASSPATH, the UDF could not be called during a query.

Bug: [IMPALA-695](#)

Issues Fixed in the 1.2.1 Release

This section lists the most significant issues fixed in Impala 1.2.1. For the full list of fixed issues, see [this report in the JIRA system](#).

Scanners use too much memory when reading past scan range

While querying a table with long column values, Impala could over-allocate memory leading to an out-of-memory error. This problem was observed most frequently with tables using uncompressed RCFile or text data files.

Bug: [IMPALA-525](#)

Resolution: Fixed in 1.2.1

Join node consumes memory way beyond mem-limit

A join query could allocate a temporary work area that was larger than needed, leading to an out-of-memory error. The fix makes Impala return unused memory to the system when the memory limit is reached, avoiding unnecessary memory errors.

Bug: [IMPALA-657](#)

Resolution: Fixed in 1.2.1

Excessive memory consumption when query tables with 1k columns (Parquet file)

Impala could encounter an out-of-memory condition setting up work areas for Parquet tables with many columns. The fix reduces the size of the allocated memory when not actually needed to hold table data.

Bug: [IMPALA-652](#)

Resolution: Fixed in 1.2.1

Issues Fixed in the 1.2.0 Beta Release

This section lists the most significant issues fixed in Impala 1.2 (beta). For the full list of fixed issues, see [this report in the JIRA system](#).

Issues Fixed in the 1.1.1 Release

This section lists the most significant issues fixed in Impala 1.1.1. For the full list of fixed issues, see [this report in the JIRA system](#).

Unexpected LLVM Crash When Querying Doubles on CentOS 5.x

Certain queries involving `DOUBLE` columns could fail with a serious error. The fix improves the generation of native machine instructions for certain chipsets.

Bug: [IMPALA-477](#)

"block size is too big" error with Snappy-compressed RCFfile containing null

Queries could fail with a "block size is too big" error, due to `NUL` values in RCFfile tables using Snappy compression.

Bug: [IMPALA-482](#)

Cannot query RC file for table that has more columns than the data file

Queries could fail if an Impala RCFfile table was defined with more columns than in the corresponding RCFfile data files.

Bug: [IMPALA-510](#)

Views Sometimes Not Utilizing Partition Pruning

Certain combinations of clauses in a view definition for a partitioned table could result in inefficient performance and incorrect results.

Bug: [IMPALA-495](#)

Update the serde name we write into the metastore for Parquet tables

The SerDes class string written into Parquet data files created by Impala was updated for compatibility with Parquet support in Hive. See [Incompatible Changes Introduced in Impala 1.1.1](#) on page 103 for the steps to update older Parquet data files for Hive compatibility.

Bug: [IMPALA-485](#)

Selective queries over large tables produce unnecessary memory consumption

A query returning a small result sets from a large table could tie up memory unnecessarily for the duration of the query.

Bug: [IMPALA-534](#)

Impala stopped to query AVRO tables

Queries against Avro tables could fail depending on whether the Avro schema URL was specified in the `TBLPROPERTIES` or `SERDEPROPERTIES` field. The fix causes Impala to check both fields for the schema URL.

Bug: [IMPALA-538](#)

Impala continues to allocate more memory even though it has exceed its mem-limit

Queries could allocate substantially more memory than specified in the `impalad -mem_limit` startup option. The fix causes more frequent checking of the limit during query execution.

Bug: [IMPALA-520](#)

Issues Fixed in the 1.1.0 Release

This section lists the most significant issues fixed in Impala 1.1. For the full list of fixed issues, see [this report in the JIRA system](#).

10-20% perf regression for most queries across all table formats

This issue is due to a performance tradeoff between systems running many queries concurrently, and systems running a single query. Systems running only a single query could experience lower performance than in early beta releases. Systems running many queries simultaneously should experience higher performance than in the beta releases.

planner fails with "Join requires at least one equality predicate between the two tables" when "from" table order does not match "where" join order

A query could fail if it involved 3 or more tables and the last join table was specified as a subquery.

Bug: [IMPALA-85](#)

CDH 5 Release Notes

Parquet writer uses excessive memory with partitions

INSERT statements against partitioned tables using the Parquet format could use excessive amounts of memory as the number of partitions grew large.

Bug: [IMPALA-257](#)

Comments in impala-shell in interactive mode are not handled properly causing syntax errors or wrong results

The impala-shell interpreter did not accept comment entered at the command line, making it problematic to copy and paste from scripts or other code examples.

Bug: [IMPALA-192](#)

Cancelled queries sometimes aren't removed from the inflight query list

The Impala web UI would sometimes display a query as if it were still running, after the query was cancelled.

Bug: [IMPALA-364](#)

Impala's 1.0.1 Shell Broke Python 2.4 Compatibility (AttributeError: 'module' object has no attribute 'field_size_limit')

The impala-shell command in Impala 1.0.1 does not work with Python 2.4, which is the default on Red Hat 5.

For the impala-shell command in Impala 1.0, the -o option (pipe output to a file) does not work with Python 2.4.

Bug: [IMPALA-396](#)

Issues Fixed in the 1.0.1 Release

This section lists the most significant issues fixed in Impala 1.0.1. For the full list of fixed issues, see [this report in the JIRA system](#).

Impala parquet scanner cannot read all data files generated by other frameworks

Impala might issue an erroneous error message when processing a Parquet data file produced by a non-Impala Hadoop component.

Bug: [IMPALA-333](#)

Resolution: Fixed

Impala is unable to query RCFile tables which describe fewer columns than the file's header

If an RCFile table definition had fewer columns than the fields actually in the data files, queries would fail.

Bug: [IMPALA-293](#)

Resolution: Fixed

Impala does not correctly substitute _HOST with hostname in --principal

The _HOST placeholder in the --principal startup option was not substituted with the correct hostname, potentially leading to a startup error in setups using Kerberos authentication.

Bug: [IMPALA-351](#)

Resolution: Fixed

HBase query missed the last region

A query for an HBase table could omit data from the last region.

Bug: [IMPALA-356](#)

Resolution: Fixed

Hbase region changes are not handled correctly

After a region in an HBase table was split or moved, an Impala query might return incomplete or out-of-date results.

Bug: [IMPALA-300](#)

Resolution: Fixed

Query state for successful create table is EXCEPTION

After a successful `CREATE TABLE` statement, the corresponding query state would be incorrectly reported as `EXCEPTION`.

Bug: [IMPALA-349](#)**Resolution:** Fixed

Double check release of JNI-allocated byte-strings

Operations involving calls to the Java JNI subsystem (for example, queries on HBase tables) could allocate memory but not release it.

Bug: [IMPALA-358](#)**Resolution:** Fixed

Impala returns 0 for bad time values in UNIX_TIMESTAMP, Hive returns NULL

Impala returns 0 for bad time values in `UNIX_TIMESTAMP`, Hive returns `NULL`.

Impala:

```
impala> select UNIX_TIMESTAMP('10:02:01') ;
impala> 0
```

Hive:

```
hive> select UNIX_TIMESTAMP('10:02:01') FROM tmp;
hive> NULL
```

Bug: [IMPALA-16](#)**Anticipated Resolution:** Fixed

INSERT INTO TABLE SELECT <constant> does not work.

`Insert INTO TABLE SELECT <constant>` will not insert any data and may return an error.

Anticipated Resolution: Fixed**Issues Fixed in the 1.0 GA Release**

Here are the major user-visible issues fixed in Impala 1.0. For a full list of fixed issues, see [this report in the public issue tracker](#).

Undeterministically receive "ERROR: unknown row batch destination..." and "ERROR: Invalid query handle" from impala shell when running union query

A query containing both `UNION` and `LIMIT` clauses could intermittently cause the `impalad` process to halt with a segmentation fault.

Bug: [IMPALA-183](#)**Resolution:** Fixed

Insert with NULL partition keys results in SIGSEGV.

An `INSERT` statement specifying a `NULL` value for one of the partitioning columns could cause the `impalad` process to halt with a segmentation fault.

Bug: [IMPALA-190](#)**Resolution:** Fixed

CDH 5 Release Notes

INSERT queries don't show completed profiles on the debug webpage

In the Impala web user interface, the profile page for an `INSERT` statement showed obsolete information for the statement once it was complete.

Bug: [IMPALA-217](#)

Resolution: Fixed

Impala HBase scan is very slow

Queries involving an HBase table could be slower than expected, due to excessive memory usage on the Impala nodes.

Bug: [IMPALA-231](#)

Resolution: Fixed

Add some library version validation logic to impalad when loading impala-lzo shared library

No validation was done to check that the `impala-lzo` shared library was compatible with the version of Impala, possibly leading to a crash when using LZO-compressed text files.

Bug: [IMPALA-234](#)

Resolution: Fixed

Workaround: Always upgrade the `impala-lzo` library at the same time as you upgrade Impala itself.

Problems inserting into tables with `TIMESTAMP` partition columns leading table metadata loading failures and failed dchecks

`INSERT` statements for tables partitioned on columns involving datetime types could appear to succeed, but cause errors for subsequent queries on those tables. The problem was especially serious if an improperly formatted timestamp value was specified for the partition key.

Bug: [IMPALA-238](#)

Resolution: Fixed

Ctrl-C sometimes interrupts shell in system call, rather than cancelling query

Pressing Ctrl-C in the `impala-shell` interpreter could sometimes display an error and return control to the shell, making it impossible to cancel the query.

Bug: [IMPALA-243](#)

Resolution: Fixed

Empty string partition value causes metastore update failure

Specifying an empty string or `NULL` for a partition key in an `INSERT` statement would fail.

Bug: [IMPALA-252](#)

Resolution: Fixed. The behavior for empty partition keys was made more compatible with the corresponding Hive behavior.

Round() does not output the right precision

The `round()` function did not always return the correct number of significant digits.

Bug: [IMPALA-266](#)

Resolution: Fixed

Cannot cast string literal to string

Casting from a string literal back to the same type would cause an “invalid type cast” error rather than leaving the original value unchanged.

Bug: [IMPALA-267](#)

Resolution: Fixed

Excessive mem usage for certain queries which are very selective

Some queries that returned very few rows experienced unnecessary memory usage.

Bug: [IMPALA-288](#)**Resolution:** Fixed

HdfsScanNode crashes in UpdateCounters

A serious error could occur for relatively small and inexpensive queries.

Bug: [IMPALA-289](#)**Resolution:** Fixed

Parquet performance issues on large dataset

Certain aggregation queries against Parquet tables were inefficient due to lower than required thread utilization.

Bug: [IMPALA-292](#)**Resolution:** Fixed

impala not populating hive metadata correctly for create table

The Impala CREATE TABLE command did not fill in the owner and tbl_type columns in the Hive metastore database.

Bug: [IMPALA-295](#)

Resolution: Fixed. The metadata was made more Hive-compatible.

impala daemons die if statestore goes down

The impalad instances in a cluster could halt when the statestored process became unavailable.

Bug: [IMPALA-312](#)**Resolution:** Fixed

Constant SELECT clauses do not work in subqueries

A subquery would fail if the SELECT statement inside it returned a constant value rather than querying a table.

Bug: [IMPALA-67](#)**Resolution:** Fixed

Right outer Join includes NULLs as well and hence wrong result count

The result set from a right outer join query could include erroneous rows containing NULL values.

Bug: [IMPALA-90](#)**Resolution:** Fixed

Parquet scanner hangs for some queries

The Parquet scanner non-deterministically hangs when executing some queries.

Bug: [IMPALA-204](#)**Resolution:** Fixed

Issues Fixed in Version 0.7 of the Beta Release

Impala does not gracefully handle unsupported Hive table types (INDEX and VIEW tables)

When attempting to load metadata from an unsupported Hive table type (INDEX and VIEW tables), Impala fails with an unclear error message.

Bug: [IMPALA-167](#)

Resolution: Fixed in 0.7

DDL statements (CREATE/ALTER/DROP TABLE) are not supported in the Impala Beta Release

Resolution: Fixed in 0.7

Avro is not supported in the Impala Beta Release

Resolution: Fixed in 0.7

Workaround: None

[Impala does not currently allow limiting the memory consumption of a single query](#)

It is currently not possible to limit the memory consumption of a single query. All tables on the right hand side of JOIN statements need to be able to fit in memory. If they do not, Impala may crash due to out of memory errors.

Resolution: Fixed in 0.7

Aggregate of a subquery result set returns wrong results if the subquery contains a 'limit' and data is distributed across multiple nodes

Aggregate of a subquery result set returns wrong results if the subquery contains a 'limit' clause and data is distributed across multiple nodes. From the query plan, it looks like we are just summing the results from each worker node.

Bug: [IMPALA-20](#)

Resolution: Fixed in 0.7

[Partition pruning for arbitrary predicates that are fully bound by a particular partition column](#)

We currently cannot utilize a predicate like "country_code in ('DE', 'FR', 'US')" to do partitioning pruning, because that requires an equality predicate or a binary comparison.

We should create a superclass of planner.ValueRange, ValueSet, that can be constructed with an arbitrary predicate, and whose isInRange(analyzer, valueExpr) constructs a literal predicate by substitution of the valueExpr into the predicate.

Bug: [IMPALA-144](#)

Resolution: Fixed in 0.7

[Issues Fixed in Version 0.6 of the Beta Release](#)

[Impala reads the NameNode address and port as command line parameters](#)

Impala reads the NameNode address and port as command line parameters rather than reading them from core-site.xml. Updating the NameNode address in the core-site.xml file does not propagate to Impala.

Severity: Low

Resolution: Fixed in 0.6 - Impala reads the namenode location and port from the Hadoop configuration files, though setting -nn and -nn_port overrides this. Users are advised not to set -nn or -nn_port.

[Queries may fail on secure environment due to impalad Kerberos ticket expiration](#)

Queries may fail on secure environment due to impalad Kerberos tickets expiring. This can happen if the Impala -kerberos_reinit_interval flag is set to a value ten minutes or less. This may lead to an impalad requesting a ticket with a lifetime that is less than the time to the next ticket renewal.

Bug: [IMPALA-64](#)

Resolution: Fixed in 0.6

[Concurrent queries may fail when Impala uses Thrift to communicate with the Hive Metastore](#)

Concurrent queries may fail when Impala is using Thrift to communicate with part of the Hive Metastore such as the Hive Metastore Service. In such a case, the error get_fields failed: out of sequence response" may occur because Impala shared a single Hive Metastore Client connection across threads. With Impala 0.6, a separate connection is used for each metadata request.

Bug: [IMPALA-48](#)

Resolution: Fixed in 0.6

impalad fails to start if unable to connect to the Hive Metastore

Impala fails to start if it is unable to establish a connection with the Hive Metastore. This behavior was fixed, allowing Impala to start, even when no Metastore is available.

Bug: [IMPALA-58](#)

Resolution: Fixed in 0.6

Impala treats database names as case-sensitive in some contexts

In some queries (including "USE database" statements), database names are treated as case-sensitive. This may lead queries to fail with an IllegalStateException.

Bug: [IMPALA-44](#)

Resolution: Fixed in 0.6

Impala does not ignore hidden HDFS files

Impala does not ignore hidden HDFS files, meaning those files prefixed with a period '.' or underscore '_'. This diverges from Hive/MapReduce, which skips these files.

Bug: [IMPALA-18](#)

Resolution: Fixed in 0.6

Issues Fixed in Version 0.5 of the Beta Release

Impala may have reduced performance on tables that contain a large number of partitions

Impala may have reduced performance on tables that contain a large number of partitions. This is due to extra overhead reading/parsing the partition metadata.

Resolution: Fixed in 0.5

Backend client connections not getting cached causes an observable latency in secure clusters

Backend impalads do not cache connections to the coordinator. On a secure cluster, this introduces a latency proportional to the number of backend clients involved in query execution, as the cost of establishing a secure connection is much higher than in the non-secure case.

Bug: [IMPALA-38](#)

Resolution: Fixed in 0.5

Concurrent queries may fail with error: "Table object has not been initialised : `PARTITIONS`"

Concurrent queries may fail with error: "Table object has not been initialised : `PARTITIONS`". This was due to a lack of locking in the Impala table/database metadata cache.

Bug: [IMPALA-30](#)

Resolution: Fixed in 0.5

UNIX_TIMESTAMP format behaviour deviates from Hive when format matches a prefix of the time value

The Impala UNIX_TIMESTAMP(val, format) operation compares the length of format and val and returns NULL if they do not match. Hive instead effectively truncates val to the length of the format parameter.

Bug: [IMPALA-15](#)

Resolution: Fixed in 0.5

CDH 5 Release Notes

Issues Fixed in Version 0.4 of the Beta Release

Impala fails to refresh the Hive metastore if a Hive temporary configuration file is removed

Impala is impacted by Hive bug [HIVE-3596](#) which may cause metastore refreshes to fail if a Hive temporary configuration file is deleted (normally located at `/tmp/hive-<user>-<tmp_number>.xml`). Additionally, the `impala-shell` will incorrectly report that the failed metadata refresh completed successfully.

Anticipated Resolution: To be fixed in a future release

Workaround: Restart the `impalad` service. Use the `impalad` log to check for metadata refresh errors.

Ipad/rpad builtin functions is not correct.

The `Ipad/rpad` builtin functions generate the wrong results.

Resolution: Fixed in 0.4

Files with .gz extension reported as 'not supported'

Compressed files with extensions incorrectly generate an exception.

Bug: [IMPALA-14](#)

Resolution: Fixed in 0.4

Queries with large limits would hang.

Some queries with large limits were hanging.

Resolution: Fixed in 0.4

Order by on a string column produces incorrect results if there are empty strings

Resolution: Fixed in 0.4

Issues Fixed in Version 0.3 of the Beta Release

All table loading errors show as unknown table

If Impala is unable to load the metadata for a table for any reason, a subsequent query referring to that table will return an `unknown table` error message, even if the table is known.

Resolution: Fixed in 0.3

A table that cannot be loaded will disappear from SHOW TABLES

After failing to load metadata for a table, Impala removes that table from the list of known tables returned in `SHOW TABLES`. Subsequent attempts to query the table returns 'unknown table', even if the metadata for that table is fixed.

Resolution: Fixed in 0.3

Impala cannot read from HBase tables that are not created as external tables in the hive metastore.

Attempting to select from these tables fails.

Resolution: Fixed in 0.3

Certain queries that contain OUTER JOINS may return incorrect results

Queries that contain OUTER JOINS may not return the correct results if there are predicates referencing any of the joined tables in the WHERE clause.

Resolution: Fixed in 0.3.

Issues Fixed in Version 0.2 of the Beta Release

Subqueries which contain aggregates cannot be joined with other tables or Impala may crash

Subqueries that contain an aggregate cannot be joined with another table or Impala may crash. For example:

```
SELECT * FROM (SELECT sum(col1) FROM some_table GROUP BY col1) t1 JOIN other_table ON  
(...);
```

Resolution: Fixed in 0.2

An insert with a limit that runs as more than one query fragment inserts more rows than the limit.

For example:

```
INSERT OVERWRITE TABLE test SELECT * FROM test2 LIMIT 1;
```

Resolution: Fixed in 0.2

Query with limit clause might fail.

For example:

```
SELECT * FROM test2 LIMIT 1;
```

Resolution: Fixed in 0.2

Files in unsupported compression formats are read as plain text.

Attempting to read such files does not generate a diagnostic.

Resolution: Fixed in 0.2

[Impala server raises a null pointer exception when running an HBase query.](#)

When querying an HBase table whose row-key is string type, the Impala server may raise a null pointer exception.

Resolution: Fixed in 0.2

Cloudera Manager 5 Release Notes

These Release Notes provide information on the new features and known issues and limitations for Cloudera Manager 5. These Release Notes also include fixed issues for releases starting from Cloudera Manager 5.0.0 beta 1.

To view the Release Notes (or other documentation) for a specific Cloudera Manager release, go to [Cloudera Documentation](#), click a major version link, and use the drop-down menu to select the release.

For information about supported operating systems, and other requirements for using Cloudera Manager, see [CDH 5 and Cloudera Manager 5 Requirements and Supported Versions](#) on page 505.

New Features and Changes in Cloudera Manager 5

The following sections describe what's new and changed in each Cloudera Manager 5 release.

What's New in Cloudera Manager 5

The following sections describe what is new in each Cloudera Manager 5 release.

What's New in Cloudera Manager 5.9.0

- **Creating Virtual Machine Images**

Documentation has been added with procedures to create virtual images of Cloudera Manager and cluster hosts. See [Creating Virtual Images of Cluster Hosts](#).

- **Security**

- **External/Cloud account configuration in Cloudera Manager**

Account configuration for access to Amazon Web Services is now available through the centralized UI menu **External Accounts**.

- **Key Trustee Server rolling restart**

Key Trustee Server now supports rolling restart.

- **Backup and Disaster Recovery**

- You can now replicate HDFS files and Hive data to and from an Amazon S3 instance. See [HDFS Replication to Amazon S3](#) and [Hive Replication To and From Amazon S3](#).

- There are some new tuning options to improve performance of HDFS replication. See [HDFS Replication Tuning](#).

- You can now download performance data about HDFS replication jobs from the **Replication Schedules** and **Replication History** pages. See [Monitoring the Performance of HDFS Replications](#).

- Hive replication now stores Hive UDFs in the Hive metastore. [Replication of Impala and Hive User Defined Functions \(UDFs\)](#).

- The user interface for creating replication schedules has been reorganized to present the configuration options on three tabs: **General**, **Resources**, and **Advanced**.

- **Uncheck Replicate Impala Metadata by default**

When creating a Hive replication schedule, the option **Replicate Impala Metadata** was checked (*true*) by default. In Cloudera Manager 5.9 and higher, the value is unchecked (*false*) by default.

- **YARN BDR enhancement**

YARN jobs now include the BDR schedule ID that launched the job so you can connect logs with existing schedules, if multiple schedules exist.

- **Resource Management**

- **Custom Cluster Utilization Reports**

Documentation has been added to create custom Cluster Utilization reports that you can export data from. See [Creating a Custom Cluster Utilization Report](#).

- **New settings for continuous scheduling**

For new installs, default values for configurations have been changed.

`yarn_scheduler_fair_continuous_scheduling_enabled` is set to `false`.

`resourcemanager_fair_scheduler_assign_multiple` is set to '`true`'. Existing settings are preserved when you upgrade from a lower version.

- **YARN historical reports by user show pool-user entity**

When Cloudera Manager manages multiple clusters, there is no per user tracking for historical applications and queries across clusters. Instead, **Historical Applications by User** and **Historical Queries by User** show applications and queries per user and pool. (A pool is associated with a specific cluster.)

- **Directory Usage Report needs export capability**

Directory usage reports can be exported as a CSV file.

- **Cloudera Manager Admin Console User Interface**

- **Service colors**

A new set of colors is used to represent each kind of service.

- **Move the table sorting icon to the right**

The table sorting icon now appears consistently on the right hand side of each column.

- **Improved Configuration Diff Display**

Changes displayed in the configuration history page are much more user friendly. For a large section of changed text, Cloudera Manager generates a diff between the old and the new and displays the diff.

When a user changes only the password, Cloudera Manager does not show the delta: both the old and the new passwords are masked out before the comparison is performed.

- **Move actions menu to the top header**

The actions menu now appears next to the entity title.

- **Move Federation and High Availability to a separate page**

The **Federation** and **High Availability** sections used to appear on the **HDFS Instances** page of an HDFS service. They have been moved to a new page called **Federation and High Availability**. There is a link from the existing **Instances** page to this new page.

- **Remove repeated heading below the second level navigation**

Subtitles below the second level navigation tabs are removed because they repeated the content in the tabs.

- **Move maintenance mode and badges to the title area**

Maintenance mode, staleness badges now appear next to the title of the entity.

- **Express wizard allows you to add Kafka**

Kafka is now listed in the custom services when you click the **Add Cluster** button.

- **Cloudera Manager API**

- **Add update_user to Python API client**

Added the `update_user()` method to the Python API client `api_client.py`.

- **Expose API endpoint to add a specific path**

New API endpoints have been added that allow users to add, list and remove Watched Directories in HDFS service.

- **Logging**

- **Include host in log file name**

Kafka log4j log files now include the host name in the format `kafka-broker-$\{host\}.log`. Similarly, MirrorMaker logs now include the host name in the format `kafka-mirrormaker-$\{host\}.log`. Due to the log file name change, when you upgrade Cloudera Manager it no longer recognizes your old log files in log search, though they are still present on disk.

- **Configuration changes to Cloudera Manager audit log**

Cloudera Manager displays the **History** and **Rollback** support for the Cloudera Manager Settings. (**Administration > Settings**). This helps you to track the changes made by an administrator so that Cloudera Support can provide better service when certain Cloudera Manager administrative settings are modified.

- **Diagnostic Bundles**

- **Show the Diagnostic Bundle Redaction Policy using the redaction config**

You can specify what information should be redacted in the diagnostic bundle in the UI using **Administration > Settings > Redaction Parameters for Diagnostic Bundles**.

- **Upgrade**

- **Report that a simple restart was performed if rolling restart could not be performed**

Informs you when a simple restart is performed instead of rolling restart on a service because rolling restart is not available.

- **Oozie**

- **Provide dump / load functionality for Oozie DB**

The **Actions** menu in the Oozie service has two new commands, **Dump Database** and **Load Database**. These commands make it easier to migrate an Oozie database to another database supported by Oozie. The **Dump Database** command exports Oozie's database to a file (configurable by **Database Dump File** setting). **Load Database** loads the file into a database.

- **Install Oozie ShareLib permissions change**

Install Oozie ShareLib Command assigns correct permissions to the uploaded libraries. This prevents breaking Oozie workflows with a custom umask setting.

- **Configuration Changes**

- **Solr zkClientTimeout option**

Added the `zkClientTimeout` parameter for ZooKeeper.

- **Add JHIST compression as a configuration option**

Added a new option for setting the file format used by an ApplicationMaster when generating the `.jhist` file.

- **Enable heap dump by default for all daemons**

Starting in version 5.9, when you configure roles that are JVM based, the `Dump Heap When Out of Memory` configuration parameter defaults to `true`. An upgrade from a pre-5.9 version maintains your pre-5.9 settings.

- **Cloudera Manager support for client-side YARN graceful decommissioning**

Adds the ability to perform a graceful decommission on YARN NodeManager roles whereby the Node Manager is not assigned new containers, and waits for any currently running applications to finish before being decommissioned unless a timeout occurs. You can configure the timeout using the *Node Manager Graceful Decommission Timeout* configuration property in the YARN Service. The default behavior has not changed, and continues to be a non-graceful decommission. Affects Cloudera Manager 5.9.0 and higher, and CDH 5.9.0 and higher.

- **Deploy Client Configuration command details page now shows stdout/stderr**

stdout and stderr log links are now shown in the UI when there is a failure while deploying client configurations.

- **Make EXTRA_RATIO configurable for Headlamp indexing**

Added the configuration parameter, *Extra Space Ratio for Indexing*, to Reports Manager. You can use the parameter to make the speed of indexing faster by allocating additional memory.

- **Configure HBase Indexer to wait longer for ZooKeeper to come up**

The default amount of time that HBase Indexer roles attempts to connect to ZooKeeper has been increased from 30 to 60 seconds. This default can be adjusted by setting a new Cloudera Manager configuration parameter, HBase Indexer ZooKeeper Session Timeout.

- **Embedded database mode improvements**

In version 5.9 and higher, Cloudera Manager can clearly identify whether or not a customer is using the embedded PostgreSQL database. Cloudera does not recommend the embedded database for production use, and requests that customers deploy production systems using an external database. The diagnostic bundles now contain information about whether or not a customer is using the embedded PostgreSQL database. Support can then reach out to customers accordingly.

If Cloudera Manager is configured to use the embedded PostgreSQL database, a yellow banner appears in the UI recommending that you upgrade to a supported external database.

- **Fix CatalogServiceClient to handle TLS connections to catalogd for UDF replication**

When Impala uses SSL, we now support TLS Connection to Catalog Server. Customers can enable replication for any Impala UDFs/Metadata (in Hive Replication) in Cloudera Manager 5.9 and higher.

- **Do not show steps that are unreachable (skipped)**

When running wizards from the Cloudera Manager Admin Console that add a cluster, add a service, perform an upgrade, and other tasks, steps do not display when they are not reachable or do not apply to the current configuration.

- **Improve Cloudera Manager provisioning performance on AWS**

Add support for resetting Cloudera Manager GUID/UUID. This is accomplished by checking the UUID file.

If Cloudera Manager finds the UUID file (`/etc/cloudera-scm-server/uuid`) and the UUID is different than the GUID in the `cm_version` table, it updates the GUID in the `cm_version` table with the contents of the UUID file and removes the UUID file.

- **DSSD**

- **Trigger HDFS rolling upgrade command for 5.8 to 5.9 in DSSD mode**

DSSD has implemented a data format change for the DSSD-DN in DHP 1.3 (equivalent of Cloudera Manager 5.9). DSSD relies on the HDFS rolling upgrade mechanism to automatically convert from old data format prior to DHP 1.3 to the new data format. This process is very similar to the HDFS metadata (rolling) upgrade process. You can roll back to the old data format before finalizing the upgrade. After the cluster is upgraded to DHP 1.3 and the data format conversion is complete, you can finalize the upgrade, and the old data format is cleaned up. Cloudera Manager issues the HDFS rolling upgrade commands automatically when upgrading a DSSD cluster from 5.8 to 5.9.

In Cloudera Manager 5.8, the DSSD DataNodes operating in the Hadoop cluster stores version and ID information on a local disk directory. The default directory is /tmp/hadoop-hdfs. This default directory may be deleted by the OS over time or when the OS is restarted. If the DSSD DataNode process is restarted after the default directory is deleted, it will not be able to locate replicas stored on the DSSD D5 appliance. This will cause HDFS service to report under-replicated or even missing HDFS blocks.

It is a best practice to change the default directory to a different value such as /var/lib/hadoop-hdfs/dssddn. The directory should be created and configured prior to initiating the installation of the Hadoop cluster:

1. Log into each host that will run the DataNode process.
2. Create the directory by executing the command:

```
mkdir -p /var/lib/hadoop-hdfs/dssddn
```

3. Change the directory ownership by executing the command:

```
chown -R hdfs:hadoop /var/lib/hadoop-hdfs
```

4. Change the directory permissions by executing the command:

```
chmod -R 700 /var/lib/hadoop-hdfs
```

5. In Cloudera Manager, configure the property `dfs.datanode.data.dir` in the **HDFS Service Advanced Configuration Snippet (Safety Valve) for hdfs-site.xml** section of the configuration page. This property must be set after the initial setup is completed, but prior to writing any data to HDFS.



Note: Changing the value of the `dfs.datanode.data.dir` property after data has been written to HDFS will result in under-replicated or lost HDFS blocks.

- **Collect additional DSSD 1.2 metrics**

Cloudera Manager collects the following new metrics from DSSD-DN. Note that these metrics are only available in DHP 1.3 and higher.

Table 3: DSSD-DN Metrics

Metric	Description
<code>hdfs_dssd_usable_capacity</code>	This reports a single-datanode-level view of the usable capacity on the D5 to which that particular DSSD-DN is connected.
<code>hdfs_dssd_used_capacity</code>	This reports a single-datanode-level view of used capacity on the D5 to which that particular DSSD-DN is connected.
<code>hdfs_dssd_object_max_number</code>	This reports a single-datanode-level view of the maximum number of blocks that can be created on the D5 to which that particular DSSD-DN is connected.
<code>hdfs_dssd_object_used_number</code>	This reports a single-datanode-level view of the number of blocks created on the D5 to which that particular DSSD-DN is connected.

- **Remove "Usable Capacity" in cluster setup wizard in DSSD mode**

The DSSD specific configuration `com.dssd.hadoop.floodds.usablecapacity` is no longer required by DPH 1.3 and is not emitted by Cloudera Manager for CDH 5.9 and higher. The **Usable Capacity** configuration no longer exists in the HDFS service and setup wizard in Cloudera Manager for CDH 5.9 and Higher.

- **Libflood CPU ID accepting Decimal and alphanumeric values**

Cloudera Manager validates the Flood CPU ID field and only allows comma separated integers or the string "all".

What's New in Cloudera Manager 5.8.3

Cloudera Manager 5.8.3 is a maintenance release with many fixed issues. See [Issues Fixed in Cloudera Manager 5.8.3](#) on page 399.

What's New in Cloudera Manager 5.8.2

Cloudera Manager 5.8.2 is a maintenance release with many fixed issues. See [Issues Fixed in Cloudera Manager 5.8.2](#) on page 400.

What's New in Cloudera Manager 5.8.1

An issue has been fixed. See [Issues Fixed in Cloudera Manager 5.8.1](#) on page 401.



Note: Although there is a Cloudera Manager 5.8.1 release, there is no synchronous CDH 5.8.1 release.

What's New in Cloudera Manager 5.8.0

- **Operating Systems** - Support for Debian 8.2.
- **Resource management and utilization** - Added support for nesting dynamic resource pools within a named pool at runtime.
- **Backup and Disaster Recovery**
 - You can now configure incremental replication for Hive Replications. See [Enabling Incremental Replication](#).
 - The **Replication Schedules** page now has a search function for finding scheduled replications.
- You can now specify a start and end time for the events that are included in manually-triggered diagnostic bundles. See [Manually Triggering Collection and Transfer of Diagnostic Data to Cloudera](#).
- **Impala**
 - Impala adds a new configuration option, `Use HDFS Rules to Map Kerberos Principals to Short Names`. Enabling this option makes Impala pickup `hadoop.security.auth_to_local` configuration from HDFS configurations and uses it for Kerberos principal-to-short-name translation. This only applies for Cloudera Manager 5.8.0 and higher and CDH 5.8.0 and higher. It only affects deployments where Impala is set up to use Kerberos as the authentication mechanism. It defaults to `false`, to preserve the behavior from earlier CDH versions. This has no impact on upgrade.
 - Enable Impala Admission Control and Enable Dynamic Resource Pools are now enabled by default. Customized configuration values are preserved during the upgrade.
 - Impala Admission Control now supports a global method for editing the Access Control List.
- **EMC DSSD D5**
 - You can now deploy multiple DSSD D5 storage appliances in a cluster using Cloudera Manager. See [Configuring Multiple DSSD D5 Appliances in a Cluster](#).
- **EMC Isilon**

Cloudera Manager 5 Release Notes

- Kerberos is now fully supported for replications between clusters using Isilon storage. You must configure a custom principal.
- **Security**
 - **Active Directory KDC**
 - Active Directory account properties are now configurable from Cloudera Manager's **Administration > Settings** page.
 - It is now possible to use Cloudera Manager to regenerate principals for clusters using an Active Directory KDC. Cloudera Manager 5.8 includes a new configuration called **Active Directory Delete Accounts on Credential Regeneration**. Enabling this will allow Cloudera Manager to automatically delete existing AD accounts and complete the regeneration process.
 - Cloudera Manager now allows you to configure the encryption types (or `enctype`) used by an Active Directory KDC to protect its data.
 - **Redaction:** In the Cloudera Manager Admin Console, Advanced Configuration Snippet parameters will now be redacted to block sensitive information such as passwords or secret keys.
 - **Sentry**
 - Cloudera Search adds support for storing permissions in the Sentry service. You can enable storing permissions in the Sentry service by [Enabling the Sentry Service for Solr](#). If you have already configured Sentry's policy file-based approach, you can migrate existing authorization settings as described in [Migrating from Sentry Policy Files to the Sentry Service](#). `solrctl` has been extended to support:
 - Migrating existing policy files to the Sentry service
 - Managing managing permissions in the Sentry service
 - Sentry adds support for securing data on Amazon S3. As a result, Sentry will now be able to secure URLs with an S3 schema.
- **YARN**
 - YARN Allowed System Users now includes `hbase` by default. This is helpful when running certain tools for HBase that need to execute MapReduce jobs.

What's New in Cloudera Manager 5.7.2

A number of issues have been fixed. See [Issues Fixed in Cloudera Manager 5.7.2](#) on page 409.

What's New in Cloudera Manager 5.7.1

A number of issues have been fixed. See [Issues Fixed in Cloudera Manager 5.7.1](#) on page 410.

What's New in Cloudera Manager 5.7.0

- **Operating Systems** - Support for:

- RHEL/CentOS 6.6, 6.7, 7.1, and 7.2
- Oracle Enterprise Linux (OEL) 7.1 and 7.2
- SUSE Linux Enterprise Server (SLES) 11 with Service Packs 2, 3, 4
- Debian: Wheezy 7.0, 7.1, and 7.8



Important: Cloudera supports RHEL 7 with the following limitations:

- Only RHEL 7.2 and 7.1 are supported. RHEL 7.0 is not supported.
- RHEL 7.1 is only supported with CDH 5.5 and higher.
- RHEL 7.2 is only supported with CDH 5.7 and higher.
- Only new installations of RHEL 7.2 and 7.1 are supported by Cloudera. For upgrades to RHEL 7.1 or 7.2, contact your OS vendor and see [Does Red Hat support upgrades between major versions of Red Hat Enterprise Linux?](#)

- **Resource management and utilization**

- Simplified and expanded resource management. The screens for YARN and Impala dynamic resource pools are now managed separately.
- Resource pools now support the `allowPreemptionFrom`, `minSharePreemptionTimeout`, and `fairSharePreemptionTimeout` attributes.
- Cluster utilization reports track usage of resources allocated using dynamic resource pools.
- The new **Directory Usage Report** now shows aggregated usage information, including quotas and file sizes, which are sortable. You can also perform multiple actions on filesystem objects. .
- Two new predicates have been added to the tsquery language: `day in` and `hour in`. These allow you to limit streams to specified days of the week and specified hours of each day, respectively.

- **Extensibility** - For more information, see [Cloudera Manager Extensions](#).

- Parcels are typed according to the OS version. The parcel extension indicates the version. The library for developing external parcels now supports an extension for RHEL 7.
- A new environment variable, `ZK_PRINCIPAL_NAME`, is now defined for CSD processes when ZooKeeper is Kerberized and has a custom principal.
- A new flag, `jvmBased`, is now available to CSD authors to indicate that a CSD role is JVM-based. This flag enables a set of JVM-related features in Cloudera Manager—for example, the ability to automatically generate a heap dump when an Out Of Memory error occurs.

- **API**

- Cloudera Manager now attempts to gracefully handle overlapping API calls.
- All distributed filesystem services (such as HDFS or Isilon) installed in a cluster can now be enumerated using the API.
- You can export the complete configuration of a CDH cluster managed by Cloudera Manager as a template, modify the template, and import the template to create a new cluster.
- The advanced configuration snippet editor now allows you to edit properties as name/value pairs. This is the default, however you can also choose to edit the snippet as XML.
- HBase now includes metrics and charts for replication. These charts are available in the Chart Library for each RegionServer.
- When you click the **Role Log** link when viewing a role, the log is opened at the current timestamp, rather than the top of the log file. This enables you to see the relevant log messages when investigating an event that occurred recently.
- A new tsquery function called `counter_delta` has been added to accurately compute the difference between consecutive data points for counter metrics.
- The `distcp` utility now supports setting records per chunk, using the `distcp.dynamic.recordsPerChunk` in an advanced configuration snippet to set the number of records (paths) in each chunk. When a value is set for `distcp.dynamic.recordsPerChunk`, other related settings, such as the maximum number of chunks tolerable, the ideal number of chunks, and the split ratio, are ignored.
- A warning is shown when upgrading with incompatible versions of Kafka and CDH. The Kafka client libraries bundled in CDH cannot communicate with an older Kafka server.
- You can override the sudo commands that Cloudera Manager agent uses by redirecting the sudo commands to a script that you write to allow or disallow certain actions.

Cloudera Manager 5 Release Notes

- **Hive**

- Hive on Spark is now supported.
- The default execution engine for Hive can now be configured, which makes it easy to run all Hive jobs on Spark.
- HiveServer2 now has a Web UI.
- Hive and HDFS replication source/target listings now work with Isilon.
- A dialog box now displays when scheduling Hive replications, reminding you to take snapshots of the Hive warehouse directory.
- The Direct SQL option is now enabled in Hive Metastore.
- When upgrading to CDH 5.7, if Hive is configured to use YARN, all Hive on Spark parameters are automatically tuned to recommended values. If Hive on Spark was previously tuned, this is skipped.

A number of issues have been fixed. See [Issues Fixed in Cloudera Manager 5.7.0](#) on page 412.

What's New in Cloudera Manager 5.6.1

A number of issues have been fixed. See [Issues Fixed in Cloudera Manager 5.6.1](#) on page 415.

What's New in Cloudera Manager 5.6.0

EMC DSSD D5 Storage Appliance Support

Support is added for using the EMC DSSD D5 storage appliance (a high-speed, low-latency storage solution based on flash media) as storage for DataNodes.

Cloudera Distribution of Apache Kafka 2.0 is now supported with Cloudera Manager

A number of issues have also been fixed. See [Issues Fixed in Cloudera Manager 5.6.0](#) on page 415.

What's New in Cloudera Manager 5.5.4

A number of issues have been fixed. See [Issues Fixed in Cloudera Manager 5.5.4](#) on page 416.

What's New in Cloudera Manager 5.5.3

An issue has been fixed. See [Issues Fixed in Cloudera Manager 5.5.3](#) on page 417.

What's New in Cloudera Manager 5.5.2



Note: If you are upgrading from a previous version of Cloudera Manager, Cloudera recommends that you upgrade to version 5.5.3 or higher.

- New Impala flags added for web server certificate files and passwords. This adds support for the `--webserver_private_key_file` and `--webserver_private_key_password_cmd` flags for the Impala Daemon, the Impala Catalog Server, and the Impala StateStore roles.

A number of issues have also been fixed. See [Issues Fixed in Cloudera Manager 5.5.2](#) on page 417.

What's New in Cloudera Manager 5.5.1

An issue has been fixed. See [Issues Fixed in Cloudera Manager 5.5.1](#) on page 420.

What's New in Cloudera Manager 5.5.0

- **Operating Systems** - Support for RHEL/CentOS 6.6 (in SE Linux mode), 6.7, and 7.1, and Oracle Enterprise Linux 7.1.



Important: Cloudera supports RHEL 7 with the following limitations:

- Only RHEL 7.1 is supported. RHEL 7.0 is not supported.
- Only new installations of RHEL 7.1 are supported by Cloudera. For upgrades to RHEL 7.1, contact your OS vendor and see [Does Red Hat support upgrades between major versions of Red Hat Enterprise Linux?](#)

- **Databases** - Supports MariaDB 5.5, Oracle 12c, and PostgreSQL 9.4.
- Selective service restart after activating parcels is supported.
- Retrying upgrade actions is supported. If a cluster upgrade command fails while in progress, you can retry a command after fixing the cause of failure. On retry, the command restarts from the command step where it failed.
- The command details page for running and recent commands has been redesigned for usability and scalability.
- Instead of serially starting all services for the first time, services that are not dependent are started in parallel. This decreases the time required to start services for the first time after creating a cluster.
- Performance has improved for service startup, client configuration deployment, and calculation of stale configurations.
- **Suppression of notifications**
 - You can suppress the warnings that Cloudera Manager issues when a configuration value is outside the recommended range or is invalid.
 - You can suppress health test warnings.

Suppression can be useful if a warning does not apply to your deployment and you no longer want to see the notification. Suppressed warnings are still retained by Cloudera Manager, and you can unsuppress the warnings at any time.

- **Multi Cloudera Manager Dashboard** - A special mode of Cloudera Manager that enables you to view monitoring data aggregated from multiple Cloudera Manager instances that manage one or more CDH clusters.
- You can decommission roles when services are completely stopped. This allows you to decommission hosts during cluster downtime.
- You can disable collection of certain domain metrics—for example, for HBase RegionServers, Kafka Brokers, and others—through new settings in the host advanced configuration snippet. This is useful in certain support situations and should only be done under the direction of Cloudera Support.
- You can configure which aggregate metrics are automatically generated. This advanced feature can be useful in certain situations to impact the monitoring workload, allowing unused or less-important aggregate metrics to be skipped. This may result in improved performance and the ability to handle larger monitoring workloads, or to retain data for a larger workload for longer. Cloudera recommends using this only under the direction of Cloudera Support.
- Alert Publisher can be configured to pass alert events to a user-defined script. Use this for integrating with other alerting systems or for custom logic (for example, to send some alerts to some people and others to other people).
- Agent minor version mismatches (5.4 to 5.5) now cause bad host health. Maintenance version mismatches (for example, 5.4.x to 5.4.y) still cause concerning host health.
- Cloudera Manager indicates if the Java version in use is too old.
- Cloudera Manager indicates if the supervisor component of the Agent needs to be restarted after an upgrade.
- Full and User Administrators can view active user sessions.
- Full Administrators and Auditors can audit failed and successful logins.
- Multiple user session logins can be disallowed.
- You can configure external authentication so that local administrator emergency access is disabled. This means that no local accounts can log in under any circumstances, including when the external system is not functioning.
- You can turn on authentication for the URLs for downloading client configuration zip files. Previously, authentication was never required.
- Passwords are no longer accessible in cleartext through the Cloudera Manager UI or in the configuration files stored on disk. There are some exceptions; see [Known Issues and Workarounds in Cloudera Manager 5](#) on page 383.

Cloudera Manager 5 Release Notes

- **HBase**

- Use a configuration option in HBase to skip region reload during rolling restart and rolling upgrade, to increase the speed of the operations.
- HBase rolling restart performance can be improved by increasing the number of Region Mover Threads. If the value of this property is 1, it can lower rolling restart speed. The Admin Console now displays this information and, if the value is 1, advises increasing it.
- HBase Thrift Server and Rest Server support TLS/SSL.

- **HDFS**

- HDFS encryption can be enabled using a wizard.
- Exposes AES as an encryption option for HDFS RPC encryption.

- **Hive**

- Hive can use TLS/SSL and Kerberos at the same time.
- When Hive is configured to use TLS/SSL, Hue is automatically configured to use that protocol when communicating with Hive. Similarly, when Impala is configured to use TLS/SSL, Hue is automatically configured to use that protocol when communicating with Impala.
- HiveServer2 supports a timeout value for idle sessions and operations. By default, it times out client sessions after a week and idle operations after three days. This helps alleviate problems with long-running sessions when using Hue.
- Cloudera Manager collects and displays various operational metrics for Hive.

- **Hue**

- Hue supports a Load Balancer role using HTTPD as a load balancer.
- You can configure certificates trusted by Hue using the TLS/SSL Truststore configuration. This replaces the REQUESTS_CA_BUNDLE advanced configuration snippet entry.
- You can specify a password that protects the Hue private key file.
- Cloudera Manager collects and displays various operational metrics for Hue. New health tests have been added for Hue as well.

- **Impala** supports TLS/SSL internally between the StateStore and the Catalog Server roles as well as Impala Daemon.

- **Kafka**

- Kafka supports rolling restart.
- Kafka displays additional broker metrics.
- Kafka exposes additional commonly configured parameters.
- Existing Kafka parameter definitions have updated descriptions, default values, and validation settings.
- The Kafka broker instance list now shows which broker is the active controller.

- **Key Trustee**

- The Key Trustee Server CSD is included in Cloudera Manager. Manual installation of the Key Trustee Server CSD is not required.
- A Key Administrator role in Cloudera Manager is used for configuring HDFS Data at Rest Encryption. Only a Key Administrator and a Full Administrator can make configuration changes to Java Keystore KMS, Key Trustee KMS, and Key Trustee Server. Configuring HDFS to use Data at Rest Encryption is also limited to the Key Administrator and Full Administrator roles. This allows organizations to keep Key Administrators and Cluster Administrators separate, which is a security best practice.
- When running Key Trustee KMS in a highly available configuration, Cloudera Manager can automatically generate the load balancer URL.

- **Sentry**

- Sentry introduces column-level access control for tables in Hive and Impala. Previously, Sentry supported privilege granularity only at the table level. You can now assign the SELECT privilege on a subset of columns in a table.

- Sentry supports Kerberos authentication for the Sentry web server.
- **Solr**
 - Solr can be configured with a load balancer in a secure environment.
 - There is a new Solr Max Connector Threads property for Solr Server in CDH 5.1.0 and higher.
 - Solr supports LDAP/AD authentication.
- **Backup and Disaster Recovery**
 - The user interface for scheduling and reviewing replications and snapshots has been improved. You can now view the history of replication jobs and subtasks more easily.
 - When specifying an HDFS replication job, you can apply exclusion filters to exclude specific files or directories.
 - You can download or send to Cloudera Support a diagnostic bundle to troubleshoot replication jobs. Bundles include logs of the replication run..
 - The performance of the file-listing phase of a replication job has been improved.
 - The performance of the initialization and running phase has been improved.
 - The following advanced configuration snippets for configuring replications have been added:
 - HDFS Replication Advanced Configuration Snippet (Safety Valve) for hadoop-env.sh
 - Hive Replication Advanced Configuration Snippet (Safety Valve) for hive-site.xml
 - HDFS Replication Advanced Configuration Snippet (Safety Valve) for yarn-site.xml
 - HDFS Replication Advanced Configuration Snippet (Safety Valve) for mapred-site.xml
 - Snapshot properties for HBase such as thread pool size can be configured in the **HBase Client Advanced Configuration Snippet (Safety Valve) for hbase-site.xml** property.
 - Hive partitions are chunked during export and import to avoid message size limitations.
 - Hive replications validate metadata on the destination Hive Metastore before copying HDFS data from the source to avoid copying errors during replication.
 - The use of snapshots to improve replications is documented.
 - The effect of network latency on replications is documented.
 - Scheduled snapshots can be disabled and re-enabled.
 - API improvements:
 - Explicit support for pausing snapshot policies
 - Failed file listing
 - Collection of diagnostic bundles for replication schedules and history

What's New in Cloudera Manager 5.4.10

A number of issues have been fixed. See [Issues Fixed in Cloudera Manager 5.4.10](#) on page 424.

What's New in Cloudera Manager 5.4.9

A number of issues have been fixed. See [Issues Fixed in Cloudera Manager 5.4.9](#) on page 425.

What's New in Cloudera Manager 5.4.8

New ability to decommission hosts with stopped services

Adds ability to decommission roles when services are completely stopped. This allows users to decommission hosts during cluster downtime.

A number of issues have also been fixed. See [Issues Fixed in Cloudera Manager 5.4.8](#) on page 426.

What's New in Cloudera Manager 5.4.7

New service-level advanced configuration snippets for Solr

The following new properties were added:

Cloudera Manager 5 Release Notes

- Solr Service Advanced Configuration Snippet (Safety Valve) for `core-site.xml`
- Solr Service Advanced Configuration Snippet (Safety Valve) for `hdfs-site.xml`

A number of issues have also been fixed. See [Issues Fixed in Cloudera Manager 5.4.7](#) on page 426.

What's New in Cloudera Manager 5.4.6

An issue has been fixed. See [Issues Fixed in Cloudera Manager 5.4.6](#) on page 427.



Note: Although there is a Cloudera Manager 5.4.6 release, there is no synchronous CDH 5.4.6 release.

What's New in Cloudera Manager 5.4.5

A number of issues have been fixed. See [Issues Fixed in Cloudera Manager 5.4.5](#) on page 427.



Note: Although there is a CDH 5.4.4 release, there is no synchronous Cloudera Manager 5.4.4 release.

What's New in Cloudera Manager 5.4.3

Rollback for CDH 4 to CDH 5 Upgrades

- Rolling back a CDH 4 to CDH 5 upgrade is now supported using Cloudera Manager.

A number of issues have been fixed. See [Issues Fixed in Cloudera Manager 5.4.3](#) on page 430.



Note: Although there is a CDH 5.4.2 release, there is no synchronous Cloudera Manager 5.4.2 release.

What's New in Cloudera Manager 5.4.1

Hue HA Improvements

- The Cloudera Manager Express and Add Service wizards allow you to add a Hue service with multiple Hue Server roles. For Kerberized clusters, the Add Service wizard automatically adds a colocated Kerberos Ticket Renewer role for each Hue Server role instance.
- When Kerberos is enabled, Cloudera Manager now checks to ensure each Hue Server role is colocated with a Kerberos Ticket Renewer role. If you forget to add a Kerberos Ticket Renewer role when adding a new Hue Server role, a configuration error is generated.

Cloudera Manager High Availability

- High availability for Cloudera Manager is now supported for 5.4.

A number of issues have also been fixed. See [Issues Fixed in Cloudera Manager 5.4.1](#) on page 434.

What's New in Cloudera Manager 5.4.0

- **OS** - Added support for RHEL 6.6 and CentOS 6.6.
- Cloudera Manager prevents installing or upgrading to a CDH version that is too new for the Cloudera Manager version. When using parcels, it prevents parcel installation. When using packages, it prevents creating services.
- Installation and add service wizards now support the Oozie database.
- New wizard for NameNode, Failover Controller, and JournalNode role migration.
- Parcel page layout redesigned in terms of layout, performance and ease of use. A new parcel per host detail view is added.
- **Configuration**

- Configuration pages use the new layout by default. The new layout is dramatically improved in terms of layout, performance, and ease of use. The existing layout is accessible via the **Switch to the classic layout** link.
- New configuration actions:
 - Configuration can now be applied to all clusters as well as for a specific cluster.
 - Several new configuration views have been added to show all non-default values across all clusters and the Cloudera Management Service, as well as differences across all clusters and multiple services of the same type.
 - One-click differences in configuration settings for a specific service across multiple clusters.

- **Support**

- Include a Cloudera support ticket with YARN application support bundles.
- Reduce the size of support bundles by specifying log data of interest to include in the bundle.

- **HDFS**

- Support for HDFS DataNode hot swap.
- Option to include replication of extended attributes during HDFS replication. HDFS ACLs will now be replicated along with permissions.

- Added support for Hive on Spark.



Important: Hive on Spark is included in CDH 5.4 and higher but is not currently supported nor recommended for production use. To try this feature, use it in a test environment until Cloudera resolves currently existing issues and limitations to make it ready for production use.

- **Security**

- Secure impersonation support for the Hue HBase app.
- Redaction of sensitive data in log files and in SQL query history.
- Support for custom Kerberos principals.
- Added commands for regenerating Kerberos keytabs at service and host levels. These commands will clear existing keytabs from affected role instances and then trigger the **Generate Credentials** command to create new keytabs.
- Kerberos support for Sqoop 2.
- Kerberos and TLS/SSL support for Flume Thrift Source and Sink.
- Solr TLS/SSL support.
- Navigator Key Trustee Server can be installed and monitored by Cloudera Manager.
- HBase Indexer integration with Sentry (File-based) for authorization.

What's New in Cloudera Manager 5.3.10

A number of issues have been fixed. See [Issues Fixed in Cloudera Manager 5.3.10](#) on page 437.

What's New in Cloudera Manager 5.3.9

A number of issues have been fixed. See [Issues Fixed in Cloudera Manager 5.3.9](#) on page 437.

What's New in Cloudera Manager 5.3.8

A number of issues have been fixed. See [Issues Fixed in Cloudera Manager 5.3.8](#) on page 438.

What's New In Cloudera Manager 5.3.7

An issue has been fixed. See [Issues Fixed in Cloudera Manager 5.3.7](#) on page 438.



Note: Although there is a Cloudera Manager 5.3.7 release, there is no synchronous CDH 5.3.7 release.

What's New in Cloudera Manager 5.3.6

A number of issues have been fixed. See [Issues Fixed in Cloudera Manager 5.3.6](#) on page 438.



Note: Although there is a CDH 5.3.5 release, there is no synchronous Cloudera Manager 5.3.5 release.

What's New in Cloudera Manager 5.3.4

A number of issues have been fixed. See [Issues Fixed in Cloudera Manager 5.3.4](#) on page 439.

What's New in Cloudera Manager 5.3.3

A number of issues have been fixed. See [Issues Fixed in Cloudera Manager 5.3.3](#) on page 439.

What's New in Cloudera Manager 5.3.2

A number of issues have been fixed. See [Issues Fixed in Cloudera Manager 5.3.2](#) on page 440.

What's New in Cloudera Manager 5.3.1

A number of issues have been fixed. See [Issues Fixed in Cloudera Manager 5.3.1](#) on page 441.

What's New in Cloudera Manager 5.3.0

- **JDK 1.8** - Cloudera Manager adds support for Oracle JDK 1.8.
- **Single user mode** - The Cloudera Manager Agent and all service processes can now be run as a single configured user in environments where running as root is not permitted.
- **CDH upgrade wizard enhanced** - The CDH upgrade wizard now supports minor and maintenance version upgrade as well as major version upgrade.
- **Oozie Sharelib** - The Oozie Sharelib can be updated without restarting the Oozie service.
- **Read-only users prevented from viewing process logs or environment** - Read-only users can no longer view the environment or logs of a process. This is to prevent read-only users from seeing potentially sensitive information.
- New icons for the KMS and Key Trustee services.
- **Data-at-rest encryption**



Important: Cloudera provides two solutions:

- **Navigator Encrypt** is production ready and available to Cloudera customers licensed for Cloudera Navigator. Navigator Encrypt operates at the Linux volume level, so it can encrypt cluster data inside and outside HDFS. Consult your Cloudera account team for more information.
- **HDFS Encryption** is production ready and operates at the HDFS directory level, enabling encryption to be applied only to HDFS folders where needed.

HDFS encryption implements transparent, end-to-end encryption of data read from and written to HDFS by creating encryption zones. An encryption zone is a directory in HDFS with every file and subdirectory in it encrypted. Use one of the following services to store, manage, and access encryption zone keys:

- **KMS (File)** - The Hadoop Key Management Server with a file-based Java keystore; maintains a single copy of keys, using simple password-based protection.
- **KMS (Navigator Key Trustee)** - An enterprise-grade key management service that replaces the file-based Java keystore and leverages the advanced key-management capabilities of Cloudera Navigator Key Trustee.

Navigator Key Trustee is designed for secure, authenticated administration and cryptographically strong storage of keys on multiple redundant servers that can be located outside the cluster.

- The Cloudera Manager Server now reports the correct number of physical cores and hyper-threading cores if hyper-threading is enabled.
- Client configurations** - Client configurations are now managed so that they are redeployed when a machine is re-imaged.



Important: The changes to client configurations affect some API calls, as follows:

- When a host ceases to have a client configuration assigned to it, Cloudera Manager will remove it, rather than leaving it behind. If a host has a client configuration assigned and the client configuration is missing, Cloudera Manager will recreate it.
- If you currently use the API command `deployClientConfig` to deploy the client configurations for a particular service, and you pass a specific set of role names to this call to narrow the set of hosts that receive the new client configuration, then you should be aware that:
 - The API command will continue to generate and deploy the client configuration only to the hosts that correspond to the specified role names.
 - Any other hosts that previously had deployed client configurations, but do not have gateway roles assigned to them, will have those client configurations removed from them. This is the new behavior.
- The behavior of the cluster level `deployClientConfig` command, and calling the service level command with no arguments, is unchanged. The command still deploys a new client configuration to all hosts with roles corresponding to the specified service or cluster.
- As this change is due to internal functional changes inside CM, it is not restricted to any new API level. The `deployClientConfig` command in all API levels is affected.

- Configuration**

- NameNode configuration** - The decommissioning parameters `dfs.namenode.replication.max-streams` and `dfs.namenode.replication.max-streams-hard-limit` are now available.
- Hue debug options** - Two service-level configuration parameters have been added to the Hue service to enable Django debug mode and debugging of internal server error responses.

What's New in Cloudera Manager 5.2.7

An issue has been fixed. See [Issues Fixed in Cloudera Manager 5.2.7](#) on page 444.



Note: Although there is a Cloudera Manager 5.2.7 release, there is no synchronous CDH 5.2.7 release.

What's New in Cloudera Manager 5.2.6

A number of issues have been fixed, see [Issues Fixed in Cloudera Manager 5.2.6](#) on page 444.

What's New in Cloudera Manager 5.2.5

A number of issues have been fixed, see [Issues Fixed in Cloudera Manager 5.2.5](#) on page 445.

What's New in Cloudera Manager 5.2.4

There are no changes for Cloudera Manager 5.2.4. It was released to provide the Cloudera Navigator fix in [What's New in Cloudera Navigator 2.1.4](#) on page 465.



Note: Although there is a CDH 5.2.3 release, there is no synchronous Cloudera Manager 5.2.3 release.

What's New in Cloudera Manager 5.2.2

- **HDFS Decommissioning** - The following decommissioning properties have been exposed in Cloudera Manager 5.2.2.
 - **Maximum number of replication threads on a Datanode** (`dfs.namenode.replication.max-streams`)
 - **Hard limit on the number of replication threads on a Datanode** (`dfs.namenode.replication.max-streams-hard-limit`)
- New icons for the KMS and Key Trustee services.

What's New in Cloudera Manager 5.2.1

This release fixes the “POODLE” vulnerability and a number of other issues. See [Issues Fixed in Cloudera Manager 5.2.1](#) on page 446.

- The YARN `yarn.nodemanager.recovery.dir` property can be configured.
- A health check indicates whether the HDFS metadata upgrade has not been finalized.

What's New in Cloudera Manager 5.2.0

- **OS and database support** - Adds support for Ubuntu Trusty (version 14.04) and PostgreSQL 9.3.
- **Services** - the following new services have been added:
 - **Isilon** - supports the EMC Isilon distributed filesystem.
 - **KMS** - the Java keystore-based key management server.
 - **Key Trustee** - the enterprise-grade key management server using Cloudera Navigator Key Trustee.
 - **Spark** - running Spark applications on YARN. The existing Spark service has been renamed Spark (Standalone).
- **Accumulo** - Kerberos authentication is now supported. If you have been using advanced configuration snippets (safety valves) to configure Kerberos with Accumulo, you may now remove those settings and have Cloudera Manager generate the principal and keytab file for you.
- **HDFS Data at Rest Encryption** -



Note: Cloudera provides the following two solutions for data at rest encryption:

- **Navigator Encrypt** - Is production ready and available for Cloudera customers licensed for Cloudera Navigator. Navigator Encrypt operates at the Linux volume level, so it can encrypt cluster data inside and outside HDFS. Talk to your Cloudera account team for more information about this capability.
- **HDFS Encryption** - Included in CDH 5.2.0 and operates at the HDFS folder level, enabling encryption to be applied only to HDFS folders where needed. This feature has several known limitations. Therefore, Cloudera does not currently support this feature in CDH 5.2 and it is not recommended for production use. To try the feature, upgrade to the latest version of CDH 5.

HDFS now implements transparent, end-to-end encryption of data read from and written to HDFS by creating encryption zones. An encryption zone is a directory in HDFS with all of its contents, that is, every file and subdirectory in it, encrypted. You can use either the **KMS** or the **Key Trustee** service to store, manage, and access encryption zone keys.

- **HBase** - Support for configuring hedged reads has been added for HBase. The default configuration is to turn hedged reads off. Cloudera Manager will emit two properties, `dfs.client.hedged.read.threadpool.size` (default: 0) and `dfs.client.hedged.read.threshold.millis` (default: 500ms) to `hbase-site.xml`.

- **ZooKeeper** - the RMI port can be configured. The port is configured using the JDK7 flag `-Dcom.sun.management.jmxremote.rmi.port`. The default value is set to be same as the JMX Agent port. Also, a special value of 0 or -1 disables the setting and a random port is used. The configuration has no effect on versions lower than Oracle JDK 7u4.
- **Cloudera Manager Agent configuration**
 - The supervisord port can now be configured in the Agent configuration `supervisord_port`. The change takes effect the next time supervisord is restarted (not simply when the Agent is restarted).
 - Added an Agent configuration `local_filesystem_whitelist` that allows configuring the list of local filesystems that should always be monitored.
- **Proxy user configuration**
 - All services' proxy user configuration properties have been moved to the HDFS service. Other services running on the cluster inherit the configuration values provided in HDFS. If you have previously configured a service to have values different from those configured in HDFS, then the proxy user configuration properties will be moved to that service's Advanced Configuration Snippet (Safety Valve) for `core-site.xml` to retain existing behavior.
 - Oozie and Solr are exceptions to this. Oozie proxy user configuration properties have been moved to **Oozie Server Advanced Configuration Snippet (Safety Valve)** for `oozie-site.xml` if they differ from HDFS. Solr proxy user configuration properties have been moved to **Solr Service Environment Advanced Configuration Snippet (Safety Valve)** if they differ from HDFS.
- **Resource management** - YARN and Llama integrated resource management and Llama high availability wizard.
- **New and changed user roles** - BDR Administrator, Cluster Administrator, Navigator Administrator, and User Administrator. The Administrator role has been renamed Full Administrator.
- **Configuration UI**
 - Cluster-wide configuration - you can view all modified settings and configure log directories, disk space thresholds, and port settings.
 - New configuration layout - the new layout provides an alternate way to view configuration pages. In the **classic** layout, pages are organized by role group and categories within the role groups. The **new** layout allows you to filter on configuration status, category, and scope. On each configuration page you can easily switch between the classic and new layout.



Important: The classic layout is the default. All the configuration procedures described in the Cloudera Manager documentation assume the classic layout.

What's New in Cloudera Manager 5.1.6

A number of issues have been fixed. See [Issues Fixed in Cloudera Manager 5.1.6](#) on page 448.

What's New in Cloudera Manager 5.1.5

A number of issue have been fixed. See [Fixed Issues in Cloudera Manager 5.1.5](#) on page 448.

What's New in Cloudera Manager 5.1.4

A number of issues have been fixed. See [Fixed Issues in Cloudera Manager 5.1.4](#) on page 448.

What's New in Cloudera Manager 5.1.3

A number of issues have been fixed. See [Fixed Issues in Cloudera Manager 5.1.3](#).

- **JDK Installation**

- Users who are adding or upgrading hosts can now choose not to install the JDK that ships with Cloudera Manager.

Cloudera Manager 5 Release Notes

What's New in Cloudera Manager 5.1.2

A number of issues have been fixed. See [Fixed Issues in Cloudera Manager 5.1.2](#).

- **New SAML configuration option**

- You can now specify the binding protocol to be used for AuthNResponses sent from the IDP to Cloudera Manager. Previously, Cloudera Manager would only use HTTP-Artifact, but it is now possible to choose HTTP-Post. HTTP-Artifact remains the default binding.

What's New in Cloudera Manager 5.1.1

An issue has been fixed. See [Issues Fixed in Cloudera Manager 5.1.1](#) on page 450.

What's New in Cloudera Manager 5.1.0



Important: Cloudera Manager 5.1.0 is no longer available for download from the Cloudera website or from archive.cloudera.com due to the JCE policy file issue described in the [Fixed Issues in Cloudera 5.1.1](#) section of the Release Notes. The download URL at archive.cloudera.com for Cloudera Manager 5.1.0 now forwards to Cloudera Manager 5.1.1 for the RPM-based distributions for Linux RHEL and SLES.

- **SSL Encryption**

- Supports several new SSL-related configuration parameters for HDFS, MapReduce, YARN and HBase, which allow you to configure and enable encrypted shuffle and encrypted web UIs for these services.
- Cloudera Manager now also supports the monitoring of HDFS, MapReduce, YARN, and HBase when SSL is enabled for these services. New configuration parameters allow you to specify the location and password of the truststore used to verify certificates in HTTPS communication with CDH services and the Cloudera Manager Server.

- **Sentry Service**

- A new Sentry service that stores the authorization metadata in an underlying relational database and allows you to use Grant/Revoke statements to modify privileges.
- You can also configure the Sentry service to allow Pig, MapReduce, and WebHCat queries access to Sentry-secured data stored in Hive.

- **Kerberos Authentication**

- Now supports a Kerberos cluster using an Active Directory KDC.
- New wizard to enable Kerberos on an existing cluster. The wizard works with both MIT KDC and Active Directory KDC.
- Ability to configure and deploy Kerberos client configuration (`krb5.conf`) on a cluster.

- **Spark Service** - added the History Server role

- **Impala** - added support for Llama ApplicationMaster High Availability

- **User Roles** - there are two new roles: Operator and Configurator that support fine-grained access to Cloudera Manager features.

- **Monitoring**

- Updates to Oozie monitoring
- New Hive metastore canary

- **UI** - The UI has been updated to improve scalability. The **Home > Status** tab can be configured to display clusters in a full or summary format. There is a new Cluster page for each cluster. The Hosts and Instances pages have added faceted filters.

What's New In Cloudera Manager 5.0.7

A number of issues have been fixed. See [Fixed Issues in Cloudera Manager 5.0.7](#) on page 453.

What's New in Cloudera Manager 5.0.6

A number of issues have been fixed. See [Fixed Issues in Cloudera Manager 5.0.6](#) on page 453.

What's New in Cloudera Manager 5.0.5

A number of issues have been fixed. See [Fixed Issues in Cloudera Manager 5.0.5](#) on page 453.

What's New in Cloudera Manager 5.0.2

A number of issues have been fixed. See [Issues Fixed in Cloudera Manager 5.0.2](#) on page 454.

What's New in Cloudera Manager 5.0.1

A number of issues have been fixed. See [Issues Fixed in Cloudera Manager 5.0.1](#) on page 454.

- **Monitoring**

- The Java Garbage Collection Duration health test for the Service Monitor, Host Monitor, and Activity Monitor has been replaced with the new Java Pause Duration health test.

What's New in Cloudera Manager 5.0.0

- **Service and Configuration Management**
 - HDFS - cache management
- **Resource Management** - Impala admission control
- **Monitoring**
 - Host disks overview
 - Impala best practices
 - HBase table statistics
 - HDFS cache statistics

What's New in Cloudera Manager 5.0.0 Beta 2

- **Service and Configuration Management**
 - HDFS
 - HDFS NFS Gateway role
 - Supports restoration of HDFS data from a snapshot
 - YARN
 - YARN Resource Manager High Availability
 - Resource pool scheduler
 - Support for Spark service
 - Support for Accumulo service
 - Support for service extensibility
 - Support to set up Oozie server High Availability
 - Granular configuration staleness UI
 - Support for setting maximum file descriptors
- **Monitoring**
 - Support for monitoring the Cloudera Search/Solr service
 - New "failed" and "killed" badges displayed for unsuccessful YARN applications
 - More attributes available for filtering displays of YARN applications and Impala queries
 - New operational reports added for HBase tables and namespaces, Impala queries, and YARN applications
 - Support for creating user-defined triggers for metrics accessible via charts/tsquery



Important: Because triggers are a new and evolving feature, backward compatibility between releases is not guaranteed at this time.

- Charting improvements
 - New table chart type
 - New options for displaying data and metadata from charts
 - Support for exporting data from charts to CSV or JSON files
- **Administrative Settings**
 - Added a new role type with limited administrator capabilities.
 - Cloudera Manager Server and all JVMs will create a heap dump if they run out of memory.
 - Configure the location of the parcel directory and specify whether and when to remove old parcels from cluster hosts.

What's New in Cloudera Manager 5.0.0 Beta 1

- **CDH Version**
 - Supports both CDH 4 and CDH 5
 - CDH 4 to CDH 5 upgrade wizard
 - Support for YARN as a production execution environment
 - MapReduce (MRv1) to YARN (MRv2) configuration import
 - YARN-based resource management for Impala 1.2
- **JDK Version** - Cloudera Manager 5 supports and installs both JDK 6 and JDK 7.
- **Resource Management**
 - Static and dynamic partitioning of resources: provides a wizard for configuring static partitioning of resources (cgroups) across core services (HBase, HDFS, MapReduce, Solr, YARN) and dynamic allocation of resources for YARN and Impala.
 - Pool, resource group, and queue administration for YARN and Impala.
 - Usage monitoring and trending.
- **Monitoring**
 - YARN service monitoring
 - YARN (MRv2) job monitoring
 - Configurable histograms of Impala query and YARN job attributes that can be used to quickly filter query and application lists
 - Scalable back-end database for monitoring metrics
 - Charting improvements
 - New chart types: histogram and heatmap
 - New scale types: logarithmic and power
 - Updates to tsquery language: new attribute values to support YARN and new functions to support new chart types
- **Extensibility**
 - Ability to manage both ISV applications and non-CDH services (for example, Accumulo, Spark, and so on)
 - Working with select ISVs as part of Beta 1
- **Single Sign-On** - Support for SAML to enable single sign-on
- **Parcels**

- Dependency enforcement to ensure incompatible parcels are not used together
- Option to not cache downloaded parcels, to save disk space
- Improved error reporting for management operations
- **Backup and Disaster Recovery (BDR)**
 - HBase and HDFS snapshots: Supports scheduling snapshots on a recurring basis.
 - Support for YARN (MRv2): Replication jobs can now run using YARN (MRv2) instead of MRv1.
 - Global replication page: All scheduled snapshots (HDFS and HBase) and replication jobs for either HDFS or Hive are shown on a single Replications page.
- **Other**
 - Global Search box
 - Several usability improvements
 - Comprehensive detection of configuration changes that require service restarts, refresh and redeployment of client configurations

Incompatible Changes in Cloudera Manager 5

The following sections describe incompatible changes in each Cloudera Manager 5 release.

Incompatible Changes Introduced in Cloudera Manager 5.5.0

- Cloudera Manager no longer supports JDK 1.6.

Incompatible Changes Introduced in Cloudera Manager 5.4.0

- The Blacklisted Products property has been removed from the Hosts > Parcels configuration.

Incompatible Changes Introduced in Cloudera Manager 5.3.0

- Oozie metrics - The Oozie metrics framework is now controlled by the **Enable The Metrics Instrumentation Service** flag, which is enabled by default. When enabled, the old 'instrumentation' REST end-point is disabled and metrics are available on the new 'metrics' REST end-point (*hostname:port/v2/admin/metrics*).

Incompatible Changes Introduced in Cloudera Manager 5.2.0

- Due to various internal changes to configuration generation, all service and client configurations will be stale after upgrade. To propagate the updates, restart the cluster and redeploy client configurations.

Incompatible Changes Introduced in Cloudera Manager 5.1.0

- The Limited Administrator role has been renamed Limited Operator. The Limited Operator role is no longer available in Cloudera Manager Express. If you upgrade a Cloudera Manager Express installation, users in the Limited Operator role will not be able to log in. A user in the Administrator role must assign the Read-Only or Administrator role to those users.

Incompatible Changes Introduced in Cloudera Manager 5.0.0

• Cloudera Manager API

- New [upgradeCdh](#) command, which upgrades CDH cluster versions. Use this command to upgrade clusters from CDH 4 to CDH 5. The `upgradeServices` command previously used to upgrade CDH cluster versions is no longer supported.
- The `hostId` field now contains a unique UUID and no longer matches the `hostName` field. When referring to a host, both `hostId` and `hostName` are accepted. However, any API clients that were previously cross-referencing host records with external information by `hostName`, but were using the `hostId` field in the API, must be updated to use the `hostName` field. Clients updated in this manner will function correctly with older versions of Cloudera Manager because the `hostName` field has always been present.

Cloudera Manager 5 Release Notes

- The `clusterName` field displayed when viewing service and role references is now an internal name and may not match the external `displayName` field of the cluster.
- CDH 5 Hue works only with the default system Python version of the operating system it is being installed on. For example, on RHEL/CentOS 6, you need Python 2.6 to start Hue.
- Cloudera Manager 5.0 includes a change to the value of the `snmpTrapOID`. Earlier releases set the value of `snmpTrapOID` (OID: .1.3.6.1.6.3.1.4.1.0) wrongly to `clouderaManagerMIBNotifications` (OID .1.3.6.1.4.1.38374.1.1.1). This is fixed in Cloudera Manager 5.0 with the correct value, which is `clouderaManagerAlert` (OID .1.3.6.1.4.1.38374.1.1.1). This change will break SNMP server setups that are configured to expect `clouderaManagerMIBNotifications`. Cloudera Manager administrators should configure their SNMP receivers to accept the corrected OID.
- The default values for the following configurations have changed to include the JVM option `-Djava.net.preferIPv4Stack=true`, which sets the preferred protocol stack to IPv4 on dual-stack machines. Any values set to the old defaults will automatically be changed to the new default when upgrading to Cloudera Manager 5.
 - MapReduce client configuration:
 - `hadoop-env.sh`: added to `HADOOP_CLIENT_OPTS`
 - `mapred-site.xml`: added to `mapred.child.java.opts`
 - YARN client configuration:
 - `hadoop-env.sh`: added to `YARN_OPTS`
 - `mapred-site.xml`: added to `yarn.app.mapreduce.am.command-opts`, `mapreduce.map.java.opts`, and `mapreduce.reduce.java.opts`
 - HDFS client configuration: `hadoop-env.sh`: added to `HADOOP_CLIENT_OPTS`
 - Hive client configuration: `hive-env.sh`: added to `HADOOP_CLIENT_OPTS`
- MapReduce health tests have been removed:
 - Job failure
 - Map backlog
 - Reduce backlog
 - Map locality

If needed, the test can be replaced with a trigger. For example:

- Looks at all the jobs that completed in the last hour and if there are more than 10% of failed jobs, change the health of the service to concerning:

```
IF (select (jobs_failed_rate * 3600) as jobs_failed,
((jobs_failed_rate + jobs_completed_rate + jobs_killed_rate) * 3600)
as all_jobs where roleType=JOBTRACKER AND serviceName=$SERVICENAME
and last(jobs_failed_rate / (jobs_failed_rate + jobs_completed_rate +
jobs_killed_rate)) >= 10 ending at $END_TIME duration "PT3600S")
DO health:concerning
```

- If there are more than 50% maps waiting than total slots available, health goes concerning.

```
IF (select waiting_maps / map_slots where roleType=JOBTRACKER and serviceName=$SERVICENAME
and last(waiting_maps / map_slots) > 50)
DO health:concerning
```

- If there are more than 50% reduce waiting than total slots available, health goes concerning.

```
IF (select waiting_reduces / reduce_slots where roleType=JOBTRACKER and
serviceName=$SERVICENAME
and last(waiting_reduces / reduce_slots) > 50)
DO health:concerning
```

- HDFS checkpointing metrics have been removed:

- end_checkpoint_num_ops
- end_checkpoint_avg_time
- start_checkpoint_num_ops
- start_checkpoint_avg_time

Incompatible Changes Introduced in Cloudera Manager 5.0.0 Beta 2

- Impala releases earlier than 1.2.1 are no longer supported.
- Some of the constants identifying health tests have changed. The following existed in Cloudera Manager 4:
 - FAILOVERCONTROLLER_FILE_DESCRIPTOR
 - FAILOVERCONTROLLER_HOST_HEALTH
 - FAILOVERCONTROLLER_LOG_DIRECTORY_FREE_SPACE
 - FAILOVERCONTROLLER_SCM_HEALTH
 - FAILOVERCONTROLLER_UNEXPECTED_EXITS

They are now:

- MAPREDUCE_FAILOVERCONTROLLER_FILE_DESCRIPTOR
- MAPREDUCE_FAILOVERCONTROLLER_HOST_HEALTH
- MAPREDUCE_FAILOVERCONTROLLER_LOG_DIRECTORY_FREE_SPACE
- MAPREDUCE_FAILOVERCONTROLLER_SCM_HEALTH
- MAPREDUCE_FAILOVERCONTROLLER_UNEXPECTED_EXITS

and

- HDFS_FAILOVERCONTROLLER_FILE_DESCRIPTOR
- HDFS_FAILOVERCONTROLLER_HOST_HEALTH
- HDFS_FAILOVERCONTROLLER_LOG_DIRECTORY_FREE_SPACE
- HDFS_FAILOVERCONTROLLER_SCM_HEALTH
- HDFS_FAILOVERCONTROLLER_UNEXPECTED_EXITS

The reason for the change is to better distinguish between MapReduce and HDFS failover controller monitoring in the health system.

Incompatible Changes Introduced in Cloudera Manager 5.0.0 Beta 1

- Services

- **Impala** - With Cloudera Manager 4.8 (released in late November 2013), only Impala 1.2.1 is supported, due to the introduction of the Impala Catalog Server. However, CDH 5.0.0 Beta 1 was released with Impala 1.2.0 (Beta). Therefore, if you upgrade from Cloudera Manager 4.8 (with Impala 1.2.1) to Cloudera Manager 5.0.0 Beta 1, and then upgrade your CDH to CDH 5.0.0 Beta 1, your version of Impala will be downgraded to Impala 1.2.0 from 1.2.1. This will result in some loss of functionality. See [New Features in Impala](#) for a list of the new features in Impala 1.2.1 that are not in Impala 1.2.0 (Beta).
- **Hive** - HiveServer 2 is a mandatory role for Hive in CDH 5.
- **Hue** - In CDH 5, Hue no longer has a Beeswax Server role. Hue now submits queries to HiveServer2.
- **HDFS** - Cloudera Manager 5 does not support NFS-mounted shared edits directories for HDFS High Availability. It only supports the Quorum Journal method for shared edits. If you upgrade from Cloudera Manager 4 with a working CDH 4 High Availability configuration that uses NFS-mounted directories, your installation will continue to work until you disable High Availability. You will not be able to re-enable High Availability with NFS-mounted directories. Furthermore, you will not be able to upgrade to CDH 5 unless you disable High Availability, and you will need to use Quorum-based storage in order to re-enable High Availability after the upgrade.

- **YARN**

- The YARN (MRv2) configuration `mapreduce.job.userlog.retain.hours` has been replaced by `yarn.log-aggregation.retain-seconds`. Any existing value in

Cloudera Manager 5 Release Notes

- `mapreduce.job.userlog.retain.hours` will be lost. However, this configuration never had any effect, so no functionality is affected.
- The following configuration parameters were removed from YARN. These never had any effect, so no functionality is affected.

 - `mapreduce.jobtracker.maxtasks.perjob`
 - `mapreduce.jobtracker.handler.count` (non-functional duplicate of `yarn.resourcemanager.resource-tracker.client.thread-count`)
 - `mapreduce.jobtracker.persist.jobstatus.active`
 - `mapreduce.jobtracker.persist.jobstatus.hours`
 - `mapreduce.job.jvm.numtasks`

- The following YARN configuration parameters were replaced. Only the YARN parameters were replaced. Old configurations will be lost, but they never had any effect so this does not affect functionality.

 - `mapreduce.jobtracker.restart.recover` replaced by `yarn.resourcemanager.recovery.enabled` (changed from Gateway to ResourceManager)
 - `mapreduce.tasktracker.http.threads` replaced by `mapreduce.shuffle.max.connections`
 - `mapreduce.jobtracker.staging.root.dir` replaced by `yarn.app.mapreduce.am.staging-dir`

- Cloudera Manager 5 sets the default YARN Resource Scheduler to FairScheduler. If a cluster was previously running YARN with the FIFO scheduler, it will be changed to FairScheduler the next time YARN restarts. The FairScheduler is only supported with CDH 4.2.1 and later, and older clusters may hit failures and need to manually change the scheduler to FIFO or CapacityScheduler. See the Known Issues section of this Release Note for information on how to change the scheduler back to FIFO or CapacityScheduler.

Changed Features and Behaviors in Cloudera Manager 5

The following sections describe what's changed in each Cloudera Manager 5 release.



Note: Rolling upgrade is not supported between CDH 4 and CDH 5. Rolling upgrade will also *not* be supported from CDH 5.0.0 Beta 2 to any later releases, and may not be supported between any future beta versions of CDH 5 and the General Availability release of CDH 5.

What's Changed in Cloudera Manager 5.8.0

- The YARN service's list of **Allowed System Users** now includes the `hbase` user by default. The reason for this change is that several essential HBase tools such as the MOB Sweeper, Import/Export tools, and CopyTable, need to interact with HBase as the `hbase` user to be able to execute MapReduce jobs.

Note that this change is only applicable to new Cloudera Manager deployments. Upgrading to Cloudera Manager 5.8 will not add the `hbase` user to the list of defaults.

What's Changed in Cloudera Manager 5.7.0

- The Navigator Metadata Server requires 192 MiB of Java PermGen space instead of 128 MiB. This is increased automatically when upgrading to Cloudera Manager 5.7.
- The default value for `hive.compute.query.using.stats` is changed to false. The reason for the change is that certain queries such as `count`, `max`, and `min` return incorrect results with this optimization on.
- By default, Hive sessions now only consider sessions with no recent activity to be idle (`hive.server2.idle.session.timeout_check_operation`) and idle session timeouts have been reduced (`hive.server2.idle.session.timeout` and `hive.server2.idle.operation.timeout`). This helps reduce the strain on HiveServer2 from too many open sessions.

- Cloudera Manager no longer automatically refreshes scheduler configurations when dynamic resource pool settings are changed. You must explicitly refresh the configurations. This allows you to schedule the changes to minimize the impact on your cluster.
- For YARN, the default number of log directories (`yarn.nodemanager.log-dirs`) has changed from 1 to be equal to the number of mount points, to prevent applications with a large number of logs from filling up a single disk.
- The default for Java Heap Size of JournalNode in Bytes is now 512 MB.
- The **Sources** page for HDFS and Hive replications has been removed. A list of sources is available from a drop-down menu when you schedule a replication.
- The number of watched directories you can specify for the Disk Usage Report is now unlimited.
- Cloudera Manager now uses a new memory allocation algorithm to allocate memory when multiple roles are installed on the same host.
- User sessions in the Cloudera Manager Admin Console now timeout after a configurable period of time of inactivity. A dialog box warns the user before automatically logging out the user.
- The **All Recent Commands** page now loads more quickly.
- The **Disk Usage** reports now have links that take users to the **Directory Usage Report** with the correct filter applied.
- When searching for hosts on the **Hosts** page, you can now filter the hosts list by entering search terms (hostname, IP address, or role) in the search box separated by commas or spaces. You can use quotes for exact matches (for example, strings that contain spaces, such as a role name) and brackets to search for ranges. Hosts that match any of the search terms are displayed.
- Isilon is now supported as a source or destination service for HDFS replications.
- For CDH 5.7 and higher if `CDH_PYTHON` is set by a Spark plug-in, `PYSPARK_PYTHON` is set to `CDH_PYTHON` in `spark-env.sh`. If you install a Python runtime parcel, such as the [Anaconda](#) parcel, Python Spark jobs run in both YARN client and YARN cluster modes are automatically configured by redeploying the Spark client configuration.

What's Changed in Cloudera Manager 5.5.0

- Removed `-XX:-CMSConcurrentMTEnabled` from the default JVM options. This setting makes the JVM run in single threaded mode. This was needed for Java 1.6_31 and lower but not for Java 1.6_32 or higher. Anybody using Java 1.6_31 or lower should upgrade to the latest recommended version of Java 1.7.

This change causes all roles to be stale after you upgrade to Cloudera Manager 5.5 and they are indicated as requiring restart in the Cloudera Manager Admin Console. However, as with any upgrade, this is a valid, functional Cloudera Manager state, and the cluster only needs to be restarted when you want the new configurations to take effect.

- `HADOOP_USER_CLASSPATH_FIRST` is now fully respected in Hadoop client configurations. After you upgrade Cloudera Manager, services display a client configuration redeployment required icon .
- For RHEL 7, the `force_start`, `fast_*`, `clean_*`, and `hard_*` commands on the `server-scm-*` services no longer work, as custom start, restart, and stop commands are not supported on systemd based distributions. These have been replaced with `*_next_*` operations, which do not trigger an immediate operation, but signal that the next invoked operation will be forced, fast, clean, or hard.
- The Cloudera EULA is now shown when using the Cloudera Manager Admin Console for the first time.
- The Home tab has been removed from the Cloudera Manager Admin Console navigation bar. You can return to the Home page Status tab by clicking the Cloudera Manager logo.
- All the icons have been refreshed to make them cleaner and easier to read.
- In 5.4.0 an externally assigned role was combined with a Cloudera Manager assigned role and the user had the union of the role privileges. As a consequence, an external user could be assigned an administrator role in Cloudera Manager and they would be an administrator regardless of the externally assigned role. Now only the externally assigned roles are respected. No roles can be assigned to an external user in Cloudera Manager and any roles for an external user in the Cloudera Manager are ignored.

As a result of this change, external users with previously-assigned Cloudera Manager roles will have their permissions modified depending on the LDAP group they belong to. To restore permissions for external users, configure the LDAP groups for these users by navigating to **Administration > Settings**, and click **Category > External Authentication** to bring up the relevant properties.

Cloudera Manager 5 Release Notes

- Cloudera Manager and CDH components support either TLS 1.0, TLS 1.1, or TLS 1.2, but not SSL 3.0. References to SSL continue only because of its widespread use in technical jargon.
- The label on the **Generate Credentials** button has been changed to **Generate Missing Credentials** to better reflect the fact that it only creates Kerberos principals that are not present yet in Cloudera Manager.
- Cloudera Manager now downloads binaries from <https://archive.cloudera.com> instead of <https://archive.cloudera.com>.
- The embedded `hbck` feature has been removed from HBase monitoring for stability reasons.
- Increased the default heap sizes for Hive roles. On clusters with sufficient memory, newly created Hive roles have these values:
 - HiveServer2 - 4 G heap, 512 M perm gen
 - Hive Metastore - 8 G heap, 512 M perm gen
 - Gateway - 2 G heap, 512 M perm gen
- By default, Oozie now purges eligible completed workflows and coordinator actions for long-running coordinator jobs.
- Oozie actions that omit the `<job-tracker>` and `<name-node>` elements (and the workflow does not define them in the `<global>` section) use the default values for the JobTracker, Resource Manager, and NameNode from Cloudera Manager in CDH 5.5 and higher.
- Increased the defaults for Oozie parameters:
 - `oozie.service.CallableQueueService.callable.concurrency` - 10
 - `oozie.service.CallableQueueService.threads` - 50
- Sqoop 2 is no longer in the default services to be created in any of the options in the installation wizard. You can choose to add it to the Custom Services option in the Installation wizard or can add it with the Add Service wizard after installation.
- For CDH 5.5.0 and higher the default values of the YARN properties `mapreduce.[map|reduce].java.opts.max.heap` and `mapreduce.[map|reduce].memory.mb` have been changed to 0, which tells YARN to automatically select a default. This helps avoid issues where either heap or memory.mb is updated, but not the other one (memory.mb should be ~30% higher than heap to allow for JVM overhead).
- The Host DNS Resolution Duration health test was removed. Its functionality is now covered in the Host DNS Resolution health test.
- The default Replication Strategy is now **Dynamic**.

What's Changed in Cloudera Manager 5.4.1

HDFS Read Throughput Impala query monitoring property is misleading

The `hbase_bytes_read_per_second` and `hdfs_bytes_read_per_second` Impala query properties have been renamed to `hbase_scanner_average_bytes_read_per_second` and `hdfs_scanner_average_bytes_read_per_second` to more accurately reflect that these properties return the average throughput of the query's HBase and HDFS scanner threads respectively. The previous names and descriptions gave the impression that these properties were the query's total HBase and HDFS throughput, which was not accurate.

What's Changed in Cloudera Manager 5.4.0

- Cloudera Manager checks the specified version of CDH before an installation and upgrade to ensure that it is compatible with Cloudera Manager before proceeding. Specifically, for Cloudera Manager 5.4 that means no version of CDH newer than 5.4.x is supported (Cloudera Manager must be upgraded before upgrading to such a version of CDH). Cloudera Manager no longer shows these "too-new" versions of CDH. The 'latest' parcel repository URL will be replaced by the 'latest_supported' repository in the parcel configuration.
- The minimum Java heap size for the Activity Monitor, Host Monitor, and Service Monitor has been changed from 50 MB to 256 MB.
- Regenerating Kerberos principals will be denied if any roles that are using those principals are running. Stop those roles and then attempt to regenerate the principals.

- In previous versions of Cloudera Manager, the 'version' attribute in tsquery had values that were integers, for example, 4 for CDH4, 5 for CDH5, -1 for Cloudera Manager. Starting in the Cloudera Manager 5.4, the values for the 'version' attribute are in release string format, for example "cdh5.0.0".
- **Hive**
 - `hive.exec.reducers.max` default value changed from 999 to 1099
 - `hive.exec.reducers.bytes.per.reducer` default value changed from 1 GB to 64 MB
 - The default heap size for the Hive CLI is increased to 1 GB.
 - The property `hive.log.explain.output` is known to create instability of Cloudera Manager Agents in some specific circumstances, specially when the hive queries generate extremely large EXPLAIN outputs. Therefore, the property has been hidden from the Cloudera Manager configuration UI. The property can still be configured through the use of advanced configuration snippets.
- **Impala** - The Impala Daemon now supports the Impala Maximum Log Files property which specifies the total number of log files per severity level that should be retained before they are deleted. By default, after upgrading to CDH 5.4 this property is set to 10, which means that Impala Daemons will only retain up to 10 log files for each severity level. Any additional files will be deleted.
- **HBase** - Moved three settings for HBase coprocessors from Main to Advanced category:
 - Service Wide > HBase Coprocessor Abort on Error: move to 'Service Wide > Advanced > HBase Coprocessor Abort on Error'
 - 'Master Default Group > HBase Coprocessor Master Classes': move to 'Master Default Group > Advanced > HBase Coprocessor Master Classes'
 - RegionServer Default Group > HBase Coprocessor Region Classes': move to 'RegionServer Default Group > Advanced > HBase Coprocessor Region Classes'

What's Changed in Cloudera Manager 5.3.2

- Turning on the internal HBase canary (not to be confused with Cloudera Manager monitoring canary) is optional. On new clusters, it will not be enabled by default. Existing clusters will continue to run the canary until it is disabled from the HBase configuration page.

What's Changed in Cloudera Manager 5.3.0

- **Cloudera Manager upgrade** - If you have any active commands running before upgrade, the server *will fail to start* after upgrade. This includes commands a user might have run and also for commands Cloudera Manager automatically triggers, either in response to a state change, or something that's on a schedule.

What's Changed in Cloudera Manager 5.2.1

- The default value of the YARN `yarn.nodemanager.recovery.dir` property has changed from `{hadoop.tmp.dir}/yarn-nm-recovery` to `/var/lib/hadoop-yarn/yarn-nm-recovery`.

What's Changed in Cloudera Manager 5.2.0

- **Rolling upgrade** - As a result of a recent change in the way DataNodes handle block deletions during a rolling upgrade ([HDFS-5907](#)), the Trash directory may grow unexpectedly while the upgrade is in progress. Deleted blocks are kept during upgrade in case you want to roll back. The blocks are cleaned up after you finalize the upgrade.
- **Agent** -
 - The `hard_stop`, `hard_restart`, and `clean_restart` commands now show a warning message about the impact of using these commands instead of performing the actions. To actually perform the actions, you use the `hard_stop_confirmed`, `hard_restart_confirmed`, and `clean_restart_confirmed` commands.
 - The default supervisord port is changed from 9001 to 19001
- YARN application attributes renamed: `slot_millis` to `slots_millis` and `fallow_slot_millis` to `fallow_slots_millis`

Cloudera Manager 5 Release Notes

What's Changed in Cloudera Manager 5.1.0

- **UI refresh for scalability**
- Revised authorization privilege model in Sentry.

What's Changed in Cloudera Manager 5.0.0

- MapReduce now inherits topology from HDFS NameNode. Topology configuration for MapReduce JobTracker was removed. The configuration was redundant and the two parameters should always have been set to the same value.
- **UI**
 - The Clusters tab no longer has Activities, Other, and Manage Resources sections.

What's Changed in Cloudera Manager 5.0.0 Beta 2

- **Product**
 - Cloudera Backup and Disaster Recovery (BDR) is now included with Cloudera Enterprise.
 - Cloudera Standard has been renamed to Cloudera Express.
- **OS and packaging**
 - The name of the Cloudera Manager embedded database package has changed from `cloudera-manager-server-db` to `cloudera-manager-server-db-2`. For details, read the upgrade and install topics for your OS.
 - Support for Ubuntu 10.04 and Debian 6.0 is deprecated.
- **HDFS** - enabling High Availability automatically enables auto-failover, unlike in Cloudera Manager 4 where enable auto-failover was a separate command.
- **HBase**
 - In CDH 5 there is no HBase canary because HBase is now monitored by a watchdog process. In CDH 4, the HBase canary is still used.
 - The RegionServer default heap size has been increased to 4GB.
- **Monitoring**
 - Chart "Views" and actions related to views have been renamed to "Dashboard".
 - Changes to how attribute filters are displayed in the Impala queries and YARN applications screens
 - The outdated configuration indicator on the Home, service, and role pages has a new graphic and now has a tooltip that displays whether a cluster refresh or restart is required. There is a new indicator for changes that require redeploying client configurations. You can click an indicator to go to the new Stale Configurations page to view and resolve the conditions that gave rise to the indicator.
 - To match the naming convention of tsquery metrics, multiword Impala query and YARN application attribute names have changed from camel case to using an underscore separator. For example `queryType` has changed to `query_type`. For backward compatibility, camel case names are still supported.
- **UI**
 - The main navigation bar in Cloudera Manager Admin Console has been reorganized. The Services tab has been replaced by a Clusters tab that contains links to individual services, which were previously under the Services tab, Activities and Reports sections, which were removed from the main bar, and a new Manage Resources section, which contains links to the new resource pools and service pools features. The All Services page has been removed.
 - The "Safety Valve" properties have been renamed "Advanced Configuration Snippet".
 - The screen for specifying assignment of roles to hosts has been redesigned for improved scalability and usability.
- **Misc**

- The `io.compression.codecs` property has moved from MapReduce to HDFS.

What's Changed in Cloudera Manager 5.0.0 Beta 1

- When CDH 5 is installed, YARN is installed by default, rather than MapReduce, and is the default execution environment. MapReduce is deprecated in CDH 5 but is fully supported for backward compatibility through CDH 5. In CDH 4, MapReduce is still the default.
- The setting for `yarn.scheduler.maximum-allocation-mb` has been increased to a default of 64GB.
- The minimum heap size for the Solr service has been increased to 200MB (from 50MB previously) to enable it to better handle collection creation.

Known Issues and Workarounds in Cloudera Manager 5

The following sections describe the current known issues in Cloudera Manager 5.

Error when distributing parcels: No such torrent

Parcel distribution might fail with an error message conforming to:

```
Error when distributing to <host>: No such torrent: <parcel_name>.torrent
```

Workaround

Log in to the <host> and remove the following file:

```
/opt/cloudera/parcel-cache/<parcel_name>.torrent
```

Hive Replication Metadata Transfer Step fails with Temporary AWS Credential Provider

Hive Replication Schedules that use Amazon S3 as the **Source** or **Destination** fail when using temporary AWS credentials. The following error displays:

```
Message: Hive Replication Metadata Transfer Step Failed -  
com.cloudera.com.amazonaws.services.s3.model.AmazonS3Exception:  
Status Code: 403, AWS Service: Amazon S3, AWS Request ID: 76D1F6A02792908A,  
AWS Error Code: null, AWS Error Message: Forbidden,  
S3 Extended Request ID:  
Xy3nAS4HSPKLA6hHKvpqReBud7M1Fhk7On0HttYGE0eKPHKwiFkTPQxEVU820Zq5d8omSrdbhCI=.
```

Hive table Views do not get restored from S3

When creating a Hive Replication schedule that copies Hive data from S3 and you select the **Reference Data From Cloud** option, Hive table Views are not restored correctly and result in a Null Pointer Exception when querying data from the view.

ACLs are not replicated when restoring Hive data from S3

ACLs are never replicated when the **Enable Access Control Lists** option in the configuration of the HDFS service is not selected the first time a Replication Schedule that replicates from S3 to Hive runs. Enabling it and re-running the restore operation does not restore the ACLs.

Snapshot diff is not working for Hive to S3 replication when data is deleted on source

If you have enabled snapshots on an HDFS folder and a Hive table uses an external file in that folder, and then you replicate that data to S3 and delete the file on the source cluster, the file is not deleted in subsequent replications to S3, even if the **Delete Permanently** option is selected.

Block agents from heartbeating to a Cloudera Manager with different UUID until agent restart

If for some reason Cloudera Manager server's identity (Cloudera Manager guid) is changed, agents drop heartbeat requests and will not follow any requested commands from Cloudera Manager server. As a result, the agents report bad health. This situation can be fixed by taking either of the following approaches:

- Restore the previous Cloudera Manager server guid

OR

- Remove the cm_guid file from each of the agents and then restart the agent.

Cloudera Manager set catalogd default jvm memory to 4G can cause out of memory error on upgrade to Cloudera Manager 5.7 or higher

After upgrading to 5.7 or higher, you might see a reduced Java heap maximum on Impala Catalog Server due to a change in its default value. Upgrading from Cloudera Manager lower than 5.7 to Cloudera Manager 5.8.2 no longer causes any effective change in the Impala Catalog Server Java Heap size.

When upgrading from Cloudera Manager 5.7 or later to Cloudera Manager 5.8.2, if the Impala Catalog Server Java Heap Size is set at the default (4GB), it is automatically changed to either 1/4 of the physical RAM on that host, or 32GB, whichever is lower. This can result in a higher or a lower heap, which could cause additional resource contention or out of memory errors, respectively.

Cloudera Manager 5.7.4 installer does not show Key Trustee KMS

A fresh install of Cloudera Manager tries to install Key Trustee KMS 5.8.2 when trying to install the latest version. You must either choose 5.7.0 as the Key Trustee KMS version, or manually provide a link to the 5.7.4 bits.

Class Not Found Error when upgrading to Cloudera Manager 5.7.2

When you upgrade to version 5.7.2 of Cloudera Manager, the client configuration for all services is marked *stale*.

Workaround:

From the **Cluster** menu, select **Deploy Client Configuration** to redeploy the client configuration.

Kerberos setup fails on Debian 8.2

This issue is due to the following Debian bug: <https://bugs.debian.org/cgi-bin/bugreport.cgi?bug=777579;msg=5;att=0>.

Workaround:

1. Log in to the host where the Cloudera Manager server is running.
2. Edit the `systemd/system/krb5-admin-server.service` file and add `/etc/krb5kdc` to the `ReadWriteDirectories` section.
3. Run the following commands:

```
systemctl daemon-reload  
sudo service krb5-admin-server restart
```

4. Generate the credentials.

Password in Cloudera Manager's db.properties file is not redacted

The `db.properties` file is managed by customers and is populated manually when the Cloudera Manager Server database is being set up for the first time. Since this occurs before the Cloudera Manager Server has even started, encrypting the contents of this file is a completely different challenge as compared to that of redacting configuration files.

Releases affected: 5.3 and higher

Cluster provisioning fails

In some cases, provisioning of a cluster may fail at the start of the process. This does not happen in all cases and is mainly noticed on RHEL 6 and especially when some hosts are reporting bad health.

Releases affected: 5.5.0-5.5.3, 5.6.0-5.6.1, 5.7.0

Releases containing the fix: 5.5.4, 5.7.1

For releases containing the fix, parcel activation and first run command now completes as expected, even when some hosts report bad health.

This issue is fixed in Cloudera Manager 5.5.4 and 5.7.1 and higher.

Cloudera Manager can run out of memory if a remote repository URL is unreachable

If one of the URLs specified in on the **Parcel Settings** page (**Hosts > Parcels > Configuration**) becomes unreachable, Cloudera Manager may run out of memory.

Workaround:

Do one of the following:

- If the URL is incorrect, enter the correct URL.
- Deselect the **Automatically Download New Parcels** setting on the **Parcel Settings** page.
- Set the value of the **Parcel Update Frequency** on the **Parcel Settings** to a large interval such as several days.

Clients can run Hive on Spark jobs even if Hive dependency on Spark is not configured

In CDH 5.7 and higher, when Hive and Spark on YARN both are configured, but Hive is not configured to depend on Spark on YARN, clients can set the execution engine to `spark` and Hive on Spark jobs will still be executed but will run in an unsupported mode. These jobs may not appear in the Spark History Server.

Workaround: Configure Hive to depend on Spark on YARN.

Known Issues for the DSSD D5 Hadoop Plugin

Default Directory for Data Directory for a DSSD DataNode Should be Reconfigured



Note: This issue applies only to clusters that use the EMC DSSD D5 storage appliance as the storage for HDFS DataNodes.

The DSSD DataNodes operating in the Hadoop cluster store version and ID information on a local disk directory. The default directory is `/tmp/hadoop-hdfs`. This default directory may be deleted by the operating system over time or when the operating system is restarted. If the DSSD DataNode process is restarted after the default directory is deleted, it will not be able to locate replicas stored on the DSSD D5 appliance. This causes the HDFS service to report under-replicated or even missing HDFS blocks.

Therefore it is best practice to change the default directory to a different value such as `/var/lib/hadoop-hdfs/dssddn`. The directory should be created and configured prior to initiating the installation of the Hadoop cluster using the following procedure:

1. Log into each host that will run the DataNode process.
2. Create the directory by executing the command:

```
mkdir -p /var/lib/hadoop-hdfs/dssddn
```

3. Change the directory ownership by executing the command:

```
chown -R hdfs:hadoop /var/lib/hadoop-hdfs
```

Cloudera Manager 5 Release Notes

4. Change the directory permissions by executing the command:

```
chmod -R 700 /var/lib/hadoop-hdfs
```

5. In Cloudera Manager, configure the `dfs.datanode.data.dir` property in the **HDFS Service Advanced Configuration Snippet (Safety Valve) for `hdfs-site.xml`** configuration property. This property must be set after the initial setup is completed but prior to writing any data to HDFS.

Note: Changing the value of the `dfs.datanode.data.dir` property after data has been written to HDFS will result in under-replicated or lost HDFS blocks.

Cloudera Manager reports that DSSD DataNode version is different from NameNode

The following warning log entry for the NameNode displays when using the DSSD Hadoop Plugin version 1.2.0 . This is expected and does not affect normal HDFS operation:

```
Jun 24, 8:12:41.846 AM WARN BlockStateChange
BLOCK* processReport: Report from the DataNode (8fa7fed1-1044-433b-a9fc-682bc08d1e25)
is unsorted.
This will cause overhead on the NameNode which needs to sort the Full BR.
Please update the DataNode to the same version of Hadoop HDFS as the NameNode
(2.6.0-cdh5.8.0).
```

HDFS caching

HDFS caching is not supported when using the DSSD D5 plugin.

Short Circuit Reads and Access Control with Impala or HBase

Enabling short-circuit reads for HBase or Impala on an HDFS cluster that uses DSSD D5 DataNodes requires that the processes associated with these applications be granted `hdfs` group membership. When short-circuit reads are enabled for Impala (for example), Impala process that act as short-circuit read clients (like `impalad`) are able to read and write all data stored in the DSSD D5. Cloudera Manager applies the `hdfs` group membership on a per-service basis, and applications that do not require short-circuit reads or for which short-circuit reads have not been enabled will have the same granularity of access control as present on a traditional HDFS cluster. Whether short-circuit reads are enabled or not, access control that is enforced by the application rather than at the file system level is identical for DSSD D5 DataNode HDFS clusters and traditional HDFS clusters.

The YARN NodeManager connectivity health test does not work for CDH 5

The NodeManager connectivity is always GOOD (green) even if the ResourceManager considers the NodeManager to be LOST or DECOMMISSIONED.

Workaround: None.

HDFS HA clusters see NameNode failures when KDC connectivity is bad

When KDC connectivity is bad, the JVM takes 30 seconds before retrying or declaring failure to connect. Meanwhile, the JournalNode write timeout (which needs KDC authentication for the first write, or under troubled connectivity), is only 20 seconds.

Workaround: In `krb5.conf`, set the `kdc_timeout` parameter value to 3 seconds. In Cloudera Manager, perform the following steps:

1. Go to **Administration > Settings > Kerberos**.
2. Add the `kdc_timeout` parameter to the **Advanced Configuration Snippet (Safety Valve) for [libdefaults] section of `krb5.conf`** property. This should give the JVM enough time to try connecting to a KDC before the JournalNode timeout.

The HDFS File browser in Cloudera Manager fails when HDFS federation is enabled

Workaround: Use the command-line `hdfs dfs` commands to directly manipulate HDFS files when federation is enabled. CDH supports HDFS federation.

Hive Metastore canary fails to drop database

The Hive Metastore canary fails to drop the database due to [HIVE-11418](#).

Workaround: Turn off the Hive Metastore canary by disabling the Hive Metastore Canary Health Test:

1. Go to the Hive service.
2. Click the **Configuration** tab.
3. Select **SCOPE > Hive Metastore Server**.
4. Select **CATEGORY > Monitoring**.
5. Deselect the **Hive Metastore Canary Health Test** checkbox for the Hive Metastore Server Default Group.
6. Click **Save Changes** to commit the changes.

Cloudera Manager upgrade fails due to incorrect Sqoop 2 path

Sqoop 2 will not start or loses data when upgrading from Cloudera Manager 5.4.0 or 5.4.1 to Cloudera Manager 5.4.3, or when upgrading Cloudera Manager 3 or 4 to Cloudera Manager 5.4.0 or 5.4.1. This is due to Cloudera Manager erroneously configuring the Derby path with "repository" instead of "repository". To upgrade Cloudera Manager, use one of the following workarounds:

- **Workaround for Upgrading from Cloudera Manager 3 or 4 to Cloudera Manager 5.4.0 or 5.4.1**
 1. Log in to your Sqoop 2 server host using SSH and move the Derby database files to the new location, usually from `/var/lib/sqoop2/repository` to `/var/lib/sqoop2/repositoy`.
 2. Start Sqoop2. If you found this problem while upgrading CDH, run the Sqoop 2 database upgrade command using the **Actions** drop-down menu for Sqoop 2.
- **Workaround for Upgrading from Cloudera Manager 5.4.0 or 5.4.1 to Cloudera Manager 5.4.3**
 1. Log in to your Sqoop 2 server host using SSH and move the Derby database files to the new location, usually from `/var/lib/sqoop2/repository` to `/var/lib/sqoop2/repositoy`.
 2. Start Sqoop2, or if you found this problem while upgrading CDH, run the Sqoop 2 database upgrade command using the **Actions** drop-down menu for Sqoop 2.

NameNode incorrectly reports missing blocks during rolling upgrade

During a rolling upgrade to any of the CDH releases listed below, the NameNode may report missing blocks after rolling back multiple DataNodes. This is caused by a race condition with block reporting between the DataNode and the NameNode. No permanent data loss occurs, but data can be unavailable for up to six hours before the problem corrects itself.

Releases affected: CDH 5.0.6, 5.1.5, 5.2.5, 5.3.3, 5.4.1, 5.4.2.

Releases containing the fix:: CDH 5.2.6, 5.3.4, 5.4.3

Workaround:

- **To avoid the problem** - Cloudera advises skipping the affected releases and installing a release containing the fix. For example, do not upgrade to CDH 5.4.2; upgrade to CDH 5.4.3 instead.
- **If you have already completed an upgrade to an affected release, or are installing a new cluster** - You can continue to run the release, or upgrade to a release that is not affected.

Using ext3 for server dirs easily hit inode limit

Using the ext3 filesystem for the Cloudera Manager command storage directory may exceed the maximum subdirectory size of 32000.

Cloudera Manager 5 Release Notes

Workaround: Either decrease the value of the **Command Eviction Age** property so that the directories are more aggressively cleaned up, or migrate to the ext4 filesystem.

Backup and disaster recovery replication does not set MapReduce Java options

Replication used for backup and disaster recovery relies on system-wide MapReduce memory options, and you cannot configure the options using the Advanced Configuration Snippet.

Kafka 1.2 CSD conflicts with CSD included in Cloudera Manager 5.4

If the Kafka CSD was installed in Cloudera Manager to 5.3 or lower, the old version must be uninstalled, otherwise it will conflict with the version of the Kafka CSD bundled with Cloudera Manager 5.4.

Workaround: Remove the Kafka 1.2 CSD before upgrading Cloudera Manager to 5.4:

1. Determine the location of the CSD directory:
 - a. Select **Administration > Settings**.
 - b. Click the **Custom Service Descriptors** category.
 - c. Retrieve the directory from the **Local Descriptor Repository Path** property.
2. Delete the Kafka CSD from the directory.

Recommission host does not deploy client configurations

The failure to deploy client configurations can result in client configuration pointing to the wrong locations, which can cause errors such as the NodeManager failing to start with "Failed to initialize container executor".

Workaround: Deploy client configurations first and then restart roles on the recommissioned host.

Hive on Spark is not supported in Cloudera Manager and CDH 5.4 and CDH 5.5

You can configure Hive on Spark, but it is not recommended for production clusters.

CDH 5 requires JDK 1.7

JDK 1.6 is not supported on any CDH 5 release, but before CDH 5.4.0, CDH libraries have been compatible with JDK 1.6. As of CDH 5.4.0, CDH libraries are no longer compatible with JDK 1.6 and **applications using CDH libraries must use JDK 1.7**.

In addition, you must upgrade your cluster to a supported version of JDK 1.7 before upgrading to CDH 5. See [Upgrading to Oracle JDK 1.7 before Upgrading to CDH 5](#) for instructions.

Upgrade wizard incorrectly upgrades the Sentry DB

There's no Sentry DB upgrade in 5.4, but the upgrade wizard says there is. Performing the upgrade command is not harmful, and taking the backup is also not harmful, but the steps are unnecessary.

Cloudera Manager does not correctly generate client configurations for services deployed using CSDs

HiveServer2 requires a Spark on YARN gateway on the same host in order for Hive on Spark to work. You must deploy Spark client configurations whenever there's a change in order for HiveServer2 to pick up the change.

CSDs that depend on Spark will get incomplete Spark client configuration. Note that Cloudera Manager does not ship with any such CSDs by default.

Workaround: Use `/etc/spark/conf` for Spark configuration, and ensure there is a Spark on YARN gateway on that host.

Solr, Oozie and HttpFS fail when KMS and TLS/SSL are enabled using self-signed certificates

When the KMS service is added and TLS/SSL is enabled, Solr, Oozie and HttpFS are not automatically configured to trust the KMS's self-signed certificate and you might see the following error.

```
org.apache.oozie.service.AuthorizationException: E0501: Could not perform authorization
operation,
sun.security.validator.ValidatorException: PKIX path building failed:
sun.security.provider.certpath.SunCertPathBuilderException:
unable to find valid certification path to requested target
```

Workaround: You must explicitly load the relevant truststores with the KMS certificate to allow these services to communicate with the KMS. To do so, edit the truststore location and password for Solr, Oozie and HttpFS (found under the HDFS service) as follows.

1. Go to the Cloudera Manager Admin Console.
2. Go to the Solr/Oozie/HDFS service.
3. Click the **Configuration** tab.
4. Search for "<service> TLS/SSL Certificate Trust Store File" and set this property to the location of truststore file.
5. Search for "<service> TLS/SSL Certificate Trust Store Password" and set this property to the password of the truststore.
6. Click **Save Changes** to commit the changes.

Cloudera Manager 5.3.1 upgrade fails if Spark standalone and Kerberos are configured

CDH upgrade fails if Kerberos is enabled and Spark standalone is installed. Spark standalone does not work in a kerberized cluster.

Workaround: To upgrade, remove the Spark standalone service first and then proceed with upgrade.

Adding Key Trustee KMS 5.4 to Cloudera Manager 5.5 displays warning

Adding the Key Trustee KMS service to a CDH 5.4 cluster managed by Cloudera Manager 5.5 displays the following message, even if Key Trustee KMS is installed:

"The following selected services cannot be used due to missing components: keytrustee-keyprovider. Are you sure you wish to continue with them?"

Workaround: Verify that the Key Trustee KMS parcel or package is installed and click **OK** to continue adding the service.

KMS and Key Trustee ACLs do not work in Cloudera Manager 5.3

ACLs configured for the KMS (File) and KMS (Navigator Key Trustee) services do not work since these services do not receive the values for `hadoop.security.group.mapping` and related group mapping configuration properties.

Workaround:

KMS (File): Add all configuration properties starting with `hadoop.security.group.mapping` from the NameNode `core-site.xml` to the KMS (File) property, **Key Management Server Advanced Configuration Snippet (Safety Valve) for core-site.xml**

KMS (Navigator Key Trustee): Add all configuration properties starting with `hadoop.security.group.mapping` from the NameNode `core-site.xml` to the KMS (Navigator Key Trustee) property, **Key Management Server Proxy Advanced Configuration Snippet (Safety Valve) for core-site.xml**.

Exporting and importing Hue database sometimes times out after 90 seconds

Executing 'dump database' or 'load database' of Hue from Cloudera Manager returns "command aborted because of exception: Command timed-out after 90 seconds". The Hue database can be exported to JSON from within Cloudera Manager. Unfortunately, sometimes the Hue database is quite large and the export times out after 90 seconds.

Cloudera Manager 5 Release Notes

Workaround: Ignore the timeout. The command should eventually succeed even though Cloudera Manager reports that it timed out.

Changing the Key Trustee Server hostname requires editing keytrustee.conf

If you change the hostname of your active or passive Key Trustee Server, you must edit the `keytrustee.conf` file. This issue typically arises if you replace an active or passive server with a server having a different hostname. If the same hostname is used on the replacement server, there are no issues.

Workaround: Use the same hostname on the replacement server.

Hosts with Impala Llama roles must also have at least one YARN role

When integrated resource management is enabled for Impala, host(s) where the Impala Llama role(s) are running must have at least one YARN role. This is because Llama requires the `topology.py` script from the YARN configuration. If this requirement is not met, you may see errors such as:

```
"Exception running /etc/hadoop/conf.cloudera.yarn/topology.py  
java.io.IOException: Cannot run program "/etc/hadoop/conf.cloudera.yarn/topology.py"
```

in the Llama role logs, and Impala queries may fail.

Workaround: Add a YARN gateway role to each Llama host that does not already have at least one YARN role (of any type).

The high availability wizard does not verify that there is a running ZooKeeper service

If one of the following is true:

- 1. ZooKeeper present and not running and the HDFS dependency on ZooKeeper dependency is not set
- 2. ZooKeeper absent

the enable high-availability wizard fails.

Workaround: Before enabling high availability, do the following:

1. Create and start a ZooKeeper service if one does not exist.
2. Go to the HDFS service.
3. Click the **Configuration** tab.
4. Select **Scope > Service-Wide**
5. Set the **ZooKeeper Service** property to the ZooKeeper service.
6. Click **Save Changes** to commit the changes.

Cloudera Manager Installation Path A fails on RHEL 5.7 due to PostgreSQL conflict

On RHEL 5.7, `cloudera-manager-installer.bin` fails due to a PostgreSQL conflict if PostgreSQL 8.1 is already installed on your host.

Workaround: Remove PostgreSQL from host and rerun `cloudera-manager-installer.bin`.

Spurious warning on Accumulo 1.6 gateway hosts

When using the Accumulo shell on a host with only an Accumulo 1.6 Service gateway role, users will receive a warning about failing to create the directory `/var/log/accumulo`. The shell works normally otherwise.

Workaround: The warning is safe to ignore.

Accumulo 1.6 service log aggregation and search does not work

Cloudera Manager log aggregation and search features are incompatible with the log formatting needed by the Accumulo Monitor. Attempting to use either the "Log Search" diagnostics feature or the log file link off of an individual service role's summary page will result in empty search results.

Severity: High

Workaround: Operators can use the Accumulo Monitor to see recent severe log messages. They can see recent log messages below the WARNING level via a given role's process page and can inspect full logs on individual hosts by looking in `/var/log/accumulo`.

Cloudera Manager incorrectly sizes Accumulo Tablet Server max heap size after 1.4.4-cdh4.5.0 to 1.6.0-cdh4.6.0 upgrade

Because the upgrade path from Accumulo 1.4.4-cdh4.5.0 to 1.6.0-cdh4.6.0 involves having both services installed simultaneously, Cloudera Manager will be under the impression that worker hosts in the cluster are oversubscribed on memory and attempt to downsize the max heap size allowed for 1.6.0-cdh4.6.0 Tablet Servers.

Severity: High

Workaround: Manually verify that the Accumulo 1.6.0-cdh4.6.0 Tablet Server max heap size is large enough for your needs. Cloudera recommends you set this value to the sum of 1.4.4-cdh4.5.0 Tablet Server and Logger heap sizes.

Accumulo installations using LZO do not indicate dependence on the GPL Extras parcel

Accumulo 1.6 installations that use LZO compression functionality do not indicate that LZO depends on the GPL Extras parcel. When Accumulo is configured to use LZO, Cloudera Manager has no way to track that the Accumulo service now relies on the GPL Extras parcel. This prevents Cloudera Manager from warning administrators before they remove the parcel while Accumulo still requires it for proper operation.

Workaround: Check your Accumulo 1.6 service for the configuration changes mentioned in the Cloudera documentation for using Accumulo with CDH prior to removing the GPL Extras parcel. If the parcel is mistakenly removed, reinstall it and restart the Accumulo 1.6 service.

Created pools are not preserved when Dynamic Resource Pools page is used to configure YARN or Impala

Pools created on demand are not preserved when changes are made using the Dynamic Resource Pools page. If the Dynamic Resource Pools page is used to configure YARN or Impala services in a cluster, it is possible to specify pool placement rules that create a pool if one does not already exist. If changes are made to the configuration using this page, pools created as a result of such rules are not preserved across the configuration change.

Workaround: Submit the YARN application or Impala query as before, and the pool will be created on demand once again.

User should be prompted to add the AMON role when adding MapReduce to a CDH 5 cluster

When the MapReduce service is added to a CDH 5 cluster, the user is not asked to add the AMON role. Then, an error displays when the user tries to view MapReduce activities.

Workaround: Manually add the AMON role after adding the MapReduce service.

Enterprise license expiration alert not displayed until Cloudera Manager Server is restarted

When an enterprise license expires, the expiration notification banner is not displayed until the Cloudera Manager Server has been restarted. The enterprise features of Cloudera Manager are not affected by an expired license.

Workaround: None.

Configurations for decommissioned roles not migrated from MapReduce to YARN

When the **Import MapReduce Configuration** wizard is used to import MapReduce configurations to YARN, decommissioned roles in the MapReduce service do not cause the corresponding imported roles to be marked as decommissioned in YARN.

Workaround: Delete or decommission the roles in YARN after running the import.

Cloudera Manager 5 Release Notes

The HDFS command Roll Edits does not work in the UI when HDFS is federated

The HDFS command Roll Edits does not work in the Cloudera Manager UI when HDFS is federated because the command does not know which nameservice to use.

Workaround: Use the API, not the Cloudera Manager UI, to execute the Roll Edits command.

Cloudera Manager reports a confusing version number if you have oozie-client, but not oozie installed on a CDH 4.4 node

In CDH versions before 4.4, the metadata identifying Oozie was placed in the client, rather than the server package. Consequently, if the client package is not installed, but the server is, Cloudera Manager will report Oozie has been present but as coming from CDH 3 instead of CDH 4.

Workaround: Either install the oozie-client package, or upgrade to at least CDH 4.4. Parcel based installations are unaffected.

Cloudera Manager does not work with CDH 5.0.0 Beta 1

When you upgrade from Cloudera Manager 5.0.0 Beta 1 with CDH 5.0.0 Beta 1 to Cloudera Manager 5.0.0 Beta 2, Cloudera Manager won't work with CDH 5.0.0 Beta 1 and there's no notification of that fact.

Workaround: None. Do a new installation of CDH 5.0.0 Beta 2.

On CDH 4.1 secure clusters managed by Cloudera Manager 4.8.1 and higher, the Impala Catalog server needs advanced configuration snippet update

Impala queries fail on CDH 4.1 when Hive "Bypass Hive Metastore Server" option is selected.

Workaround: Add the following to Impala catalog server advanced configuration snippet for `hive-site.xml`, replacing `Hive_Metastore_Server_Host` with the host name of your Hive Metastore Server:

```
<property>
<name>hive.metastore.local</name>
<value>false</value>
</property>
<property>
<name>hive.metastore.uris</name>
<value>thrift://Hive_Metastore_Server_Host:9083</value>
</property>
```

Rolling Upgrade to CDH 5 is not supported.

Rolling upgrade between CDH 4 and CDH 5 is not supported. Incompatibilities between major versions means rolling restarts are not possible. In addition, rolling upgrade will *not* be supported from CDH 5.0.0 Beta 1 to any later releases, and may not be supported between *any future beta versions* of CDH 5 and the General Availability release of CDH 5.

Workaround: None.

Error reading .zip file created with the Collect Diagnostic Data command.

After collecting Diagnostic Data and using the Download Diagnostic Data button to download the created zip file to the local system, the zip file cannot be opened using the FireFox browser on a Macintosh. This is because the zip file is created as a Zip64 file, and the unzip utility included with Macs does not support Zip64. The zip utility must be version 6.0 or later. You can determine the zip version with `unzip -v`.

Workaround: Update the unzip utility to a version that supports Zip64.

After JobTracker failover, complete jobs from the previous active JobTracker are not visible.

When a JobTracker failover occurs and a new JobTracker becomes active, the new JobTracker UI does not show the completed jobs from the previously active JobTracker (that is now the standby JobTracker). For these jobs the "Job Details" link does not work.

Severity: Med

Workaround: None.

After JobTracker failover, information about rerun jobs is not updated in Activity Monitor.

When a JobTracker failover occurs while there are running jobs, jobs are restarted by the new active JobTracker by default. For the restarted jobs the Activity Monitor will not update the following: 1) The start time of the restarted job will remain the start time of the original job. 2) Any Map or Reduce task that had finished before the failure happened will not be updated with information about the corresponding task that was rerun by the new active JobTracker.

Severity: Med

Workaround: None.

Installing on AWS, you must use private EC2 hostnames.

When installing on an AWS instance, and adding hosts using their public names, the installation will fail when the hosts fail to heartbeat.

Severity: Med

Workaround:

Use the Back button in the wizard to return to the original screen, where it prompts for a license.

Rerun the wizard, but choose "Use existing hosts" instead of searching for hosts. Now those hosts show up with their internal EC2 names.

Continue through the wizard and the installation should succeed.

If HDFS uses Quorum-based Storage without HA enabled, the SecondaryNameNode cannot checkpoint.

If HDFS is set up in non-HA mode, but with Quorum-based storage configured, the `dfs.namenode.edits.dir` is automatically configured to the Quorum-based Storage URI. However, the SecondaryNameNode cannot currently read the edits from a Quorum-based Storage URI, and will be unable to do a checkpoint.

Severity: Medium

Workaround: Add to the NameNode's advanced configuration snippet the `dfs.namenode.edits.dir` property with both the value of the Quorum-based Storage URI as well as a local directory, and restart the NameNode. For example,

```
<property> <name>dfs.namenode.edits.dir</name>
<value>qjournal://jn1HostName:8485;jn2HostName:8485;jn3HostName:8485/journalhdfs1,file:///dfs/edits</value>
</property>
```

Changing the rack configuration may temporarily cause mis-replicated blocks to be reported.

A rack re-configuration will cause HDFS to report mis-replicated blocks until HDFS rebalances the system, which may take some time. This is a normal side-effect of changing the configuration.

Severity: Low

Workaround: None

Cannot use '/' as a mount point with a Federated HDFS Nameservice.

A Federated HDFS Service does not support nested mount points, so it is impossible to mount anything at '/'. Because of this issue, the root directory will always be read-only, and any client application that requires a writeable root directory will fail.

Severity: Low

Workaround:

1. In the CDH 4 HDFS Service > Configuration tab of the Cloudera Manager Admin Console, search for "nameservice".

2. In the Mountpoints field, change the mount point from "/" to a list of mount points that are in the namespace that the Nameservice will manage. (You can enter this as a comma-separated list - for example, "/hbase, /tmp, /user" or by clicking the plus icon to add each mount point in its own field.) You can determine the list of mount points by running the command hadoop fs -ls / from the CLI on the NameNode host.

Historical disk usage reports do not work with federated HDFS.

Severity: Low

Workaround: None.

(CDH 4 only) Activity monitoring does not work on YARN activities.

Activity monitoring is not supported for YARN in CDH 4.

Severity: Low

Workaround: None

HDFS monitoring configuration applies to all Nameservices

The monitoring configurations at the HDFS level apply to all Nameservices. So, if there are two federated Nameservices, it's not possible to disable a check on one but not the other. Likewise, it's not possible to have different thresholds for the two Nameservices.

Severity: Low

Workaround: None

Supported and Unsupported Replication Scenarios and Limitations

See [Data Replication](#).

Restoring snapshot of a file to an empty directory does not overwrite the directory

Restoring the snapshot of an HDFS file to an HDFS path that is an empty HDFS directory (using the Restore As action) will result in the restored file present inside the HDFS directory instead of overwriting the empty HDFS directory.

Workaround: None.

HDFS Snapshot appears to fail if policy specifies duplicate directories.

In an HDFS snapshot policy, if a directory is specified more than once, the snapshot appears to fail with an error message on the Snapshot page. However, in the HDFS Browser, the snapshot is shown as having been created successfully.

Severity: Low

Workaround: Remove the duplicate directory specification from the policy.

Hive replication fails if "Force Overwrite" is not set.

The Force Overwrite option, if checked, forces overwriting data in the target metastore if there are incompatible changes detected. For example, if the target metastore was modified and a new partition was added to a table, this option would force deletion of that partition, overwriting the table with the version found on the source. If the Force Overwrite option is not set, recurring replications may fail.

Severity: Med

Workaround: Set the Force Overwrite option.

Cloudera Manager set cataloged default JVM memory to 4G can cause an out of memory error during upgrade to Cloudera Manager 5.7 and higher

When upgrading from Cloudera Manager 5.7 or higher to Cloudera Manager 5.8.2, if the Impala Catalog Server Java Heap Size is set at the default (4GB), it is automatically changed to either 1/4 of the physical RAM on that host, or 32GB, whichever is lower. This can result in a higher or a lower heap, which could cause additional resource contention or out of memory errors, respectively.

Issues Fixed in Cloudera Manager 5

The following sections describe issues fixed in each Cloudera Manager 5 release.

Issues Fixed in Cloudera Manager 5.9

[Impala load_catalog_in_background configuration should be set to *false* by default in Cloudera Manager](#)

Changed default value for `load_catalog_in_background` to *false*. This fixes catalog scalability issues observed in Impala since v2.2.

[Oozie first run fails with custom principals](#)

Fixed an issue in Cloudera Manager 5.8.1 where the first time Oozie is run it fails if using Kerberos custom principals.

[Cluster export fails when service configuration is invalid](#)

Fixed an issue in the export cluster template code path where it was failing because of stale configuration settings in the Cloudera Manager database. This can occur when configurations are deprecated on older CDH releases.

[Add gateway role to Kafka and move dependencies' client configs into Kafka's client configs](#)

Cloudera recommends that you place your Kafka host in the same logical cluster as your Sentry host. However, if you deploy Kafka as a separate logical cluster, you can deploy a dummy Sentry service on Kafka's logical cluster with an override for `Sentry-site.xml` to point to the Sentry service on first logical cluster, and then it can be turned off. This is a workaround that allows Cloudera Manager to generate the appropriate Sentry client configurations for Kafka.

[Sentry Upgrade command is not retriable](#)

If the Sentry Upgrade command fails, you can now retry the command.

[Impala Breakpad script does not convert exponentials into decimals and leads to errors](#)

Fixed an issue where the Impala Breakpad script fails if you try to collect more than 10 MB of dumps from a single role.

[HDFS-S3: Incremental backup support](#)

HDFS-S3 replication supports incremental backups using snapshot diffs. HDFS-S3 replication does not support incremental restore using snapshot diff.

[Fix css in navigation](#)

Fixed several small issues:

1. Do not show role link again on the role instances page.
2. Do not bold the decommission state.
3. Do not use long display name for hosts; use short display name.
4. Show the role link on the role health test page.
5. Show *Today* rather than the long time label.
6. Show only month, date when the year is the same.
7. Make the icons in navigation line up.
8. Make the icons in navigation not so wide.

Cloudera Manager 5 Release Notes

9. Change all line-height settings to multiples of 4.

10 Add back the border below the title.

Fix sorting on the instances page

Fixed the initial sort order on the instances page.

Nightly Charts are broken due to gridster refactoring

There is a permission issue on the Cluster Status page where the grid is never enabled, and on the View Page where the grid is always enabled for all users. The former change allows the user to customize the Cluster status page. The latter change ensures that users do not get errors when customizing specific view pages.

Redact s3 credential properties in MR by default.

AWS S3 credentials, if specified in the job configuration by setting `fs.s3a.access.key` and `fs.s3a.secret.key`, were shown as clear text in MapReduce UIs. The credentials are now redacted by default in all MapReduce UIs.

Latest Cloudera Manager Java client does not work on older Cloudera Manager

Fixed an issue in Cloudera Manager client versions 5.8, 5.8.1, and 5.8.2 where `ApiCollectDiagnosticDataArguments` was incompatible with Cloudera Manager versions lower than 5.8.

Comment for command retry in the cm_api Python github needs to be updated

Fixed the API command documentation to include the `canRetry` attribute. Added a new method, `ApiCommand.getCanRetry()` and deprecated the method `ApiCommand.isCanRetry()`.

Defaults to 1st Impala service in the cluster when there are multiple Impalas

Fixed an issue that prevented the display of Impala usage on the **Cluster Utilization** page when there are two Impala services in the same cluster. The UI shows usage for the first Impala service only.

Staleness check page popped up 30s ~ 60s after clicking the icon.

Fixed an issue where a stale configuration page would take a lot of time to load for a large cluster.

Oozie points to older sharelib even after running sharelib install command

Fixed the Install Oozie Share Lib action so that the Oozie service is informed that there is a new shared library installed. This eliminates the need for a separate manual restart.

HDFS File Browser exposes contents to all users

HDFS File browser can now only be viewed by users with the roles BDR Admin, Cluster Admin, or Full Admin. Previously, any Cloudera Manager user, including read-only users, could browse the file names in HDFS.

Cloudera Manager set catalogd default jvm memory to 4G can cause out of memory error on upgrade to Cloudera Manager 5.7+

After upgrading to 5.7 or later, you might see a reduced Java heap maximum on Impala Catalog Server due to a change in its default value. Upgrading from Cloudera Manager lower than 5.7 to Cloudera Manager 5.8.2 no longer causes any effective change in the Impala Catalog Server Java Heap size.

When upgrading from Cloudera Manager 5.7 or later to Cloudera Manager 5.8.2, if the Impala Catalog Server Java Heap Size is set at the default (4GB), it is automatically changed to either 1/4 of the physical RAM on that host, or 32GB, whichever is lower. This can result in a higher or a lower heap, which could cause additional resource contention or out of memory errors, respectively.

Agent uses USER env var to run everything if it is set

Cloudera Manager Server now runs services using the user ID with which that single-user mode is configured. If single-user mode is configured on the agents to use the user `cloudera-scm`, then an attempt to run Hive as the user `hive` fails.

In LZO enabled clusters, the cluster utilization feature fails

YARN usage aggregation job now runs successfully on clusters where LZO compression is being used.

HDFS Snapshot policy is selecting unhealthy host to run on

When selecting a role to run HDFS Snapshot command, Cloudera Manager selects a non-decommissioned host in Active status. Also, hosts in maintenance mode have a lower priority than the ones in active state.

Discourage CSDs from producing client config files that pollute the root dir

For CSD development, there are now deprecation warnings in the CLI validation tool and the Maven validation plugin. These indications provide guidance for future behavior changes in the support of CSD services in Cloudera Manager. A deprecation warning is shown when applicable, but does not affect the outcome of the validation.

Ensure that one-off processes cannot re-execute.

One-off requests, such as "Inspect Hosts," do not re-execute.

Hive Replication Command should update copy the Serde properties correctly

Hive Replication now replicates the Serde Properties and also copies corresponding HDFS file.

Cloudera Manager complains "Unable to parse the input XML" when you enter an working XML for the "Fair Scheduler XML Advanced Configuration Snippet (Safety Valve)"

Passing a legal XML value to Fair Scheduler XML Advanced Configuration Snippet (Safety Valve) no longer causes a parsing error.

If files excluded by exclusion filters are renamed, they are not replicated

Customers using Incremental HDFS replication had an issue where if an excluded file (through exclusion filters) is renamed to an included file, the new file was not copied to the destination cluster. This issue is resolved as part of this fix.

Proper fix to handle parcel activation falsely succeeding when host health is failing

Fixed an issue related to the first run failing when some of the hosts are in bad health. This fix involves adding an extra stop to wait for hosts to report correct the parcel version before proceeding.

Cloudera Manager is reporting incorrect Configured Capacity of HDFS in a multi-D5 (Cloudera Manager/CDH5.8) environment

DSSD correctly reports the usable capacity, and HDFS Configured Capacity correctly shows the sum of usable capacity from individual DSSD D5 appliances.

Deferred cgroup removal might not be waiting at all

This fix provides some wait time before the SCM agent attempts to remove control groups (cgroups) again after initial failure.

SMON should only fall back to Oozie's instrumentation endpoint if the metrics endpoint gives a 503

SMON blindly fell back to Oozie's instrumentation endpoint if it had any problems with the metrics endpoint. Now, SMON falls back only if the metrics endpoint returns a 503 error code, which is the only case it could possibly have success with instrumentation endpoint.

Support bundle missing client configuration deployment logs

Client configuration deployment logs are now collected as a part of diagnostics support bundle, which were unintentionally removed in Cloudera Manager 5.3.

Cloudera Manager 5 Release Notes

Use concurrency option when uploading the sharelib to be faster

You can increase the number of threads used to upload the Oozie sharelib to HDFS. This significantly reduces the time it takes for this operation. This feature is available starting in CDH 5.9.0.

Host inspector incorrectly warns about kernel version "2.6.32-504.16.2"

Before this fix, the host inspector incorrectly warned about kernel version "2.6.32-504.16.2" as "non-recommended"

Agent orphan cleanup removes process dir from in flight process

With this fix, the preventative steps in TSB-181 are no longer required.

Move default HDFS data dir for DSSD DN to a safer directory

Cloudera Manager 5.9 automatically creates the DSSD DataNode metadata directory and emits the `dfs.datanode.data.dir` configuration for DSSD DataNode.

Rolling upgrade fails if services stopped

Rolling upgrade will no longer fail if services are stopped.

Redact Content from Flume configuration

Enabled redaction of sensitive information from Flume configuration.

Redaction of sensitive information from diagnostic bundles at creation time

Cloudera Manager is designed to transmit certain diagnostic data (or *bundles*) to Cloudera. The Cloudera Support team uses diagnostic bundles to reproduce, debug, and address technical issues for customers. Cloudera support discovered that potentially sensitive data might be included in diagnostic bundles and transmitted to Cloudera. Notwithstanding any possible transmission, such sensitive data is not used by Cloudera for any purpose. Cloudera has taken the following actions:

- Modified Cloudera Manager so that known sensitive data is redacted from the bundles before transmission to Cloudera.
- Updated Cloudera CDH components to remove logging and output of known potentially sensitive properties and configurations.

Huge memory leak if the HDFS File Browser UI remains open for an extended time

If the HDFS File Browser were kept open for long periods of time, it could cause a memory leak that would cause Cloudera Manager to crash. This fix addresses the issue.

Evaluate impact of disabling second-level cache on performance at scale

Disabled second level hibernate cache, which was causing the cache to become "stale" under load. There is no significant difference in Cloudera Manager performance after disabling the cache.

Express wizard failing to install ZooKeeper

Installing CDH using Packages on Debian 8.2 through the Cloudera Manager wizard could install the incorrect CDH binaries, leading to failures in starting the cluster. Certain packages are available in both the default repository and the Cloudera repository, which could result in the wizard using the wrong one. The Cloudera Manager wizard now ensures the correct binaries are selected. Workaround: Use the parcels format or manually install the correct packages.

Empty archive files should not be created for impala role diagnostics

Support Bundles started collecting Impala minidumps bundles in Cloudera Manager 5.8.0 and higher. It was generating an empty TAR file if no bundles were found. With this fix, no empty TAR files are generated.

Cloudera Manager no longer shows "this step is expected to fail" when enabling HDFS-HA

Cloudera Manager now shows the correct label when performing HA.

If total_space_bytes is very large, heartbeats fail

Fixed an issue where the heartbeat fails with a host that has a mount point backed by cloud storage, such as AWS.

Enabling cgroups requires "impala" group even if there is no Impala service

Fixed the presumption that any default-named group exists in hosts. YARN cgroup containers do not require the Impala user/group to be present in hosts.

OOMKiller script not works for Impala Catalog

Fixed a bug where OutOfMemory errors in the Catalog Server might lead to killing multiple Java processes, including other roles on the same host.

Issues Fixed in Cloudera Manager 5.8.3

YARN historical reports by user shows pool-user entity

When Cloudera Manager manages multiple clusters, there is no per user tracking for historical applications and queries across clusters. Instead, **Historical Applications by User** and **Historical Queries by User** show applications and queries per user and pool. (A pool is associated with a specific cluster.)

If total_space_bytes is very large, heartbeats fail

Fixes heartbeat failure with a host that has a mount point backed by cloud storage, such as AWS.

OPSAPS-29327 Add config for hive.metastore.server.max.message.size

You can configure `hive.metastore.max.message.size` using **Max Message Size for Hive MetaStore**. The default setting is 100MB. This can cause staleness during a Cloudera Manager upgrade.

CM blocks from HS2 enabling both LDAP and Kerberos authentication

HS2 now supports LDAP and Kerberos authentication on the same instance for CDH 5.7.0 and higher.

Backport OPSAPS-36100 Debian 8.2 packages cdh install fix to C5.8

Installing CDH using Packages on Debian 8.2 through the Cloudera Manager wizard could install the incorrect CDH binaries, leading to failures in starting the cluster. Certain packages are available in both the default repository and the Cloudera repository, which caused the installer to get confused and pick the wrong one. Cloudera Manager's wizard now ensures the correct binaries are selected.

Impala load_catalog_in_background config should set to "false" by default in CM

Changed default value for `load_catalog_in_background` to false. This fixes catalog scalability issues observed in Impala since v2.2.

[api] Latest CM Java client does not work on older CM

Fixed an issue in Cloudera Manager client versions 5.8, 5.8.1, and 5.8.2 where `ApiCollectDiagnosticDataArguments` was incompatible with Cloudera Manager versions lower than 5.8.

Empty archive files should not be created for Impala role diagnostics

Support Bundles started collecting Impala minidump bundles with Cloudera Manager 5.8 and higher. If no bundles were found, Cloudera Manager generated an empty TAR file. With this fix, no empty TAR files are generated.

OOMKiller script not works for Impala Catalog

Fixed a bug where OutOfMemory errors in the Catalog Server could lead to killing multiple java processes including other roles on the same host.

Cloudera Manager 5 Release Notes

Sentry Upgrade command is not retriable

If the Sentry Upgrade command fails, you can now retry the command.

Support non-public schemas

Cloudera Manager used to work with only 'public' schema in PostgreSQL. With this change, Cloudera Manager supports custom schema names. Cloudera Manager relies on the schema search path in PostgreSQL, which can be set for a user, database, and so on. <https://www.postgresql.org/docs/current/static/ddl-schemas.html>

Issues Fixed in Cloudera Manager 5.8.2

Improve advanced configuration snippet redaction to encompass cloud provider credentials and other access tokens

The redaction of potentially sensitive parameters in advanced configuration snippets is extended to those commonly used by the Azure Data Lake.

If files excluded by exclusion filters are renamed, they are not replicated

Customers using Incremental HDFS replication had an issue where if an excluded file (through exclusion filters) is renamed to an 'included' file, the new file is still not copied to the destination cluster. This issue is resolved as part of this fix.

Hive Replication Command should update copy the Serde properties correctly

After this fix, Hive Replication replicates the Serde Properties and also copies corresponding HDFS file.

Increase default Solr watchdog Timeout value

Solr server initialization can take up to 60 secs to complete. During this time interval, Solr server does not respond to the solr watchdog requests. This can result in solr watchdog terminating the Solr server process. The default timeout duration for watchdog is increased to 70 secs.

Add service changes YARN settings

In Cloudera Manager 5.7, adding any new service to your cluster can cause the YARN setting for mapreduce.job.reduces to change unexpectedly. Adding a service no longer causes this problem.

OPSAPS-29327 Add config for hive.metastore.server.max.message.size

Hive Metastore max message size can be configured now using **Max Message Size for Hive MetaStore**. It defaults to 100MB. It can cause staleness for customers on Cloudera Manager upgrade.

XSS in Kerberos activation

In lower releases, there was an XSS vulnerability on the Kerberos page. This is now fixed.

Upload deployment.json fails when it contains replication info

While trying to migrate Cloudera Manager using deployment.json file with existing Hive replication, schedules used to fail. This has been fixed in this release.

Cloudera Manager set cataloged default jvm memory to 4G may cause oom on upgrade to Cloudera Manager 5.7+

After upgrading to 5.7 or later, customers could see a reduced Java heap maximum on Impala Catalog Server, due to a change in its default value. Upgrading from Cloudera Manager < 5.7 to Cloudera Manager 5.8.2 no longer sees any effective change in the Impala Catalog Server Java Heap size.

Oozie first run fails with custom principals

In 5.8.1, Oozie first run fails if using kerberos custom principals. This is now fixed.

Hive Replication shows "Dry Run" incorrectly

The earlier known issue that running Hive Replication shows "Dry Run" in status message is fixed now.

HDFS Snapshot policy is selecting unhealthy host to run on

Policy for selecting a role to run HDFS Snapshot command: We select a non-decommissioned host that is in Active status. Also, hosts in maintenance mode have a lower priority than hosts in active state.

Cluster export fails when service configuration is invalid

The export cluster template code path was failing because of stale configuration in the Cloudera Manager database. Having a stale configuration in the database is possible. This could happen when configurations are deprecated on older CDH releases.

Agent orphan cleanup removes process dir from in flight process

With this fix, the preventative steps in TSB-181 are no longer required.

Fix CatalogServiceClient to handle TLS connections to catalogd for UDF replication

When Impala uses SSL, Cloudera supports TLS Connection to Catalog Server. Customers are able to enable replication for any Impala UDFs/Metadata (in Hive Replication).

Redact Content from Flume config

Enabled redacting sensitive information from Flume configuration.

Host inspector incorrectly warns about kernel version "2.6.32-504.16.2"

Host inspector incorrectly warns about kernel version "2.6.32-504.16.2" as "non-recommended."

Impala Breakpad script does not convert exponentials into decimals and leads to errors

Impala Breakpad script failure that happens when trying to collect more than 10 MB of dumps from a single role is fixed.

Oozie points to older sharelib even after running sharelib install command

After an "Install Oozie Share Lib" action, the Oozie service is informed that there is a new shared lib installed. This eliminates the need for a separate manual restart.

Issues Fixed in Cloudera Manager 5.8.1

CDH upgrade from 5.7.x to 5.8.x fails when Sentry gateway role is enabled

If the Sentry Gateway role is configured on any hosts of a CDH 5.7.x cluster, the upgrade process to CDH 5.8.x fails.

Upgrades to CDH 5.8.x now complete successfully.

Issues Fixed in Cloudera Manager 5.8.0

Kafka MirrorMaker unable to start due to KAFKA_HOME not being set

Kafka MirrorMaker would not start when Kafka is installed using packages. This occurred because `KAFKA_HOME` was not set to the correct default when starting MirrorMaker. This issue affected Cloudera Manager 5.4.0 and higher with Kafka 1.4.0 and higher.

Kafka unable to start due to misconfigured security.inter.broker.protocol when Kerberos is enabled

Kafka would not start when Kerberos is enabled and the default value of `security.inter.broker.protocol` was not changed. This occurred because Kafka tried to use the same port for `SASL_PLAINTEXT` and `PLAINTEXT`. By default, Cloudera Manager now infers the protocol based on the security settings.

This issue affected Cloudera Manager 5.5.2 and higher with Kafka 2.0.0 and higher.

Cloudera Manager 5 Release Notes

Upgrading to Cloudera Manager 5.7.1 or higher upgrades currently configured values to `INFERRRED` unless SSL/TLS is enabled and the values are currently either `PLAINTEXT` or `SASL_PLAINTEXT`. This does not cause any change in behavior.

Child commands for deleting or adding a nameservice show stack trace

In an existing HDFS deployment with high availability, when you try to add or delete a nameservice and attempt to view the progress of the child commands, a stack trace is triggered if some of the child commands have not yet run. This fix eliminates the stack trace and informs you that the child commands have not yet been run.

Setting owner of a file in Isilon fails

On Isilon systems, the owner that the file is being changed to must be present on the system. In general cases, the user is not present, so this command fails with an error message suggesting that the user is not part of the supergroup. This fix addresses the issue by not failing the command.

Handle drop-recreate partition efficiently

With this change, properties of entities (database, table, partition, index) are not updated by default. You can choose to update properties by setting `REPLICATE_PARAMETERS=true` in **Hive Replication Environment Advanced Configuration Snippet (Safety Valve)**.

HiveReplicationCmdArgs.update is not accessible using Cloudera Manager

In Hive replication, you can choose to update one or more of the entities `INDICES`, `PARAMETERS`, `PARTITIONS`, and `PRIVILEGES` adding the following instruction in **Hive Replication Environment Advanced Configuration Snippet (Safety Valve)**.

```
PROPERTIES_TO_UPDATE=INDICES,PARAMETERS,PARTITIONS,PRIVILEGES
```

Operation log directory should be configurable and monitored in Cloudera Manager

HiveServer has two new properties for configuration of operation logging.

Property	Default
Enable HiveServer2 Operations Logging	true
HiveServer2 Operations Log Directory	/var/logs/hive/operation_logs

Kerberos should use non-person objects when creating principals in Active Directory

You can now configure Active Directory account properties. You can use custom values for `objectClasses` to configure accounts, including non-person objects.

Changes to `yarn.nodemanager.remote-app-log-dir` not picked up from gateway

YARN log aggregation did not work when `yarn.nodemanager.remote-app-log-dir` was configured to a non-default location. Now this value is emitted in YARN client configuration, ensuring clients logs go to the proper location.

Key Trustee KMS should use round-robin configuration when Key Trustee server uses High Availability

If the Key Trustee server is configured with High Availability, the Key Management Service needs to use round-robin DNS.

`AD_ACCOUNT_PREFIX` should not be required

Active Directory Account Prefix was improperly implemented as a required configuration for Security/ Kerberos. The use of this configuration is now optional.

Proper fix to handle parcel activation falsely succeeding when host health is failing

Fixed an issue related to the first run failing when some of the hosts are in bad health. This fix involves adding an extra stop to wait for hosts to report the correct parcel version before proceeding.

Document Kerberos + Isilon support in Disaster Recovery

BDR is now supported on Isilon (including on clusters secured with Kerberos).

Separation of authentication and authorization coprocessor configs in HBase

HBase Secure Bulkload is now enabled for all CDH5.5 and higher clusters, regardless of whether Kerberos is enabled. Also fixed related issue where clusters with authentication (kerberos) but not authorization failed in Hbase-related MapReduce jobs.

Cloudera Manager - BDR UI - Generate replication diagnostics data. Failed due to java.lang.NullPointerException.

Trying to collect diagnostic data for a Hive replication schedule caused a Java stack trace to be shown on the page. This fix shows an error message instead of displaying a Java stack trace.

Support DSSD 1.2: move flood volume name param spec to role level

The Flood Volume Name parameter can now be set to distinct values for each DSSD DataNode role or role configuration group. Prior to release 5.8, a single Flood Volume Name setting applied to all HDFS roles.

Diagnostic collection fails on a failed BDR job

Failed BDR jobs no longer report errors on "Collect diagnostic data" actions.

Redact Advanced Configuration Snippets in the UI that contain secrets

Cloudera Manager 5.8 supports redaction of Advanced Configuration Snippet parameters in UI configuration pages. Redaction is based on matching keywords defined as *sensitive*, and detected within the contents of the Advanced Configuration Snippet text. Users who can edit the parameter still see the sensitive words, but users without edit privileges see only the redacted contents.

Cloudera Manager - snapshot - take snapshot - needs reset of illegal character when fixed

Fixed an issue in the Take Snapshot dialog where the validation of the snapshot name could become "stuck" and prevent execution of the operation.

Spark standalone does not come up if HDFS is not available

The Spark standalone service now works without an HDFS service. This requires Spark services to show up as stale and require a restart after upgrade to Cloudera Manager 5.7.1 and higher.

Cloudera Manager - add hdfs nameservice - gets stack trace

In an existing HDFS High Availability setup, when the user tries to add or delete a nameservice and, in the process, attempts to get progress on the child commands, a stack trace is triggered if some of the child commands have not yet run. This fix eliminates the stack trace and informs the user that the child commands have not yet run.

BDR UI: Add Search back - It is difficult to find the desired replication schedule in the Cloudera Manager 5.5 UI

The Replications page provides enhanced search functionality to filter schedules by any specified text. Searches occur within any of these schedule fields: HDFS paths, database names and table names.

BDR Replication Schedule page is slow to load

Replication schedules page load performance is improved.

Cloudera Manager 5 Release Notes

Replication DistCp - setOwner behavior on Isilon causing failures

On Isilon systems, the owner to which the file is being changed must be present on the system. In general cases, the user is not present, so this command tends to fail with an error message suggesting that the user is not part of the supergroup. This fix addresses the issue by not failing the command.

Remove deprecated DSSD metrics

Starting from DHP 1.2 (DSSD's parcel release for CDH 5.8), DSSD DataNode no longer reports the metrics `hdfs_dssd_crop_ops` – `hdfs_dssd_failed_heartbeat_ops` and `hdfs_dssd_successful_heartbeat_ops`. This change removes these metrics from Cloudera Manager. You can no longer pull and plot these metrics in Cloudera Manager 5.8.

Unable to start Hue on cluster that's using Kerberos and Isilon

Hue service can now start with Isilon if Kerberos is enabled.

Enable database notifications from Hive

Hive now has the property **Enable Stored Notifications in Database**. When set, Hive logs DDL notifications in Hive Metastore.

The `hbase` user should be whitelisted by default in the list of allowed system users to launch YARN applications

YARN Allowed System Users now includes `hbase` by default. This is helpful when running certain tools for HBase that need to execute MapReduce jobs.

Allow HDFS Balancer to login with keytab

When running an HDFS rebalance command on a kerberized cluster with a large amount of data, it could take enough time to complete that authentication would expire and cause an error. Leveraging a new capability in HDFS, the rebalance command is now able to use a keytab file to automatically renew authentication before it expires.

Add support for delete tables and databases deleted on the source

Cloudera Manager now supports deletion of entities from a target Hive database when those entities are deleted from source Hive database during Hive incremental replication.

Default MapReduce option should be YARN

The default MapReduce service for a new replication schedule is YARN.

BDR Administrator cannot enable snapshots though the role says it can

BDR Administrator authority message is now more accurate: Create replication schedules and snapshot policies.

Hive incremental replication enhancements

Cloudera Manager 5.8 includes Hive incremental replication support.

Cloudera Manager API allows users with read-only privileges to list all Cloudera Manager users

A security bug in the Cloudera Manager API allowed users with read-only permissions to view all the existing users in Cloudera Manager. This issue is now fixed.

Decide the value for TTL for DbNotifications

You can configure Time-to-live for Database Notifications in Hive for notifications present in the NOTIFICATION_LOG. The default is 2 days.

Make the cluster menu sticky

Previously, the Clusters Menu expanded the first cluster by default. As the user expanded or collapsed the accordion, it remembered the configuration for the current session. When the user goes to the services, roles, or host of another cluster, it maintained the configuration of the previously expanded cluster (which might or might not match). In Cloudera

Manager 5.7.1 and higher, Cloudera Manager records the last relevant cluster when the user visits a cluster, service, role or host page, and expands that cluster in the Clusters menu by default.

NPE on Impala Admission Control page if Memory Limit is not set

In lower releases, if the configuration Impala Daemon Memory Limit is not set, the Impala Admission Control page throws a NullPointerException. This is now fixed.

Change Impala service configs for admission control

Enable Impala Admission Control and **Enable Dynamic Resource Pools** are now enabled by default. Customized configuration values are preserved during upgrade.

Disaster Recovery on Isilon breaks with Kerberos

BDR on Isilon storage is now supported with kerberized clusters.

Breakpad Crash Reporting for Impala

Support bundles now collect "minidumps" from Impala that help to debug Impala crash issues. As part of this functionality, Cloudera Manager exposes two properties for three roles (Impalad, Catalog Server, Statestore Server).

Property	Description
Breakpad Dump Dir	This determines where the "minidumps" are temporarily available.
Max Breakpad Dump Files	This determines the maximum number of files stored in <code>dump_dir</code> (this limits the amount of storage allocated to Impala dump files).

s3 protocol and scheme should not be pruned during Hive metadata export

With this change, the cloud HDFS path remains as it is after replication.

YARN mapreduce.shuffle.max.connections is a NodeManager setting and not a client setting

`mapreduce.shuffle.max.connections` was emitted to files of YARN clients instead of the NodeManager. It is now correctly emitted only for the NodeManager.

Kafka unable to start due to listener misconfiguration when Kerberos is enabled

Kafka would not start when Kerberos was enabled and the `security.inter.broker.protocol` default was not changed. This occurred because Kafka would try to use the same port for SASL_PLAINTEXT and PLAINTEXT. By default, Kafka now infers the protocol based on the security settings. This issue affected Cloudera Manager 5.5.2 and higher with Kafka 2.0.0 and higher. Upgrading to Cloudera Manager 5.7.1 and higher upgrades currently configured values to INFERRED unless SSL / TLS is enabled and the values are currently either PLAINTEXT or SASL_PLAINTEXT. This does not cause any change in behavior.

Chart Builder not showing graphs on IE9

In Cloudera Manager 5.7, charts would not render in the Chart Builder page on IE9. This issue is fixed in Cloudera Manager 5.8.

Deb 8.2 support for Cloudera Manager

Cloudera Manager is now supported on Debian 8.2.

ResourceManager would not start if NodeManager were down during the start phase of the restart cycle

All YARN roles are stopped and started together when the service stop or start command is issued with CDH 5.2 and higher. If CDH version is lower than 5.2, the previous behavior of stopping ResourceManagers before NodeManagers and starting them after NodeManagers stays the same.

Cloudera Manager 5 Release Notes

[Default TLS keystore location for HTTPFS is on non-persistent disk](#)

The default location for the HTTPFS TLS / SSL keystore was `/var/run/hadoop-httpfs/.keystore`, which could be deleted upon machine reboot. Newly created clusters now have an empty default. When upgrading to Cloudera Manager 5.7.1 or higher, the old value is maintained. There should be no disruption on upgrade, but Cloudera Manager presents a warning that the keystore is in a dangerous location. To fix this problem, move the files to a safe path on the new host, then update the configuration in Cloudera Manager to point to the new path.

[Active Directory principals created without AES 128/256 bit cause job failures if cluster is configured for AES](#)

It is now possible to configure encryption types for Active Directory setups in Cloudera Manager, using a new property on the Kerberos configuration page, **Kerberos Encryption Types**.

[MR2 counter limits in the Cloudera Manager YARN page should populate all MR2 config files](#)

The MapReduce2 property `mapreduce.job.counters.max` was not included in the configuration for the JobHistory Server, which could cause jobs to fail if there were too many counters. This might happen despite increasing the limit configured in Cloudera Manager. The property is now included in the JobHistory Server configuration, in addition to the related property `mapreduce.job.counters.groups.max`.

[BDR: support use of a custom principal](#)

Cloudera Manager now supports custom Kerberos principals for BDR.

[Add QueryMonitoring chart to Impala service charts](#)

A new chart on the Impala service page shows query duration for completed Impala queries.

[Add symbols to Cloudera Manager generated Active Directory passwords](#)

Cloudera Manager now allows Active Directory password complexity to be configured using a new security configuration on the UI. The following table lists the default values.

Name	Value
length	12
minLowerCaseLetters	2
minUpperCaseLetters	2
minDigits	2
minSpaces	0
minSpecialChars	0
specialChars	? . ! \$ % ^ * () - _ + = ~

You can modify these values and regenerate Active Directory credentials to create new passwords.

[Regenerate principals should delete from Active Directory, too](#)

It is now possible to regenerate principals for Active Directory setups. This involves deletion of existing accounts and regeneration of new principals. Since some customers might not want to do this through Cloudera Manager, starting with Cloudera Manager 5.8 you can enable the new property `Active Directory Delete Accounts on Credential Regeneration`. Regenerating credentials with this setting enabled automatically deletes existing accounts and completes the regeneration. This setting is disabled by default. If disabled, regeneration of principals throws an error message saying that deletion of accounts is required. Cloudera Manager needs the new configuration to be set in order to delete accounts automatically.

[XSS in Kerberos activation](#)

In lower releases, there was an XSS vulnerability on the Kerberos page. This is now fixed.

XSS in host addition

In lower releases, there was an XSS vulnerability on the Add Hosts page. This is now fixed.

XSS in Host Templates

In lower releases, there was an XSS vulnerability on the Host Templates page. This is now fixed.

Make it more obvious when phone home is off

In lower releases, it was not obvious in the Send Diagnostics dialog whether data would be sent back to Cloudera. The dialog is enhanced to make this information more visible.

Allow ad hoc sub-pools to be created within existing pools

Cloudera Manager supports the nestedUserQueue feature in the UI. You no longer have to use the safety valve to specify nestedUserQueues. This means you can make all jobs without specified user queues go to root.<YOUR_POOL>. <username> or root.<YOUR_POOL>. <primaryGroup> from the UI.

Add an option to duplicate resource pool configs

Selecting **Clone** lets you create a new pool from the settings of an existing pool.

Impala Admission Control root pool should have configurable ACLs

Impala Admission Control now supports a global way of editing ACLs.

Do not create host templates when create cluster wizard is run

During cluster creation, host templates are no longer created automatically.

The modal height is not calculated correctly

In lower releases, the modal dialog is sometimes too tall. This is now fixed.

new required fields Key Management Server Proxy Group created if there is more than 1 KMS instance

When adding two Key Trustee KMS roles during the initial setup wizard, sometimes these roles were assigned to different groups. The required configuration was set for some but not all of these groups, causing errors. This is now fixed.

Links should not be shown on print out

When you print pages from Cloudera Manager, the printout no longer displays link URLs.

Allow customization of ldap account properties for AD-Kerberos setup

The Active Directory account properties objectClasses and accountExpires are now configurable from the Kerberos Configuration UI page.

Support bundle has no way to enforce that a certain time is guaranteed to be in bundle

Support bundles can now be collected for a certain time range. Users are also able to get an estimate of the bundle size before collecting the diagnostic data. Only role logs collected as a part of the support bundle are supported by this item in Cloudera Manager 5.8. This item is not available for scheduled support bundles.

Fix Spark CSD to keep client config files in subdir

The Spark CSD was modified to avoid conflicts with other CSDs that depend on it, and causes the Spark service to show up as stale on upgrade to Cloudera Manager 5.7.1 and higher.

Cloudera Manager Agent clears out JN data directories that leads to HDFS not restarting

On RHEL 7 class systems, certain configuration actions intended to be executed only once during enablement of HDFS HA might be re-executed when you request a system shutdown or reboot. This can result in data loss. This issue is

Cloudera Manager 5 Release Notes

fixed in Cloudera Manager 5.8, but a hard restart is required on RHEL 7-class systems (including Oracle and CentOS variants) for the fix to take effect. This can be performed after upgrade in a scheduled, rolling manner.

Hue should not use the embedded sqlite db by default when possible

Hue now has built-in support for PostgreSQL by taking advantage of the system library python-psycopg2. In addition to this library, Hue also includes the system libraries listed in the following table.

System	Library
CentOS 5	Not supported
CentOS 6 and 7	postgresql-libs
Ubuntu 10.04	libpq5, python-egenix-mxdatetime and python-central
Ubuntu 12.04 and 14.04	libpq5
SLES 11 sp2	libpq5

Issues Fixed in Cloudera Manager 5.7.5

The following issues are fixed in Cloudera Manager 5.7.5.

[OOMKiller script does not work for Impala Catalog](#)

Fixed a bug where OutOfMemory errors in the Catalog Server might lead to killing multiple Java processes, including other roles on the same host.

Issues Fixed in Cloudera Manager 5.7.4

The following issues are fixed in Cloudera Manager 5.7.4.

[Agent orphan cleanup removes process dir from in flight process](#)

With this fix, the preventative steps in TSB-181 are no longer required.

[YARN historical reports by user shows pool-user entity](#)

When Cloudera Manager manages multiple clusters, there is no per user tracking for historical applications and queries across clusters. Instead, **Historical Applications by User** and **Historical Queries by User** show applications and queries per user and pool. (A pool is associated with a specific cluster.)

[Host inspector incorrectly warns about kernel version "2.6.32-504.16.2"](#)

Host inspector no longer warns that kernel version "2.6.32-504.16.2" as "non-recommended."

[Fix CatalogServiceClient to handle TLS connections to catalogd for UDF replication](#)

When Impala uses SSL, Cloudera supports TLS Connection to Catalog Server. You can enable replication for any Impala UDFs/Metadata (in Hive Replication).

[If total_space_bytes is really big, heartbeats fail](#)

Fixes heartbeat failure with a host that has a mount point backed by cloud storage such as AWS.

[Oozie points to older sharelib even after running sharelib install command](#)

After an "Install Oozie Share Lib" action, the Oozie service is informed that there is a new shared lib installed. This eliminates the need for a separate manual restart.

[Cloudera Manager blocks from HS2 enabling both LDAP and Kerberos auth](#)

HS2 supports LDAP and Kerberos authentication on the same instance for CDH 5.7.0 or higher. Previously, this was considered an error.

Hive Replication Command should update copy the Serde properties correctly

Hive Replication replicates the Serde Properties and also copies corresponding hdfs file.

OOMKiller script does not work for Impala Catalog

Fixed a bug where OutOfMemory errors in the Catalog Server could lead to killing multiple java processes, including other roles on the same host.

Issues Fixed in Cloudera Manager 5.7.2

Unable to start Hue on cluster that's using Kerberos and Isilon

Hue service can now start with Isilon if Kerberos is enabled.

BDR UI: Add Search back - It is difficult to find the desired replication schedule in the Cloudera Manager 5.5 UI

The Replications page provides enhanced search functionality to filter schedules by any specified text. Searches occur within any of these schedule fields: HDFS paths, database names and table names.

Handle drop-recreate partition efficiently

With this change, properties of entities (database, table, partition, index) are not updated by default. You can choose to update properties by setting REPLICATE_PARAMETERS=true in **Hive Replication Environment Advanced Configuration Snippet (Safety Valve)**.

Hue static directory is browsable

The static web directory in Hue is no longer indexed and browsable when the Hue Load Balancer is used.

XSS in Kerberos activation

In lower releases, there was an XSS vulnerability on the Kerberos page. This is now fixed.

XSS in host addition

In lower releases, there was an XSS vulnerability on the Add Hosts page. This is now fixed.

XSS in Host Templates

In lower releases, there was an XSS vulnerability on the Host Templates page. This is now fixed.

Cloudera Manager Agent clears out JN data directories that leads to HDFS not restarting

On RHEL 7 class systems, certain configuration actions intended to be executed only once during enablement of HDFS HA might be re-executed when you request a system shutdown or reboot. This can result in data loss. This issue is fixed in Cloudera Manager 5.7.2, but a hard restart is required on RHEL 7-class systems (including Oracle and CentOS variants) for the fix to take effect. This can be performed after upgrade in a scheduled, rolling manner.

Separation of authentication and authorization coprocessor configs in HBase

HBase Secure Bulkload is now enabled for all CDH5.5 and higher clusters, regardless of whether Kerberos is enabled. Also fixed related issue where clusters with authentication (kerberos) but not authorization failed in Hbase-related MapReduce jobs.

Adding a new service changes mapreduce.job.reduces setting

Adding a new service in Cloudera Manager no longer changes the YARN setting mapreduce.job.reduces.

BDR: Hive replication status always shows "(Dry Run)"

Running a Hive replication schedule in BDR no longer displays "(Dry Run)" in the status.

Cloudera Manager 5 Release Notes

BDR Replication Schedule page is slow to load

Replication schedules page load performance is improved.

HDFS Snapshot running on unavailable nodes

When selecting a host on which to run the **HDFS Snapshot** command, Cloudera Manager now excludes unavailable hosts, such as hosts in maintenance mode or decommissioned hosts.

Migrating Cloudera Manager using deployment.json fails if replication schedules are configured

Migrating Cloudera Manager using `deployment.json` no longer fails if replication schedules are configured.

Files excluded from replication are not replicated if they are renamed

If a file excluded from replication by an exclusion filter is renamed, it is now replicated properly.

Issues Fixed in Cloudera Manager 5.7.1

Spark standalone does not come up if HDFS is not available

The Spark standalone service now works without an HDFS service. This requires Spark services to show up as stale and require a restart after upgrade to Cloudera Manager 5.7.1 and higher.

Kafka unable to start due to listener misconfiguration when Kerberos is enabled

Kafka would not start when Kerberos was enabled and the `security.inter.broker.protocol` default was not changed. This occurred because Kafka would try to use the same port for SASL_PLAINTEXT and PLAINTEXT. By default, Kafka now infers the protocol based on the security settings. This issue affected Cloudera Manager 5.5.2 and higher with Kafka 2.0.0 and higher. Upgrading to Cloudera Manager 5.7.1 and higher upgrades currently configured values to INFERRRED unless SSL / TLS is enabled and the values are currently either PLAINTEXT or SASL_PLAINTEXT. This does not cause any change in behavior.

Default TLS keystore location for HTTPS is on non-persistent disk

The default location for the HTTPS TLS / SSL keystore was `/var/run/hadoop-https/.keystore`, which could be deleted upon machine reboot. Newly created clusters now have an empty default. When upgrading to Cloudera Manager 5.7.1 or higher, the old value is maintained. There should be no disruption on upgrade, but Cloudera Manager presents a warning that the keystore is in a dangerous location. To fix this problem, move the files to a safe path on the new host, then update the configuration in Cloudera Manager to point to the new path.

Fix Spark CSD to keep client config files in subdir

The Spark CSD was modified to avoid conflicts with other CSDs that depend on it, and causes the Spark service to show up as stale on upgrade to Cloudera Manager 5.7.1 and higher.

Cloudera Manager HDFS usage reports do not include Inode references

Lower versions of Cloudera Manager HDFS usage reports do not include Inode references. As a result, usage reports underreported HDFS directory sizes and data used by users and groups in certain circumstances where HDFS snapshots were used.

Kafka MirrorMaker unable to start due to KAFKA_HOME not being set

Kafka MirrorMaker would not start when Kafka is installed using packages. This occurred because `KAFKA_HOME` was not set to the correct default when starting MirrorMaker. This issue affected Cloudera Manager 5.4.0 and higher with Kafka 1.4.0 and higher.

Authentication errors occur due to missing SAML metadata

- If you use SAML for external authentication and were on Cloudera Manager 5.5.0 and higher prior to this upgrade and if you used to notice an error screen while logging out, then this upgrade will fix that issue.

- If you use SAML for external authentication and were on Cloudera Manager 5.5.0 and higher prior to this upgrade and if you did not notice any error screen while logging out, then you will most likely see an error screen while logging out after this upgrade. In order to fix that, you can follow either of these steps:
 1. Update the metadata file in your IdP with the new file from <Cloudera Manager Server>/saml/metadata
 2. Change SAML Entity Alias under **Administration > Setting "clouderaManager"** to " and restart Cloudera Manager.

Child commands for deleting or adding a nameservice show stack trace

In an existing HDFS deployment with high availability, when you try to add or delete a nameservice and attempt to view the progress of the child commands, a stack trace is triggered if some of the child commands have not yet run. This fix eliminates the stack trace and informs you that the child commands have not yet been run.

HiveServer2 Web UI did not use SSL when Kerberos was enabled

SSL configuration for the HiveServer2 Web UI is now used regardless of whether Kerberos is in use.

Clusters menu expands to last cluster viewed

Previously, the **Clusters** menu expanded the first cluster by default, and as you expand or collapse the menu, Cloudera Manager remembers that cluster for the session. However, when you go to services, roles, or hosts of another cluster, Cloudera Manager does not remember the other cluster and shows the previously expanded cluster instead.

In release 5.7.1 and higher, Cloudera Manager remembers the last cluster viewed by a user and expands that cluster in the **Clusters** menu by default.

Impala does not throw null pointer exception when memory limit is not set

If the configuration property **Impala Daemon Memory Limit** was not set, the **Impala Admission Control** page threw a null pointer exception.

HDFS rolling restart fails after CDH upgrade

In previous Cloudera Manager releases, when one of a pair of highly available NameNodes was down, it was possible for rolling restart or rolling upgrade commands to fail with an error message incorrectly describing the state of the NameNode as "Busy." The error message now correctly identifies the state of the NameNode (typically "Stopped" in this situation).

The Expand Range option did not work for some charts

The **Expand range to fill all values** option in **Chart Builder** now works for all charts.

Kafka unable to start due to misconfigured security.inter.broker.protocol when Kerberos is enabled

Kafka would not start when Kerberos is enabled and the default value of `security.inter.broker.protocol` was not changed. This occurred because Kafka tried to use the same port for `SASL_PLAINTEXT` and `PLAINTEXT`. By default, Cloudera Manager now infers the protocol based on the security settings.

This issue affected Cloudera Manager 5.5.2 and higher with Kafka 2.0.0 and higher.

Upgrading to Cloudera Manager 5.7.1 or higher upgrades currently configured values to `INFERRED` unless SSL/TLS is enabled and the values are currently either `PLAINTEXT` or `SASL_PLAINTEXT`. This does not cause any change in behavior.

Oozie JVM heap metrics not reported in Chart Builder for some services

Oozie JVM metrics are now available and display on the role page. They can also be accessed through **Chart Builder** using the `oozie_memory_heap_used` and `oozie_memory_total_max` metrics.

Cloudera Manager 5 Release Notes

Spark CSD modified

The Spark CSD was modified to avoid conflicts with other CSDs that depend on it. This change causes the Spark service to require a restart when upgrading to Cloudera Manager 5.7.1.

Poorly formed Advanced Configuration Snippets cause null pointer exception with diagnostic bundles

Certain poorly formed Advanced Configuration Snippets could cause a Null Pointer Exception when uploading diagnostic bundles and setting up a Cloudera Manager peer.

Setting owner of a file in Isilon fails

On Isilon systems, the owner that the file is being changed to must be present on the system. In general cases, the user is not present, so this command fails with an error message suggesting that the user is not part of the supergroup. This fix addresses the issue by not failing the command.

The Install Oozie ShareLib command is now visible to users with the Configurator role.

Default location for TLS Keystore for HTTPFS is nonpersistent

The default location for the HTTPFS TLS / SSL keystore was `/var/run/hadoop-httpfs/.keystore`, which could be deleted when the host reboots. Newly created clusters now have an empty default instead of one that could be deleted. When upgrading to Cloudera Manager 5.7.1 or higher, the old value is maintained, so there is no disruption on upgrade. However, Cloudera Manager warns that the path is in a dangerous location. To fix this problem, move the files to a safe path on that host, and then update the configuration in Cloudera Manager to point to the new path.

Collecting diagnostic bundle displayed Java stack trace

Collecting a diagnostic bundle for a Hive replication schedule caused a Java stack trace to be shown on the page. This fix shows an error message instead of throwing a Java stack trace.

Unable to stop Cloudera Manager Agent on SLES 11

Fixes TSB-144.

Running the restart or stop service commands failed to stop the Agent.

Error creating bean

Occasionally, some users encountered the message `Error creating bean with name 'newServiceHandlerRegistry'` in the Cloudera Manager Admin Console. This issue has been resolved.

Impala JVM heap size is configurable

The JVM heap size of the Impala catalog server can be configured now using the **Java Heap Size of Catalog Server in Bytes** property. The property defaults to 4 GB, and like all memory parameters may require tuning.

Issues Fixed in Cloudera Manager 5.7.0

Plain-text passwords sent to users

An authenticated user could request and receive documents that included passwords they were permitted to modify. This information was sent as plain text. Passwords are no longer included in request responses.

Cloudera Manager forces Solr shutdown before operations complete

When stopping the Solr service, Cloudera Manager would forcibly stop the service after a period of time. This could result in Solr cores not coming up cleanly. Cloudera Manager now waits for all operations to complete on the Solr server before exiting.

Re-enabling Kerberos fails due to duplicate roles

As part of enabling Kerberos, a role is created. Attempts to re-enable Kerberos failed because the role already existed from when Kerberos was enabled before. When Kerberos is enabled, Cloudera Manager reuses any required roles if they already exist.

Processes used incorrect `HOME` variable

CDH service processes and third-party (CSD) processes used an incorrect `HOME` variable. These process now run with the `HOME` environment variable set to the correct home directory based on the process user.

Stale Kerberos configuration reported after deploying Kerberos client configuration

After making Kerberos configuration changes through **Administration > Settings > Kerberos**, and **Manage krb5.conf** is enabled, the configuration issue 'Cluster has stale Kerberos configuration' might have displayed and might not have disappeared after running **Cluster > cluster_name > Deploy Kerberos Client Configuration**. The deploying Kerberos client configuration action now waits for pending staleness checks before identifying stale hosts on which to apply Kerberos configuration.

Workaround: Make a different, non-Kerberos edit to a configuration, and save that change. Revert that change immediately afterward.

Time range settings on report pages are incorrect

Time range settings were incorrect. They are now set correctly on report pages.

Add Service Wizard fails to set Hive on Spark performance tuning parameters

Automatic configuration for Hive on Spark performance tuning parameters did not run when adding a Hive service to an existing cluster. In the past, these tuning parameters ran when adding a cluster that contained Hive, YARN, and Spark on YARN. The rules now run as long as Hive depends on YARN in both the add service and add cluster wizards.

Setting empty values for LDAP Base DN and LDAP Domain produces errors

Setting empty values for the LDAP Base DN and LDAP Domain for Impala resulted in errors. These settings are now handled without errors.

Restart of deleted roles does not wait until deleted roles are identified

Restart and rolling restart commands did not always wait until all deleted roles were identified. As a result, services that had not yet been identified as requiring a restart were not restarted. Restart and rolling restart commands now wait for staleness checks to complete, if the user requests that only stale services be restarted. This avoids a race between staleness computation checking and how services are determined to be stale and thus restarted.

Miscellaneous problems with the Replication user interface

This release includes several fixes to the replication user interface including the following:

- The **Last Run** column incorrectly sorted dates. Dates are now correctly sorted.
- Collecting diagnostic data for failed Hive Replication commands failed with an error. This data collection now succeeds.
- Finished schedules were shown as running. Schedules now accurately reflect their state.
- Scheduled time was incorrectly translated between browser time and server time. Scheduled time is now correctly translated.
- The **Actions** menu would be disabled while a replication schedule was running, which blocked changing future runs configurations. The **Actions** menu is now enabled for replication schedules with commands running.

Misleading error occurs while deploying client configuration

The error "There is already a pending command on this entity" was shown if a command was in process for a service and an attempt is made to run the "Deploy client configuration" command. This error is no longer shown.

Cloudera Manager 5 Release Notes

Gathering task progress produced null pointer exceptions

Gathering progress information on a variety of tasks could result in null pointer exceptions. For example, this could occur when decommissioning a host. Progress information is now gathered as expected.

User interface elements are hidden when windows are resized

Some windows could be resized so buttons were not visible. For example, this could happen with the Create Replication dialog box. The user interface now handles resizing as expected.

HDFS data transfers uses 3DES despite configuration

The configuration information for using the AES/CTR/NoPadding cipher suite for HDFS data transfers was incomplete. As a result, traffic was encrypted with the much slower 3DES algorithm. The correct configuration is now included with the HDFS client.

Solr keystore passwords are no longer presented in clear text

Because of Tomcat restrictions, Solr keystore passwords were sent as clear text on the machines running the service. They are now redacted.

Removing a host from a cluster removes all Kerberos client configuration

Removing a host from a cluster automatically removed all Kerberos client configuration from any other hosts still in the cluster. Now, when one host is removed from a cluster, any Kerberos client configuration on other hosts is unaffected.

Hive replication import fails to include some information

During the Hive replication import phase, schema and location information was not consistently populated. This information is now populated as expected.

Restarts attempt to deploy client configuration when action is not supported

After upgrading a parcel, restart or rolling restart steps automatically attempted to deploy the client configuration, even if that was not supported by the cluster configuration. Now, client configurations are deployed only if the cluster configuration supports this action.

Some pages load very slowly

Some pages, such as the **All Recent Commands** page, may take over 30 seconds to load. This process has been optimized, reducing load times.

`kt_renewer` not automatically created when Hue is added to a Kerberos-enabled cluster

When the Hue service was added to a Kerberos-enabled cluster, a single corresponding `kt_renewer` was not created. The process of adding the Hue service now includes checking if a keytab renewer is required, and creating one if it is.

Some jobs never run due to memory configuration

It was possible to configure the YARN NodeManager with an amount of available memory less than the amount of memory available to the YARN container. In such a case, a job might never find a NodeManager that meets the memory requirements. The system now ensures that at least one YARN container is configured with an equal or greater amount of memory than the YARN NodeManager value.

Replication schedule API not compatible with older versions

Clients using the version 10 replication schedule API did not work as expected with instances of Cloudera Manager using version 11 of the API. This meant that clients from Cloudera Manager 5.4.0 and lower did not work as expected with servers running Cloudera Manager 5.5.0 and higher. Clients and servers using these different API versions now function as expected.

Issues Fixed in Cloudera Manager 5.6.1

Scheme and location not filled in consistently during Hive replication import

In previous releases, Hive replication import phase did not consistently fill in scheme and location information. This information is now filled in as expected.

Cloudera Manager HDFS usage reports do not include Inode references

Lower versions of Cloudera Manager HDFS usage reports do not include Inode references. As a result, usage reports underreported HDFS directory sizes and data used by users and groups in certain circumstances where HDFS snapshots were used.

Kafka MirrorMaker unable to start due to KAFKA_HOME not being set

Kafka MirrorMaker would not start when Kafka is installed using packages. This occurred because KAFKA_HOME was not set to the correct default when starting MirrorMaker. This issue affected Cloudera Manager 5.4.0 and higher with Kafka 1.4.0 and higher.

kt_renewer not automatically created when Hue is added to a Kerberos-enabled cluster

When the Hue service was added to a Kerberos-enabled cluster, a single corresponding kt_renewer was not created. The process of adding the Hue service now includes checking if a keytab renewer is required, and creating one if it is.

Authentication errors occur due to missing SAML metadata

- If you use SAML for external authentication and were on Cloudera Manager 5.5.0 and higher prior to this upgrade and if you used to notice an error screen while logging out, then this upgrade will fix that issue.
- If you use SAML for external authentication and were on Cloudera Manager 5.5.0 and higher prior to this upgrade and if you did not notice any error screen while logging out, then you will most likely see an error screen while logging out after this upgrade. In order to fix that, you can follow either of these steps:
 1. Update the metadata file in your IdP with the new file from <Cloudera Manager Server>/saml/metadata
 2. Change SAML Entity Alias under **Administration > Setting** "clouderaManager" to " and restart Cloudera Manager.

Oozie JVM heap metrics not reported in Chart Builder for some services

Oozie JVM metrics are now available and display on the role page. They can also be accessed through **Chart Builder** using the oozie_memory_heap_used and oozie_memory_total_max metrics.

Issues Fixed in Cloudera Manager 5.6.0

CDH upgrade fails if the GPL Extras parcel in use

CDH and GPL Extras parcel versions must match exactly. When upgrading CDH, by default Cloudera Manager validates the dependency between the new CDH parcel and existing GPL Extras parcel. Since the dependency is not satisfied, the check returns an error and the upgrade fails.

Workaround:

1. Deactivate parcel dependency checking:
 - a. Select **Administration > Settings**.
 - b. Search for **Validate Parcel Relations**.
 - c. Deselect the checkbox.
 - d. Click **Save Changes** to commit the changes.
2. Deactivate the GPL Extras parcel.
3. Download, distribute, and activate the GPL Extras parcel that matches the CDH upgrade version.
4. Upgrade CDH.

5. Reactivate parcel dependency checking.

Issues Fixed in Cloudera Manager 5.5.5

Issues Fixed in Cloudera Manager 5.5.4

[Cluster provisioning fails](#)

In some cases, provisioning of a cluster may fail at the start of the process. This does not happen in all cases and is mainly noticed on RHEL 6 and especially when some hosts are reporting bad health.

Releases affected: 5.5.0-5.5.3, 5.6.0-5.6.1, 5.7.0

Releases containing the fix: 5.5.4, 5.7.1

For releases containing the fix, parcel activation and first run command now completes as expected, even when some hosts report bad health.

This issue is fixed in Cloudera Manager 5.5.4 and 5.7.1 and higher.

[Scheme and location not filled in consistently during Hive replication import](#)

In previous releases, Hive replication import phase did not consistently fill in scheme and location information. This information is now filled in as expected.

[Kafka MirrorMaker unable to start due to KAFKA_HOME not being set](#)

Kafka MirrorMaker would not start when Kafka is installed using packages. This occurred because KAFKA_HOME was not set to the correct default when starting MirrorMaker. This issue affected Cloudera Manager 5.4.0 and higher with Kafka 1.4.0 and higher.

[Cloudera Manager HDFS usage reports do not include Inode references](#)

Lower versions of Cloudera Manager HDFS usage reports do not include Inode references. As a result, usage reports underreported HDFS directory sizes and data used by users and groups in certain circumstances where HDFS snapshots were used.

[New Hive tables fail to replicate when Sentry Sync is enabled](#)

When Sentry Sync is enabled, new Hive tables failed to replicate. Replication now occurs as expected.

[kt_renewer not automatically created when Hue is added to a Kerberos-enabled cluster](#)

When the Hue service was added to a Kerberos-enabled cluster, a single corresponding kt_renewer was not created. The process of adding the Hue service now includes checking if a keytab renewer is required, and creating one if it is.

[Authentication errors occur due to missing SAML metadata](#)

- If you use SAML for external authentication and were on Cloudera Manager 5.5.0 and higher prior to this upgrade and if you used to notice an error screen while logging out, then this upgrade will fix that issue.
- If you use SAML for external authentication and were on Cloudera Manager 5.5.0 and higher prior to this upgrade and if you did not notice any error screen while logging out, then you will most likely see an error screen while logging out after this upgrade. In order to fix that, you can follow either of these steps:
 1. Update the metadata file in your IdP with the new file from <Cloudera Manager Server>/saml/metadata
 2. Change SAML Entity Alias under **Administration > Setting** "clouderaManager" to " and restart Cloudera Manager.

[HDFS rolling restart fails after CDH upgrade](#)

In previous Cloudera Manager releases, when one of a pair of highly available NameNodes was down, it was possible for rolling restart or rolling upgrade commands to fail with an error message incorrectly describing the state of the

NameNode as "Busy." The error message now correctly identifies the state of the NameNode (typically "Stopped" in this situation).

Oozie JVM heap metrics not reported in Chart Builder for some services

Oozie JVM metrics are now available and display on the role page. They can also be accessed through **Chart Builder** using the `oozie_memory_heap_used` and `oozie_memory_total_max` metrics.

Issues Fixed in Cloudera Manager 5.5.3

Users using external LDAP authentication with no local Cloudera Manager user role explicitly set may default to the read-only role when upgrading to Cloudera Manager 5.5.2

When upgrading to Cloudera Manager 5.5.2, customers who have non-read-only roles configured through LDAP, and have not explicitly set Cloudera Manager local roles, may lose their Cloudera Manager privileges set by LDAP.

Releases affected: Cloudera Manager 5.5.2

Users affected: Customers who use LDAP for Cloudera Manager user authorization and have upgraded Cloudera Manager from a version lower than 5.5.0 to Cloudera Manager 5.5.2. For example:

- **May be affected:** Install Cloudera Manager 5.3 -> Upgrade to Cloudera Manager 5.5.1 -> Upgrade to Cloudera Manager 5.5.2
- **Unaffected:** Install Cloudera Manager 5.5.1 -> Upgrade to Cloudera Manager 5.5.2

Users not affected:

- Customers who installed Cloudera Manager 5.5.0 and higher and upgraded.
- Customers who use Cloudera Manager local role authorization, regardless of upgrade path and version.

Severity: High

Action required: If you have upgraded to Cloudera Manager 5.5.2 and cannot log in with proper permissions, do the following:

1. Resolve any conflicting user authorization permissions between LDAP and Cloudera Manager local permissions.
2. Contact Cloudera Support for further instructions if you cannot resolve conflicting LDAP and Cloudera Manager user permissions.

If you have not yet upgraded to Cloudera Manager 5.5.2, and are using LDAP user authorization:

1. Before upgrading, resolve any conflicting user authorization permissions between LDAP and Cloudera Manager local permissions.
2. Upgrade to Cloudera Manager 5.5.3 or higher.

Issues Fixed in Cloudera Manager 5.5.2



Note: If you are upgrading from a previous version of Cloudera Manager, Cloudera recommends that you upgrade to version 5.5.3 or higher.

Oozie and HttpFS keystore passwords are no longer presented in clear text

Because of Tomcat restrictions, the Oozie and HttpFS keystore passwords were sent as clear text on the machines running the services. They are now hidden.

Cross-site scripting vulnerability using malformed strings in the Parcel Remote URLs list

An attacker could set a malformed string in the Parcel Remote URLs list in the database and trigger the attack when a user accesses the Administration Settings page. This attack is now prevented.

Cloudera Manager 5 Release Notes

Starting/stopping roles for Flume instance succeeds but displays nothing in popup

In Cloudera Manager 5.5.0, running a Flume start/stop service command would succeed, but display an empty popup. This is now fixed.

Role process commands missing stderr and stdout in command details

In Cloudera Manager 5.5, certain commands did not show links to stderr or stdout in the Cloudera Manager UI even if they were executed recently. These could still be found in /var/run/cloudera-scm-agent/process/ on that host. In Cloudera Manager 5.5.2 stderr and stdout should appear as they did before.

Note that links to stderr and stdout for commands may disappear if another command is run on that role. This is expected. The logs can still be found in /var/run/cloudera-scm-agent/process/ on that host.

Updating the Hive NameNode location multiple times could lead to data corruption

Multiple updates to the Hive NameNode location could cause Hive Metastore database corruption. Issuing the same command multiple times no longer produces problems.

Cloudera Manager skips NameNode logs in the diagnostic bundle

Scheduled diagnostic bundles did not include recent role logs. Diagnostic bundles collected manually (not scheduled) worked as expected. Now, scheduled diagnostic bundles include the latest NameNode logs.

Kafka 2.0 fails to deploy on large clusters as reserved.broker.max.id defaults to 1000

Large Kafka clusters would not start when Cloudera Manager-generated broker IDs exceeded the value set by reserved.broker.max.id. The default value of broker.id.generation.enable has now been set to false to disable the reserved.broker.max.id configuration property and avoid collisions.

Cloudera Manager fails to propagate HBase coprocessors to the gateway nodes

Cloudera Manager does not propagate HBase coprocessors to the gateway nodes. As a result, tools that depend on the HBase security subsystem, such as the loadIncrementalHFiles tool, do not use security features, even in secure environments.

Workaround: Add the following properties to the **HBase Client Advanced Configuration Snippet (Safety Valve) for hbase-site.xml** and restart all HBase clients:

```
<property>
  <name>hbase.coprocessor.region.classes</name>
  <value>org.apache.hadoop.hbase.security.token.TokenProvider,org.apache.hadoop.hbase.security.access.SecureBulkLoadEndpoint</value>
</property>
```

HDFS nameservices API call returns incorrect HA role status

The HDFS nameservices API returned incorrect active/standby status information for HA roles. API calls made about roles that were active might return standby status and roles that were in standby status might return active status. Information about host statuses is now accurate.

Hive requires the hive principal for HiveServer2 host as well as load balancer

An issue with HiveServer2 missing its principal and keytab when Hive is load-balanced has been fixed.

More descriptive error message for service trying to start on a decommissioned host

Cloudera Manager now displays a more descriptive error message when it skips the "Start" command because all roles are started or decommissioned or on a decommissioned host.

UI shows repeated errors when loading replications page

Issue fixed with loading the replication page when a previous replication command fails without launching a MapReduce job.

[Allow option for external users to be assigned roles in the local database](#)

In Cloudera Manager 5.5, role assignments for external users was disabled, which caused upgrade issues. The fix rolls back the change but instead of using a union approach, implements the following precedence rules:

- If a user is assigned a role in Cloudera Manager, this local role is used.
- Otherwise, a user's LDAP group association determines the user role.

[Clean up usercache directories on migration from unsecure to secure mode](#)

Fixes an issue that led to YARN jobs failing after migration from unsecure to secure mode.

[Spark REST API does not work when parcels are used](#)

The REST API for retrieving data from a live Spark UI or from the Spark History Server has been fixed.

In secure clusters, DataNode fails to start when `dfs.data.transfer.protection` is set and DataNode ports are changed to unprivileged ports

Before this fix, the only way to run the DataNode on unprivileged ports (port number > 1024) in a Kerberized cluster with DataNode Data Transfer Protection enabled, was to use single-user mode. Now this configuration works for both regular and single-user mode installs.

Both Hadoop SSL and DataNode Data Transfer Protection are still required for unprivileged DataNode ports to work in a Kerberized cluster. This configuration is supported only in CDH 5.2 and higher.

[New validation warning for non-recommended secure DataNode configurations in CDH 5.2 and higher](#)

On a Kerberos-enabled cluster running CDH 5.2 and higher, there are two recommended DataNode configurations. Use SASL/TLS with DataNode Data Transfer Protection enabled to encrypt the connection, or use only privileged ports to communicate. The supported combinations of HDFS configuration properties follow:

- **Security through SASL/TLS (preferred):**
 - DataNode Data Transfer Protection - Enabled
 - Hadoop TLS/SSL - Enabled
 - DataNode Transceiver Port - Non-privileged (that is, port number ≥ 1024)
 - Secure DataNode Web UI Port (TLS/SSL) - Non-privileged (that is, port number ≥ 1024)
- **Security through privileged ports:**
 - DataNode Data Transfer Protection - Disabled
 - Hadoop TLS/SSL - Disabled
 - DataNode Transceiver Port - Privileged (that is, port number < 1024)
 - DataNode HTTP Web UI Port - Privileged (that is, port number < 1024)

Any configuration other than these results in a validation warning or error from Cloudera Manager. In particular, the following configuration, which is allowed by HDFS but is not recommended, results in a (dismissible) validation warning:

- **(Not Recommended) Security without enabling DataNode Data Transfer Protection:**
 - DataNode Data Transfer Protection - Disabled
 - Hadoop TLS/SSL - Enabled
 - DataNode Transceiver Port - Privileged (that is, port number < 1024)
 - Secure DataNode Web UI Port (TLS/SSL) - Non-privileged (that is, port number ≥ 1024)

All configurations other than the three listed result in Cloudera Manager displaying a validation error.

[Sensitive environment parameters not redacted for CSDs](#)

Passwords in environment variables for CSDs are now redacted.

Cloudera Manager 5 Release Notes

Update Apache Commons Collections library in Cloudera Manager due to major security vulnerability

The Apache Commons Collections library has been upgraded to 3.2.2 to fix a critical security vulnerability.

Remove plaintext keystore password from /api/v6/cm/config

With the addition of the JVM parameter, `\-Dcom.cloudera.api.redaction=true`, sensitive configuration values are redacted from the API.

JVM parameter to redact passwords now redacts the password salt and hash

When API redaction is turned on using the JVM argument `-Dcom.cloudera.api.redaction=true`, it also redacts the user's pwHash and pwSalt values. Passwords for Cloudera Manager Peers are also redacted.

Oozie keystore and truststore passwords now redacted

Oozie's Java keystore and truststore passwords are no longer sent in clear text on the command line.

Several Replication UI fixes

- The Replication UI **Last Run** column now sorts correctly based on dates.
- Collecting diagnostic data for failed Hive Replication commands no longer fails.
- Finished schedules are no longer shown as running when the user is watching the page.
- Scheduled time now accurately translates between browser time and server time.
- Actions menu is now enabled for replication schedules with commands running. Previously, it would be disabled while a replication schedule was running, which blocked changing the configuration for future runs.

Fix Java detection

Java detection in the "Components" view for a host was fixed to account for Java versions installed using symlinks in `/usr/java` (such as `/usr/java/default`). A similar fix was made to the host inspector's Java detection logic.

When Host Monitor is stopped, cluster/services status in Cloudera Manager API returns Good instead of N/A or Unknown

If the Host Monitor is down, the service details page will still be able to present non-host related health status.

Issues Fixed in Cloudera Manager 5.5.1

Apache Commons Collections deserialization vulnerability

Cloudera has learned of a potential security vulnerability in a third-party library called the [Apache Commons Collections](#). This library is used in products distributed and supported by Cloudera ("Cloudera Products"), including core Apache Hadoop. The Apache Commons Collections library is also in widespread use beyond the Hadoop ecosystem. At this time, no specific attack vector for this vulnerability has been identified as present in Cloudera Products.

In an abundance of caution, we are currently in the process of incorporating a version of the Apache Commons Collections library with a fix into the Cloudera Products. In most cases, this will require coordination with the projects in the Apache community. One example of this is tracked by [HADOOP-12577](#).

The Apache Commons Collections potential security vulnerability is titled "Arbitrary remote code execution with InvokerTransformer" and is tracked by [COLLECTIONS-580](#). MITRE has not issued a CVE, but related [CVE-2015-4852](#) has been filed for the vulnerability. CERT has issued [Vulnerability Note #576313](#) for this issue.

Releases affected: CDH 5.5.0, CDH 5.4.8 and lower, CDH 5.3.8 and lower, CDH 5.2.8 and lower, CDH 5.1.7 and lower, Cloudera Manager 5.5.0, Cloudera Manager 5.4.8 and lower, Cloudera Manager 5.3.8 and lower, and Cloudera Manager 5.2.8 and lower, Cloudera Manager 5.1.6 and lower, Cloudera Navigator 2.4.0, Cloudera Navigator 2.3.8 and lower.

Users affected: All

Impact: This potential vulnerability may enable an attacker to execute arbitrary code from a remote machine without requiring authentication.

Immediate action required: Upgrade to Cloudera Manager 5.5.1 and CDH 5.5.1, Cloudera Manager 5.4.9 and CDH 5.4.9, Cloudera Manager 5.3.9 and CDH 5.3.9, and Cloudera Manager 5.2.9 and CDH 5.2.9, and Cloudera Manager 5.1.7 and CDH 5.1.7.

Issues Fixed in Cloudera Manager 5.5.0

Setting the HBase WAL provider to the default in Cloudera Manager causes the RegionServer process to fail to start

Setting the HBase WAL provider to the default using Cloudera manager erroneously sets the value of `hbase.wal.provider` to `default`, when it should be set to `defaultProvider`. This causes the RegionServer process to fail to start.

Workaround: Do not use Cloudera Manager to set the WAL provider to the default. Instead, add the following properties to the **RegionServer Advanced Configuration Snippet (Safety Valve) for `hbase-site.xml`** and restart all HBase clients.

```
<property>
  <name>hbase.wal.provider</name>
  <value>defaultProvider</value>
</property>
```

Incorrect path format causes access permission failure

The file path format used in sequence file `ing was incorrect. In replications involving a large number of files and when the **Delete Policy** for the replication is set to **Delete to trash** or **Delete Permanently**, distcp uses the local file system to save the intermediate result of sequence file sorting. An incorrect path format causes access permission failure.

Out-of-memory exception for Hive replication

Hive replication throws an out-of-memory exception when exporting a table with a large number of partitions.

Incorrect value emitted into `hbase-site.xml`

When configuring an HBase WAL provider with the "HBase Default" option, Cloudera Manager emits an incorrect value into `hbase-site.xml` and HBase reports an error.

Failed replication no longer fail silently

When a replication is attempted between secure and insecure clusters, the replication reports an error and fails silently.

The Replication Schedules page displays error after failed replication

A dialog box that shows a Java exception displays in the Replication Schedules page after a failed replication and the page is inaccessible.

Cloudera Manager now allows you to use '/' in cluster names

In previous versions, this resulted in problems during replication because '/' was treated as an URL path.

HDFS replication fails because of improper snapshot directory handling

Replications fail when converting a snapshot path to a regular path when the grandparent of the source directory is snapshottable.

Renewal time limits and lifetime limits are removed for Kerberos tickets

Snapshots that take longer than 30 minutes failed because the Kerberos tickets for the snapshots expired too soon . The renewal and lifetime limits have been replaced with the system default lifetime and renewal limits, instead of 30 minutes.

Replication schedules fails to catch configuration errors during creation

The schedule configuration is not validated when it is created, which causes errors during replication.

Cloudera Manager 5 Release Notes

Audit events now include the schedule ID

The replication ID has been added to the schedule creation, update, and deletion audit events.

OutOfMemory error when running HDFS replication

HDFS replication fails with an OutOfMemory exception (in `stderr.log`) when an HDFS replication job replicates a large number of files.

Hive replication does not preserve user and group names for the database directory

File permissions of the database directory that maps to a Hive database are not preserved.

Exception in configuration history after changing host configuration

An IllegalStateException is reported in the History and Rollback page after changing host configuration.

Host Inspector no longer expects the impala user to be in the hdfs group

To conform to security best practices, the impala user should not be in the hdfs group.

The DataNode Refresh Data Directories command fails on secure clusters

DataNode refresh failed on Kerberized clusters.

Topology should only include hosts with Hadoop daemons

A topology map is created for the services HDFS, YARN, and MapReduce. In releases earlier than 5.5.0, hosts with only gateway roles for these services were also added to the topology map. This might erroneously mark many configurations as stale, requiring restarts or refreshes to resolve the staleness.

Agent host_id property is erroneously set to hostname

Previously, attempting to set the `listening_hostname` property in the `Agent config.ini` file (which is not normally necessary) changed the Agent's host ID to use this hostname, instead of the normal value.

JAVA_HOME override setting does not affect component list

The component list for a host is not affected by a custom `JAVA_HOME` setting.

Monitoring performance issues on large, busy clusters

A number of fixes have been made to improve Service Monitor and Host Monitor performance, particularly problems manifesting as large, regular garbage-collection pauses, on large, busy clusters.

JAVA_HOME does not get passed to Agents

Client configuration deployment fails to locate Java in the custom `JAVA_HOME` environment variable, if that is specified via host configuration.

Job History Server Retaining Logs in Secure Clusters

The Job History Server fails to delete old logs from HDFS on secure clusters, with the error "Failed to specify server's Kerberos principal name."

On Kerberized clusters, incorrect values reported for DataNode process metrics

The DataNode process is incorrectly monitored on Kerberized clusters.

During kt_renewer kinit call authentication may fail

On Kerberized clusters, Cloudera Manager might periodically report monitoring failures due to authentication errors when attempting to communicate with various role web servers to collect metrics. This is due to synchronization issues between the Agent's calls to kinit and attempts to perform Kerberos authentication with the role web servers. This

has been fixed by adding logic to retry requests for metrics in the Agent a number of times on authentication failures. These retries will be logged but should not result in health failures and alerts.

Client configurations are incorrectly marked as stale when a host is rebooted

Rebooting a host can causes the client configurations for gateway roles on that host to be marked as stale even though they are up to date.

Rolling restart fails when inheriting inappropriate JVM properties

Rolling restart inherits custom HBase RegionServer JVM properties and can fail when those properties are inappropriate for non-daemon JVMs.

Clusters can fail if Java 1.6 is installed

JAVA_HOME is set to Java 1.6 if installed, even if version 1.8 is also installed.

Start should clearly indicate when it fails due to a missing parcel

Previously a missing parcel was logged only in the Agent log. The Cloudera Manager Admin Console now indicates that a required parcel is missing when starting a role or deploying client configurations.

Client configuration deployment timeout set too large for large number of hosts

The deploy client configuration timeout value was set to according to the number of hosts. This caused a problem when there are large number of hosts. The tasks to deploy client configurations are run concurrently so there was no need to wait that long. The timeout value was changed to a fixed value.

Two HBase metrics charts display "No Data"

HBase IPC metrics were not collected on CDH 5.4 and higher HBase due to a metric name change.

Agent operating system detection logic fixed on updated Oracle Enterprise Linux 6.x

An Oracle Enterprise Linux 6.x update to its `python-l1bs` unexpectedly changed the output of the Agent's operating system detection logic, which caused problems with parcel distribution and other issues.

ZooKeeper `jute.maxbuffer` emitted into configuration instead of JVM arguments

The ZooKeeper `jute.maxbuffer` property was emitted into `zoo.cfg` instead of in the JVM arguments. It is now passed to the JVM through the process environment variable `ZOOKEEPER_SERVER_OPTS` and takes effect correctly.

High `-Xms` set for Hive clients

`-Xms` is set equal to `-Xmx` for all Hive clients which causes the Java runtime to reserve `-Xms` memory even if the client does not need it. This particularly affects machines with low resources. The `-Xms` setting has been removed so that the Java runtime does not assign `-Xmx` memory at the start. Instead, it starts with a much lower Java heap and increases it if needed.

Service autoconfiguration set Kafka memory to 0

The Kafka service was initially configured with 0 memory because the autoconfiguration rules did not respect specified units. This is now correctly set to a value between 50 and 1024 MiB depending on available memory on the host. This issue affected any third-party CSD-based services using memory parameters with units other than "bytes".

Topics with a period not reporting metrics

Topics with a period (".") in the name would not show metrics in Cloudera Manager.

Sensitive information in Cloudera Manager diagnostic support bundles

Cloudera Manager is designed to transmit certain diagnostic data (or "bundles") to Cloudera. These diagnostic bundles are used by the Cloudera support team to reproduce, debug, and address technical issues for our customers. Cloudera internally discovered a potential vulnerability in this feature, which could cause any sensitive data stored as "advanced

Cloudera Manager 5 Release Notes

configuration snippets (ACS)" (formerly called "safety valves") to be included in diagnostic bundles and transmitted to Cloudera. Notwithstanding any possible transmission, such sensitive data is not used by Cloudera for any purpose.

Cloudera has taken the following actions: (1) Modified Cloudera Manager so that it no longer transmits advanced configuration snippets containing the sensitive data, and (2) Modified Cloudera Manager TLS/SSL configuration to increase the protection level of the encrypted communication.

Cloudera strives to follow and also help establish best practices for the protection of customer information. In this effort, we continually review and improve our security practices, infrastructure, and data-handling policies.

Users affected:

- Users storing sensitive data in advanced configuration snippets

Impact: Possible transmission of sensitive data

CVE: CVE-2015-6495

Immediate Action Required:

- Upgrade Cloudera Manager to one of the following releases: Cloudera Manager 5.5.0, 5.4.6, 5.3.7, 5.2.7, 5.1.6, 5.0.7, 4.8.6

[Cloudera Management Service can fail with a large Flume configuration file](#)

When a Flume configuration file is large, calling its Kerberos credentials with regex can cause the Cloudera Management Service to time out and fail. In addition, the Cloudera Manager Server uses 100% of the CPU and the UI hangs.

[Charts built on simple select statements can return partial results](#)

Charts built on queries in the form <select metric> for service or role metrics might not filter entities properly and might report hitting the stream limit.

Issues Fixed in Cloudera Manager 5.4.11

[Cloudera Manager HDFS usage reports do not include Inode references](#)

Lower versions of Cloudera Manager HDFS usage reports do not include Inode references. As a result, usage reports underreported HDFS directory sizes and data used by users and groups in certain circumstances where HDFS snapshots were used.

[Kafka MirrorMaker unable to start due to KAFKA_HOME not being set](#)

Kafka MirrorMaker would not start when Kafka is installed using packages. This occurred because `KAFKA_HOME` was not set to the correct default when starting MirrorMaker. This issue affected Cloudera Manager 5.4.0 and higher with Kafka 1.4.0 and higher.

Issues Fixed in Cloudera Manager 5.4.10

[Scheme and location not filled in consistently during Hive replication import](#)

In previous releases, Hive replication import phase did not consistently fill in scheme and location information. This information is now filled in as expected.

[Having hive.compute.query.using.stats enabled by default produced incorrect results for some queries that used stats only](#)

By default, `hive.compute.query.using.stats` was enabled. This produced incorrect results for some queries that used stats only. This setting is now disabled by default.

[YARN jobs fail after enabling Kerberos authentication or selecting Always Use Container Executor](#)

After Kerberos security is enabled on a cluster or Always Use Container Executor is selected, YARN jobs failed. This occurred because the contents of any previously existing YARN User Cache directory could not be overridden after

security was enabled. YARN jobs now complete as expected after a change in Kerberos security or usage of Container Executor.

Scheduled diagnostic bundles do not include recent role logs

Diagnostic bundles collected manually included all expected logs, but bundles collected on a schedule did not include role logs. Scheduled diagnostic bundles now include all expected logs.

Issues Fixed in Cloudera Manager 5.4.9

Apache Commons Collections deserialization vulnerability

Cloudera has learned of a potential security vulnerability in a third-party library called the [Apache Commons Collections](#). This library is used in products distributed and supported by Cloudera (“Cloudera Products”), including core Apache Hadoop. The Apache Commons Collections library is also in widespread use beyond the Hadoop ecosystem. At this time, no specific attack vector for this vulnerability has been identified as present in Cloudera Products.

In an abundance of caution, we are currently in the process of incorporating a version of the Apache Commons Collections library with a fix into the Cloudera Products. In most cases, this will require coordination with the projects in the Apache community. One example of this is tracked by [HADOOP-12577](#).

The Apache Commons Collections potential security vulnerability is titled “Arbitrary remote code execution with InvokerTransformer” and is tracked by [COLLECTIONS-580](#). MITRE has not issued a CVE, but related [CVE-2015-4852](#) has been filed for the vulnerability. CERT has issued [Vulnerability Note #576313](#) for this issue.

Releases affected: CDH 5.5.0, CDH 5.4.8 and lower, CDH 5.3.8 and lower, CDH 5.2.8 and lower, CDH 5.1.7 and lower, Cloudera Manager 5.5.0, Cloudera Manager 5.4.8 and lower, Cloudera Manager 5.3.8 and lower, and Cloudera Manager 5.2.8 and lower, Cloudera Manager 5.1.6 and lower, Cloudera Navigator 2.4.0, Cloudera Navigator 2.3.8 and lower.

Users affected: All

Impact: This potential vulnerability may enable an attacker to execute arbitrary code from a remote machine without requiring authentication.

Immediate action required: Upgrade to Cloudera Manager 5.5.1 and CDH 5.5.1, Cloudera Manager 5.4.9 and CDH 5.4.9, Cloudera Manager 5.3.9 and CDH 5.3.9, and Cloudera Manager 5.2.9 and CDH 5.2.9, and Cloudera Manager 5.1.7 and CDH 5.1.7.

Cross-site scripting vulnerability using malformed strings in the parcel remote URL list

An attacker could set a malformed string in parameters that consist of a list of strings and trigger the attack when a user accessed the corresponding configuration page in classic layout mode. This attack is now prevented.

Cross-site scripting vulnerability using malformed host template name

An attacker could set a malformed host template name in the backend database and trigger the attack when a user applies the host template. This attack is now prevented.

Snapshot policies with names with special characters not handled as expected

Snapshot policies with names containing special characters such as #, \$, ?, or % were not handled as expected. These snapshot policies were not consistently found because the special characters in their names were parsed incorrectly. Snapshot policies with names containing special characters are now handled as expected.

Kafka MirrorMaker fails to find messages if ZooKeeper root directory is changed

When the ZooKeeper root directory is changed, the corresponding value in ZK_QUORUM that is passed to Kafka MirrorMaker processes is not updated. In that case, MirrorMaker fails to find messages. Changes to ZooKeeper root are now propagated properly, resulting in MirrorMaker finding messages.

Updating the Hive NameNode location multiple times could lead to data corruption

Multiple updates to the Hive NameNode location could cause Hive metastore database corruption. Issuing the same command multiple times no longer produces problems.

Cloudera Manager 5 Release Notes

Cloudera Manager monitors subject to excessive garbage-collection workload

The Cloudera Manager Service Monitor and Cloudera Manager Host Monitor wrote aggregate timeseries data in a way that resulted in significant garbage-collection workloads. Writes are now split based on metric threshold counts, resulting in lower garbage-collection loads.

Kafka MirrorMaker fails to start because of missing settings

Kafka MirrorMaker provided no defaults for Destination Broker List, Topic Whitelist, and Topic Blacklist, and no way to set these values in the wizard. These values can now be set in the wizard when adding a new instance.

Cloudera Manager dry-run replication history shows unexpected values

BDR replication in dry-run mode should show the number of files that would be copied and the number of bytes those files would comprise if the same job were executed without the dry-run option. Dry-run mode showed the number of replicable files accessed up to a maximum of 1024 files and showed the total number of bytes those files comprise, up to 512 bytes per file.

The results of dry-runs now show the actual number of source files and their composite bytes that would be covered in the replication schedule. These categories are labeled replicable files and replicable bytes.

Issues Fixed in Cloudera Manager 5.4.8

Clusters can fail if Java 1.6 is installed

JAVA_HOME is set to Java 1.6 if installed even if 1.8 is also installed.

OutOfMemory error when running HDFS replication

HDFS replication fails with an OutOfMemory exception (in stderr.log) when an HDFS replication job replicates a large number of files.

Cloudera Management Service can fail with a large Flume configuration file

When a Flume configuration file is large, calling its Kerberos credentials with regex can cause the Cloudera Management Service to timeout and fail. In addition, the Cloudera Manager Server uses too much CPU (100%) and the UI hangs.

Client configurations are incorrectly marked as stale when a host is rebooted

Rebooting a host can cause the client configurations for gateway roles on that host to be marked as stale even though they are actually up to date.

Issues Fixed in Cloudera Manager 5.4.7

Operating system detection logic for Oracle Enterprise Linux 6 breaks parcel distribution

Fixes the operating system detection logic on updated Oracle Enterprise Linux 6 systems. An Oracle update to its python-libs logic unexpectedly changed the output of the Agent's operating system detection logic, which caused problems with parcel distribution and other issues.

ZooKeeper jute.maxbuffer property emitted into the wrong file

The ZooKeeper jute.maxbuffer property is emitted into zoo.cfg instead of in the JVM arguments. It is now passed to the JVM through the process environment variable ZOOKEEPER_SERVER_OPTS and takes effect correctly.

Create user API call is allowed for user with insufficient permissions

Using the "create a user" API call, a user who normally could not create users is able to create a read-only user account. The API call now respects the permissions.

[Spark Authentication property is propagated to the wrong client configuration](#)

Beginning with Cloudera Manager 5.4.6, the **Spark Authentication** configuration property is correctly propagated to client configurations.

[Agent hostname in config.ini is changed to wrong value](#)

Attempting to set the `listening_hostname` property in the Agent's `config.ini` file (which is not normally necessary) changes the Agent's host ID to use this hostname, instead of the normal value. The host ID is now left unchanged, as expected.

[Cloudera Manager monitors the incorrect process for DataNode](#)

On Kerberized clusters, Cloudera Manager monitors the wrong process as the DataNode. That has been fixed. For customers using Kerberized HDFS, Cloudera Manager reports incorrect statistics in some areas (memory, file descriptor, CPU usage, I/O, and networking, but not HDFS statistics). There is a small impact to health monitoring because of this issue. For customers using the stacks collection feature on a Kerberized DataNode and where `jstack` collection was enabled, this issue kills the parent `jsvc` process of the DataNode and leaves the DataNode up, but causes Cloudera Manager to report the process as dead.

[Issues Fixed in Cloudera Manager 5.4.6](#)

[Sensitive Information in Cloudera Manager Diagnostic Support Bundles](#)

Cloudera Manager is designed to transmit certain diagnostic data (or "bundles") to Cloudera. These diagnostic bundles are used by the Cloudera support team to reproduce, debug, and address technical issues for our customers. Cloudera internally discovered a potential vulnerability in this feature, which could cause any sensitive data stored as "advanced configuration snippets (ACS)" (formerly called "safety valves") to be included in diagnostic bundles and transmitted to Cloudera. Notwithstanding any possible transmission, such sensitive data is not used by Cloudera for any purpose.

Cloudera has taken the following actions: (1) modified Cloudera Manager so that it no longer transmits advanced configuration snippets containing the sensitive data, and (2) modified Cloudera Manager TLS/SSL configuration to increase the protection level of the encrypted communication.

Cloudera strives to follow and also help establish best practices for the protection of customer information. In this effort, we continually review and improve our security practices, infrastructure, and data handling policies.

Users affected:

- Users storing sensitive data in advanced configuration snippets

Impact: Possible transmission of sensitive data

CVE: CVE-2015-6495

Immediate Action Required:

- Upgrade Cloudera Manager to one of the following releases: Cloudera Manager 5.4.6, 5.3.7, 5.2.7, 5.1.6, 5.0.7, 4.8.6

[Issues Fixed in Cloudera Manager 5.4.5](#)

[Cancel Impala Query attempts to connect via TLS/SSL despite TLS/SSL being disabled](#)

In Impala queries, if you select Cancel for any query, you see a small "internal error" at the top of the query list. This occurs because an attempt to connect via TLS/SSL is performed even though Impala does not have TLS/SSL enabled.

[Cloudera Manager displays a spurious validation warning about the Cloudera Management Service truststore](#)

Cloudera Manager incorrectly warns that Cloudera Management Service daemons will use HTTPS for communication with either Cloudera Manager or CDH services, even if no Cloudera Management Service truststore is in use.

Cloudera Manager 5 Release Notes

Aggregation of Work attributes

Cloudera Manager now correctly aggregates Work attributes such as YARN applications or Impala query duration.

Hue Solr Indexer

Cloudera Manager now creates the correct configuration required to create a collection.

Cloudera Manager incorrectly reports “Not finalized” status for rolling upgrade

When performing a rolling upgrade from a version of CDH lower than 5.4 to CDH 5.4, and the HDFS rolling upgrade is finalized, Cloudera Manager incorrectly reports the status as not finalized. This is an error in reporting only and does not affect HDFS functionality.

Validation errors not visible from service-level configuration pages

Validation errors and warnings were only visible when accessing the individual instance-level configuration pages. This has been fixed.

Add Role Instances wizard does not work when initialized using the Cloudera Manager search box

You can now start the Add Role Instances wizard by searching for "<service name> Add Role" in the Cloudera Manager Admin Console search box.

Cloudera Manager now allows you to use '/' in cluster names

In previous versions, this resulted in problems during replication because '/' was treated as an URL path.

Expose HBase multi-WAL configuration properties in Cloudera Manager

The properties were added and are now being written to the `hbase-site.xml` file.

Custom Kerberos principals now handled correctly during Solr startup

The Solr custom Kerberos principal is now initialized properly during Solr server startup.

Added a check in the Upgrade and Kerberos wizards to make sure Spark-standalone is not enabled

Spark Standalone does not work in clusters with Kerberos authentication. Spark on YARN supports Kerberos and is recommended over Spark Standalone. Either disable Kerberos or remove Spark Standalone before upgrading.

Fixed link for Reports when YARN high availability is enabled

The Reports link in HA mode would result in a 404 error.

Removed bogus failure when deploying client configuration

Deploying client configuration would sometimes fail because Cloudera Manager could not locate `JAVA_HOME`. This is not a valid failure because deploying client configuration does not require Java.

Added `core-site.xml` to Sentry's classpath

Previously, `core-site.xml` was only added to Sentry's configuration folder, but not the classpath.

Improved memory usage in serializing objects and writing them to support bundles

Performance improvements that require less memory were made for the creation of bundles.

New property to enable suppressing INFO-level log messages from NameNode

You can now use the **NameNode Block State Change Logging Threshold** property to suppress INFO-level block state change log messages from the NameNode.

Improved advice for clock offset health test

The way the health of the host's NTP daemon is determined was changed recently, which caused some cases where the related health test (host clock offset) failed without a warning. Information on this change was added to Cloudera Manager.

Cloudera Manager displays warning about using RHEL 6 with Transparent Huge Pages (THP)

The THP algorithm was broken in certain variants of RHEL 6.2 and above. Cloudera Manager now displays a warning if THP is enabled for all RHEL 6 and above.

Agent gets no logs if the last log4j event is larger than the max-size specified

If the `byte_limit (max-size)` specified by Cloudera Manager during log retrieval was smaller than the last log4j event to be collected, the Agent skipped the complete event and return nothing. This behavior was modified so pick the first N bytes (N = `max-size`) are picked from the log4j event and return a partial log4j event.

Agent log retrieval does not always honor timeouts

Cloudera Manager Agents no longer enter an infinite loop during log retrieval.

Cloudera Manager Agent missing log messages

The default timeout for displaying log entries (`../logs/search` and `../logs/context`) has been increased to 60 seconds.

Fixed cross-site scripting vulnerability

A cross-site scripting vulnerability was discovered and fixed in Cloudera Manager.

New property added for ResourceManager high availability failover

The ZooKeeper session timeout property `yarn.resourcemanager.zk-timeout-ms` was added, and its default value is 1 minute.

Set maximum value for YARN mapreduce.jobhistory.max-age-ms to 10 years

Cloudera Manager would previously display a validation error when the value was greater than 60 days.

Added warning in upgrade wizard regarding dropped support for symlinks in CDH 5

This fix added a warning about removing HDFS symlinks when upgrading from CDH 4 to CDH 5.

Refresh Data Directories command no longer fails on secure clusters

The `DataNodeRefreshCommand` now sets `SCM_KERBEROS_PRINCIPAL` in the environment of the command process, which causes `hdfs.sh` to do a kinit. Before this change, a manual kinit was required.

New Sentry Synchronization Path Prefixes added in NameNode configuration are not enforced correctly

Any new path prefixes added in the NameNode configuration are not correctly enforced by Sentry. The ACLs are initially set correctly, however they would be reset to the old default after some time interval.

Workaround: Set the following property in **Sentry Service Advanced Configuration Snippet (Safety Valve)** and **Hive Metastore Server Advanced Configuration Snippet (Safety Valve)** for `hive-site.xml`:

```
<property>
<name>sentry.hdfs.integration.path.prefixes</name>
<value>/user/hive/warehouse, ADDITIONAL_DATA_PATHS</value>
</property>
```

where `ADDITIONAL_DATA_PATHS` is a comma-separated list of HDFS paths where Hive data will be stored. The value should be the same value as `sentry.authorization-provider.hdfs-path-prefixes` set in the `hdfs-site.xml` on the NameNode.

Cloudera Manager 5 Release Notes

Fixed NullPointerException on health tests' Details page

The health tests **Details** page threw a NullPointerException because it was referring to a deprecated metric name.

Improved Service Monitor Canary check to see if HTable is disabled

Without this check, Service Monitor would fail due to too many ZooKeeper connection messages leaking into the Service Monitor log. This resulted in resource and allocation pressures on the Service Monitor.

Cloudera Manager no longer retains unnecessary references to HTables

Retaining too many unnecessary references to HTable was using up too much memory, especially when working with a large number of tables.

Sqoop 2 failure in Kerberized clusters fixed

Cloudera Manager was using the wrong authentication package and picking up the wrong configuration properties for Sqoop 2 authentication with Kerberos.

Fixed Solr server startup error

The Solr server would not start due to insufficient space for the shared memory file.

Added HBase Canary security configuration properties

Enabling the HBase canary on a secure cluster would fail. The new properties now let Cloudera Manager specify the canary's Kerberos principal and keytab in the `hbase-site.xml` deployed at the RegionServers.

Issues Fixed in Cloudera Manager 5.4.3

Improve Impala queries coordinator node metrics handling

For Impala queries that returned very few rows, Cloudera Manager could fail to report information such as HDFS I/O metrics on the **Impala Query Monitoring** and **Query Detail** pages. The discrepancy was typically relatively small because those queries often did very little work.

Performance issues when changing configurations on HDFS

Fixed a performance issue where HDFS configuration pages responded slowly.

Type in Cloudera Manager metrics reference

The word "Concerning" was misspelled in many metrics reference pages.

Issues with Navigator field audit_log_max_file_size

The log4j appender changed from `RollingFileAppender` to `RollingFileWithoutDeleteAppender`.

The Isilon client configuration core-site.xml file does not contain proxy users

The parameters are available in the Cloudera Manager Admin Console, but the configurations are not emitted in the `core-site.xml` file.

Solr gateway role should not have a log4j.properties advanced configuration snippet

The Solr gateway role does not have a `log4j.properties` file.

The Cloudera Manager Agent force_start's hard stop commands did not set all invariants

This resulted in NPE being reported in Cloudera Manager logs when accessing active and recent command operations.

Configuration staleness icons appear to be enabled for users in read-only role

When moused over, the icons change to a hand indicating that they are active. However, users in the read-only role cannot act on changed configurations.

Setting `yarn.resourcemanager.am.max-retries` throws error

Observed when setting the Maximum Number of Attempts for MapReduce Jobs and then setting ApplicationMaster Maximum Attempts, which also sets `yarn.resourcemanager.am.max-retries`.

Cloudera Manager reports the wrong value for Impala bytes read from cache

Instead of cached bytes it reported the value of short circuit bytes.

Fixed cross-site scripting vulnerabilities

A variety of possible cross-site scripting vulnerabilities have been fixed.

Location of Number of rows drop-down changed

On pages where multiple rows display, the drop-down menu where users select the number of rows to display on a page now appears at the bottom of all lists.

Minimum allowed value change for YARN property

The **Max Shuffle Connections** property now allows a value of 0, which indicates no limit on the number of connections.

Upgrade error

A bug was fixed that prevented upgrades from CDH 4.7.1 to CDH 5.4.3.

Change to Parcels page

On the Parcels page, the first cluster in the list is now automatically selected by default.

All Password Input Fields do not allow auto complete

All password input fields in Cloudera Manager do not allow auto complete.

TLS Keystore Configuration Error

It is no longer possible to delete the values of the **Path to TLS Keystore File** and **Keystore Password** properties and save them while the **Use TLS Encryption for Admin Console** property is enabled.

Host configuration properties and Agent restart messages

Some host configuration properties no longer incorrectly state that an Agent restart is required.

More detailed error messages for failed role migration

If there is a failure validating the NameNode or JournalNode data directories while migrating roles, Cloudera Manager now displays detailed error information, including error codes.

New property to configure Oozie shared library upload timeout

To prevent timeouts due to slow disks or networks, a new Oozie property, **Oozie Upload ShareLib Command Timeout**, has been added to set the timeout.

New Cluster-Wide Configuration Pages

The following new Cluster-Wide configuration pages have been added:

- Databases
- Local Data Directories
- Local Data Files
- Navigator Settings
- Service Dependencies

To access these pages in Cloudera Manager, select **Cluster > Cluster Name > Configuration**.

Cloudera Manager 5 Release Notes

Naming of Health Tests

The names of some Health Tests have changed to use consistent capitalization.

Impala Monitoring Queries for Per-node peak memory

Impala queries that report per-node peak memory were incorrect when the value is zero.

Enable Hive on Spark Property

The description of the **Enable Hive on Spark** property has been updated to remind the user that the **Enable Spark on YARN** property must also be selected.

Role Trigger property in Flume

Setting a value for the Flume **Role Triggers** property no longer causes validation warnings.

Restart of Service Monitor leaves files that can fill the disk

Restarts of the Service Monitor no longer leave extraneous copies of files that unnecessarily take up disk space.

HiveServer 2 properties omit Java options

Setting any of the following properties no longer causes Java options to be omitted:

- **Allow URIs in Database Policy File**
- **HiveServer2 TLS/SSL Certificate Trust Store File**
- **HiveServer2 TLS/SSL Certificate Trust Store Password**

CDH Parcel distribution reports HTTP 503 errors

Cloudera Manager no longer displays HTTP 503 errors during distribution of the CDH parcel to a large cluster.

Diagnostic bundle reports incorrect status for SELinux

Diagnostic bundles sometimes reported SELinux as disabled when it was enabled. The bundle now reports the correct status.

Hue configuration warnings do not link to correct page

On the Cloudera Manager page that displays Hue configuration issues, the links now take the user to the correct page where the user can correct the configuration.

Date display in Cloudera Manager log viewer

The month and date have been added before the time value in logs displayed in Cloudera Manager.

Disabling Hive Metastore Canary Test

When you disable the Hive Metastore health test by deselecting the **Hive Metastore Canary Health** property, the Hive Canary is now also disabled.

Agent failure when TLS 1.0 is disabled

If TLS 1.0 is disabled, the Agent now tries to negotiate the connection using TLS 1.1 or TLS 1.2.

Slowness when displaying details of a stale configuration

The details page now displays more quickly when a user clicks on the Stale Configuration icon.

Slowness observed when accessing replication page in Cloudera Manager

When you access the replication page in Cloudera Manager, the page responds slowly due to a large number of replication history records. The number of displayed historical records has been changed from 100 to 20.

Log Searches for Cloudera Manager Server

Searching the Cloudera Manager Server logs now works as expected.

Failed TLS Configuration and Cloudera Manager Restart

If the TLS configuration has errors, Cloudera Manager now falls back to non-TLS operation when restarting.

New headers added

New headers have been added to Cloudera Manager HTTP response headers to protect against vulnerabilities.

Hive Logging property restored

The `Enable Explain Logging (hive.log.explain.output)` property was removed in an earlier release and is now included in the configurations.

Hive Metastore Update NameNodes Command

A 150 second timeout was removed from the Update Hive Metastore NameNodes command to prevent timeouts on deployments that use Hive extensively.

Kafka Parcel Installation

Cloudera Manager now correctly detects the Kafka version for parcel installation.

Agent restart failure

In a condition where an Agent restart was required due to a Hive configuration change and a subsequent disk failure, the Agent now restarts as expected.

Error message wording

Some Cloudera Manager error messages referred to Cloudera Manager as “CM”. These messages now use the full name “Cloudera Manager”.

Oozie metrics failures

Retrieval of Oozie metrics sometimes fails due to timeout issues which are now resolved.

NameNode Role Migration Failures

When a NameNode role migration fails due to the destination role data directories being non-empty or having incorrect permissions, you no longer need to complete the migration manually. An error message displays and you can now correct the problem and re-run the command.

AWS S3 HBase configuration property renamed to Amazon S3

Several configuration properties for HBase have been renamed from AWS S3 to Amazon S3, in order to use the correct product name.

NodeManager Host Resources page display for the NodeManager Recovery Directory

The NodeManager Recovery Directory now displays on the NodeManager host resources page.

Host Inspector page now includes link to Show Inspector Results

The Host Inspector page now displays a link to a page that displays detailed results.

Initialization Script Improvements

The Cloudera Manager Agent initialization script now checks correctly for running processes.

Default Value for Hue parameter changed

The default value for the Hue `cherrypy_server_threads` property has been changed from 10 to 50.

Cloudera Manager 5 Release Notes

Express Installation Wizard Package Installation Page CDH Version

The Express Installation Wizard Package installation page no longer allows the user to proceed without selecting a CDH version.

Host Component page display

The Host Component page now displays the package version for the KMS Trustee Key Provider.

Installation Wizard hangs during package installation

The Installation Wizard hangs during a CDH package installation and the status displays as "Acquiring Installation Lock". A bug was fixed where the Agent incorrectly failed to release a lock until the Agent is restarted.

Minimum allocation violation not caught by Cloudera Manager

NodeManager did not start because Cloudera Manager did not correctly validate memory and CPU settings against their minimum values.

Impala core dump directories are now configurable

Three new properties that specify the location of core dump directories have been added to the Impala configurations:

- Catalog Server Core Dump Directory
- Impala Daemon Core Dump Directory
- StateStore Core Dump Directory

Type in Sqoop DB path suffix (`SqoopParams.DERBY_SUFFIX`)

Sqoop 2 appears to lose data when upgrading to CDH 5.4. This is due to Cloudera Manager erroneously configuring the Derby path with "repositoy" instead of "repository". The correct path name is now used.

Agent fails when retrieving log files with very long messages

When searching or retrieving large log files using the Agent, the Agent no longer consumes near 100% CPU until it is restarted. This can also happen when the collect host statistics command is issued.

Automated Solr TLS/SSL configuration may fail silently

Cloudera Manager 5.4.1 offers simplified TLS/SSL configuration for Solr. This process uses a `solrctl` command to configure the `urlSchemeSolr` cluster property. The `solrctl` command produces the same results as the Solr REST API call `/solr/admin/collections?action=CLUSTERPROP&name=urlScheme&val=https`. For example, the call might appear as:

```
https://example.com:8983/solr/admin/collections?action=CLUSTERPROP&name=urlScheme&val=https
```

Cloudera Manager automatically executes this command during Solr service startup. If this command fails, the Solr service startup now reports an error.

Removing the default value of a property fails

For example, when you access the **Automatically Downloaded Parcels** property on the following page: **Home > Administration > Settings** and remove the default `CDH` value, the following error message displays: "Could not find config to delete with template name: `parcel_autodownload_products`". This error has been fixed.

Issues Fixed in Cloudera Manager 5.4.1

`distcp` default configuration memory settings overwrite MapReduce settings

Replication used for backup and disaster recovery does not correctly set the MapReduce Java options, and you cannot configure them. In release 5.4.1, Cloudera Manager uses the MapReduce gateway configuration to determine the Java options for replication jobs. Replication job settings cannot be configured independently of MapReduce gateway configuration. See [Backup and disaster recovery replication does not set MapReduce Java options](#) on page 388.

Oozie high availability plug-in is now configured by Cloudera Manager

In CDH 5.4.0, Oozie added a new HA plugin that allows all of the Oozie servers to synchronize their Job ID assignments and prevent collisions. Cloudera Manager 5.4.0 did not configure this new plugin; Cloudera Manager 5.4.1 now does so.

HDFS read throughput Impala query monitoring property is misleading

The `hbase_bytes_read_per_second` and `hdfs_bytes_read_per_second` Impala query properties have been renamed to `hbase_scanner_average_bytes_read_per_second` and `hdfs_scanner_average_bytes_read_per_second` to more accurately reflect that these properties return the average throughput of the query's HBase and HDFS scanner threads, respectively. The previous names and descriptions indicated that these properties were the query's total HBase and HDFS throughput, which was not accurate.

Enabling wildcarding in a secure environment causes NameNode to fail to start

In a secure cluster, if you use a wildcard for the NameNode's RPC or HTTP bind address, the NameNode fails to start. For example, `dfs.namenode.http-address` must be a real, routable address and port, not `0.0.0.0.port`. In Cloudera Manager, the "Bind NameNode to Wildcard Address" property must not be enabled. This should affect you only if you are running a secure cluster and your NameNode needs to bind to multiple local addresses.

Bug: [HDFS-4448](#)

Workaround: Disable the "Bind NameNode to Wildcard Address" property found on the Configuration tab for the NameNode role group.

Support for adding Hue with high availability

The Express and Add Service wizards now allow users to define multiple Hue service roles. If Kerberos is enabled, a co-located KT Renewer role is automatically added for each Hue server row.

Parameter validation fails with more than one Hue role

When you add a second Hue role to a cluster, the error message "Failed parameter validation" displays.

Cross-site scripting vulnerabilities

Various cross-site scripting vulnerabilities were fixed.

Clicking the "Revert to default" icon stores the default value as a user-defined value in the new configuration pages

Cloudera Manager 5.4.1 fixes an issue in which saving an empty configuration value causes the value to be replaced by the default value. The empty value is now saved instead of the default value.

Spurious validation warning and missing validations when multiple Hue Server roles are present

When multiple Hue Server roles are created for a single Hue Service, Cloudera Manager displays a spurious validation warning for Hue with the label "Failed parameter validation." The Cloudera Manager Server log may also contain exception messages of the form:

```
2015-03-30 17:15:45,077 WARN
ActionablesProvider-0:com.cloudera.cmf.service.ServiceModelValidatorImpl:
Parameter validation failed java.lang.IllegalArgumentException: There is more than one
role with roletype: HUE_SERVER [...] {
```

These messages do not correspond to actual validation warnings and can be ignored. However, some validations normally performed are skipped when this spurious warning is generated, and should be done manually. Specifically, if Hue's authentication mechanism is set to LDAP, the following configuration should be validated:

1. The Hue **LDAP URL** property must be set.
2. For CDH 4.4 and lower, set one (but not both) of the following two Hue properties: **NT Domain** or **LDAP Username Pattern**.

Cloudera Manager 5 Release Notes

3. For CDH 4.5 and higher, if the Hue property **Use Search Bind Authentication** is selected, exactly one of the two Hue properties **NT Domain** and **LDAP Username Pattern** must be set, as described in step 2 above.

Logging of command unavailable message improved

When a command is unavailable, the error messages are now more descriptive.

Client configuration logs no longer deleted by the Agent

If the Agent fails to deploy a new client configuration, the client log file is no longer deleted by the agent. The Agent saves the log file and appends new log entries to the saved log file.

HDFS role migration requires certain HDFS roles to be running

Before using the Migrate Roles wizard to migrate HDFS roles, you must ensure that the following HDFS roles are running as described:

- A majority of the JournalNodes in the JournalNode quorum must be running. With a quorum size of three JournalNodes, for example, at least two JournalNodes must be running. The JournalNode on the source host need not be running, as long as a majority of all JournalNodes are running.
- When migrating a NameNode and co-located Failover Controller, the other Failover Controller (that is, the one that is not on the source host) must be running. This is true whether or not a co-located JournalNode is being migrated as well, in addition to the NameNode and Failover Controller.
- When migrating a JournalNode by itself, at least one NameNode / Failover Controller co-located pair must be running.

HDFS role migration requires automatic failover to be enabled

Migration of HDFS NameNode, JournalNode, and Failover Controller roles through the Migrate Roles wizard is only supported when HDFS automatic failover is enabled. Otherwise, it causes a state in which both NameNodes are in standby mode.

HDFS/Hive replication fails when replicating to target cluster that runs CDH 4 and has Kerberos enabled

Workaround: None.

Issues Fixed in Cloudera Manager 5.4.0

Proxy Configuration in Single User Mode is Fixed

In single user mode, all services are using the same user to proxy other users in an unsecure cluster, which is the user that is running all the CDH processes on the cluster. To restrict that user so that it can proxy other users from only certain hosts and only certain groups, configure the **YARN Proxy User Hosts** and **YARN Proxy User Groups** properties in the HDFS service. The setting here supersedes all other proxy user configurations in single user mode.

The Parcels page allows access to the patch release notes

Clicking the icon with an "i" in a blue circle next to a parcel shows the release notes.

Monitoring Fails on Impala Catalog Server with TLS/SSL Enabled

When enabling TLS/SSL for Impala web servers (`webserver_certificate_file`), Cloudera Manager does not emit `use_ssl` in the `cloudera-monitor.properties` file for the Catalog Server. Other services (Impala Daemon and StateStore) are configured correctly. This causes monitoring to fail for the Catalog Server even though it is working as expected.

Default Value Changed for `hive.exec.reducer.bytes.per.reducer`

To improve performance, the default value for the `hive.exec.reducer.bytes.per.reducer` property has been changed from 1 GB to 64 MB. If this value has been customized, the customized value is retained during an upgrade. If the old default of 1 GB was not changed, the value will be updated to 64MB during an upgrade.

Issues Fixed in Cloudera Manager 5.3.10

Scheme and location not filled in consistently during Hive replication import

In previous releases, Hive replication import phase did not consistently fill in scheme and location information. This information is now filled in as expected.

YARN jobs fail after enabling Kerberos authentication or selecting Always Use Container Executor

After Kerberos security is enabled on a cluster or Always Use Container Executor is selected, YARN jobs failed. This occurred because the contents of any previously existing YARN User Cache directory could not be overridden after security was enabled. YARN jobs now complete as expected after a change in Kerberos security or usage of Container Executor.

Issues Fixed in Cloudera Manager 5.3.9

Apache Commons Collections deserialization vulnerability

Cloudera has learned of a potential security vulnerability in a third-party library called the [Apache Commons Collections](#). This library is used in products distributed and supported by Cloudera (“Cloudera Products”), including core Apache Hadoop. The Apache Commons Collections library is also in widespread use beyond the Hadoop ecosystem. At this time, no specific attack vector for this vulnerability has been identified as present in Cloudera Products.

In an abundance of caution, we are currently in the process of incorporating a version of the Apache Commons Collections library with a fix into the Cloudera Products. In most cases, this will require coordination with the projects in the Apache community. One example of this is tracked by [HADOOP-12577](#).

The Apache Commons Collections potential security vulnerability is titled “Arbitrary remote code execution with InvokerTransformer” and is tracked by [COLLECTIONS-580](#). MITRE has not issued a CVE, but related [CVE-2015-4852](#) has been filed for the vulnerability. CERT has issued [Vulnerability Note #576313](#) for this issue.

Releases affected: CDH 5.5.0, CDH 5.4.8 and lower, CDH 5.3.8 and lower, CDH 5.2.8 and lower, CDH 5.1.7 and lower, Cloudera Manager 5.5.0, Cloudera Manager 5.4.8 and lower, Cloudera Manager 5.3.8 and lower, and Cloudera Manager 5.2.8 and lower, Cloudera Manager 5.1.6 and lower, Cloudera Navigator 2.4.0, Cloudera Navigator 2.3.8 and lower.

Users affected: All

Impact: This potential vulnerability may enable an attacker to execute arbitrary code from a remote machine without requiring authentication.

Immediate action required: Upgrade to Cloudera Manager 5.5.1 and CDH 5.5.1, Cloudera Manager 5.4.9 and CDH 5.4.9, Cloudera Manager 5.3.9 and CDH 5.3.9, and Cloudera Manager 5.2.9 and CDH 5.2.9, and Cloudera Manager 5.1.7 and CDH 5.1.7.

Cloudera Manager monitors subject to excessive garbage-collection workload

The Cloudera Manager Service Monitor and Cloudera Manager Host Monitor wrote aggregate timeseries data in a way that resulted in significant garbage-collection workloads. Writes are now split based on metric threshold counts, resulting in lower garbage-collection loads.

Cloudera Manager dry-run replication history shows unexpected values

BDR replication in dry-run mode should show the number of files that would be copied and the number of bytes those files would comprise if the same job were executed without the dry-run option. Dry-run mode showed the number of replicable files accessed up to a maximum of 1024 files and showed the total number of bytes those files comprise, up to 512 bytes per file.

The results of dry-runs now show the actual number of source files and their composite bytes that would be covered in the replication schedule. These categories are labeled replicable files and replicable bytes.

Cloudera Manager 5 Release Notes

Updating the Hive NameNode location multiple times could lead to data corruption

Multiple updates to the Hive NameNode location could cause Hive metastore database corruption. Issuing the same command multiple times no longer produces problems.

Issues Fixed in Cloudera Manager 5.3.8

Rolling Restart fails when inheriting inappropriate JVM properties

Rolling Restart inherits custom HBase RegionServer JVM properties and can fail when those properties are inappropriate for non-daemon JVMs.

Exception thrown when viewing host configuration change

When ConfigContext is initialized to both cluster and host, it defaults to cluster. If the context is a host, this can cause the ConfigTableUtil class to throw an IllegalStateException.

Issues Fixed in Cloudera Manager 5.3.7

Sensitive Information in Cloudera Manager Diagnostic Support Bundles

Cloudera Manager is designed to transmit certain diagnostic data (or “bundles”) to Cloudera. These diagnostic bundles are used by the Cloudera support team to reproduce, debug, and address technical issues for our customers. Cloudera internally discovered a potential vulnerability in this feature, which could cause any sensitive data stored as “advanced configuration snippets (ACS)” (formerly called “safety valves”) to be included in diagnostic bundles and transmitted to Cloudera. Notwithstanding any possible transmission, such sensitive data is not used by Cloudera for any purpose.

Cloudera has taken the following actions: (1) modified Cloudera Manager so that it no longer transmits advanced configuration snippets containing the sensitive data, and (2) modified Cloudera Manager TLS/SSL configuration to increase the protection level of the encrypted communication.

Cloudera strives to follow and also help establish best practices for the protection of customer information. In this effort, we continually review and improve our security practices, infrastructure, and data handling policies.

Users affected:

- Users storing sensitive data in advanced configuration snippets

Severity: High

Impact: Possible transmission of sensitive data

CVE: CVE-2015-6495

Immediate Action Required:

- Upgrade Cloudera Manager to one of the following releases: Cloudera Manager 5.4.6, 5.3.7, 5.2.7, 5.1.6, 5.0.7, 4.8.6

Issues Fixed in Cloudera Manager 5.3.6

Cloudera Manager reports the wrong value for Impala bytes read from cache

Instead of cached bytes, it reported the value of short-circuit bytes.

Cancel Impala Query attempts to connect via TLS/SSL despite TLS/SSL being disabled

In Impala queries, if you select Cancel for any query, you get a small "internal error" at the top of the query list. This is because an attempt to connect via TLS/SSL is done even though Impala does not have TLS/SSL enabled.

The Cloudera Manager displays a spurious validation warning about the Cloudera Management Service truststore

Cloudera Manager incorrectly warns that Cloudera Management Service daemons will use HTTPS for communication with either Cloudera Manager or CDH services even if no Cloudera Management Service truststore is in use.

Issues Fixed in Cloudera Manager 5.3.4

Slowness observed when accessing replication page in Cloudera Manager

When you access the replication page in Cloudera Manager, the page responds slowly due to a large number of replication history records. The number of displayed historical records has been changed from 100 to 20.

Cloudera Manager overwrites the krb5.conf file after disabling management by Cloudera Manager

When a cluster has been configured to enable Kerberos by clicking the **Manage krb5.conf through Cloudera Manager** button, Cloudera Manager writes out a `krb5.conf` file. If a user subsequently disables this feature by disabling the **Manage krb5.conf through Cloudera Manager** option and then restarting Cloudera Manager and the Agent, Cloudera Manager overwrites the existing `krb5.conf` file.

Header injection vulnerability in internal logging call

An internal call to a servlet with a malformed logger name created information that could be used in a cross-site scripting attack.

Graph for "Total Containers Running Across NodeManagers" displays high values

The graph **Total Containers Running Across NodeManagers**, which displays on the YARN status page, displayed incorrect high values.

Clicking the "Revert to default" icon stores the default value as a user-defined value in the new configuration pages

Saving an empty configuration value causes the value to be replaced by the default value. The empty value is now saved instead of the default value.

Some Operational Reports do not return results

The following reports do not return results:

- Overpopulated Directories
- Large Directories
- Custom Reports where the **Replication** parameter is set to 0.

Xalan has been upgraded to version 2.7.2

To address a vulnerability identified by [CVE-2014-0107](#), Xalan has been updated to version 2.7.2.

Setting threshold values of -1 or -2 not accepted in new configuration layout pages

When you set threshold values in various properties, you could not select **Specify** and then enter `-1` to indicate "Any" or `-2` to indicate "Never". You can now enter `-1` or `-2`.

Rolling upgrade arguments are reversed

The following arguments for rolling upgrade are reversed:

- Sleep seconds
- Failure threshold

Issues Fixed in Cloudera Manager 5.3.3

`hive.metastore.client.socket.timeout` default value changed to 60

The default value of the `hive.metastore.client.socket.timeout` property has changed to 60 seconds.

TLS/SSL Enablement property name changes

The property `hadoop.ssl.enabled` is deprecated. Cloudera Manager has been updated to use either `dfs.http.policy` or `yarn.http.policy` properties instead.

Cloudera Manager 5 Release Notes

[Changing the Service Monitor Client Config Overrides property requires restart](#)

Cloudera Manager no longer requires you to restart your cluster after changing the **Service Monitor Client Config Overrides** property for a service.

[Cluster name changed from specified name to "cluster" after upgrade](#)

After updating to a new release, Cloudera Manager replaces the specified cluster name with `cluster`. Cloudera Manager now uses the correct cluster name.

[Configuration without host_id in upgrade DDL causes upgrade problems](#)

A client configuration row in the database DDL did not set `host_id`, causing upgrade problems. Cloudera Manager now catches this condition before upgrading.

[hive.log.explain.output property is hidden](#)

The property `hive.log.explain.output` is known to create instability of Cloudera Manager Agents in some specific circumstances, especially when the hive queries generate extremely large EXPLAIN outputs. Therefore, the property has been hidden from the Cloudera Manager configuration screens. You can still configure the property through the use of advanced configuration snippets.

[Slow staleness calculation can lead to ZooKeeper data loss when new servers are added](#)

In Cloudera Manager 5.x, starting new ZooKeeper Servers shortly after adding them can cause ZooKeeper data loss when the number of new servers exceeds the number of old servers.

[Spark and Spark \(standalone\) services fail to start if you upgrade to CDH 5.2.x parcels from an older CDH package](#)

Spark and Spark standalone services fail to start if you upgrade to CDH 5.2.x parcels from an older CDH package.

Workaround: After upgrading rest of the services, uninstall the old CDH packages, and then start the Spark service.

[Deploy client configuration across cluster after upgrade from Cloudera Manager 4.x to 5.3](#)

Following a 4.x -> 5.3 upgrade, you must deploy client configuration across the entire cluster before deleting any gateway roles, any services, or any hosts. Otherwise the existing 4.x client configurations may be left registered and orphaned on the hosts where they were deployed, requiring you to manually intervene to delete them.

[Oozie health bad when Oozie is HA, cluster is kerberized, and Cloudera Manager and CDH are upgraded](#)

Oozie health will go bad if high availability is enabled in a kerberized cluster with Cloudera Manager 5.0 and CDH 5.0 and Cloudera Manager and CDH are then upgraded to 5.1 or higher.

Workaround: Disable Oozie HA and then re-enable HA again.

[HDFS/Hive replication fails when replicating to target cluster that runs CDH 4.0 and has Kerberos enabled](#)

Workaround: None.

Issues Fixed in Cloudera Manager 5.3.2

[The Review Changes page sometimes hangs](#)

The **Review Changes** page hangs due to the inability to handle the "File missing" scenario.

[High volume of TGT events against AD server with "bad token" messages](#)

A fix has been made to how Kerberos credential caching is handled by management services, resulting in a reduction in the number of Kerberos Ticket Granting Ticket (TGT) requests from the cluster to a KDC. This would have been noticed as "Bad Token" messages being seen in high volume in KDC logging and unnecessarily causing re-authentication by management services.

Accumulo missing kinit when running with Kerberos

Cloudera Manager is unable to run Accumulo when `hostname` command does not return FQDN of hosts.

HiveServer2 leaks threads when using impersonation

For CDH 5.3 and higher, Cloudera Manager will configure HiveServer2 to use the HDFS cache even when impersonation is on. For earlier CDH, there were bugs with the cache when impersonation was in use, so it is still disabled.

Deploying client configurations fails if there are dead hosts present in the cluster

If there are hosts in the cluster where the Cloudera Manager agent heartbeat is not working, then deploying client configurations does not work. Starting with Cloudera Manager 5.3.2, such hosts are ignored while deploying client configurations. When the issues with the host are fixed, Cloudera Manager will show those hosts as having stale client configurations, at which point you can redeploy them.

Health test monitors free space available on the wrong filesystem

The Cloudera Manager Health Test to monitor free space available for the Cloudera Manager Agent's process directory monitors space on the wrong filesystem. It should monitor the `tmpfs` that the Cloudera Manager Agent creates, but instead monitors the Cloudera Manager Agent working directory.

Starting ZooKeeper Servers from Service or Instance page fails

Stopped ZooKeeper servers cannot be started from the Service or Instance page, but only from the Role page of the server using the `start` action for the role.

Flume Metrics page does not render agent metrics

Starting in Cloudera Manager 5.3, some or all Flume component data was missing from the Flume Metrics Details page.

Broken link to help pages on Chart Builder page

The help icon (question mark) on the Chart Builder page returns a 404 error.

Import MapReduce configurations to YARN now handles NodeManager vcores and memory

Running the wizard to import MapReduce configurations to YARN will now populate `yarn.nodemanager.resource.cpu-vcores` and `yarn.nodemanager.resource.memory-mb` correctly based on equivalent MapReduce configuration.

Issues Fixed in Cloudera Manager 5.3.1

Deploy client configuration across cluster after upgrade from Cloudera Manager 4.x to 5.3

Following a 4.x -> 5.3 upgrade, you must deploy client configuration across the entire cluster before deleting any gateway roles, any services, or any hosts. Otherwise the existing 4.x client configurations may be left registered and orphaned on the hosts where they were deployed, requiring you to manually intervene to delete them.

Deploy client configuration across cluster after upgrade from Cloudera Manager 4.x to 5.3

Following a 4.x -> 5.3 upgrade, you must deploy client configuration across the entire cluster before deleting any gateway roles, any services, or any hosts. Otherwise the existing 4.x client configurations may be left registered and orphaned on the hosts where they were deployed, requiring you to manually intervene to delete them.

Oozie health bad when Oozie is HA, cluster is kerberized, and Cloudera Manager and CDH are upgraded

Oozie health will go bad if high availability is enabled in a kerberized cluster with Cloudera Manager 5.0 and CDH 5.0 and Cloudera Manager and CDH are then upgraded to 5.1 or higher.

Workaround: Disable Oozie HA and then re-enable HA again.

Cloudera Manager 5 Release Notes

Deploy client configuration no longer fails after 60 seconds

When configuring a gateway role on a host that already contains a role of the same type—for example, an HDFS gateway on a DataNode—the deploy client configuration command no longer fails after 60 seconds.

service cloudera-scm-server force_start now works

After deleting services, the Cloudera Manager Server log no longer contains foreign key constraint failure exceptions

When using Isilon, Cloudera Manager now sets mapred_submit_replication correctly

When EMC Isilon storage is used, there is no DataNode, so you cannot set `mapred_submit_replication` to a number smaller than or equal to the number of DataNodes in the network. Cloudera Manager now does the following when setting `mapred_submit_replication`:

- If using HDFS, sets to a minimum of 1 and issues a warning when greater than the number of DataNodes
- If using Isilon, sets to 1 and does not check against the number of DataNodes

The Cloudera Manager Agent now sets the file descriptor ulimit correctly on Ubuntu

During upgrade, bootstrapping the standby NameNode step no longer fails with standby NameNode connection refused when connecting to active NameNode

Deploy krb5.conf now also deploys it on hosts with Cloudera Management Service roles

Cloudera Manager allows upgrades to unknown CDH maintenance releases

Cloudera Manager 5.3.0 supports any CDH release less than or equal to 5.3, even if the release did not exist when Cloudera Manager 5.3.0 was released. For packages, you cannot currently use the upgrade wizard to upgrade to such a release. This release adds a custom CDH field for the package case, where you can type in a version that did not exist at the time of the Cloudera Manager release.

impalad memory limit units error in EnableLlamaRMCommand

The `EnableLlamaRMCommand` sets the value of the `impalad` memory limit to equal the NM container memory value. But the latter is in MB, and the former is in bytes. Previously, the command did not perform the conversion; this has been fixed.

Running MapReduce v2 jobs are now visible using the Application Master view

In the Application view, selecting **Application Master** for a MRv2 job previously resulted in no action.

Deleting services no longer results in foreign key constraint exceptions

The Cloudera Manager Server log previously showed several foreign key constraint exceptions that were associated with deleted services. This has been fixed.

HiveServer2 keystore and LDAP group mapping passwords are no longer exposed in client configuration files

The HiveServer2 keystore password and LDAP group mapping passwords were emitted into the client configuration files. This exposed the passwords in plain text in a world-readable file. This has been fixed.

A cross-site scripting vulnerability in Cloudera Management Service web UIs fixed

The high availability wizard now sets the HDFS dependency on ZooKeeper

Workaround: Before enabling high availability, do the following:

1. Create and start a ZooKeeper service if one does not exist.
2. Go to the HDFS service.
3. Click the **Configuration** tab.
4. Select **HDFS Service-Wide**.
5. Select **Category > Main**.

6. Locate the **ZooKeeper Service** property or search for it by typing its name in the Search box. Select the ZooKeeper service you created.

If more than one role group applies to this configuration, edit the value for the appropriate role group. See [Modifying Configuration Properties Using Cloudera Manager](#).

7. Click **Save Changes** to commit the changes.

BDR no longer assumes superuser is common if clusters have the same realm

If source and destination clusters are in the same Kerberos realm, Cloudera Manager assumed that superuser of the destination is also the superuser on the source cluster. However, HDFS can be configured so that this is not the case.

Issues Fixed in Cloudera Manager 5.3.0

Setting the default umask in HDFS fails in new configuration layout

Setting the default umask in the HDFS configuration section to 002 in the new configuration layout displays an error: "Could not parse: Default Umask : Could not parse parameter 'dfs_umaskmode'. Was expecting an octal value with a leading 0. Input: '2'", preventing the change from being submitted.

Workaround: Submit the change using the classic configuration layout.

Spark and Spark (standalone) services fail to start if you upgrade to CDH 5.2.x parcels from an older CDH package

Spark and Spark standalone services fail to start if you upgrade to CDH 5.2.x parcels from an older CDH package.

Workaround: After upgrading rest of the services, uninstall the old CDH packages, and then start the Spark service.

Fixed MapReduce Usage by User reports when using an Oracle database backend

Setting the default umask in HDFS fails in new configuration layout

Setting the default umask in the HDFS configuration section to 002 in the new configuration layout displays an error: "Could not parse: Default Umask : Could not parse parameter 'dfs_umaskmode'. Was expecting an octal value with a leading 0. Input: '2'", preventing the change from being submitted.

Workaround: Submit the change using the classic configuration layout.

Enabling Integrated Resource Management for Impala sets Impala Daemon Memory Limit Incorrectly

The Enable Integrated Resource Management command for Impala (available from the **Actions** pull-down menu on the Impala service page) sets the Impala Daemon Memory Limit to an unusably small value. This can cause Impala queries to fail.

Workaround 1: Upgrade to Cloudera Manager 5.3.

Workaround 2:

1. Run the Enable Integrated Resource Management wizard up to the **Restart Cluster** step. Do not click **Restart Now**.
2. Click on the **leave this wizard** link to exit the wizard without restarting the cluster.
3. Go to the YARN service page. Click **Configuration**, expand the category **NodeManager Default Group**, and click **Resource Management**.
4. Note the value of the Container Memory property.
5. Go to the Impala service page and click **Configuration**. Type `impala daemon memory limit` into the search box.
6. Set the value of the Impala Daemon Memory Limit property to the value noted in step 4 above.
7. Restart the cluster.

Rolling restart and upgrade of Oozie fails if there is a single Oozie server

Rolling restart and upgrade of Oozie fails if there is only a single Oozie server. Cloudera Manager will show the error message "There is already a pending command on this role."

Cloudera Manager 5 Release Notes

Workaround: If you have a single Oozie server, do a normal restart.

Allow "Started but crashed" processes to be restarted by a Start command

In Cloudera Manager 5.3, it is now possible to restart a crashed process with the **Start** command and not just the **Restart** command.

Add dependency from Agent to Daemons package to yum

In Cloudera Manager 5.3, an explicit dependency has been added from the Agent package to the Daemons package so that upgrading Cloudera Manager 5.2.0 or later to Cloudera Manager 5.3 causes the agent to be upgraded as well. Previously, the Cloudera Manager installer always installed both packages, but this is now enforced at the package dependency level as well.

Issues Fixed in Cloudera Manager 5.2.7

Sensitive Information in Cloudera Manager Diagnostic Support Bundles

Cloudera Manager is designed to transmit certain diagnostic data (or “bundles”) to Cloudera. These diagnostic bundles are used by the Cloudera support team to reproduce, debug, and address technical issues for our customers. Cloudera internally discovered a potential vulnerability in this feature, which could cause any sensitive data stored as “advanced configuration snippets (ACS)” (formerly called “safety valves”) to be included in diagnostic bundles and transmitted to Cloudera. Notwithstanding any possible transmission, such sensitive data is not used by Cloudera for any purpose.

Cloudera has taken the following actions: (1) modified Cloudera Manager so that it no longer transmits advanced configuration snippets containing the sensitive data, and (2) modified Cloudera Manager TLS/SSL configuration to increase the protection level of the encrypted communication.

Cloudera strives to follow and also help establish best practices for the protection of customer information. In this effort, we continually review and improve our security practices, infrastructure, and data handling policies.

Users affected:

- Users storing sensitive data in advanced configuration snippets

Severity:

Impact: Possible transmission of sensitive data

CVE: CVE-2015-6495

Immediate Action Required:

- Upgrade Cloudera Manager to one of the following releases: Cloudera Manager 5.4.6, 5.3.7, 5.2.7, 5.1.6, 5.0.7, 4.8.6

Issues Fixed in Cloudera Manager 5.2.6

Cloudera Manager Agent may become slow or get stuck when responding to commands and when sending heartbeats to Cloudera Manager Server

This issue can occur when Cloudera Navigator auditing is turned on. The auditing code reads audit logs and sends them to the Audit Server. It acquires a lock to protect the list of roles being audited. The same list is also modified by the Cloudera Manager Agent's main thread when a role is started or stopped. If the Audit thread takes too much time to send audits to the Audit Server (which can happen if there is backlog of audit logs), it starves the main Agent thread. This causes the main Agent thread to not send heartbeats and to not respond to commands from the Cloudera Manager Server.

Problems restarting the NameNode

The NameNode would not restart due to a blocked thread that was processing audit events.

Cloudera Manager reports the wrong value for Impala bytes read from cache

Instead of cached bytes, it reported the value of short-circuit bytes.

Directory operational reports do not return results

The following reports now return results:

- Overpopulated Directories
- Large Directories
- Custom reports where **Replication=0**

Cloudera Manager uses Xalan 2.7.2

The version of Xalan used in Cloudera Manager has been upgraded to version 2.7.2 to address a possible vulnerability.

Cluster name changed to "cluster" after upgrade

After upgrading CDH, the display name of the cluster no longer changes to "cluster".

Cloudera Manager monitors the wrong directory for space threshold warnings

Cloudera Manager was monitoring the disk space thresholds using the `/var/run/cloudera-scm-agent` directory. Cloudera Manager now monitors the correct directory: `/var/run/cloudera-scm-agent/process`.

Default timeout value for the Hive MetaStore changed to 60 seconds

The default value in the Hive **Service Monitor Client Config Overrides** property for the `hive.metastore.client.socket.timeout` property is now 60.

Changing client configuration overrides

Changes made to client configuration overrides did not take effect until the service was restarted.

Issues Fixed in Cloudera Manager 5.2.5

Slow staleness calculation can lead to ZooKeeper data loss when new servers are added

In Cloudera Manager 5, starting new ZooKeeper Servers shortly after adding them can cause ZooKeeper data loss when the number of new servers exceeds the number of old servers.

Permissions set incorrectly on YARN Keytab files

Permissions on YARN Keytab files for NodeManager were set incorrectly to allow read access to any user.

Issues Fixed in Cloudera Manager 5.2.2

Impalad memory limit units error in EnableLlamaRMCommand has been fixed

The `EnableLlamaRMCommand` sets the value of the impalad memory limit to equal the NM container memory value. But the latter is in MB, and the former is in bytes. Previously, the command did not perform the conversion; this has been fixed.

Fixed MapReduce Usage by User reports when using an Oracle database backend

HiveServer2 keystore and LDAP group mapping passwords are no longer exposed in client configuration files

The HiveServer2 keystore password and LDAP group mapping passwords were emitted into the client configuration files. This exposed the passwords in plain text in a world-readable file. This has been fixed.

Running MapReduce v2 jobs are now visible using the Application Master view

In the Application view, selecting **Application Master** for a MRv2 job previously resulted in no action.

Cloudera Manager 5 Release Notes

Deleting services no longer results in foreign key constraint exceptions

The Cloudera Manager Server log previously showed several foreign key constraint exceptions that were associated with deleted services. This has been fixed.

Issues Fixed in Cloudera Manager 5.2.1

"POODLE" vulnerability on TLS/SSL enabled ports

The POODLE (Padding Oracle On Downgraded Legacy Encryption) attack takes advantage of a cryptographic flaw in the obsolete TLS/SSLv3 protocol, after first forcing the use of that protocol. The only solution is to disable TLS/SSLv3 entirely. This requires changes across a wide variety of components of CDH and Cloudera Manager in 5.2.0 and all earlier versions. Cloudera Manager 5.2.1 provides these changes for Cloudera Manager 5.2.0 deployments. All Cloudera Manager 5.2.0 users should upgrade to 5.2.1 as soon as possible. For more information, see the [Cloudera Security Bulletin](#).

Can use the log4j advanced configuration snippet to override the default audit logging configuration even if not using Navigator

In Cloudera Manager 5.2.0 only, it was not possible to use the log4j advanced configuration snippet to override the default audit logging configuration when Navigator was not being used.

Cloudera Manager now collects metrics for CDH 5.0 DataNodes and NameNodes

A number of NameNode and DataNode charts show no data and a number of NameNode and DataNode health checks show unknown results. Metric collection for CDH 5.1 roles is unaffected.

Workaround: None.

The Reports Manager and Event Server Thrift servers no longer crash on HTTP requests

HTTP queries against the Reports Manager and Event Server Thrift server would earlier cause it to crash with out-of-memory exception.

Replication commands now use the correct JAVA_HOME if an override has been provided for it

ZooKeeper connection leaks from HBase clients in Service Monitor have been fixed

When a parcel is activated, user home directories are now created with umask 022 instead of using the "useradd" default 077

Issues Fixed in Cloudera Manager 5.2.0

Bug in openssl-1.0.1e-15 disrupts TLS/SSL communication between Cloudera Manager Agents and CDH services

This issue was observed in TLS/SSL-enabled clusters running CentOS 6.4 and 6.5, where the Cloudera Manager Agent failed when trying to communicate with CDH services. You can see the bug report [here](#).

Workaround: Upgrade to openssl-1.0.1e-16.el6_5.7.x86_64.

Alternatives database points to client configurations of deleted service

In the past, if you created a service, deployed its client configurations, and then deleted that service, the client configurations lived in the alternative database, with a possibly high priority, until cleaned up manually. Now, for a given "alternatives path" (for example /etc/hadoop/conf) if there exist both "live" client configurations (ones that would be pushed out with deploy client configurations for active services) and ones that have been "orphaned" client configurations (the service they correspond to has been deleted), the orphaned ones will be removed from the alternatives database. In other words, to trigger cleanup of client configurations associated with a deleted service you must create a service to replace it.

The YARN property ApplicationMaster Max Retries has no effect in CDH 5

The issue arises because `yarn.resourcemanager.am.max-retries` was replaced with `yarn.resourcemanager.am.max-attempts`.

Workaround:

1. Add the following to **ResourceManager Advanced Configuration Snippet for `yarn-site.xml`**, replacing `MAX_ATTEMPTS` with the desired maximum number of attempts:

```
<property>
<name>yarn.resourcemanager.am.max-attempts</name><value>MAX_ATTEMPTS</value>
</property>
```

2. Restart the ResourceManager(s) to pick up the change.

The Spark History Server does not start when Kerberos authentication is enabled.

The Spark History Server does not start when managed by a Cloudera Manager 5.1 instance when Kerberos authentication is enabled.

Workaround:

1. Go to the Spark service.
2. Expand the **Service-Wide > Advanced** category.
3. Add the following configuration to the **History Server Environment Advanced Configuration Snippet** property:

```
SPARK_HISTORY_OPTS=-Dspark.history.kerberos.enabled=true \
-Dspark.history.kerberos.principal=principal \
-Dspark.history.kerberos.keytab=keytab
```

where `principal` is the name of the Kerberos principal to use for the History Server, and `keytab` is the path to the principal's keytab file on the local filesystem of the host running the History Server.

Hive replication issue with TLS enabled

Hive replication will fail when the source Cloudera Manager instance has TLS enabled, even though the required certificates have been added to the target Cloudera Manager's trust store.

Workaround: Add the required Certificate Authority or self-signed certificates to the default Java trust store, which is typically a copy of the cacerts file named `jssecacerts` in the `$JAVA_HOME/jre/lib/security/` path of your installed JDK. Use keytool to import your private CA certificates into the `jssecacert` file.

The Spark Upload Jar command fails in a secure cluster

The Spark **Upload Jar** command fails in a secure cluster.

Workaround: To run Spark on YARN, manually upload the Spark assembly jar to HDFS `/user/spark/share/lib`. The Spark assembly jar is located on the local filesystem, typically in `/usr/lib/spark/assembly/lib` or `/opt/cloudera/parcels/CDH/lib/spark/assembly/lib`.

Clients of the JobHistory Server Admin Interface Require Advanced Configuration Snippet

Clients of the JobHistory server administrative interface, such as the `mapred hsadmin` tool, may fail to connect to the server when run on hosts other than the one where the JobHistory server is running.

Workaround: Add the following to both the **MapReduce Client Advanced Configuration Snippet for `mapred-site.xml`** and the **Cluster-wide Advanced Configuration Snippet for `core-site.xml`**, replacing `JOBHISTORY_SERVER_HOST` with the hostname of your JobHistory server:

```
<property>
<name>mapreduce.history.admin.address</name>
<value>JOBHISTORY_SERVER_HOST:10033</value>
</property>
```

Issues Fixed in Cloudera Manager 5.1.6

Sensitive Information in Cloudera Manager Diagnostic Support Bundles

Cloudera Manager is designed to transmit certain diagnostic data (or “bundles”) to Cloudera. These diagnostic bundles are used by the Cloudera support team to reproduce, debug, and address technical issues for our customers. Cloudera internally discovered a potential vulnerability in this feature, which could cause any sensitive data stored as “advanced configuration snippets (ACS)” (formerly called “safety valves”) to be included in diagnostic bundles and transmitted to Cloudera. Notwithstanding any possible transmission, such sensitive data is not used by Cloudera for any purpose.

Cloudera has taken the following actions: (1) modified Cloudera Manager so that it no longer transmits advanced configuration snippets containing the sensitive data, and (2) modified Cloudera Manager TLS/SSL configuration to increase the protection level of the encrypted communication.

Cloudera strives to follow and also help establish best practices for the protection of customer information. In this effort, we continually review and improve our security practices, infrastructure, and data handling policies.

Users affected:

- Users storing sensitive data in advanced configuration snippets

Severity: High

Impact: Possible transmission of sensitive data

CVE: CVE-2015-6495

Immediate Action Required:

- Upgrade Cloudera Manager to one of the following releases: Cloudera Manager 5.4.6, 5.3.7, 5.2.7, 5.1.6, 5.0.7, 4.8.6

Cloudera Manager Agent may become slow or get stuck when responding to commands and when sending heartbeats to Cloudera Manager Server

This issue can occur when Cloudera Navigator auditing is turned on. The auditing code reads audit logs and sends them to the Audit Server. It acquires a lock to protect the list of roles being audited. The same list is also modified by the Cloudera Manager Agent's main thread when a role is started or stopped. If the Audit thread takes too much time to send audits to the Audit Server (which can happen if there is backlog of audit logs), it starves the main Agent thread. This causes the main Agent thread to not send heartbeats and to not respond to commands from the Cloudera Manager Server.

Fixed Issues in Cloudera Manager 5.1.5

Slow staleness calculation can lead to ZooKeeper data loss when new servers are added

In Cloudera Manager 5, starting new ZooKeeper Servers shortly after adding them can cause ZooKeeper data loss when the number of new servers exceeds the number of old servers.

Permissions set incorrectly on YARN Keytab files

Permissions on YARN Keytab files for NodeManager were set incorrectly to allow read access to any user.

Fixed Issues in Cloudera Manager 5.1.4

“POODLE” vulnerability on TLS/SSL enabled ports

The POODLE (Padding Oracle On Downgraded Legacy Encryption) attack takes advantage of a cryptographic flaw in the obsolete TLS/SSLv3 protocol, after first forcing the use of that protocol. The only solution is to disable TLS/SSLv3 entirely. This requires changes across a wide variety of components of CDH and Cloudera Manager. Cloudera Manager 5.1.4 provides these changes for Cloudera Manager 5.1.x deployments. All Cloudera Manager 5.1.x users should upgrade to 5.1.4 as soon as possible. For more information, see the [Cloudera Security Bulletin](#).

Issues Fixed in Cloudera Manager 5.1.3

Improved speed and heap usage when deleting hosts on cluster with long history

Speed and heap usage have been improved when deleting hosts on clusters that have been running for a long time.

When there are multiple clusters, each cluster's topology files and validation for legal topology is limited to hosts in that cluster

When there are multiple clusters, each cluster's topology files and validation for legal topology is limited to hosts in that cluster. Most commands will now fail up front if the cluster's topology is invalid.

The size of the statement cache has been reduced for Oracle databases

For users of Oracle databases, the size of the statement cache has been reduced to help with memory consumption.

Improvements to memory usage of "cluster diagnostics collection" for large clusters.

Memory usage of "cluster diagnostics collection" has been improved for large clusters.

Issues Fixed in Cloudera Manager 5.1.2

If a NodeManager that is used as ApplicationMaster is decommissioned, YARN jobs will hang

Jobs can hang on NodeManager decommission due to a race condition when continuous scheduling is enabled.

Workaround:

1. Go to the YARN service.
2. Expand the **ResourceManager Default Group > Resource Management** category.
3. Uncheck the **Enable Fair Scheduler Continuous Scheduling** checkbox.
4. Click **Save Changes** to commit the changes.
5. Restart the YARN service.

Could not find a healthy host with CDH 5 on it to create HiveServer2 error during upgrade

When upgrading from CDH 4 to CDH 5, if no parcel is active then the error message "Could not find a healthy host with CDH5 on it to create HiveServer2" displays. This can happen when transitioning from packages to parcels, or if you explicitly deactivate the CDH 4 parcel (which is not necessary) before upgrade.

Workaround: Wait 30 seconds and retry the upgrade.

AWS installation wizard requires Java 7u45 to be installed on Cloudera Manager Server host

Cloudera Manager 5.1 installs Java 7u55 by default. However, the AWS installation wizard does not work with Java 7u55 due to a bug in the jClouds version packaged with Cloudera Manager.

Workaround:

1. Stop the Cloudera Manager Server.

```
sudo service cloudera-scm-server stop
```

2. Uninstall Java 7u55 from the Cloudera Manager Server host.
3. Install Java 7u45 (which you can download from
<http://www.oracle.com/technetwork/java/javase/downloads/java-archive-downloads-javase7-521261.html#jdk-7u45-oth-JPR>)
on the Cloudera Manager Server host.
4. Start the Cloudera Manager Server.

```
sudo service cloudera-scm-server start
```

5. Run the AWS installation wizard.



Note: Due to a bug in Java 7u45 (http://bugs.java.com/bugdatabase/view_bug.do?bug_id=8014618), TLS/SSL connections between the Cloudera Manager Server and Cloudera Manager Agents and between the Cloudera Management Service and CDH processes break intermittently. If you do not have TLS/SSL enabled on your cluster, there is no impact.

The YARN property ApplicationMaster Max Retries has no effect in CDH 5

The issue arises because `yarn.resourcemanager.am.max-retries` was replaced with `yarn.resourcemanager.am.max-attempts`.

Workaround:

1. Add the following to **ResourceManager Advanced Configuration Snippet for yarn-site.xml**, replacing `MAX_ATTEMPTS` with the desired maximum number of attempts:

```
<property>
<name>yarn.resourcemanager.am.max-attempts</name><value>MAX_ATTEMPTS</value>
</property>
```

2. Restart the ResourceManager(s) to pick up the change.

(BDR) Replications can be affected by other replications or commands running at the same time

Replications can be affected by other replications or commands running at the same time, causing replications to fail unexpectedly or even be silently skipped sometimes. When this occurs, a `StaleObjectException` is logged to the Cloudera Manager logs. This is known to occur even with as few as four replications starting at the same time.

Issues Fixed in Cloudera Manager 5.1.1

Checking "Install Java Unlimited Strength Encryption Policy Files" During Add Cluster or Add/Upgrade Host Wizard on RPM based distributions if JDK 7 or above is pre-installed will cause Cloudera Manager and CDH to fail

If you have manually installed Oracle's official JDK 7 or 8 rpm on a host (or hosts), and check the **Install Java Unlimited Strength Encryption Policy Files** checkbox in the Add Cluster or Add Host wizard when installing Cloudera Manager on that host (or hosts), or when upgrading Cloudera Manager to 5.1, Cloudera Manager installs JDK 6 policy files, which will prevent any Java programs from running against that JDK. Additionally, if this situation does apply, Cloudera Manager/CDH will also choose that particular Java as the default to run against, meaning that Cloudera Manager/CDH fail to start, throwing the following message in logs: `Caused by: java.lang.SecurityException: The jurisdiction policy files are not signed by a trusted signer!`.

Workaround: Do not select the **Install Java Unlimited Strength Encryption Policy Files** checkbox during the aforementioned wizards. Instead download and install them manually, following the instructions on Oracle's website.

- JDK 7 Instructions: <http://www.oracle.com/technetwork/java/javase/downloads/jce-7-download-432124.html>
- JDK 8 Instructions: <http://www.oracle.com/technetwork/java/javase/downloads/jce8-download-2133166.html>



Note:

To return to the default limited strength files, reinstall the original Oracle rpm:

- yum - yum reinstall jdk
- zypper - zypper in -f jdk
- rpm - rpm -iv --replacepkgs *filename*, where *filename* is `jdk-7u65-linux-x64.rpm` or `jdk-8u11-linux-x64.rpm`)

Issues Fixed in Cloudera Manager 5.1.0



Important: Cloudera Manager 5.1.0 is no longer available for download from the Cloudera website or from archive.cloudera.com due to the JCE policy file issue described in the [Issues Fixed in Cloudera Manager 5.1.1](#) on page 450 section of the Release Notes. The download URL at archive.cloudera.com for Cloudera Manager 5.1.0 now forwards to Cloudera Manager 5.1.1 for the RPM-based distributions for Linux RHEL and SLES.

Changes to property for `yarn.nodemanager.remote-app-log-dir` are not included in the JobHistory Server `yarn-site.xml` and Gateway `yarn-site.xml`

When "Remote App Log Directory" is changed in YARN configuration, the property `yarn.nodemanager.remote-app-log-dir` are not included in the JobHistory Server `yarn-site.xml` and Gateway `yarn-site.xml`.

Workaround: Set **JobHistory Server Advanced Configuration Snippet (Safety Valve) for `yarn-site.xml`** and **YARN Client Advanced Configuration Snippet (Safety Valve) for `yarn-site.xml`** to:

```
<property>
<name>yarn.nodemanager.remote-app-log-dir</name>
<value>/path/to/logs</value>
</property>
```

Secure CDH 4.1 clusters have Hue and Impala share the same Hive

In a secure CDH 4.1 cluster, Hue and Impala cannot share the same Hive instance. If "Bypass Hive Metastore Server" is disabled on the Hive service, then Hue will not be able to talk to Hive. Conversely, if "Bypass Hive Metastore" enabled on the Hive service, then Impala will have a validation error.

Severity: High

Workaround: Upgrade to CDH 4.2.

The command history has an option to select the number of commands, but does not always return the number you request

Workaround: None.

Hue does not support YARN ResourceManager High Availability

Workaround: Configure the Hue Server to point to the active ResourceManager:

1. Go to the Hue service.
2. Click the **Configuration** tab.
3. Select **Scope > Hue or Hue Service-Wide**.
4. Select **Category > Advanced**.
5. Locate the **Hue Server Advanced Configuration Snippet (Safety Valve) for `hue_safety_valve_server.ini`** property or search for it by typing its name in the Search box.
6. In the **Hue Server Advanced Configuration Snippet for `hue_safety_valve_server.ini`** field, add the following:

```
[hadoop]
[[ yarn_clusters ]]
[[[default]]]
resourcemanager_host=<hostname of active ResourceManager>
resourcemanager_api_url=http://<hostname of active resource manager>:<web port of active
resource manager>
proxy_api_url=http://<hostname of active resource manager>:<web port of active resource
manager>
```

The default web port of Resource Manager is 8088.

7. Click **Save Changes** to have these configurations take effect.

Cloudera Manager 5 Release Notes

8. Restart the Hue service.

Cloudera Manager does not support encrypted shuffle.

Encrypted shuffle has been introduced in CDH 4.1, but it is not currently possible to enable it through Cloudera Manager.

Severity: Medium

Workaround: None.

Hive CLI does not work in CDH 4 when "Bypass Hive Metastore Server" is enabled

Hive CLI does not work in CDH 4 when "Bypass Hive Metastore Server" is enabled.

Workaround: Configure Hive and disable the "Bypass Hive Metastore Server" option.

Alternatively, an approach can be taken that will cause the "Hive Auxiliary JARs Directory" to not work, but will enable basic Hive commands to work. Add the following to "Gateway Client Environment Advanced Configuration Snippet for `hive-env.sh`," then re-deploy the Hive client configuration:

```
HIVE_AUX_JARS_PATH=
AUX_CLASSPATH=/usr/share/java/mysql-connector-java.jar:/usr/share/java/oracle-connector-java.jar:$(/find
/usr/share/cmflib/postgresql-jdbc.jar 2> /dev/null | tail -n 1)
```

Incorrect Absolute Path to topology.py in Downloaded YARN Client Configuration

The downloaded client configuration for YARN includes the `topology.py` script. The location of this script is given by the `net.topology.script.file.name` property in `core-site.xml`. But the `core-site.xml` file downloaded with the client configuration has an incorrect absolute path to `/etc/hadoop/...` for `topology.py`. This can cause clients that run against this configuration to fail (including Spark clients run in yarn-client mode, as well as YARN clients).

Workaround: Edit `core-site.xml` to change the value of the `net.topology.script.file.name` property to the path where the downloaded copy of `topology.py` is located. This property must be set to an absolute path.

search_bind_authentication for Hue is not included in .ini file

When `search_bind_authentication` is set to `false`, Cloudera Manager does not include it in `hue.ini`.

Workaround: Add the following to the Hue Service Advanced Configuration Snippet (Safety Valve) for `hue_safety_valve.ini`:

```
[desktop]
[[ldap]]
search_bind_authentication=false
```

Erroneous warning displayed on the HBase configuration page on CDH 4.1 in Cloudera Manager 5.0.0

An erroneous "Failed parameter validation" warning is displayed on the HBase configuration page on CDH 4.1 in Cloudera Manager 5.0.0

Severity: Low

Workaround: Use CDH4.2 or higher, or ignore the warning.

Host recommissioning and decommissioning should occur independently

In large clusters, when problems appear with a host or role, administrators may choose to decommission the host or role to fix it and then recommission the host or role to put it back in production. Decommissioning, especially host decommissioning, is slow, hence the importance of parallelization, so that host recommissioning can be initiated before decommissioning is done.

Fixed Issues in Cloudera Manager 5.0.7

Sensitive Information in Cloudera Manager Diagnostic Support Bundles

Cloudera Manager is designed to transmit certain diagnostic data (or “bundles”) to Cloudera. These diagnostic bundles are used by the Cloudera support team to reproduce, debug, and address technical issues for our customers. Cloudera internally discovered a potential vulnerability in this feature, which could cause any sensitive data stored as “advanced configuration snippets (ACS)” (formerly called “safety valves”) to be included in diagnostic bundles and transmitted to Cloudera. Notwithstanding any possible transmission, such sensitive data is not used by Cloudera for any purpose.

Cloudera has taken the following actions: (1) modified Cloudera Manager so that it no longer transmits advanced configuration snippets containing the sensitive data, and (2) modified Cloudera Manager TLS/SSL configuration to increase the protection level of the encrypted communication.

Cloudera strives to follow and also help establish best practices for the protection of customer information. In this effort, we continually review and improve our security practices, infrastructure, and data handling policies.

Users affected:

- Users storing sensitive data in advanced configuration snippets

Severity: High

Impact: Possible transmission of sensitive data

CVE: CVE-2015-6495

Immediate Action Required:

- Upgrade Cloudera Manager to one of the following releases: Cloudera Manager 5.4.6, 5.3.7, 5.2.7, 5.1.6, 5.0.7, 4.8.6

Cloudera Manager Agent may become slow or get stuck when responding to commands and when sending heartbeats to Cloudera Manager Server

This issue can occur when Cloudera Navigator auditing is turned on. The auditing code reads audit logs and sends them to the Audit Server. It acquires a lock to protect the list of roles being audited. The same list is also modified by the Cloudera Manager Agent's main thread when a role is started or stopped. If the Audit thread takes too much time to send audits to the Audit Server (which can happen if there is backlog of audit logs), it starves the main Agent thread. This causes the main Agent thread to not send heartbeats and to not respond to commands from the Cloudera Manager Server.

Fixed Issues in Cloudera Manager 5.0.6

Slow staleness calculation can lead to ZooKeeper data loss when new servers are added

In Cloudera Manager 5, starting new ZooKeeper Servers shortly after adding them can cause ZooKeeper data loss when the number of new servers exceeds the number of old servers.

Fixed Issues in Cloudera Manager 5.0.5

“POODLE” vulnerability on TLS/SSL enabled ports

The POODLE (Padding Oracle On Downgraded Legacy Encryption) attack takes advantage of a cryptographic flaw in the obsolete TLS/SSLv3 protocol, after first forcing the use of that protocol. The only solution is to disable TLS/SSLv3 entirely. This requires changes across a wide variety of components of CDH and Cloudera Manager. Cloudera Manager 5.0.5 provides these changes for Cloudera Manager 5.0.x deployments. All Cloudera Manager 5.0.x users should upgrade to 5.0.5 as soon as possible. For more information, see the [Cloudera Security Bulletin](#).

Issues Fixed in Cloudera Manager 5.0.2

Cloudera Manager Impala Query Monitoring does not work with Impala 1.3.1

Impala 1.3.1 contains changes to the runtime profile format that break the Cloudera Manager Query Monitoring feature. This leads to exceptions in the Cloudera Manager Service Monitor logs, and Impala queries no longer appear in the Cloudera Manager UI or API. The issue affects Cloudera Manager 5.0 and 4.6 - 4.8.2.

Workaround: None. The issue will be fixed in Cloudera Manager 4.8.3 and Cloudera Manager 5.0.1. To avoid the Service Monitor exceptions, turn off the Cloudera Manager Query Monitoring feature by going to **Impala Daemon > Monitoring** and setting the Query Monitoring Period to 0 seconds. Note that the Impala Daemons must be restarted when changing this setting, and the setting must be restored once the fix is deployed to turn the query monitoring feature back on. Impala queries will then appear again in Cloudera Manager's Impala query monitoring feature.

Issues Fixed in Cloudera Manager 5.0.1

Upgrade from Cloudera Manager 5.0.0 beta 1 or beta 2 to Cloudera Manager 5.0.0 requires assistance from Cloudera Support

Contact Cloudera Support before upgrading from Cloudera Manager 5.0.0 beta 1 or beta 2 to Cloudera Manager 5.0.0.

Workaround: Contact Cloudera Support.

Failure of HDFS Replication between clusters with YARN

HDFS replication between clusters in different Kerberos realms fails when using YARN if the target cluster is CDH 5.

Workaround: Use MapReduce (MRv1) instead of YARN.

If installing CDH 4 packages, the Impala 1.3.0 option does not work because Impala 1.3 is not yet released for CDH 4.

If installing CDH 4 packages, the Impala 1.3.0 option listed in the install wizard does not work because Impala 1.3.0 is not yet released for CDH 4.

Workaround: Install using parcels (where the unreleased version of Impala does not appear), or select a different version of Impala when installing with packages.

When updating dynamic resource pools, Cloudera Manager updates roles but may fail to update role information displayed in the UI

When updating dynamic resource pools, Cloudera Manager automatically refreshes the affected roles, but they sometimes get marked incorrectly as running with outdated configurations and requiring a refresh.

Workaround: Invoke the **Refresh Cluster** command from the cluster actions drop-down menu.

Upgrade of secure cluster requires installation of JCE policy files

When upgrading a secure cluster via Cloudera Manager, the upgrade initially fails due to the JDK not having Java Cryptography Extension (JCE) unlimited strength policy files. This is because Cloudera Manager installs a copy of the Java 7 JDK during the upgrade, which does not include the unlimited strength policy files. To ensure that unlimited strength functionality continues to work, install the unlimited strength JCE policy files immediately after completing the Cloudera Manager Upgrade Wizard and before taking any other actions in Cloudera Manager.

Workaround: Install the unlimited strength JCE policy files immediately after completing the Cloudera Manager Upgrade Wizard and before taking any other action in Cloudera Manager.

The Details page for MapReduce jobs displays the wrong id for YARN-based replications

The **Details** link for MapReduce jobs is wrong for YARN-based replications.

Workaround: Find the job id in the link and then go to the **YARN Applications** page and look for the job there.

Reset non-default HDFS File Block Storage Location Timeout value after upgrade from CDH 4 to CDH 5

During an upgrade from CDH 4 to CDH 5, if the HDFS File Block Storage Locations Timeout was previously set to a custom value, it will now be set to 10 seconds or the custom value, whichever is higher. This is required for Impala to start in CDH 5, and any value under 10 seconds is now a validation error. This configuration is only emitted for Impala and no services should be adversely impacted.

Workaround: None.

HDFS NFS gateway works only on RHEL and similar systems

Because of a bug in native versions of `portmap/rpcbind`, the HDFS NFS gateway does not work out of the box on SLES, Ubuntu, or Debian systems if you install CDH from the command-line, using packages. It does work on supported versions of RHEL-compatible systems on which `rpcbind-0.2.0-10.el6` or later is installed, and it does work if you use Cloudera Manager to install CDH, or if you start the gateway as root.

Bug: [731542](#) (Red Hat), [823364](#) (SLES), [594880](#) (Debian)

Workarounds and caveats:

- On Red Hat and similar systems, make sure `rpcbind-0.2.0-10.el6` or later is installed.
- On SLES, Debian, and Ubuntu systems, do one of the following:
 - Install CDH using Cloudera Manager; or
 - As of CDH 5.1, start the NFS gateway as root; or
 - [Start the NFS gateway without using packages](#); or
- You can use the gateway by running `rpcbind` in insecure mode, using the `-i` option, but keep in mind that this allows anyone from a remote host to bind to the portmap.

Sensitive configuration values exposed in Cloudera Manager

Certain configuration values that are stored in Cloudera Manager are considered sensitive, such as database passwords. These configuration values should be inaccessible to non-administrator users, and this is enforced in the Cloudera Manager Administration Console. However, these configuration values are not redacted when they are read through the API, possibly making them accessible to users who should not have such access.

Gateway role configurations not respected when deploying client configurations

Gateway configurations set for gateway role groups other than the default one or at the role level were not being respected.

Documentation reflects requirement to enable at least Level 1 encryption before enabling Kerberos authentication

Cloudera Security documentation now indicates that before enabling Kerberos authentication you should first enable at least Level 1 encryption.

HDFS NFS gateway does not work on all Cloudera-supported platforms

The NFS gateway cannot be started on some Cloudera-supported platforms.

Workaround: None. Fixed in Cloudera Manager 5.0.1.

Replace `YARN_HOME` with `HADOOP_YARN_HOME` during upgrade

If `yarn.application.classpath` was set to a non-default value on a CDH 4 cluster, and that cluster is upgraded to CDH 5, the classpath is not updated to reflect that `$YARN_HOME` was replaced with `$HADOOP_YARN_HOME`. This will cause YARN jobs to fail.

Workaround: Reset `yarn.application.classpath` to the default, then re-apply your classpath customizations if needed.

Cloudera Manager 5 Release Notes

Insufficient password hashing in Cloudera Manager

In versions of Cloudera Manager earlier than 4.8.3 and earlier than 5.0.1, user passwords are only hashed once. Passwords should be hashed multiple times to increase the cost of dictionary based attacks, where an attacker tries many candidate passwords to find a match. The issue only affects user accounts that are stored in the Cloudera Manager database. User accounts that are managed externally (for example, with LDAP or Active Directory) are not affected.

In addition, because of this issue, Cloudera Manager 4.8.3 cannot be upgraded to Cloudera Manager 5.0.0. Cloudera Manager 4.8.3 must be upgraded to 5.0.1 or later.

Workaround: Upgrade to Cloudera Manager 5.0.1.

Upgrade to Cloudera Manager 5.0.0 from SLES older than Service Pack 3 with PostgreSQL older than 8.4 fails

Upgrading to Cloudera Manager 5.0.0 from SUSE Linux Enterprise Server (SLES) older than Service Pack 3 will fail if the embedded PostgreSQL database is in use and the installed version of PostgreSQL is less than 8.4.

Workaround: Either migrate away from the embedded PostgreSQL database (use MySQL or Oracle) or upgrade PostgreSQL to 8.4 or greater.

MR1 to MR2 import fails on a secure cluster

When running the MR1 to MR2 import on a secure cluster, YARN jobs will fail to find container-executor.cfg.

Workaround: Restart YARN after the import.

After upgrade from CDH 4 to CDH 5, Oozie is missing workflow extension schemas

After an upgrade from CDH 4 to CDH 5, Oozie does not pick up the new workflow extension schemas automatically. User will need to update oozie.service.SchemaService.wf.ext.schemas manually and add the schemas added in CDH 5: shell-action-0.3.xsd, sqoop-action-0.4.xsd, distcp-action-0.2.xsd, oozie-sla-0.1.xsd, oozie-sla-0.2.xsd. Note: None of the existing jobs will be affected by this bug, only new workflows that require new schemas.

Workaround: Add the new workflow extension schemas to Oozie manually by editing oozie.service.SchemaService.wf.ext.schemas.

Issues Fixed in Cloudera Manager 5.0.0

HDFS replication does not work from CDH 5 to CDH 4 with different realms

HDFS replication does not work from CDH 5 to CDH 4 with different realms. This is because authentication fails for services in a non-default realm via the WebHdfs API due to a JDK bug. This has been fixed in JDK6-u34 (b03) and in JDK7.

Workaround: Use JDK 7 or upgrade JDK6 to at least version u34.

The Sqoop Upgrade command in Cloudera Manager may report success even when the upgrade fails

Workaround: Do one of the following:

- 1. Click the Sqoop service and then the **Instances** tab.
 - 2. Click the Sqoop server role then the **Commands** tab.
 - 3. Click the **stdout** link and scan for the Sqoop Upgrade command.
- In the All Recent Commands page, select the **stdout** link for latest Sqoop Upgrade command.

Verify that the upgrade did not fail.

Cannot restore a snapshot of a deleted HBase table

If you take a snapshot of an HBase table, and then delete that table in HBase, you will not be able to restore the snapshot.

Severity: Med

Workaround: Use the "Restore As" command to recreate the table in HBase.

Stop dependent HBase services before enabling HDFS Automatic Failover.

When enabling HDFS Automatic Failover, you need to first stop any dependent HBase services. The Automatic Failover configuration workflow restarts both NameNodes, which could cause HBase to become unavailable.

Severity: Medium

New schema extensions have been introduced for Oozie in CDH 4.1

In CDH 4.1, Oozie introduced new versions for Hive, Sqoop and workflow schema. To use them, you must add the new schema extensions to the Oozie SchemaService Workflow Extension Schemas configuration property in Cloudera Manager.

Severity: Low

Workaround: In Cloudera Manager, do the following:

1. Go to the CDH 4 **Oozie** service page.
2. Go to the **Configuration** tab, **View and Edit**.
3. Search for "Oozie Schema". This should show the **Oozie SchemaService Workflow Extension Schemas** property.
4. Add the following to the **Oozie SchemaService Workflow Extension Schemas** property:

```
shell-action-0.2.xsd  
hive-action-0.3.xsd  
sqoop-action-0.3.xsd
```

5. Save these changes.

YARN Resource Scheduler user FairScheduler rather than FIFO.

Cloudera Manager 5.0.0 sets the default YARN Resource Scheduler to FairScheduler. If a cluster was previously running YARN with the FIFO scheduler, it will be changed to FairScheduler next time YARN restarts. The FairScheduler is only supported with CDH4.2.1 and later, and older clusters may hit failures and need to manually change the scheduler to FIFO or CapacityScheduler.

Severity: Medium

Workaround: For clusters running CDH 4 prior to CDH 4.2.1:

1. Go the YARN service Configuration page
2. Search for "scheduler.class"
3. Click in the Value field and select the schedule you want to use.
4. Save your changes and restart YARN to update your configurations.

Resource Pools Summary is incorrect if time range is too large.

The Resource Pools Summary does not show correct information if the Time Range selector is set to show 6 hours or more.

Severity: Medium

Workaround: None.

When running the MR1 to MR2 import on a secure cluster, YARN jobs will fail to find container-executor.cfg

Workaround: Restart YARN after the import steps finish. This causes the file to be created under the YARN configuration path, and the jobs now work.

Cloudera Manager 5 Release Notes

[When upgrading to Cloudera Manager 5.0.0, the "Dynamic Resource Pools" page is not accessible](#)

When upgrading to Cloudera Manager 5.0.0, users will not be able to directly access the "Dynamic Resource Pools" page. Instead, they will be presented with a dialog saying that the Fair Scheduler XML Advanced Configuration Snippet is set.

Workaround:

1. Go to the YARN service.
2. Click the **Configuration** tab.
3. Select **Scope > Resource Manager or YARN Service-Wide**.
4. Select **Category > Advanced**.
5. Locate the **Fair Scheduler XML Advanced Configuration Snippet** property or search for it by typing its name in the Search box.
6. Copy the value of the **Fair Scheduler XML Advanced Configuration Snippet** into a file.
7. Clear the value of **Fair Scheduler XML Advanced Configuration Snippet**.
8. Recreate the desired Fair Scheduler allocations in the **Dynamic Resource Pools** page, using the saved file for reference.

[New Cloudera Enterprise licensing is not reflected in the wizard and license page](#)

Workaround: None.

[The AWS Cloud wizard fails to install Spark due to missing roles](#)

Workaround: Do one of the following:

- Use the Installation wizard.
- Open a new window, click the Spark service, click on the **Instances** tab, click **Add**, add all required roles to Spark. Once the roles are successfully added, click the **Retry** button in the Installation wizard.

[Spark on YARN requires manual configuration](#)

Spark on YARN requires the following manual configuration to work correctly: modify the YARN Application Classpath by adding /etc/hadoop/conf, making it the very first entry.

Workaround: Add /etc/hadoop/conf as the first entry in the YARN Application classpath.

[Monitoring works with Solr and Sentry only after configuration updates](#)

Cloudera Manager monitoring does not work out of the box with Solr and Sentry on Cloudera Manager 5. The Solr service is in Bad health, and all Solr Servers have a failing "Solr Server API Liveness" health check.

Severity: Medium

Workaround: Complete the configuration steps below:

1. Create "HTTP" user and group on all machines in the cluster (with `useradd 'HTTP'` on RHEL-type systems).
2. The instructions that follow this step assume there is no existing Solr Sentry policy file in use. In that case, first create the policy file on /tmp and then copy it over to the appropriate location in HDFS that Solr Servers check. If there is already a Solr Sentry policy in use, it must be modified to add the following [group] / [role] entries for 'HTTP'. Create a file (for example, /tmp/cm-authz-solr-sentry-policy.ini) with the following contents:

```
[groups]
HTTP = HTTP
[roles]
HTTP = collection = admin->action=query
```

3. Copy this file to the location for the "Sentry Global Policy File" for Solr. The associated config name for this location is `sentry.solr.provider.resource`, and you can see the current value by navigating to the **Sentry** sub-category in the **Service Wide** configuration editing workflow in the Cloudera Manager UI. The default value for this entry is `/user/solr/sentry-provider.ini`. This refers to a path in HDFS.

- 4.** Check if you have entries in HDFS for the parent(s) directory:

```
sudo -u hdfs hadoop fs -ls /user
```

- 5.** You may need to create the appropriate parent directories if they are not present. For example:

```
sudo -u hdfs hadoop fs -mkdir /user/solr/sentry
```

- 6.** After ensuring the parent directory is present, copy the file created in step 2 to this location, as follows:

```
sudo -u hdfs hadoop fs -put /tmp/cm-authz-solr-sentry-policy.ini  
/user/solr/sentry/sentry-provider.ini
```

- 7.** Ensure that this file is owned/readable by the solr user (this is what the Solr Server runs as):

```
sudo -u hdfs hadoop fs -chown solr /user/solr/sentry/sentry-provider.ini
```

- 8.** Restart the Solr service. If both Kerberos and Sentry are being enabled for Solr, the MGMT services also need to be restarted. The Solr Server liveness health checks should clear up once SMON has had a chance to contact the servers and retrieve metrics.

[Out-of-memory errors may occur when using the Reports Manager](#)

Out-of-memory errors may occur when using the Cloudera Manager Reports Manager.

Workaround: Set the value of the "Java Heap Size of Reports Manager" property to at least the size of the HDFS filesystem image (`fsimage`) and restart the Reports Manager.

[Applying license key using Internet Explorer 9 and Safari fails](#)

Cloudera Manager is designed to work with IE 9 and above and Safari. However the file upload widget used to upload a license currently does not work with IE 9 or Safari. Therefore, installing an enterprise license does not work.

Workaround: Use another supported browser.

Issues Fixed in Cloudera Manager 5.0.0 Beta 2

[The Sqoop Upgrade command in Cloudera Manager may report success even when the upgrade fails](#)

Workaround: Do one of the following:

- **1.** Click the Sqoop service and then the **Instances** tab.
- **2.** Click the Sqoop server role then the **Commands** tab.
- **3.** Click the **stdout** link and scan for the Sqoop Upgrade command.
- In the All Recent Commands page, select the **stdout** link for latest Sqoop Upgrade command.

Verify that the upgrade did not fail.

[The HDFS Canary Test is disabled for secured CDH 5 services.](#)

Due to a bug in Hadoop's handling of multiple RPC clients with distinct configurations within a single process with Kerberos security enabled, Cloudera Manager will disable the HDFS canary test when security is enabled so as to prevent interference with Cloudera Manager's MapReduce monitoring functionality.

Severity: Medium

Workaround: None

[Not all monitoring configurations are migrated from MR1 to MR2.](#)

When MapReduce v1 configurations are imported for use by YARN (MR2), not all of the monitoring configuration values are currently migrated. Users may need to reconfigure custom values for properties such as thresholds.

Cloudera Manager 5 Release Notes

Severity: Medium

Workaround: Manually reconfigure any missing property values.

"Access Denied" may appear for some features after adding a license or starting a trial.

After starting a 60-day trial or installing a license for Enterprise Edition, you may see an "access denied" message when attempting to access certain Enterprise Edition-only features such as the Reports Manager. You need to log out of the Admin Console and log back in to access these features.

Severity: Low

Workaround: Log out of the Admin Console and log in again.

Hue must set impersonation on when using Impala with impersonation.

When using Impala with impersonation, the `impersonation_enabled` flag must be present and configured in the `hue.ini` file. If impersonation is enabled in Impala (in other words, if Impala is using Sentry) then this flag must be set **true**. If Impala is not using impersonation, it should be set **false** (the default).

Workaround: Set advanced configuration snippet value for `hue.ini` as follows:

1. Go to the **Hue Service Configuration Advanced Configuration Snippet for `hue_safety_valve.ini`** under the Hue service Configuration settings, **Service-Wide > Advanced** category.
2. Add the following, then uncomment the setting and set the value True or False as appropriate:

```
#####
# Settings to configure Impala
#####

[impala]
...
# Turn on/off impersonation mechanism when talking to Impala
## impersonation_enabled=False
```

Cloudera Manager Server may fail to start when upgrading using a PostgreSQL database.

If you're upgrading to Cloudera Manager 5.0.0 beta 1 and you're using a PostgreSQL database, the Cloudera Manager Server may fail to start with a message similar to the following:

```
ERROR [main:dbutil.JavaRunner@57] Exception while executing
com.cloudera.cmf.model.migration.MigrateConfigRevisions
java.lang.RuntimeException: java.sql.SQLException: Batch entry <xxx> insert into REVISIONS
(REVISION_ID, OPTIMISTIC_LOCK_VERSION, USER_ID, TIMESTAMP, MESSAGE) values (...) was aborted. Call getNextException to see the cause.
```

Workaround: Use `psql` to connect directly to the server's database and issue the following SQL command:

```
alter table REVISIONS alter column MESSAGE type varchar(1048576);
```

After that, your Cloudera Manager server should start up normally.

Issues Fixed in Cloudera Manager 5.0.0 Beta 1

After an upgrade from Cloudera Manager 4.6.3 to 4.7, Impala does not start.

After an upgrade from Cloudera Manager 4.6.3 to 4.7 when Navigator is used, Impala will fail to start because the Audit Log Directory property has not been set by the upgrade procedure.

Severity: Low.

Workaround: Manually set the property to `/var/log/impalad/audit`.

Cloudera Navigator 2 Data Management Release Notes

These release notes provide information on the new and changed features, known issues, and fixed issues for Cloudera Navigator.

New Features and Changes in Cloudera Navigator 2 Data Management

The following sections describe what's new and changed in each Cloudera Navigator 2 release.

New Features in Cloudera Navigator 2

The following sections describe what's new in each Cloudera Navigator 2 release.

What's New in Cloudera Navigator 2.8.0

- Support for purging metadata for Hive and Impala select queries, and YARN, Sqoop, and Pig operations, has been added. See [Purging Metadata for HDFS Entities, Hive and Impala Select Queries, and YARN, Sqoop, and Pig Operations](#).
- A new role for editing metadata has been introduced: Custom Metadata Administrator has been added to User Roles. Now there is a role for editing managed metadata, and a new role for editing custom metadata. This enables data stewards and curators to restrict access to editing managed metadata, while enabling end users to tag data sets with custom metadata. See [User Roles](#).
- You can configure display of inputs and outputs in the entity Details page. See [Configuring Display of Inputs and Outputs](#).

A number of issues have been fixed. See [Issues Fixed in Cloudera Navigator 2.8.0](#) on page 469.

What's New in Cloudera Navigator 2.7.3

An issue has been fixed. See [Issues Fixed in Cloudera Navigator 2.7.3](#) on page 470.

What's New in Cloudera Navigator 2.7.2

A number of issues have been fixed. See [Issues Fixed in Cloudera Navigator 2.7.2](#) on page 471.

What's New in Cloudera Navigator 2.7.1

An issue has been fixed. See [Issues Fixed in Cloudera Navigator 2.7.1](#) on page 471.

What's New in Cloudera Navigator 2.7.0

- Platform enhancements**
 - Added support for assigning managed metadata through policies. Only single-valued managed metadata of type Text is currently supported. See [Metadata Policies](#).
 - Added support for purging physical operations corresponding to logical operations. See [Performing Actions on Entities](#).
- Cross Site Request Forgery (CSRF)** protection is enabled by default. If you currently use cookies for authentication in API requests with a non-GET method (for example, the POST method), you must also obtain CSRF tokens for the request, or disable CSRF protection.

You can disable CSRF protection by setting `nav.disable_api_csrf_security=true`.

A number of issues have been fixed. See [Issues Fixed in Cloudera Navigator 2.7.0](#) on page 471

Cloudera Navigator 2 Data Management Release Notes

What's New in Cloudera Navigator 2.6.5

A number of issues have been fixed. See [Issues Fixed in Cloudera Navigator 2.6.5](#) on page 472

What's New in Cloudera Navigator 2.6.4

There are no new features or fixed issues.

What's New in Cloudera Navigator 2.6.3

There are no new features or fixed issues.

What's New in Cloudera Navigator 2.6.2

A number of issues have been fixed. See [Issues Fixed in Cloudera Navigator 2.6.2](#) on page 473.

What's New in Cloudera Navigator 2.6.1

A number of issues have been fixed. See [Issues Fixed in Cloudera Navigator 2.6.1](#) on page 473.

What's New in Cloudera Navigator 2.6.0

- **Platform enhancements**

- Added support for defining new types of business metadata.
- Enhanced entity details with type-specific information and behavior.
- Added support for filtering lineage.
- Added support for purging the metadata store.
- Added support for securing audit messages sent to Kafka.

A number of issues have been fixed. See [Issues Fixed in Cloudera Navigator 2.6.0](#) on page 474.

What's New in Cloudera Navigator 2.5.1



Note: In the Cloudera Navigator 2.5.1 release, there are no new features or fixed issues.

What's New in Cloudera Navigator 2.5.0

An issue has been fixed. See [Issues Fixed in Cloudera Navigator 2.5.0](#) on page 475.

What's New in Cloudera Navigator 2.4.4

Several issues have been fixed. See [Issues Fixed in Cloudera Navigator 2.4.4](#) on page 475.

What's New in Cloudera Navigator 2.4.3



Note: In the Cloudera Navigator 2.4.3 release, there are no new features or fixed issues.

What's New in Cloudera Navigator 2.4.2



Note: If you are upgrading from a previous version of Cloudera Navigator, Cloudera recommends that you upgrade to version 2.4.3 or higher.

A number of issues have been fixed. See [Issues Fixed in Cloudera Navigator 2.4.2](#) on page 475.

What's New in Cloudera Navigator 2.4.1

An issue has been fixed. See [Issues Fixed in Cloudera Navigator 2.4.1](#) on page 477.

What's New in Cloudera Navigator 2.4.0

- **Platform enhancements**

- New entity details page.
- HDFS metadata and audit analytics.
- Command actions, integrated with policies and search, for archiving and purging files.
- Redesigned policies page for scalability.
- Publishing of audit events to Kafka topics. Kafka audit event publishing does not support authorization and Cloudera does not recommend its use in production.
- Navigator SDK for metadata and lineage augmentation.
- Significant scale improvements. Large lineage graphs render faster and consume less memory.

- **Expanded service coverage**

- Hive metadata: support for extended attributes.
- Hive on Spark metadata and lineage.



Important: Like Spark metadata and lineage, Hive on Spark metadata and lineage is not supported or recommended for production use. By default it is disabled. To try this feature, use it in a test environment until Cloudera resolves currently existing issues and limitations to make it ready for production use.

- Oozie metadata and lineage added support for the `hive2` action, which Cloudera recommends that you use in preference to the `hive` action.
- Hue auditing (CDH 5.5.0 and higher): added login and logout, user and group events.
- Cloudera Manager auditing: login and logout events.
- Navigator Metadata Server auditing: successful and failed login events.
- Added filtering of audit events for Sentry, Solr, Impala, and Navigator Metadata Server.

Also see [What's New in Cloudera Manager 5.5.0](#) on page 362.

What's New in Cloudera Navigator 2.3.10



Note: In the Cloudera Navigator 2.3.10 release, there are no new features or fixed issues.

What's New in Cloudera Navigator 2.3.9

An issue has been fixed. See [Issues Fixed in Cloudera Navigator 2.3.9](#) on page 478.

What's New in Cloudera Navigator 2.3.8

Several issues have been fixed. See [Issues Fixed in Cloudera Navigator 2.3.8](#) on page 479.



Note: In the Cloudera Navigator 2.3.5, 2.3.6, and 2.3.7 releases, there are no new features or fixed issues.



Note: Although there is a CDH 5.4.4 release, there is no synchronous Cloudera Navigator 2.3.4 release.

Cloudera Navigator 2 Data Management Release Notes

What's New in Cloudera Navigator 2.3.3

Several issues have been fixed. See [Issues Fixed in Cloudera Navigator 2.3.3](#) on page 479.



Note: Although there is a CDH 5.4.2 release, there is no synchronous Cloudera Navigator 2.3.2 release.

What's New in Cloudera Navigator 2.3.1

- Navigator self audit events have been enhanced with additional information such as the names of audit reports and policies
- Performance and stability improvements

Also, a number of issues have been fixed. See [Issues Fixed in Cloudera Navigator 2.3.1](#) on page 480.

What's New in Cloudera Navigator 2.3.0

- **Platform enhancements**

- Redesigned metadata search provides autocomplete, enhanced filtering, and saving searches.
- Added support for SAML for single sign-on.

- **Expanded service coverage**

- Added Impala (CDH 5.4 and higher) metadata and lineage
- Added Cloudera Search (CDH 5.4 and higher) auditing
- Added auditing for Navigator Metadata Server activity, such as audit views, metadata searches, and policy editing
- Added support for inferring the schema of HDFS Avro and Parquet entities
- Added Spark (CDH 5.4 and higher) metadata and lineage.



Important: Spark metadata and lineage is not supported or recommended for production use. By default it is disabled. To try this feature, use it in a test environment until Cloudera resolves currently existing issues and limitations to make it ready for production use.

What's New in Cloudera Navigator 2.2.10



Note: In the Cloudera Navigator 2.2.10 release, there are no new features or fixed issues.

What's New in Cloudera Navigator 2.2.9

An issue has been fixed. See [Issues Fixed in Cloudera Navigator 2.3.9](#) on page 478.



Note: In the Cloudera Navigator 2.2.6, 2.2.7, and 2.2.8 releases, there are no new features or fixed issues.



Note: Although there is a CDH 5.3.5 release, there is no synchronous Cloudera Navigator 2.2.5 release.

What's New in Cloudera Navigator 2.2.4

An issue has been fixed. See [Issues Fixed in Cloudera Navigator 2.2.4](#) on page 481.

What's New in Cloudera Navigator 2.2.3

An issue has been fixed. See [Issues Fixed in Cloudera Navigator 2.2.3](#) on page 481.

What's New in Cloudera Navigator 2.2.2

An issue has been fixed. See [Issues Fixed in Cloudera Navigator 2.2.2](#) on page 481.

What's New in Cloudera Navigator 2.2.1

A number of issues have been fixed. See [Issues Fixed in Cloudera Navigator 2.2.1](#) on page 481.

What's New in Cloudera Navigator 2.2.0

- **Policies** are generally available and are always enabled. Policy properties now support Java expressions.
- **Search** - Search functionality now includes autocomplete.

What's New in Cloudera Navigator 2.1.6

An issue has been fixed. See [Issues Fixed in Cloudera Navigator 2.1.6](#) on page 482.

What's New in Cloudera Navigator 2.1.5

A number of issues have been fixed. See [Issues Fixed in Cloudera Navigator 2.1.5](#) on page 482.

What's New in Cloudera Navigator 2.1.4

An issue has been fixed. See [Issues Fixed in Cloudera Navigator 2.1.4](#) on page 482.



Note: There was no Cloudera Navigator 2.1.3 release.

What's New in Cloudera Navigator 2.1.2

A number of issues have been fixed. See [Issues Fixed in Cloudera Navigator 2.1.2](#) on page 482.

What's New in Cloudera Navigator 2.1.1

A number of issues have been fixed. See [Issues Fixed in Cloudera Navigator 2.1.1](#) on page 483.

What's New in Cloudera Navigator 2.1.0

- **Auditing Component**
 - New auditing UI featuring saved audit reports. The Navigator auditing UI is no longer available from Cloudera Manager. Instead, auditing is integrated with lineage, discovery, and the policy engine. The UI is provided by the Metadata Server, which is now required for the auditing component.
 - Sentry auditing now includes Sentry commands issued from Impala.
- **Metadata Component**
 - Search results contain a type appropriate link to a Hue browser.
 - HDFS directories and files - File Browser
 - Hive database and tables - Metastore Manager
 - MapReduce, YARN, Pig - Job Browser
 - **Policies** - support rules for modifying metadata and sending notifications when entities are extracted.



Note: Policies is a beta feature that is disabled by default.

- **Security**

Cloudera Navigator 2 Data Management Release Notes

- **Role-Based Access Control** - support assigning groups to roles that constrain access to Navigator features
- **Authentication** - LDAP and Active Directory authentication of Navigator users
- **TLS/SSL** - enable TLS/SSL for encrypted communication
- **API**
 - Version changed to v3
 - Supports auditing and policies

What's New in Cloudera Navigator 2.0.5

An issue was fixed. See [Issues Fixed in Cloudera Navigator 2.0.5](#) on page 483.



Note: There is no Cloudera Navigator 2.0.4 release.

What's New in Cloudera Navigator 2.0.3

A number of issues have been fixed. See [Issues Fixed in Cloudera Navigator 2.0.3](#) on page 483.

What's New in Cloudera Navigator 2.0.2

An issue was fixed. See [Issues Fixed in Cloudera Navigator 2.0.2](#) on page 483.

What's New in Cloudera Navigator 2.0.1

- Masking of personally identifiable information (PII) in query strings that appear in audit events and lineage. Enabled by default.
- REST API support for registering custom metadata for entities before they appear in Navigator.

A number of issues have been fixed. See [Issues Fixed in Cloudera Navigator 2.0.1](#) on page 483.

What's New in Cloudera Navigator 2.0.0

- **Auditing Component**
 - Added support for auditing the Sentry service
 - Added support for publishing audit logs to syslog
- **Metadata Component**
 - Newly designed Query Builder with faceted filtering
 - Simplified Pig lineage
 - Added support for Sqoop and Oozie lineage
 - Many performance and stability improvements
- **Security** - includes Navigator Encrypt and Navigator Key Trustee, formerly known as Gazzang zNcrypt and Gazzang zTrustee. These features provide enterprise-grade encryption and key management. For information on these features, see the [Cloudera Security Datasheet](#) and contact your account team.

Changed Features in Cloudera Navigator 2

The following sections describe what's changed in each Cloudera Navigator 2 release.

What's Changed in Cloudera Navigator 2.7.0

Previously, Cloudera Navigator captured the underlying MapReduce and YARN operations for Hive, Sqoop, and Pig operations, and stored them as physical operations. As of release 2.7.0, Cloudera Navigator removes previously captured physical operations, operation executions, and associated relations during purge. Removing these physical operations increases the overall scalability of Cloudera Navigator.

During the upgrade to Cloudera Navigator 2.7, Cloudera Navigator removes previously captured physical operations, operation executions, and associated relations.

What's Changed in Cloudera Navigator 2.6.0

- Queries generated using filters now use the [Lucene Query Parser](#) , + , - Boolean operator syntax.
- Hive on Spark metadata extraction is now supported.
- The Navigator Metadata Server requires 192 MiB of Java PermGen space instead of 128 MiB. This is increased automatically when upgrading to Cloudera Manager 5.7.

What's Changed in Cloudera Navigator 2.4.0

- Cloudera Navigator no longer supports JDK 1.6.
- In the Administration tab, after roles have been assigned to at least one group, **Groups with Navigator Roles** is the default selection.

What's Changed in Cloudera Navigator 2.3.3

- In the Search UI, facet values with the count of 0 are not displayed.

What's Changed in Cloudera Navigator 2.3.0

- The PII masking regular expression is superseded by log file redaction in CDH 5.4.

What's Changed in Cloudera Navigator 2.2.0

- **Metadata Component** - Policies created with Cloudera Navigator 2.1 (containing the Beta version policy engine) are not retained when upgrading to Cloudera Navigator 2.2.

What's Changed in Cloudera Navigator 2.1.0

- **Auditing Component**
 - HDFS audit events generated by the `impala` user are discarded by default.

What's Changed in Cloudera Navigator 2.0.1

- **Metadata Component**
 - The REST API version changed to v2.

What's Changed in Cloudera Navigator 2.0.0

- **Auditing Component**
 - HDFS audit events generated by the `solr` user are discarded by default.

Known Issues and Workarounds in Cloudera Navigator 2 Data Management

The following sections describe the current known issues in Cloudera Navigator 2.

User might see a "not authorized" message when logging in

The Navigator UI saves the state of the last URL accessed when you log out, and takes you to the same page on the next login. If two different users log in to Navigator using the same browser tab, the state of the first user applies to the second. If the second user does not have permissions to that page, that user receives an error message.

Cloudera Navigator 2 Data Management Release Notes

Workaround: Close the browser tab and log in on a new tab. The state is cleared, and the access error message does not appear.

Spark metadata extraction and lineage is unsupported and disabled by default

Purge specifications for Navigator

Policies cannot use cluster names in queries. Cluster name is a derived attribute and cannot be used as-is.

Workaround: When setting move actions for Cloudera Navigator, if there is only one cluster known to the Navigator instance, remove the `clusterName` clause.

If there is more than one cluster known to the Navigator instance, replace `clusterName` with `sourceId`. To get the `sourceId`, issue a query in this format:

```
curl '<nav-url>/api/v9/entities/?query=type%3Asource&limit=100&offset=0'
```

Use the identity of the matching HDFS service for this cluster as the `sourceId`.

[When creating or updating a managed metadata property, selected class is not shown on the review screen](#)

When creating a new managed metadata property, there is no information about class on the review screen, but the class (as selected in the previous screen) is still saved together with the property. When updating an existing managed metadata property, there is only information about the existing class(es) on the review screen, but no information about any additional class to be added.

Workaround: None.

[Lineage returns deleted entities even if the Hide deleted entities option is selected](#)

[Purge appears suspended while extraction is running](#)

When you issue a purge command, it does not start if extractors are running. During this time, the maintenance page indicates that maintenance tasks are not running. Once the extraction is complete and purge starts, it shows the status of the purge operations.

[Audit logs are not drained when audited process is stopped](#)

If an audited role is deleted or migrated to a different host and there are pending audits that are waiting to be transferred to Audit Server, then those audits may not get transferred. There are pending audits when audits cannot be transferred either because Audit Server is down or is unreachable because of network issue. So during role migration ensure that Audit Server is in healthy state to make sure all audited actions make to Audit Server.

[Spurious errors about missing database connectors are reported in the Metadata Server log file](#)

Workaround: Ignore the errors.

[Audit CSV has extra columns and is missing some data](#)

When you export audits to CSV, Sentry data is not visible in the generated CSV file. Also, some of the column names show up twice (Operation Text, Database Name, Object Type, and so on.), but the data only shows up in one of the columns.

Workaround: Export audits to JSON format to see Sentry data.

[Metadata component in Cloudera Navigator 1.2 \(included with Cloudera Manager 5.0\) cannot be upgraded to 2.0](#)

Cloudera does not provide an upgrade path from the Navigator Metadata Server that was a beta release in Cloudera Navigator 1.2 to the Cloudera Navigator 2 release. If you are upgrading from Cloudera Navigator 1.2 (included with Cloudera Manager 5.0), you must perform a clean install of Cloudera Navigator 2.

Workaround:

1. Delete the Navigator Metadata Server role.
2. Remove the contents of the Navigator Metadata Server storage directory.
3. Add the Navigator Metadata Server role according to the process described in [Adding the Navigator Metadata Server](#).
4. Clear the cache of any browser that used the 1.2 release of the Navigator Metadata component. Otherwise, you may see errors in the Navigator Metadata UI.

The Hive extractor does not handle all Hive statements

The Hive extractor does not handle the following cases:

- Table generating functions
- Lateral views
- Transform clauses
- Regular expression in select clause

If a query involves any of the above, lineage will not be complete for that Hive query.

Workaround: None.**The IP address in a Hue service audit log shows as "unknown"**

The IP address in a Hue service audit log shows as "unknown".

Severity: Low**Workaround:** None.**Hive service configuration impact on Hue service auditing**

If the audit configuration for a Hive service is changed, Beeswax must be restarted to pick up the change in the Hue service audit log.

Severity: Low**Workaround:** None.**Hive service configuration in auditing component**

For Hive services, the auditing component does not support the "Shutdown" option for the "Queue Policy" property.

Severity: Low**Workaround:** None.

Issues Fixed in Cloudera Navigator 2 Data Management

The following sections describe the issues fixed in each Cloudera Navigator 2 release.

Issues Fixed in Cloudera Navigator 2.8.0

Last access time displayed for Hive tables is incorrect

Navigator no longer displays incorrect access time information for Hive.

Deleted items not shown by default when Deleted filter added

Showing deleted items when the Deleted filter was not enabled by default. Deleted items are now shown by default when the Deleted filter is applied.

Cloudera Navigator 2 Data Management Release Notes

No message shown when no Groups with Navigator roles are found

A message is now displayed in Navigator when no Groups with Navigator roles are found.

Hive started and ended time fields not shown in results

Previously, started and ended time fields were not displayed when selected. These fields are now displayed in results.

Maximum Complexity error message value incorrect

The value in the Maximum Complexity Exceeded error message has been corrected (>3000).

List of properties in audit filter not sorted

Audit filter properties are now sorted alphabetically.

Unable to exit Lineage Full Screen mode from Search Results page

Audit results do not display seconds or fractions seconds

Audit result display granularity has increased to show seconds and fractions of seconds in results.

Search on lineage page does not work for children

Lineage children are now shown when selected in a lineage search.

Create property UI does not use class display names

The create property user interface now displays the correct class display name instead of the abbreviated classname.

Regex entered cannot be confirmed as correct in managed metadata

Navigator now includes regex test validation to managed metadata creation.

Create namespace does not check against reserved names

The Navigator UI now rejects a reserved namespace when creating a new namespace.

Issues Fixed in Cloudera Navigator 2.7.3

Incomplete query strings in Navigator metadata

Storing more than 32K characters in Solr causes an exception. Navigator now truncates any field longer than 32K to prevent the exception.

Issues Fixed in Cloudera Navigator 2.7.2

Long query text alters the display of the search box

Popup error in UI in clicking on hive table entity's schema entity

Lineage Options and Search items are missing

Google analytics might not load if Navigator is served over https/ssl

Issues Fixed in Cloudera Navigator 2.7.1

Oozie and Yarn metadata is not visible for some jobs

Issues Fixed in Cloudera Navigator 2.7.0

Linker causing high memory consumption

AbstractSolrManager.query incorrectly set autoAdjustBatchSize parameter to true. This caused Navigator to always retrieve all remaining results for a query in the second batch request, resulting in high memory usage.

CheckDeletedEntity::start method should only act on entities that are not deleted

The CheckDeletedEntity::start method uses the *:***** query to get entities to mark as deleted. This adversely affects performance and can cause an upgrade to enter an infinite loop if the server crashes or is restarted during upgrade.

Managed metadata does not show up in lineage diagram

User specified relations should be marked correctly in lineage

User-defined relations are now properly marked in lineage view.

Lineage export returns a server error if direction is not included

Policies Command actions placement differ between edit and view

Delete VIEW to FILE relations

Logical physical relationships from Hive to YARN should be removed

popover2 binding is positioned incorrectly

Remove physical operations from Navigator metadata

For existing data, Cloudera Navigator performs an upgrade to remove existing physical operations, operation executions, and relations that involve them. The upgrade process starts with logical-physical relations to operations and operation executions. Navigator removes the relations themselves, their physical endpoints, and any relations connected to those physical endpoints.

Cloudera Navigator 2 Data Management Release Notes

Show parent when applicable on entity details page

Entity cannot be deselected in Lineage

Increase zoom out capability in lineage

API/interactive/entities fails when managed metadata properties exist but do not yet have values

Entity Details page edit metadata does not show namespace in managed metadata

Actions menu should not show Move for deleted entities

No Operation/OperationExecution extracted for Hive and HDFS

numPartitions attribute for Hive should be suppressed

Inconsistent size and block size units

Search dropdown width should be the same as the search input field width

Heap for Navigator Metadata Server needs to be increased

Improve labels for Analytics top users and top commands filter

CSRF security not enabled by default

CSRF security is now enabled by default.

Policy earliest start time should be current time

Policy editor line wrapping divides words in the middle

Date format is different for facets, breadcrumbs, and full query text

Improve rendering for parent elements

"created" date is later than "lastAccessed" date

Added support for managed metadata in the Policy Editor

Fixed issue with Navigator ignoring all /hbase/ subpaths aside of WALs and oldWALs

Issues Fixed in Cloudera Navigator 2.6.5

Out of memory during file purge

Navigator runs out of memory during purge of files.

Hive parser issues errors

The Hive parser issued errors during a query in the following cases:

- Multiple struct attributes in a query
- Constants are projected as part of a subquery
- Functions involving constants with an alias

Incomplete query strings in Navigator metadata

Storing more than 32K characters in Solr causes an exception. Navigator now truncates any field longer than 32K to prevent the exception.

Issues Fixed in Cloudera Navigator 2.6.2

Linker causes high memory consumption

Navigator Metadata Server fails with nav.batch and a high amount of heap

`java.lang.IllegalArgumentException` error causes Navigator to crash

Class not visible on Managed Metadata review screen

The fields Class and Additional class are not visible on the Review screen when creating or updating a Managed Metadata property.

Google analytics might not load when Navigator is served over HTTPS/SSL

Column type for HBase qualifier value in Navigator Audit database should be LONGBLOB

Issues Fixed in Cloudera Navigator 2.6.1

Lineage returns deleted entities even if the Hide deleted entities option is selected

Browser sessions could be unexpectedly logged out

When multiple browser tabs are open to the same Metadata Server, an inactive tab can log you out even though you are active in another tab.

Navigator Metadata Server does not correctly index /hbase/ subpaths

Navigator disabled by default the metadata indexing of `/hbase/WALs` and `/hbase/oldWALs` and any entry under `/hbase/*` entirely.

Previously, `/hbase/WALs` and `/hbase/oldWALs` rules were not effective because they were masked by the rule `/ . *`, which was matching everything under `/hbase`. This has been corrected. Now only nested hidden files and directories under `/hbase` are discarded.

Issues with additional Group facet values in the Search page

If a new value is added to the subfacet `SourceType:HDFS->Group`, the 'Clear all' actions selects all the values in the Group facet, although they are not included in the query.

The input field expands outside the facet panel.

Adding a new (not suggested) value to the Group facet in Firefox leaves the suggestion drop-down open.

Logical Physical relationships unexpectedly shown in lineage

HDFS directories are displayed in Hive lineage.

Cloudera Navigator 2 Data Management Release Notes

[Lineage rendering sometimes throws NPE when latest operation is selected](#)

[ClassCastException thrown when a custom entity is encountered](#)

[Mismatch in number of files between Search and Search from Analytics due to deleted files](#)
Search queries invoked from the Analytics page now show the correct number of files.

[On load, the originating entity might not be visible in lineage](#)

If the lineage graph is tall or wide the originating entity might not be initially visible.

[Browser support warning was not working](#)

Internet Explorer 11 was not in the list of supported browsers.

[Lineage could fail or display partial results when related to Workflows entities](#)

Oozie workflows can have multiple parents however lineage cannot display this use case. Workflows are no longer returned for an operation. To get workflow detail, you must launch lineage on the workflow itself.

[Metadata UI could be unreliable when viewing entities with no name](#)

If an entity has no value in the name or originalName properties, they are now set to "(no name)".

[User logging would in some cases be presented with the login screen again](#)

[Lineage load error banner could show on subsequent lineages](#)

Navigator displays two warnings when lineage load fails. The yellow banner displays until dismissed or the page is reloaded even if it does not apply to the current lineage.

[Missing managed metadata property details](#)

Regular expressions and other details are not listed in the detailed managed metadata property view.

[Alignment of logical physical relations does not follow data flow](#)

Logical physical relations should be shown such that they follow data flow. For example, if downstream you hit a logical entity first then it should be logical -> physical. If you hit physical first it should be physical -> logical.

[Input validation could in certain cases show errors after the error was fixed](#)

Native input validation messages on properties in the Edit Metadata dialog keep displaying, for example when working with regular expression patterns.

[Combo box display issues in Internet Explorer](#)

The combo box scroll bar in search facets does not work in Internet Explorer. Clicking it closes the drop-down.

[Hue link missing](#)

When the Cluster Name facet is used, the Hue link is missing in search results.

Issues Fixed in Cloudera Navigator 2.6.0

[Lineage performance suffers when more than 10000 relations are extracted](#)

If more than 10000 relations must be traversed for a lineage diagram, performance suffers. This can occur when there are thousands of files in a directory or hundreds of columns in a table.

In 2.3, the Navigator Search UI is missing the User Defined Property filter

The Navigator Search UI has changed between 2.2 and 2.3 with respect to the User Defined Property filter. In 2.3 it is no longer present.

Workaround: In the Search box type: `up_propertyName:value`. For example, `up_myName:myValue`.

Displaying the lineage of an entity may consume too much memory and the Metadata Server may run out of memory

Memory leaks have been fixed. However, the Navigator Metadata Server embeds Solr server inside its process, and its memory requirement is proportional to the data that is being indexed. For recommendations on how to configure memory, see [Navigator Metadata Server Memory Sizing Recommendations](#).

After upgrade to 2.4.x, only entries related to deleted files are visible

Issues Fixed in Cloudera Navigator 2.5.0

After upgrade to 2.4.x only entries related to deleted files are visible

Issues Fixed in Cloudera Navigator 2.4.4

Lineage does not render on MacOS using Chrome 48.0.2564.109

After upgrade to 2.4.x only entries related to deleted files are visible

Issues Fixed in Cloudera Navigator 2.4.2



Note: If you are upgrading from a previous version of Cloudera Navigator, Cloudera recommends that you upgrade to version 2.4.3 or higher.

Searching for strings that start with forward slash ignored sometimes

Searching for a string that starts with a forward slash (/) in conjunction with another filter/facet or followed by AND and another search term will result in the string being ignored completely.

Workaround: Enclose the string starting with forward slash in double quotes, for example `"/some/path"` AND `sourceType:hdfs`

After upgrade, entities are incorrectly marked as deleted

The issue is fixed. However, if you are upgrading from Cloudera Navigator 2.4.0 or 2.4.1 to 2.4.2, after shutting down the Navigator Metadata Server as instructed in the upgrade documentation, you must additionally perform the following steps:

1. Edit the HDFS extractor state file under the Navigator storage directory (by default, `/var/lib/cloudera-scm-navigator`), in the `extractorState` directory. The HDFS extractor state file contains a JSON object with an attribute `upgrades`, which has a list of upgrades and their statuses. For example:

```
{
  "nextTxId": 11394669, "lastKnownTransactionTime": 1453931555151, "layoutVersion": -60,
  "namespaceId": 1613692955, "nextExtractionRunId": "df61cf41183dec8947078c228ec505e6##7453",
  "upgrades": [
    {"taskType": "SET_MISSING_ATTRIBUTES", "status": "SUCCEEDED", "startTime": 1453484370390, "endTime": 1453484370410, "attempts": 1},
    {"taskType": "UNDELETE", "status": "SUCCEEDED", "startTime": 1453484370391, "endTime": 1453484370418, "attempts": 1},
    {"taskType": "BLOCK_SIZE", "status": "SUCCEEDED", "startTime": 1453484370391, "endTime": 1453484370427, "attempts": 1},
    {"taskType": "REPLICATION_COUNT", "status": "SUCCEEDED", "startTime": 1453484370391, "endTime": 1453484370434, "attempts": 1}
  ]
}
```

Under the `upgrades` list, change the `status` attribute of `taskType: UNDELETE` to FAILED. For example, change:

```
{"taskType": "UNDELETE", "status": "SUCCEEDED", "startTime": 1453484370391, "endTime": 1453484370418, "attempts": 1}
```

to

```
{"taskType": "UNDELETE", "status": "FAILED", "startTime": 1453484370391, "endTime": 1453484370418, "attempts": 1}
```

2. If your Navigator instance manages multiple HDFS instances, repeat for each HDFS extractor state file.

On a newly installed cluster, Hive lineage does not work

When Cloudera Manager starts, it does not automatically create the directory `/var/log/hive/lineage`. Because the Agent cannot detect this directory, it fails to set a watch on it. HiveServer2 creates the directory, but Hive metadata extracted to the directory is never read by Agent and sent to Navigator Metadata Server.

Workaround: Do one of the following:

- In the Cloudera Manager Admin Console UI, restart HiveServer2.
- To avoid HiveServer2 downtime, restart the Cloudera Manager Agent. On the host where the HiveServer2 role is running:

```
sudo service cloudera-scm-agent restart
```

Entity landing page schema overflow hides name

If the schema type is very long, it can hide the name of the schema entity.

When Navigator uses the Cloudera Manager authenticator and Cloudera Manager is TLS enabled, authentication fails

When Navigator is configured to use the Cloudera Manager DB for authentication, and Cloudera Manager is TLS enabled, when a user attempts authentication, certificate validation fails with:

```
Caused by: sun.security.validator.ValidatorException: PKIX path building failed:  
sun.security.provider.certpath.SunCertPathBuilderException: unable to find valid  
certification path to requested target
```

Analytics UI needs to filter out deleted files

The Analytics UI was counting deleted files.

Policies created from the Analytics UI were not working

Policies created from analytics UI were not working because it filtered on `clusterName`. Now it filters on `sourceld`.

There is no indication that the Entity Details page is not completely loaded

"Technical Metadata is displayed immediately, so it is not clear that the Entity Details page has not completed loading. A loading spinner now displays on slow loads.

Tables might not load on Entity Details page

Tables might not load, depending on whether the schema finished loading before or after tables finish loading.

Searching for strings that start with forward slash ignored sometimes

Searching for a string that starts with a forward slash (/) in conjunction with another filter/facet or followed by `AND` and another search term results in the string being ignored completely.

Workaround: Enclose the string starting with a forward slash in double quotes, for example `"/some/path"` `AND` `sourceType:hdfs`.

Issues Fixed in Cloudera Navigator 2.4.1

Apache Commons Collections Deserialization Vulnerability

Cloudera has learned of a potential security vulnerability in a third-party library called the [Apache Commons Collections](#). This library is used in products distributed and supported by Cloudera (“Cloudera Products”), including core Apache Hadoop. The Apache Commons Collections library is also in widespread use beyond the Hadoop ecosystem. At this time, no specific attack vector for this vulnerability has been identified as present in Cloudera Products.

In an abundance of caution, we are currently in the process of incorporating a version of the Apache Commons Collections library with a fix into the Cloudera Products. In most cases, this will require coordination with the projects in the Apache community. One example of this is tracked by [HADOOP-12577](#).

The Apache Commons Collections potential security vulnerability is titled “Arbitrary remote code execution with InvokerTransformer” and is tracked by [COLLECTIONS-580](#). MITRE has not issued a CVE, but related [CVE-2015-4852](#) has been filed for the vulnerability. CERT has issued [Vulnerability Note #576313](#) for this issue.

Releases affected: CDH 5.5.0, CDH 5.4.8 and lower, Cloudera Manager 5.5.0, Cloudera Manager 5.4.8 and lower, Cloudera Navigator 2.4.0, Cloudera Navigator 2.3.8 and lower

Users affected: All

Severity (Low/Medium/High): High

Impact: This potential vulnerability may enable an attacker to execute arbitrary code from a remote machine without requiring authentication.

Immediate action required: Upgrade to Cloudera Navigator 2.4.1.

Issues Fixed in Cloudera Navigator 2.4.0

After HDFS upgrade, some changes to HDFS entities may not appear in Navigator

After an HDFS upgrade, Navigator might not detect changes to HDFS entities, such as move, rename, and delete operations, that were recorded only in the HDFS edit logs before the upgrade. This may cause an inconsistent view of HDFS entities between Navigator and HDFS.

Workaround: None.

After upgrading TLS/SSL-enabled Navigator to 5.4.0, Navigator starts in non-TLS/SSL

After upgrading an TLS/SSL-enabled Navigator to 5.4.x, Navigator will start in non-TLS/SSL mode. Before the 5.4.x release of Navigator, TLS/SSL for Navigator was turned on using an advanced configuration snippet. In release 5.4.x, this option is exposed in the Cloudera Manager configuration UI. After upgrade to 5.4.x, this setting overrides the setting in the advanced configuration snippet, disabling TLS/SSL.

Workaround: Remove the setting from the advanced configuration snippet and specify it in the configuration UI directly.

Navigator login using Cloudera Manager credentials does not work if Cloudera Manager has TLS enabled

Memory leak when creating Impala lineage for very big queries (more than 32K characters)

Disabling auditing in Hive causes exception in HiveServer2

If you disable auditing for Hive, but do not restart the Hive service, HiveServer2 logs the following exception:

```
Hive Internal Error: com.cloudera.navigator.shaded.avro.AvroRuntimeException(Field
serviceName type:STRING pos:0 does not accept null values)
```

Cloudera Navigator 2 Data Management Release Notes

Lineage is not collected for Hive on Spark jobs

Handle HDFS root as special case when generating Hue link

Entity description does not allow multiple lines

Expanding entity causes unintuitive lineage layout

Selected Pig field not shown in lineage

When you view lineage for a Hive field, the Hive parent is expanded so you see the field. However, if there is a field of a Pig table inside a Pig operation execution, only the operation execution is expanded, not the table. The field is not visible and it is not obvious to which entity it belongs.

Pig inside Oozie not shown in lineage view

Page headers missing after selecting "Show Full Page" in lineage diagram

Facet does not visually retain selection after refresh

After reloading a page, some selections are not rendered as selected even though the search string itself is still correct.

Link and parent Hive table not displayed

If you launch a job with two input directories that are also both Hive tables, the Hive table is shown for one of the directories but not the other. If you expand the other directory, the Hive table is shown.

Links disappear when expanding Pig job run by Oozie

Expanding a directory folder obscures the Hive table it is linked to

Source filter missing from filter list

Subfacets should be enabled and disabled based on values in facet type

Tags and key-value pairs should be sorted in the policy editor

Create Policy link should not be available for a policy viewer role

Impala auditing does not support filtering during event capture

Impala auditing does not support filtering during event capture.

Severity: Low

Workaround: None.

Issues Fixed in Cloudera Navigator 2.3.9

Apache Commons Collections Deserialization Vulnerability

Cloudera has learned of a potential security vulnerability in a third-party library called the [Apache Commons Collections](#). This library is used in products distributed and supported by Cloudera (“Cloudera Products”), including core Apache Hadoop. The Apache Commons Collections library is also in widespread use beyond the Hadoop ecosystem. At this time, no specific attack vector for this vulnerability has been identified as present in Cloudera Products.

In an abundance of caution, we are currently in the process of incorporating a version of the Apache Commons Collections library with a fix into the Cloudera Products. In most cases, this will require coordination with the projects in the Apache community. One example of this is tracked by [HADOOP-12577](#).

The Apache Commons Collections potential security vulnerability is titled “Arbitrary remote code execution with InvokerTransformer” and is tracked by [COLLECTIONS-580](#). MITRE has not issued a CVE, but related [CVE-2015-4852](#) has been filed for the vulnerability. CERT has issued [Vulnerability Note #576313](#) for this issue.

Releases affected: CDH 5.5.0, CDH 5.4.8 and lower, Cloudera Manager 5.5.0, Cloudera Manager 5.4.8 and lower, Cloudera Navigator 2.4.0, Cloudera Navigator 2.3.8 and lower

Users affected: All

Severity (Low/Medium/High): High

Impact: This potential vulnerability may enable an attacker to execute arbitrary code from a remote machine without requiring authentication.

Immediate action required: Upgrade to Cloudera Navigator 2.3.9.

Issues Fixed in Cloudera Navigator 2.3.8

Disabling auditing in Hive causes exception in HiveServer2

If you disable auditing for Hive, but do not restart the Hive service, HiveServer2 logs the following exception:

```
Hive Internal Error: com.cloudera.navigator.shaded.avro.AvroRuntimeException(Field  
serviceName type:STRING pos:0 does not accept null values)
```

NPE during YARN extraction

The extractor no longer throws a NPE during YARN extraction.

Issues Fixed in Cloudera Navigator 2.3.3

Upgrade to 2.3.1 does not create column "NAME" in NAVMS_AUDIT_EVENTS

The upgrade process neglected to include a script to create the "NAME" column.

Navigator Search UI hangs due to blocked log4j thread

The Metadata Server log4j logger blocked due to a defect in how old logs are deleted. The logger has been replaced and the Cloudera Manager Agent deletes old log files.

Expanding Pig lineage can cause overlap

When you expand a Pig lineage diagram, relationship lines can sometimes obstruct one another.

Audit reports subject to cross-site scripting

Navigator does not do input validation on report names, which allows users to enter arbitrary code in this field. The next time a report name is viewed, the script is executed.

Session timeout results in a redirect using HTTP basic authentication

Upon session timeout, Navigator redirects to HTTP basic authentication instead of redirecting to the main login page.

“POODLE” vulnerability on TLS/SSL enabled ports

The POODLE (Padding Oracle On Downgraded Legacy Encryption) attack takes advantage of a cryptographic flaw in the obsolete SSLv3 protocol, after first forcing the use of that protocol. The only solution is to disable SSLv3 entirely. This requires changes across a wide variety of components of Cloudera Navigator in 2.3.x and all earlier versions. Cloudera Navigator 2.3.3 provides these changes for Cloudera Navigator 2.3.x deployments. All Cloudera Navigator users should upgrade to 2.3.3 as soon as possible. For more information, see the [Cloudera Security Bulletin](#).

Cloudera Navigator 2 Data Management Release Notes

LDAP group searches appear slow because of web UI behavior

Searching for LDAP groups in the web UI can appear slow because every time a character changes in the search field, an onchange Javascript event triggers a new LDAP search.

Solr audit fields are incorrectly shown with capital letters in the web UI

Solr audit fields such as **sub_operation** and **entity_id** are rendered with capital letters and spaces such as **Sub Operation** or **Entity ID**.

Issues Fixed in Cloudera Navigator 2.3.1

Link disappears when expanded

Sometimes when you expand a lineage diagram, a link disappears.

Operations on canary files should be filtered out

Hive views have same icon as Hive fields in lineage

Facet count should show (0) instead of (-) when there are no matching entities

Facet counts show (-) when the Metadata Server does not return a value. It should show (0).

Workaround: None.

Column lineage does not display

Column lineage does not display when its parent is automatically expanded.

Workaround: Redisplay the lineage by clicking the parent entity.

Kite dataset extraction fails when HDFS HA is enabled

Workaround: None.

Exporting audit reports to CSV does not work

You can export audits to JSON with a limit of 10,000 audits.

Sentry auditing does not work if the Python version is lower than 2.5

Sqoop sub-operations don't display in schema view of lineage

Workaround: None.

Issues Fixed in Cloudera Navigator 2.3.0

There were no issues fixed in Cloudera Navigator 2.3.0.

Issues Fixed in Cloudera Navigator 2.2.9

Apache Commons Collections Deserialization Vulnerability

Cloudera has learned of a potential security vulnerability in a third-party library called the [Apache Commons Collections](#). This library is used in products distributed and supported by Cloudera (“Cloudera Products”), including core Apache Hadoop. The Apache Commons Collections library is also in widespread use beyond the Hadoop ecosystem. At this time, no specific attack vector for this vulnerability has been identified as present in Cloudera Products.

In an abundance of caution, we are currently in the process of incorporating a version of the Apache Commons Collections library with a fix into the Cloudera Products. In most cases, this will require coordination with the projects in the Apache community. One example of this is tracked by [HADOOP-12577](#).

The Apache Commons Collections potential security vulnerability is titled “Arbitrary remote code execution with InvokerTransformer” and is tracked by [COLLECTIONS-580](#). MITRE has not issued a CVE, but related [CVE-2015-4852](#) has been filed for the vulnerability. CERT has issued [Vulnerability Note #576313](#) for this issue.

Releases affected: CDH 5.5.0, CDH 5.4.8 and lower, Cloudera Manager 5.5.0, Cloudera Manager 5.4.8 and lower, Cloudera Navigator 2.4.0, Cloudera Navigator 2.3.8 and lower, Cloudera Navigator 2.2.8 and lower.

Users affected: All

Severity (Low/Medium/High): High

Impact: This potential vulnerability may enable an attacker to execute arbitrary code from a remote machine without requiring authentication.

Immediate action required: Upgrade to Cloudera Navigator 2.2.9.

Issues Fixed in Cloudera Navigator 2.2.4

“POODLE” vulnerability on TLS/SSL enabled ports

The POODLE (Padding Oracle On Downgraded Legacy Encryption) attack takes advantage of a cryptographic flaw in the obsolete SSLv3 protocol, after first forcing the use of that protocol. The only solution is to disable SSLv3 entirely. This requires changes across a wide variety of components of Cloudera Navigator in 2.2.x and all earlier versions. Cloudera Navigator 2.2.4 provides these changes for Cloudera Navigator 2.2.x deployments. All Cloudera Navigator users should upgrade to 2.2.4 as soon as possible. For more information, see the [Cloudera Security Bulletin](#).

Issues Fixed in Cloudera Navigator 2.2.3

Navigator Audit Server reports invalid null characters in HBase audit events when using the PostgreSQL database

Navigator Audit Server reports invalid null characters in HBase audit events when using the PostgreSQL database. HBase allows null characters in qualifiers, so now Navigator escapes them.

Oozie extractor throws too many Boolean clauses exception

Issues Fixed in Cloudera Navigator 2.2.2

The audit reports UI now returns results when there are a large number of audit records

The audit reports UI was not returning results when there were a large number of audit records matching a particular time period, especially when the period included multiple days. The UI is now also much more responsive.

Issues Fixed in Cloudera Navigator 2.2.1

Browser autocomplete no longer enabled before authentication

Form fields before authentication in the application have auto-complete enabled. Any user using the same computer would be able to see information entered by a previous user.

Navigator Web UI no longer exposes paths to directory listing/forceful browsing

The web server is configured to display the list of files contained in this directory. This is not recommended because the directory may contain files that are not normally exposed through links on the web site.

Cloudera Navigator 2 Data Management Release Notes

[Navigator Audit Server no longer throws OOM for very long Impala queries](#)

Issues Fixed in Cloudera Navigator 2.2.0

[If Hue is added after Navigator, search results do not have links to Hue](#)

In the Metadata UI, search results contain links to an appropriate application in Hue. However, if you add a Hue service after Navigator roles, there will be no links to Hue.

Workaround:

1. Set the cluster's display name and name properties to be the same:

- a. Get cluster's name and display name using following API: `http://hostname:7180/api/v6/clusters`.
- b. In the Cloudera Manager Admin Console, at the right of the cluster name, click the down arrow and select **Rename Cluster**. Set the cluster display name to match its name.

2. Restart the Navigator Metadata server.

Issues Fixed in Cloudera Navigator 2.1.6

[When auditing is enabled, the Cloudera Manager Agent may become slow or get stuck when responding to commands and when sending heartbeats to Cloudera Manager Server](#)

This issue can occur when Cloudera Navigator auditing is turned on. The auditing code reads audit logs and sends them to the Audit Server. It acquires a lock to protect the list of roles being audited. The same list is also modified by the Cloudera Manager Agent's main thread when a role is started or stopped. If the Audit thread takes too much time to send audits to the Audit Server (which can happen if there is backlog of audit logs), it starves the main Agent thread. This causes the main Agent thread to not send heartbeats and to not respond to commands from the Cloudera Manager Server.

Issues Fixed in Cloudera Navigator 2.1.5

[Navigator Audit Server reports invalid null characters in HBase audit events when using the PostgreSQL database](#)

Navigator Audit Server reports invalid null characters in HBase audit events when using the PostgreSQL database. HBase allows null characters in qualifiers, so now Navigator escapes them.

[Oozie extractor throws too many Boolean clauses exception](#)

Issues Fixed in Cloudera Navigator 2.1.4

[The audit reports UI now returns results when there are a large number of audit records](#)

The audit reports UI was not returning results when there were a large number of audit records matching a particular time period, especially when the period included multiple days. The UI is now also much more responsive.

Issues Fixed in Cloudera Navigator 2.1.2

[Search results in Navigator now have links to Hue](#)

In the Metadata UI, search results contain links to an appropriate application in Hue. The links may be missing for either of the following reasons:

- Hue was added after Navigator Metadata server was started.
- The cluster name and display name are different.

Workaround:

1. Set the cluster's display name and name properties to be the same:

- a. Get cluster's name and display name using following API: `http://hostname:7180/api/v6/clusters`.
 - b. In the Cloudera Manager Admin Console, at the right of the cluster name, click the down arrow and select **Rename Cluster**. Set the cluster display name to match its name.
2. Restart the Navigator Metadata server.

Browser autocomplete no longer enabled before authentication

Form fields before authentication in the application have auto-complete enabled. Any user using the same computer would be able to see information entered by a previous user.

Navigator Web UI no longer exposes paths to directory listing/forceful browsing

The web server is configured to display the list of files contained in this directory. This is not recommended because the directory may contain files that are not normally exposed through links on the web site.

Navigator Audit Server no longer throws OOM for very long Impala queries

Issues Fixed in Cloudera Navigator 2.1.1

LDAP lookups in Active Directory to resolve group membership are now working

Dropping a Hive table and creating a view with same name or vice versa no longer raises an error

HDFS extraction now works after upgrading CDH from 5.1 to 5.2

Setting a property in the Hue advanced configuration snippet no longer throws a "too many Boolean clauses" error in Navigator Metadata

Issues Fixed in Cloudera Navigator 2.0.5

Memory leak in Navigator Audit Server due to error during batch operations

Issues Fixed in Cloudera Navigator 2.0.3

Dropping a Hive table and creating a view with same name or vice versa no longer raises an error

Setting a property in the Hue advanced configuration snippet no longer throws a "too many Boolean clauses" error in Navigator Metadata

Issues Fixed in Cloudera Navigator 2.0.2

HBase auditing initialization failure can prevent region opening indefinitely

Issues Fixed in Cloudera Navigator 2.0.1

Hive View to Table lineage is missing

The lineage of the underlying tables does not appear in the lineage view.

Workaround: Launch lineage on the underlying tables directly.

The "allowed" query selector is missing from the audit REST API

Queries such as `http://hostname:7180/api/v7/audits?maxResults=10&query=allowed==false` are now supported.

Cloudera Navigator 2 Data Management Release Notes

Workaround: None.

Metadata in metadata files is not processed

Workaround: None.

Lineage does not work when launched on a field

When you launch lineage on a field (Hive column or Pig field, or a Sqoop sub-operation), the UI displays the initial graph properly. However if you expand the parent item (Hive table in case of a Hive column), then things start disappearing from the lineage diagram.

Workaround: None.

When you specify an end date for the `created` property, no results are returned.

Workaround: Clear the end date control or specify an end date of `TO+*%5D%22%7D`.

Navigating back to the parent entity in a Pig lineage diagram sometimes displays the error: Cannot read property 'x' of undefined.

Workaround: None.

Issues Fixed in Cloudera Navigator 2.0.0

The last accessed time for Hive table is incorrect.

Workaround: None.

Pig job that has relations with self is unreadable in lineage view.

The Metadata UI currently does not handle situation where there is a data-flow relation between elements that are also related via parent-child relation.

Workaround: None.

If auditing is enabled, during an upgrade from Cloudera Manager 4.6.3 to 4.7, the Impala service won't start.

Impala auditing requires the Audit Log Directory property to be set, but the upgrade process fails to set the property.

Workaround: Do one of the following:

- Stop the Cloudera Navigator Audit server.
- Ensure that you have a Cloudera Navigator license and manually set the property to `/var/log/impalad/audit`.

Empty box appears over the list of results after adding a tag to a file

When tags are added to an entity, in some cases a white box remains after pressing **Enter**.

Workaround: Refresh the page to remove the artifact.

Issues Fixed in Cloudera Navigator 1.2.0

Certain complex multi-level lineages, such as directory/file and database/table, may not be fully represented visually.

Workaround: None.

Cloudera Navigator Key Trustee Server Release Notes

These release notes provide information on the new and changed features, known issues, and fixed issues for Cloudera Navigator Key Trustee Server.

New Features and Changes in Cloudera Navigator Key Trustee Server

The following sections describe what's new and changed in each Cloudera Navigator Key Trustee Server release.



Note: Releases 5.4.3 and 5.4.9 are the only maintenance releases of Key Trustee Server 5.4.

There is no Release 5.5.1, 5.5.3, or 5.5.4 of Key Trustee Server.

There is no Release 5.6 of Key Trustee Server.

There is no Release 5.7.1, 5.7.2, 5.7.3, or 5.7.4 of Key Trustee Server.

There is no Release 5.8.1, 5.8.2, or 5.8.3 of Key Trustee Server.

What's New in Cloudera Navigator Key Trustee Server

The following sections describe what's new in each Cloudera Navigator Key Trustee Server release.

What's New in Cloudera Navigator Key Trustee Server 5.9.0

- Key Trustee Server supports RHEL 6.8.
- Running the [Key Trustee Server backup script](#) now backs up necessary hardware security module (HSM) configuration files.
- Key Trustee Server supports rolling restart in Cloudera Manager.

What's New in Cloudera Navigator Key Trustee Server 5.8.0

- When adding the parcel-based Key Trustee Server service for the first time, Cloudera Manager automatically backs up Key Trustee Server locally and schedules ongoing hourly local backups using cron.
- The Key Trustee Server backup script (`ktbackup.sh`) adds a new option, `--roll`, which specifies the number of backups to retain.

An issue has also been fixed. For more information, see [Issues Fixed in Cloudera Navigator Key Trustee Server 5.8.0](#) on page 488.

What's New in Cloudera Navigator Key Trustee Server 5.7.0

- A backup script (`ktbackup.sh`) is included with Key Trustee Server.
- Parcel-based Key Trustee Server logs an error message at startup if the `keytrustee.conf` file is malformed.
- Error logging is improved for connection and certificate errors with Key Trustee Server connecting to Key HSM.

A number of issues have also been fixed. See [Issues Fixed in Cloudera Navigator Key Trustee Server 5.7.0](#) on page 488.

What's New in Cloudera Navigator Key Trustee Server 5.5.2

- The `ktadmin` command has a new `--passphrase` option to allow migration of existing keys from a Key Trustee Server with a password-protected private key to an HSM.

An issue has also been fixed. See [Issues Fixed in Cloudera Navigator Key Trustee Server 5.5.2](#) on page 488.

What's New in Cloudera Navigator Key Trustee Server 5.5.0

- Key Trustee Server supports RHEL 7.

Cloudera Navigator Key Trustee Server Release Notes

- Key Trustee Server supports password-protected SSL certificates.
- TLS/SSL certificate file paths are configurable in Cloudera Manager.

A number of issues have also been fixed. See [Issues Fixed in Cloudera Navigator Key Trustee Server 5.5.0](#) on page 489.

What's New in Cloudera Navigator Key Trustee Server 5.4.9

An issue has been fixed. See [Issues Fixed in Cloudera Navigator Key Trustee Server 5.4.9](#) on page 491.

What's New in Cloudera Navigator Key Trustee Server 5.4.3

A number of issues have been fixed. See [Issues Fixed in Cloudera Navigator Key Trustee Server 5.4.3](#) on page 491.

What's New in Cloudera Navigator Key Trustee Server 5.4.0

- Key Trustee Server can be installed and managed using Cloudera Manager.
- Existing keys stored in Key Trustee Server can be migrated to a Hardware Security Module (HSM) using Navigator Key HSM.

Changed Features and Behaviors in Cloudera Navigator Key Trustee Server

The following sections describe what's changed in each Cloudera Navigator Key Trustee Server release.

What's Changed in Cloudera Navigator Key Trustee Server 5.8.0

- Attempting to start the `keytrusteed` service using the `service` command when Key Trustee Server is already running now reports success with a message indicating that the service is already running.

What's Changed in Cloudera Navigator Key Trustee Server 5.7.0

The `ktadmin keyhsm` command can be used to update Key Trustee Server certificate information for Key HSM

The `ktadmin keyhsm` command can be used to update Key Trustee Server certificate information for Key HSM instead of manually modifying `keytrustee.conf`.

What's Changed in Cloudera Navigator Key Trustee Server 5.5.0

Key Trustee Server can be installed using Cloudera Manager without a CSD

Key Trustee Server no longer requires a Custom Service Descriptor (CSD) file.

What's Changed in Cloudera Navigator Key Trustee Server 5.4.0

All processes run as a single user

All Key Trustee Server processes now run as a single non-root user (`keytrustee` by default).

Key Trustee Server processes use a single port

Key Trustee Server processes (with the exception of the backing PostgreSQL database) now use a single port (11371 by default).

Known Issues and Workarounds in Cloudera Navigator Key Trustee Server



Warning:

Interrupting deposit migration from Key Trustee Server to Key HSM can result in lost data

Workaround: Do not interrupt deposit migration to Key HSM.

Upgraded passive Key Trustee Server fails to start due to incorrect ownership of recovery.conf

Passive Key Trustee Servers upgraded from Key Trustee Server 3.8.x or lower fail to start with the following error:

```
WARNING:root:stdout pg_basebackup: directory "/var/lib/pgsql/9.3/keytrustee" exists but
is not empty
:
Traceback (most recent call last):
  File "/usr/bin/ktadmin", line 484, in <module>
    main()
  File "/usr/bin/ktadmin", line 473, in main
    init_slave()
  File "/usr/bin/ktadmin", line 349, in init_slave
    pgsetup.base_backup(ARGS.pg_rootdir, ARGS.master, PKG, run_as=ARGS.postgres_user)
  File "/usr/lib/python2.6/site-packages/keytrustee/server/setup/postgres.py", line 206,
in base_backup
    run([pg_basebackup, '-D', dest, '--host=%s' % master_ip, '--port=%d' % port,
'--username=%s' % db_user], run_as=run_as)
  File "/usr/lib/python2.6/site-packages/keytrustee/util.py", line 145, in run
    raise subprocess.CalledProcessError(p.returncode, cmd)
subprocess.CalledProcessError: Command '['/usr/pgsql-9.3/bin/pg_basebackup', '-D',
'/var/lib/pgsql/9.3/keytrustee', '--host=kt01.example.com', '--port=5432',
"--username=keytrustee']' returned non-zero exit status 1
```

Workaround: Change the owner and group of /var/lib/pgsql/9.3/keytrustee/recovery.conf to postgres:

```
$ sudo chown postgres:postgres /var/lib/pgsql/9.3/keytrustee/recovery.conf
```

Key Trustee Server PKCS8 private key cannot communicate with Key HSM

If its private key is in PKCS8 format, Key Trustee Server cannot communicate with Key HSM.

Workaround: Convert the Key Trustee Server private key to raw RSA format.

Key Trustee Server backup script fails if PostgreSQL versions lower than 9.3 are installed

If PostgreSQL versions lower than 9.3 are installed on the Key Trustee Server host, the ktbackup.sh script fails with an error similar to the following:

```
pg_dump: server version: 9.3.11; pg_dump version: 9.2.14
pg_dump: aborting because of server version mismatch
```

Workaround: Uninstall the lower PostgreSQL version.

Changing the Key Trustee Server database port in Cloudera Manager does not work

Changing the Key Trustee Server database port in Cloudera Manager has no effect.

Workaround: None. Use the default port of 11381

Key migration fails when password-protected certificates are stored in a non-default location

Key migration from Key Trustee Server to Key HSM fails when using password-protected certificates in a non-default location.

Workaround: Use the --passphrase option with the ktadmin keyhsm command to prompt for the password.

Issues Fixed in Cloudera Navigator Key Trustee Server

The following sections describe issues fixed in each Cloudera Navigator Key Trustee Server release.



Note: Releases 5.4.3 and 5.4.9 are the only maintenance releases of Key Trustee Server 5.4.

There is no Release 5.5.1, 5.5.3, or 5.5.4 of Key Trustee Server.

There is no Release 5.6 of Key Trustee Server.

There is no Release 5.7.1, 5.7.2, 5.7.3, or 5.7.4 of Key Trustee Server.

There is no Release 5.8.1, 5.8.2, or 5.8.3 of Key Trustee Server.

Issues Fixed in Cloudera Navigator Key Trustee Server 5.9.0

For new features in Key Trustee Server 5.9.0, see [New Features and Changes in Cloudera Navigator Key Trustee Server](#) on page 485.

Issues Fixed in Cloudera Navigator Key Trustee Server 5.8.0

Key Trustee Server logs superfluous error during registration

When you register a client to the Key Trustee Server, the following error message appears in the Key Trustee Server log:

```
Traceback (most recent call last):
  File "/usr/lib/python2.6/site-packages/keytrustee/server/webapp.py", line 2151, in
    lookup
    return hkp.lookup(g.hkpdb, app.config['LOCAL_FINGERPRINT'])
  File "/usr/lib/python2.6/site-packages/keytrustee/server/hkp.py", line 72, in lookup
    raise wzex.BadRequest("Operation not specified.")
BadRequest: 400: Operation not specified.
```

Issues Fixed in Cloudera Navigator Key Trustee Server 5.7.0

Email messages from Key Trustee Server use hostname instead of fully qualified domain name

Email sent from Key Trustee Server includes a link that uses the Key Trustee Server short name instead of the fully qualified domain name.

Host Inspector displays the wrong version of Key Trustee Server

Host Inspector displays the wrong Key Trustee Server version, showing the Key Trustee KMS version instead.

Key Trustee Server supports weak RC4 ciphers for TLS

Key Trustee Server supports weak RC4 ciphers for TLS connections, potentially allowing clients to negotiate a weaker connection.

The ktadmin keyhsm command fails with M2Crypto error

Running the `ktadmin keyhsm --server https://khsm01.example.com:9090 --trust` command fails with the following error:

```
M2Crypto.X509.X509Error: 140405990356800:error:0906D06C:PEM routines:PEM_read_bio:no
start line:pem_lib.c:703:Expecting: CERTIFICATE
```

Issues Fixed in Cloudera Navigator Key Trustee Server 5.5.2

Key Trustee Server with password-protected private key cannot communicate with Key HSM

If its private key is password-protected, Key Trustee Server cannot communicate with Key HSM.

Issues Fixed in Cloudera Navigator Key Trustee Server 5.5.0

Key Trustee Server missing from Components page

Key Trustee Server is not listed in the **Components** page (**Hosts > Hostname > Components**) of the host it is installed on.

Passive Database role fails to restart after enabling synchronous replication

After enabling synchronous replication, the Passive Database role fails to start when restarting the Key Trustee Server service in Cloudera Manager.

Key Trustee Server using CherryPy allows SSLv3

Key Trustee Server using CherryPy as the web backend allows SSLv3 connections, which are considered insecure.

Key Trustee Server sends wrong command in email notification

When you register a client, Key Trustee Server sends an email notification with a command to import the Key Trustee Server GPG key. The command uses the `hkps` protocol and port 80. The correct protocol is `hkps` and the correct port is 11371.

Enabling synchronous replication fails if `python-ordereddict` is not installed

On Key Trustee Server hosts where the `python-ordereddict` package is not installed, enabling synchronous replication fails, and the Passive Database role fails to start. The `stderr.log` file for the `DB_ACTIVE-setup` process (`/var/run/cloudera-scm-agent/process/ID-keytrustee_server-DB_ACTIVE-setup/logs/stderr.log`) contains the following error:

```
Traceback (most recent call last):
  File "aux/prop2x.py", line 14, in <module>
    from ordereddict import OrderedDict
ImportError: No module named ordereddict
```

Key Trustee Server does not respond to requests

On systems with low entropy, local fingerprint generation fails during setup, but Cloudera Manager reports no issues with the Key Trustee Server service. Subsequent attempts to use Key Trustee Server fail with an error similar to the following in the Key Trustee Server log:

```
2015-06-03 11:44:17,232 - keytrustee.server.webapp - ERROR - 'None' is not a valid
qualified fingerprint
Traceback (most recent call last):
  File
"/opt/cloudera/parcels/KEYTRUSTEE_SERVER-5.4.0-1.keytrustee5.4.0.p0.204/lib/python2.6/
site-packages/keytrustee-5.4.0-py2.6.egg/keytrustee/server/webapp.py", line 1917, in
index
    server.response = server.respond()
  File
"/opt/cloudera/parcels/KEYTRUSTEE_SERVER-5.4.0-1.keytrustee5.4.0.p0.204/lib/python2.6/
site-packages/keytrustee-5.4.0-py2.6.egg/keytrustee/server/webapp.py", line 1783, in
respond
    return Response(self.get_fingerprint(), **g.resp_kwargs)
  File
"/opt/cloudera/parcels/KEYTRUSTEE_SERVER-5.4.0-1.keytrustee5.4.0.p0.204/lib/python2.6/
site-packages/keytrustee-5.4.0-py2.6.egg/keytrustee/server/webapp.py", line 1338, in
get_fingerprint
    key_info = g.gpg.get_key_info(str(app.config['LOCAL_FINGERPRINT']))
  File
"/opt/cloudera/parcels/KEYTRUSTEE_SERVER-5.4.0-1.keytrustee5.4.0.p0.204/lib/python2.6/
site-packages/keytrustee-5.4.0-py2.6.egg/keytrustee/gnupg.py", line 415, in get_key_info

    search_string = fp_or_id(fingerprint, keyid)
  File
```

```
"/opt/cloudera/parcels/KEYTRUSTEE_SERVER-5.4.0-1.keytrustee5.4.0.p0.204/lib/python2.6/
site-packages/keytrustee-5.4.0-py2.6.egg/keytrustee/gnupg.py", line 817, in fp_or_id
    result = unqualified_fp(fingerprint)
  File
"/opt/cloudera/parcels/KEYTRUSTEE_SERVER-5.4.0-1.keytrustee5.4.0.p0.204/lib/python2.6/
site-packages/keytrustee-5.4.0-py2.6.egg/keytrustee/gnupg.py", line 761, in unqualified_fp
    _, _, fp = read_qualified_fp(fp)
  File
"/opt/cloudera/parcels/KEYTRUSTEE_SERVER-5.4.0-1.keytrustee5.4.0.p0.204/lib/python2.6/
site-packages/keytrustee-5.4.0-py2.6.egg/keytrustee/gnupg.py", line 770, in
read_qualified_fp
    raise ValueError("'%s' is not a valid qualified fingerprint" % (s))
ValueError: 'None' is not a valid qualified fingerprint
```

Cannot register a client with Key Trustee Server after registering GPG public key

After creating an organization, the organization administrator receives a link to register a GPG key. If the administrator registers the GPG key before any clients are registered with the organization, all subsequent registrations with that organization fail.

Starting Key Trustee Server with --daemonize option fails

Starting Key Trustee Server with the command `keytrustee-server start --daemonize` fails with the following error:

```
2015-04-01 05:33:26,843 - cherrypy.error - ERROR - [01/Apr/2015:05:33:26] ENGINE Error
  in 'start' listener <bound method Daemonizer.start of
<cherrypy.process.plugins.Daemonizer object at 0x1e71490>>
Traceback (most recent call last):
  File "/usr/lib/python2.6/site-packages/cherrypy/process/wspbus.py", line 197, in
publish
    output.append(listener(*args, **kwargs))
  File "/usr/lib/python2.6/site-packages/cherrypy/process/plugins.py", line 380, in
start
    si = open(self.stdin, "r")
IOError: [Errno 2] No such file or directory: '/tmp/stdout'
```

Starting Key Trustee Server when it is already running updates `keytrustee.pid` file

If you try to start Key Trustee Server when it is already running, the `/var/lib/keytrustee/.keytrustee/keytrustee.pid` file is updated with the new process ID (PID). The new process fails to start, but the `keytrustee.pid` retains the new PID, and service commands fail with the following error:

```
Traceback (most recent call last):
  File "/usr/bin/keytrustee-server", line 136, in <module>
    main()
  File "/usr/bin/keytrustee-server", line 131, in main
    status()
  File "/usr/bin/keytrustee-server", line 104, in status
    pid_file = ProcessIDInfo(ARGUMENTS.pidfile)
  File "/usr/lib/python2.6/site-packages/keytrustee/process.py", line 8, in __init__
    raise ex.KeytrusteeError('Invalid PID path: %s' % path)
keytrustee.exceptions.KeytrusteeError: Invalid PID path:
/var/lib/keytrustee/.keytrustee/keytrustee.pid
```

Starting Key Trustee Server while the port is in use fails silently

If you try to start Key Trustee Server while the port is in use (for example, due to a stale process or another process using the port), the process appears to start successfully, but ends shortly after with no indication to the user. The Key Trustee Server log contains the following error:

```
Traceback (most recent call last):
  File "/usr/lib/python2.6/site-packages/cherrypy/process/wspbus.py", line 235, in start
    self.publish('start')
  File "/usr/lib/python2.6/site-packages/cherrypy/process/wspbus.py", line 215, in publish
    raise exc
ChannelFailures: IOError("Port 11371 not free on 'keytrustee01.example.com'",)
```

Starting Key Trustee Server as a user other than keytrustee fails

Starting Key Trustee Server as a user other than keytrustee (for example, as root or another super user) fails.

Issues Fixed in Cloudera Navigator Key Trustee Server 5.4.9

Cannot register a client with Key Trustee Server after registering GPG public key

After creating an organization, the organization administrator receives a link to register a GPG key. If the administrator registers the GPG key before any clients are registered with the organization, all subsequent registrations with that organization fail.

Issues Fixed in Cloudera Navigator Key Trustee Server 5.4.3

Key Trustee Server logs error GnuPG operation exited with return code=2

The Key Trustee Server log (`keytrustee.log`) includes error messages such as the following:

```
2014-12-30 04:23:22,166 - keytrustee.gnupg - ERROR - GnuPG operation exited with return
code=2:
[GNUPG: ] ENC_TO 9375BEFD0BEC8C12 1 0
[GNUPG: ] GOOD_PASSPHRASE
[GNUPG: ] BEGIN_DECRYPTION
[GNUPG: ] PLAINTEXT 62 1419942201
[GNUPG: ] PLAINTEXT_LENGTH 230
[GNUPG: ] ERRSIG DF46C6427B4466F3 1 10 00 1419942201 9
[GNUPG: ] NO_PUBKEY DF46C6427B4466F3
[GNUPG: ] DECRYPTION_OKAY
[GNUPG: ] END_DECRYPTION
```

Package conflict: python-jinja2

Installing or upgrading Key Trustee Server on RHEL fails on hosts where `python-jinja2` is installed.

The `keytrustee-orgtool add` command fails with timeout: timed out error on CentOS 6.6

Adding an organization with the `keytrustee-orgtool add` command on CentOS 6.6 fails with the following error:

```
Exception in thread Thread-1:
Traceback (most recent call last):
  File "/usr/lib64/python2.6/threading.py", line 532, in __bootstrap_inner
    self.run()
  File "/usr/lib64/python2.6/threading.py", line 484, in run
    self._target(*self._args, **self._kwargs)
  File "/usr/lib/python2.6/site-packages/keytrustee/server/util.py", line 363, in
send_mail
    smtp = smtplib.SMTP('localhost', timeout=timeout)
  File "/usr/lib64/python2.6/smtplib.py", line 239, in __init__
    (code, msg) = self.connect(host, port)
```

Cloudera Navigator Key Trustee Server Release Notes

```
| File "/usr/lib64/python2.6/smtplib.py", line 296, in connect  
|     (code, msg) = self.getreply()  
+ File "/usr/lib64/python2.6/smtplib.py", line 337, in getreply  
|     line = self.file.readline()  
|     File "/usr/lib64/python2.6/socket.py", line 450, in readline  
|         data = self._sock.recv(self._rbufsize)  
|     timeout: timed out
```

Passive Database role sometimes fails to start when starting the Key Trustee Server service

Starting or restarting the Key Trustee Server service attempts to start the Active Database role and Passive Database role. If the Active Database has not completed starting up when the Passive Database attempts to start, the Passive Database fails to start.

For parcel-based installations, client and path environments must be manually configured

Users cannot use Key Trustee Server command-line utilities without configuring the path and setting environment variables.

Initializing high availability Key Trustee Servers fails if SSH communication fails

Initializing the active Key Trustee Server fails if the active Key Trustee Server cannot reach the passive Key Trustee Server over the default SSH port (22).

Enabling synchronous replication for high availability Key Trustee Servers fails

Running the `ktadmin enable-synchronous-replication` command does not properly configure synchronous replication in Key Trustee Server 5.4.0.

Issues Fixed in Cloudera Navigator Key Trustee Server 5.4.0

Cannot configure Key HSM after upgrade

After upgrading Key Trustee Server and Key HSM, the `ktadmin keyhsm --server khsm01.example.com --trust` command fails with the following error:

```
| TypeError: coercing to Unicode: need string or buffer, NoneType found
```

Cloudera Navigator Key HSM Release Notes

These release notes provide information on the new and changed features, known issues, and fixed issues for Cloudera Navigator Key HSM.

New Features and Changes in Cloudera Navigator Key HSM

The following sections describe what's new and changed in each Cloudera Navigator Key HSM release.

What's New in Cloudera Navigator Key HSM

The following sections describe what's new in each Cloudera Navigator Key HSM release.

What's New in Cloudera Navigator Key HSM 1.8.0

- Key HSM supports RHEL 6.8.
- The Key HSM keystore now uses a randomly-generated password. To specify a password, set `keyhsm.keystore.password.set` to `yes` in the `application.properties` file.

What's New in Cloudera Navigator Key HSM 1.7.0

- Key HSM can be started in debug mode with more verbose logging.
- Key HSM logging has been improved.
- A new HTTP endpoint has been added to test hardware security module (HSM) connectivity and functionality.

What's New in Cloudera Navigator Key HSM 1.6.0

- Log size and rollover limits are configurable.

A number of issues have also been fixed. See [Issues Fixed in Cloudera Navigator Key HSM 1.6.0](#) on page 494.

What's New in Cloudera Navigator Key HSM 1.5.1

An issue has been fixed. See [Issues Fixed in Cloudera Navigator Key HSM 1.5.1](#) on page 495.

What's New in Cloudera Navigator Key HSM 1.5.0

- Key HSM supports RHEL 7.
- Key HSM restricts key names to universally allowed character sets to prevent problems with key migration.

A number of issues have also been fixed. See [Issues Fixed in Cloudera Navigator Key HSM 1.5.0](#) on page 495.

What's New in Cloudera Navigator Key HSM 1.4.0

- Key HSM can now be attached to a Key Trustee Server instance with existing deposits.
- Support for SafeNet KeySecure
- KeySecure can be configured for HTTPS during setup

A number of issues have also been fixed. See [Issues Fixed in Cloudera Navigator Key HSM 1.4.0](#) on page 495.

Known Issues and Workarounds in Cloudera Navigator Key HSM

Keys with certain special characters cannot be migrated from Key Trustee Server to Key HSM

If any existing key names in Key Trustee Server use special characters other than hyphen (-), period (.), or underscore (_), or begin with non-alphanumeric characters, the migration to Key HSM fails.

Workaround: Decrypt any data using the affected key names, and re-encrypt it using a new key name without special characters, and retry the migration.

Upgrading Key HSM removes init script and binary

Upgrading Key HSM from 1.4.x to 1.5.x and higher removes the Key HSM init script and /usr/bin/keyhsm binary.

Workaround: Reinstall Key HSM:

```
$ sudo yum reinstall keytrustee-keyhsm
```

Key HSM cannot trust Key Trustee Server certificate if it has extended attributes

Key HSM cannot trust the Key Trustee Server certificate if it has extended attributes, and therefore cannot integrate with Key Trustee Server.

Workaround: Import the Key Trustee Server certificate to the Key HSM trust store using Java keytool instead of the keyhsm trust command.

Key HSM cannot integrate with SafeNet KeySecure HSM over SSL

Key HSM cannot trust the Key Trustee Server certificate if it has extended attributes, and therefore cannot integrate with Key Trustee Server.

Workaround: Import the Key Trustee Server certificate to the Key HSM trust store using Java keytool instead of the keyhsm trust command.

Issues Fixed in Cloudera Navigator Key HSM

The following sections describe issues fixed in each Cloudera Navigator Key HSM release.

Issues Fixed in Cloudera Navigator Key HSM 1.8.0

For new features in Key HSM 1.8.0, see [What's New in Cloudera Navigator Key HSM 1.8.0](#) on page 493.

Issues Fixed in Cloudera Navigator Key HSM 1.7.0

Using Key HSM with SafeNet KeySecure over TLS fails

When Key HSM is integrated with SafeNet KeySecure over TLS, Key HSM stops communicating with KeySecure after a period of time.

Issues Fixed in Cloudera Navigator Key HSM 1.6.0

Key HSM fails to start if the HSM contains a large number of keys

If the HSM stores a large number (hundreds) of keys, Key HSM fails to start with a timeout error similar to the following:

```
SEVERE: Timeout attempting to start services.  
All services available: : [ Failed ]
```

Key HSM logs do not roll over

Key HSM logs do not roll over, resulting in a log file that perpetually grows larger.

Issues Fixed in Cloudera Navigator Key HSM 1.5.1

Upgrading Key HSM removes init script

Upgrading Key HSM from 1.4.x to 1.5.x removes the Key HSM init script and `/usr/bin/keyhsm` binary. With this fix, future upgrades (from 1.5.1 to a higher release) do not experience this issue. To resolve the issue after upgrading to 1.5.x, reinstall Key HSM (`yum reinstall keytrustee-keyhsm`).

Issues Fixed in Cloudera Navigator Key HSM 1.5.0

Interrupting key migration from Key Trustee Server to Key HSM can result in lost data

Interrupting key migration (for example, using `Ctrl+C`) when integrating Key Trustee Server with Key HSM can result in data loss.

Issues Fixed in Cloudera Navigator Key HSM 1.4.0

Inserting deposits to Key HSM with Luna HSM logs stack trace

Inserting deposits to Key HSM using a Luna HSM logs the following stack trace to `keyhsm.log`:

```
SEVERE: -----Extended reason-----
Nov 21, 2014 4:39:11 PM com.cloudera.app.display.logging.AppLogger logInfo
INFO: com.safenetinc.luna.LunaCryptokiException.ThrowNew(LunaCryptokiException.java:66)
com.safenetinc.luna.LunaAPI.CheckSessionState(Native Method)
com.safenetinc.luna.LunaSession.isLoggedIn(LunaSession.java:149)
com.safenetinc.luna.LunaSlot.isLoggedIn(LunaSlot.java:155)
com.safenetinc.luna.LunaSlotManager.isLoggedIn(LunaSlotManager.java:585)
[...]
org.apache.tomcat.util.net.NioEndpoint$SocketProcessor.doRun(NioEndpoint.java:1720)
org.apache.tomcat.util.net.NioEndpoint$SocketProcessor.run(NioEndpoint.java:1679)
java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1145)
java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:615)
org.apache.tomcat.util.threads.TaskThread$WrappingRunnable.run(TaskThread.java:61)
java.lang.Thread.run(Thread.java:745)
```

service keyhsm trust only works with absolute path

Running `service keyhsm trust cert_file` fails if `cert_file` is a relative path.

Key Trustee KMS Release Notes

These release notes provide information on the new and changed features, known issues, and fixed issues for Key Trustee KMS.

New Features in Key Trustee KMS

The following sections describe what's new in each Key Trustee KMS release.



Note: Release 5.4.3 is the only maintenance release of Key Trustee KMS 5.4.

Release 5.5.4 is the only maintenance release of Key Trustee KMS 5.5.

There is no Release 5.6 of Key Trustee KMS.

There is no Release 5.7.2 or 5.7.3 of Key Trustee KMS.

There is no Release 5.8.1 or 5.8.3 of Key Trustee KMS.

What's New in Key Trustee KMS

The following sections describe what's new in each Key Trustee KMS release.

What's New in Key Trustee KMS 5.9.0

- Support for RHEL 6.8
- Support for OEL 6.8
- Support for Debian 8.4
- Support for SLES 12 SP1

What's New in Key Trustee KMS 5.8.2

- An issue has been fixed. See [Issues Fixed in Key Trustee KMS 5.8.2](#) on page 498.

What's New in Key Trustee KMS 5.8.0

- Support for Debian 8.2
- When adding the parcel-based Key Trustee KMS service for the first time, Cloudera Manager automatically backs up Key Trustee KMS locally.
- The Key Trustee KMS backup script (`ktbackup.sh`) adds a new option, `--roll`, which specifies the number of backups to retain.
- When Key Trustee Server is configured for high availability, adding the Key Trustee KMS service in Cloudera Manager automatically configures round robin DNS load balancing.

What's New in Key Trustee KMS 5.7.4

- An issue has been fixed. See [Issues Fixed in Key Trustee KMS 5.7.4](#) on page 498.

What's New in Key Trustee KMS 5.7.1

- An issue has been fixed. See [Issues Fixed in Key Trustee KMS 5.7.1](#) on page 498.

What's New in Key Trustee KMS 5.7.0

- Support for RHEL 7.2
- Support for OEL 5.11, 7.2
- Support for SLES 11 SP4

- Support for Debian 7.8
- A backup script (`ktbackup.sh`) is included with Key Trustee KMS.

What's New in Key Trustee KMS 5.5.4

- A number of issues have been fixed. See [Issues Fixed in Key Trustee KMS 5.5.4](#) on page 499.

What's New in Key Trustee KMS 5.5.0

- Rolling restart works with Key Trustee KMS high availability.
- When running Key Trustee KMS in a highly available configuration, Cloudera Manager can automatically generate the load balancer URL.

What's New in Key Trustee KMS 5.4.3

- A passive Key Trustee Server can be added to the Key Trustee KMS configuration using Cloudera Manager.
- An issue has also been fixed. See [Issues Fixed in Key Trustee KMS 5.4.3](#) on page 499.

Known Issues and Workarounds in Key Trustee KMS

Adding Key Trustee KMS 5.4 to Cloudera Manager 5.5 displays warning

Adding the Key Trustee KMS service to a CDH 5.4 cluster managed by Cloudera Manager 5.5 displays the following message, even if Key Trustee KMS is installed:

"The following selected services cannot be used due to missing components: keytrustee-keyprovider. Are you sure you wish to continue with them?"

Workaround: Verify that the Key Trustee KMS parcel or package is installed and click **OK** to continue adding the service.

KMS and Key Trustee ACLs do not work in Cloudera Manager 5.3

ACLs configured for the KMS (File) and KMS (Navigator Key Trustee) services do not work since these services do not receive the values for `hadoop.security.group.mapping` and related group mapping configuration properties.

Workaround:

KMS (File): Add all configuration properties starting with `hadoop.security.group.mapping` from the NameNode `core-site.xml` to the KMS (File) property, **Key Management Server Advanced Configuration Snippet (Safety Valve) for core-site.xml**

KMS (Navigator Key Trustee): Add all configuration properties starting with `hadoop.security.group.mapping` from the NameNode `core-site.xml` to the KMS (Navigator Key Trustee) property, **Key Management Server Proxy Advanced Configuration Snippet (Safety Valve) for core-site.xml**.

The Key Trustee KMS service fails to start if the Trust Store is configured without also configuring the Keystore

If you configure the Key Trustee KMS service **Key Management Server Proxy TLS/SSL Certificate Trust Store File** and **Key Management Server Proxy TLS/SSL Certificate Trust Store Password** parameters without also configuring the **Key Management Server Proxy TLS/SSL Server JKS Keystore File Location** and **Key Management Server Proxy TLS/SSL Server JKS Keystore File Password** parameters, the Key Trustee KMS service does not start.

Workaround: Configure all Trust Store and Keystore parameters.

Key Trustee KMS Release Notes

Key Trustee KMS backup script fails if PostgreSQL versions lower than 9.3 are installed

If PostgreSQL versions lower than 9.3 are installed on the Key Trustee KMS host, the `ktbackup.sh` script fails with an error similar to the following:

```
pg_dump: server version: 9.3.11; pg_dump version: 9.2.14
pg_dump: aborting because of server version mismatch
```

Workaround: Uninstall the lower PostgreSQL version.

Issues Fixed in Key Trustee KMS

The following sections describe issues fixed in each Key Trustee KMS release.



Note: Release 5.4.3 is the only maintenance release of Key Trustee KMS 5.4.

Release 5.5.4 is the only maintenance release of Key Trustee KMS 5.5.

There is no Release 5.6 of Key Trustee KMS.

There is no Release 5.7.2 or 5.7.3 of Key Trustee KMS.

There is no Release 5.8.1 or 5.8.3 of Key Trustee KMS.

Issues Fixed in Key Trustee KMS 5.9.0

For new features in Key Trustee KMS 5.9.0, see [What's New in Key Trustee KMS 5.9.0](#) on page 496.

Issues Fixed in Key Trustee KMS 5.8.2

KMS does not fail over to Passive Key Trustee Server in some network failure scenarios

In some situations, if the Active Key Trustee Server is unreachable on the network, Key Trustee KMS does not fail over to the Passive Key Trustee Server.

Issues Fixed in Key Trustee KMS 5.8.0

For new features in Key Trustee KMS 5.8.0, see [What's New in Key Trustee KMS 5.8.0](#) on page 496.

Issues Fixed in Key Trustee KMS 5.7.4

KMS does not fail over to Passive Key Trustee Server in some network failure scenarios

In some situations, if the Active Key Trustee Server is unreachable on the network, Key Trustee KMS does not fail over to the Passive Key Trustee Server.

Issues Fixed in Key Trustee KMS 5.7.1

KMS ACLs read from wrong file

The UNDELETE and PURGE ACL entries were being read from `kms-site.xml` instead of `kms-acls.xml`.

Issues Fixed in Key Trustee KMS 5.7.0

For new features in Key Trustee KMS 5.7.0, see [What's New in Key Trustee KMS 5.7.0](#) on page 496.

Issues Fixed in Key Trustee KMS 5.5.4

Key Trustee KMS configuration file and keys are stored in a volatile location

If the Key Trustee KMS 5.5.0 parcel is deactivated, any existing GPG keys are also deactivated. If the parcel is then reactivated, new GPG keys (used to create an authenticated and private communication channel with the Key Trustee Server) are generated. The existing GPG keys that were in use before the deactivation are not lost; however, they become inactive. If remedial action is not taken before deactivation, this can result in a loss of access to HDFS Encryption Zone keys generated with the older set of GPG keys. This in turn leads to loss of access to all data in all encryption zones. As long as the Key Trustee KMS parcel directory is not deleted, access can be restored. Assistance from Cloudera Support may be required. See [TSB 2016-121](#) for more information (requires login to the Cloudera Support Portal).

KMS ACLs read from wrong file

The UNDELETE and PURGE ACL entries were being read from `kms-site.xml` instead of `kms-acls.xml`.

Issues Fixed in Key Trustee KMS 5.5.0

Key Trustee KMS intermittently fails to start due to permission issues

Key Trustee KMS sometimes fails to start with the following error:

```
java.io.IOException: TrusteeKeyProvider initialization failed.
  at
com.cloudera.keytrustee.TrusteeKeyProvider.createInitialClient(TrusteeKeyProvider.java:212)

  at com.cloudera.keytrustee.TrusteeKeyProvider.<init>(TrusteeKeyProvider.java:114)
  at com.cloudera.keytrustee.TrusteeKeyProvider.<init>(TrusteeKeyProvider.java:86)
[...]
Caused by: java.io.FileNotFoundException:
/var/lib/kms-keytrustee/keytrustee/.keytrustee/keytrustee.conf (Permission denied)
  at java.io.FileOutputStream.open(Native Method)
  at java.io.FileOutputStream.<init>(FileOutputStream.java:221)
  at java.io.FileOutputStream.<init>(FileOutputStream.java:171)
  at
com.cloudera.keytrustee.impl.ClientFactoryImpl.createNewClient(ClientFactoryImpl.java:129)

... 30 more
```

Key Trustee KMS starts up successfully even if it cannot contact the Key Trustee Server

Key Trustee KMS starts successfully when it cannot connect to the Key Trustee Server, instead of properly failing to start.

Issues Fixed in Key Trustee KMS 5.4.3

Host Component page display

The Host Component page now displays the package version for the KMS Trustee Key Provider.

Cloudera Navigator Encrypt Release Notes

These release notes provide information on the new and changed features, known issues, and fixed issues for Cloudera Navigator Encrypt.

New Features and Changes in Cloudera Navigator Encrypt

The following sections describe what's new and changed in each Cloudera Navigator Encrypt release.

What's New in Cloudera Navigator Encrypt

The following sections describe what's new in each Cloudera Navigator Encrypt release.

What's New in Cloudera Navigator Encrypt 3.10.0

- Several new commands have been added:
 - `navencrypt-collect` displays environment information for troubleshooting. For more information, see [Collecting Navigator Encrypt Environment Information](#).
 - `navencrypt-move --list-categories` lists existing categories.
 - `navencrypt restore-control-file` restores the `/etc/navencrypt/control` file from Key Trustee Server.
 - `navencrypt status --integrity` validates the Navigator Encrypt configuration.
- Access control list (ACL) policy files now support comments.

What's New in Cloudera Navigator Encrypt 3.9.0

- Support for RHEL 7.2
- Support for SLES 11 SP4
- Support for Debian 7.8
- Support for Ubuntu 14.04.3

An issue has also been fixed. See [Issues Fixed in Cloudera Navigator Encrypt 3.9.0](#) on page 501.

What's New in Cloudera Navigator Encrypt 3.8.0

- Support for RHEL 7.1
- Support for automatically mounting loop devices on boot
- Automatic failover for high availability Key Trustee Servers
- The ability to restore deleted local mount encryption keys (MEKs) from Key Trustee Server by UUID

What's New in Cloudera Navigator Encrypt 3.7.1

- A number of issues have been fixed. See [Issues Fixed in Cloudera Navigator Encrypt 3.7.1](#) on page 502.

What's New in Cloudera Navigator Encrypt 3.7.0

- Migration utility to migrate zNcrypt to Cloudera Navigator Encrypt

What's Changed in Cloudera Navigator Encrypt

The following sections describe what's changed in each Cloudera Navigator Encrypt release.

What's Changed in Cloudera Navigator Encrypt 3.8.0

- The kernel module is renamed from `zncrypt-kernel-module` to `navencrypt-kernel-module`.

Known Issues and Workarounds in Cloudera Navigator Encrypt

The `nav encrypt status --integrity` **command does not detect loop devices properly**

Running the `nav encrypt status --integrity` command on a system using a loop device incorrectly reports that the backing file does not exist.

Workaround: None.

The Navigator Encrypt kernel module does not build on Ubuntu 14.04.04

The Navigator Encrypt kernel module does not build on Ubuntu 14.04.04 (kernel version 4.2).

Workaround: None.

Using screen utility with Cloudera Navigator Encrypt does not work

Running Cloudera Navigator Encrypt commands within the Linux `screen` utility does not work correctly.

Workaround: None.

The `nav encrypt exec` **command does not work**

Using `nav encrypt exec` to run a single command without having a matching ACL rule does not work.

Workaround: Add a temporary ACL rule to allow the command you want to use.

Sophos antivirus is not compatible with Cloudera Navigator Encrypt

Kernel panics have been observed on machines running both Sophos antivirus and Cloudera Navigator Encrypt.

Workaround: None. Do not use Sophos antivirus in conjunction with Cloudera Navigator Encrypt.

Cloudera Navigator Encrypt crashes in Docker container

Cloudera Navigator Encrypt crashes when run in a Docker container.

Workaround: None. Do not use Cloudera Navigator Encrypt with Docker.

Issues Fixed in Cloudera Navigator Encrypt

The following sections describe issues fixed in each Cloudera Navigator Encrypt release.

Issues Fixed in Cloudera Navigator Encrypt 3.10.0

See [What's New in Cloudera Navigator Encrypt 3.10.0](#) on page 500 for new features in Cloudera Navigator Encrypt 3.10.0.

Issues Fixed in Cloudera Navigator Encrypt 3.9.0

The software-signing GPG key is missing from the repository tarball for Navigator Encrypt

The GPG key used to sign the Navigator Encrypt packages is missing from the repository tarball, and must be imported from `archive.gazzang.com`.

Cloudera Navigator Encrypt Release Notes

Issues Fixed in Cloudera Navigator Encrypt 3.8.0

Encrypting or decrypting a file or folder with a space in the name generates extraneous message

Running `nav encrypt-move` to encrypt or decrypt a file or folder with a space in the name generates the following message at the command line:

```
/usr/sbin/nav encrypt-move: line 233: [: /root/encrypt: binary operator expected
```

This message can be safely ignored.

Running `nav encrypt-move` twice on the same file does not fail properly

Running `nav encrypt-move` twice on the same file succeeds on files with long paths with a message similar to the following:

```
ln: creating symbolic link '/path/to/file/to/encrypt' to  
'/path/to/mountpoint/category/path/to/file/to/encrypt': File exists
```

Running `nav encrypt-prepare --undo` removes all similar entries from `/etc/nav encrypt/ztab`

Running `nav encrypt-prepare --undo /path/to/mountpoint` removes all similar entries from `/etc/nav encrypt/ztab`. For example, if you have mountpoints named `/mnt/data`, `/mnt/data1`, and `/mnt/data2`, running `nav encrypt-prepare --undo /mnt/data` removes `/mnt/data`, `/mnt/data1`, and `/mnt/data2` from the Navigator Encrypt mount table (`/etc/nav encrypt/ztab`).

Encrypting or decrypting in parallel fails

Running `nav encrypt-move` simultaneously in multiple terminals results in failure to encrypt or decrypt all of the specified data, with the following error message:

```
keyctl_unlink: Required key not available
```

No data is lost, because the source data is not removed, but you must re-run the failed operations sequentially.

Restarting `nav encrypt-mount` service after running `nav encrypt-prepare` sometimes fails

Running service `nav encrypt-mount restart` after `nav encrypt-prepare` sometimes fails with the following error message:

```
WARNING: Another nav encrypt-mount start process is currently running.
```

Issues Fixed in Cloudera Navigator Encrypt 3.7.1

Pressing **Ctrl + C** during rule creation corrupts rule file

Pressing **Ctrl + C** before rule addition completes results in a corrupted rule file. Attempting to add further ACL rules fails with the following error:

```
[ERROR] Cannot parse ACL format: ACL header is not found: Did you type an incorrect key?
```

Intermittent Key Trustee Server communication errors

Intermittently, communication with the Key Trustee Server fails with an error similar to the following:

```
[error] UnicodeDecodeError: 'utf8' codec can't decode byte 0x80 in position 1: invalid start byte
```

The `navencrypt-move --per-file` option works only if the source and destination are on the same device

Using the `navencrypt-move --per-file` fails if the source and destination are on different devices.

Issues Fixed in Cloudera Navigator Encrypt 3.7.0

Kernel module reported by SLES as tainted kernel

The Navigator Encrypt kernel module is identified as a [tainted kernel](#) by SLES.

Navigator Encrypt does not always detect all processes accessing data to encrypt

Navigator Encrypt sometimes fails to detect all running processes that are accessing data that you want to encrypt.

Using `navencrypt-move` improperly fills storage device completely

Running `navencrypt-move` to encrypt a directory which contains the encryption mount point results in an infinite loop that fills the storage device.

Version and Download Information

Version and download information for Cloudera Manager, CDH, Impala, and Search can be found in the HTML documentation on the website at [Cloudera Documentation](#). Select the release version number and go to the HTML version of the Release Guide.

CDH 5 and Cloudera Manager 5 Requirements and Supported Versions

In an enterprise data hub, Cloudera Manager and CDH interact with several products such as Apache Accumulo, Apache Impala, Hue, Cloudera Search, and Cloudera Navigator. This guide provides information about which major and minor release version of a product is supported with which release version of CDH and Cloudera Manager.

Compatibility across different release versions of Cloudera Manager and CDH must be taken into account, especially when carrying out install/upgrade procedures.

JDK compatibility also varies across different Cloudera Manager and CDH versions. Certain versions of CDH 5 are compatible with both JDK 7 and JDK 8. In such cases, ensure all your services are deployed on the same major version. For example, you should not run Hadoop on JDK 7 while running Sqoop on JDK 8. Additionally, since Cloudera does not support mixed environments, all nodes in your cluster must be running the same major JDK version.

After installing each entity, upgrade to the latest patch version and apply any other appropriate updates. An available update might be specific to the operating system on which it is installed. For example, if you are using CentOS in your environment, you could choose 6 as the major version and 4 as the minor version to indicate that you are using CentOS 6.4. After installing this operating system, apply all relevant CentOS 6.4 upgrades and patches. In some cases, such as some browsers, a minor version might not be listed.

Each product matrix contains at least a subset of the following fields:

- **Feature:** This column lists notable new features that have been included in a particular release. For products/releases that do not have this column, refer the respective Release Notes for detailed information.
- **Lowest supported Cloudera Manager version:** Specifies the earliest version of Cloudera Manager that supports a product release version.
- **Lowest supported CDH version:** Specifies the earliest version of CDH that supports a product release version. The Cloudera Search and Impala matrices are an exception to this since each release is only compatible with a specific CDH release.
- **Lowest supported Impala version:** This column might not apply to all products, for example, Cloudera Search.
- **Lowest supported Search version:** This column might not apply to all products, for example, Cloudera Impala.
- **Integrated into CDH:** This column specifies whether a particular release is shipped with CDH or available independently. This field may not apply to all products, for example, Cloudera Navigator.

CDH Requirements for Cloudera Manager

The Cloudera Manager minor version must always be equal to or greater than the CDH minor version. Older versions of Cloudera Manager might not support features in newer versions of CDH. For example, to upgrade to CDH 5.7.1 you must first upgrade to Cloudera Manager 5.7.0.

CDH and Cloudera Manager Supported Operating Systems

CDH 5.x provides 64-bit packages for RHEL-compatible, SLES, Ubuntu, and Debian systems as listed below. Review the following list of exceptions before you proceed.

- Cloudera does not support CDH cluster deployments using hosts in Docker containers.
- Cloudera supports RHEL 7 with the following limitations:
 - Only RHEL 7.2 and 7.1 are supported. RHEL 7.0 is not supported.
 - Red Hat currently supports only upgrades from Red Hat Enterprise Linux 6 to Red Hat Enterprise Linux 7 for specific/targeted use cases only. Contact your OS vendor and review [What are the supported use cases for upgrading to RHEL 7?](#)

CDH 5 and Cloudera Manager 5 Requirements and Supported Versions

- Cloudera Enterprise is supported on platforms with Security-Enhanced Linux (SELinux) enabled. However, Cloudera does not support use of SELinux with Cloudera Navigator. Cloudera is not responsible for policy support nor policy enforcement. If you experience issues with SELinux, contact your OS provider.



Note: All CDH hosts that make up a logical cluster need to run on the same major OS release to be covered by Cloudera Support. Cloudera Manager needs to run on the same OS release as one of the CDH clusters it manages, to be covered by Cloudera Support. The risk of issues caused by running different minor OS releases is considered lower than the risk of running different major OS releases. Cloudera recommends running the same minor release cross-cluster, because it simplifies issue tracking and supportability.

CDH and Cloudera Manager 5.9.x Supported Operating Systems

Operating System	Version
Red Hat Enterprise Linux (RHEL)-compatible	
RHEL (+ SELinux mode in available versions)	7.2, 7.1, 6.8, 6.7, 6.6, 6.5, 6.4, 5.11, 5.10, 5.7
CentOS (+ SELinux mode in available versions)	7.2, 7.1, 6.8, 6.7, 6.6, 6.5, 6.4, 5.11, 5.10, 5.7
Oracle Enterprise Linux (OEL) with Unbreakable Enterprise Kernel (UEK)	7.2 (UEK R2), 7.1, 6.8 (UEK R3), 6.7 (UEK R3), 6.6 (UEK R3), 6.5 (UEK R2, UEK R3), 6.4 (UEK R2), 5.11, 5.10, 5.7
SLES	
SUSE Linux Enterprise Server (SLES)	12 with Service Pack 1, 11 with Service Pack 4, 11 with Service Pack 3, 11 with Service Pack 2
Hosts running Cloudera Manager Agents must use SUSE Linux Enterprise Software Development Kit 11 SP1 .	
Ubuntu/Debian	
Ubuntu	Trusty 14.04 - Long-Term Support (LTS) Precise 12.04 - Long-Term Support (LTS)
Debian	Jessie 8.4, 8.2 Wheezy 7.8, 7.1, 7.0

CDH and Cloudera Manager 5.8.x Supported Operating Systems

Operating System	Version
Red Hat Enterprise Linux (RHEL)-compatible	
RHEL (+ SELinux mode in available versions)	7.2, 7.1, 6.7, 6.6, 6.5, 6.4, 5.10, 5.7
CentOS (+ SELinux mode in available versions)	7.2, 7.1, 6.7, 6.6, 6.5, 6.4, 5.10, 5.7
Oracle Enterprise Linux (OEL) with Unbreakable Enterprise Kernel (UEK)	7.2 (UEK R2), 7.1, 6.7 (UEK R3), 6.6 (UEK R3), 6.5 (UEK R2, UEK R3), 6.4 (UEK R2), 5.11, 5.10, 5.7

Operating System	Version
SLES	
SUSE Linux Enterprise Server (SLES)	11 with Service Pack 4, 11 with Service Pack 3, 11 with Service Pack 2
Hosts running Cloudera Manager Agents must use SUSE Linux Enterprise Software Development Kit 11 SP1 .	
Ubuntu/Debian	
Ubuntu	Trusty 14.04 - Long-Term Support (LTS) Precise 12.04 - Long-Term Support (LTS)
Debian	Jessie 8.2, Wheezy 7.8, Wheezy 7.1, Wheezy 7.0

CDH and Cloudera Manager 5.7.x Supported Operating Systems

Operating System	Version
Red Hat Enterprise Linux (RHEL)-compatible	
RHEL (+ SELinux mode in available versions)	7.2, 7.1, 6.7, 6.6, 6.5, 6.4, 5.10, 5.7
CentOS (+ SELinux mode in available versions)	7.2, 7.1, 6.7, 6.6, 6.5, 6.4, 5.10, 5.7
Oracle Enterprise Linux (OEL) with Unbreakable Enterprise Kernel (UEK)	7.2 (UEK R2), 7.1, 6.7 (UEK R3), 6.6 (UEK R3), 6.5 (UEK R2, UEK R3), 6.4 (UEK R2), 5.11, 5.10, 5.7 (UEK R2)
SLES	
SUSE Linux Enterprise Server (SLES)	11 with Service Pack 4, 11 with Service Pack 3, 11 with Service Pack 2
Hosts running Cloudera Manager Agents must use SUSE Linux Enterprise Software Development Kit 11 SP1 .	
Ubuntu/Debian	
Ubuntu	Trusty 14.04 - Long-Term Support (LTS) Precise 12.04 - Long-Term Support (LTS)
Debian	Wheezy 7.8, Wheezy 7.1, Wheezy 7.0

CDH and Cloudera Manager 5.6.x Supported Operating Systems

Operating System	Version
Red Hat Enterprise Linux (RHEL)-compatible	
RHEL (+ SELinux mode in available versions)	7.1, 6.7, 6.6, 6.5, 6.4, 5.10, 5.7
CentOS (+ SELinux mode in available versions)	7.1, 6.7, 6.6, 6.5, 6.4, 5.10, 5.7
Oracle Enterprise Linux (OEL) with Unbreakable Enterprise Kernel (UEK)	7.1, 6.7 (UEK R3), 6.6 (UEK R3), 6.5 (UEK R2, UEK R3), 6.4 (UEK R2), 5.11, 5.10, 5.7 (UEK R2)

CDH 5 and Cloudera Manager 5 Requirements and Supported Versions

Operating System	Version
SLES	
SUSE Linux Enterprise Server (SLES)	11 with Service Pack 3, 11 with Service Pack 2
Hosts running Cloudera Manager Agents must use SUSE Linux Enterprise Software Development Kit 11 SP1 .	
Ubuntu/Debian	
Ubuntu	Trusty 14.04 - Long-Term Support (LTS) Precise 12.04 - Long-Term Support (LTS)
Debian	Wheezy 7.1, Wheezy 7.0

CDH and Cloudera Manager 5.5.x Supported Operating Systems

Operating System	Version
Red Hat Enterprise Linux (RHEL)-compatible	
RHEL (+ SELinux mode in available versions)	7.1, 6.7, 6.6, 6.5, 6.4, 5.7
CentOS (+ SELinux mode in available versions)	7.1, 6.7, 6.6, 6.5, 6.4, 5.7
Oracle Enterprise Linux (OEL) with Unbreakable Enterprise Kernel (UEK)	7.1, 6.7 (UEK R3), 6.6 (UEK R3), 6.5 (UEK R2, UEK R3), 6.4 (UEK R2), 5.6 (UEK R2)
SLES	
SUSE Linux Enterprise Server (SLES)	11 with Service Pack 3, 11 with Service Pack 2
Hosts running Cloudera Manager Agents must use SUSE Linux Enterprise Software Development Kit 11 SP1 .	
Ubuntu/Debian	
Ubuntu	Trusty 14.04 - Long-Term Support (LTS) Precise 12.04 - Long-Term Support (LTS)
Debian	Wheezy 7.1, Wheezy 7.0

CDH and Cloudera Manager 5.4.x Supported Operating Systems

Operating System	Version
Red Hat Enterprise Linux (RHEL)-compatible	
RHEL (+ SELinux mode in available versions)	6.6, 6.5, 6.4, 5.10, 5.7
CentOS (+ SELinux mode in available versions)	6.6, 6.5, 6.4, 5.10, 5.7
Oracle Enterprise Linux (OEL) with Unbreakable Enterprise Kernel (UEK)	6.6 (UEK R3), 6.5 (UEK R2, UEK R3), 6.4 (UEK R2), 5.6 (UEK R2)
SLES	
SUSE Linux Enterprise Server (SLES)	11 with Service Pack 3, 11 with Service Pack 2

Operating System	Version
Hosts running Cloudera Manager Agents must use SUSE Linux Enterprise Software Development Kit 11 SP1 .	
Ubuntu/Debian	
Ubuntu	Trusty 14.04 - Long-Term Support (LTS) Precise 12.04 - Long-Term Support (LTS)
Debian	Wheezy 7.1, Wheezy 7.0

CDH and Cloudera Manager 5.3.x Supported Operating Systems

Operating System	Version
Red Hat Enterprise Linux (RHEL)-compatible	
RHEL (+ SELinux mode in available versions)	6.5, 6.4, 5.7
CentOS (+ SELinux mode in available versions)	6.5, 6.4, 5.7
Oracle Enterprise Linux (OEL) with Unbreakable Enterprise Kernel (UEK)	6.5 (UEK R2, UEK R3), 6.4 (UEK R2), 5.6 (UEK R2)
SLES	
SUSE Linux Enterprise Server (SLES)	11 with Service Pack 3, 11 with Service Pack 2
Hosts running Cloudera Manager Agents must use SUSE Linux Enterprise Software Development Kit 11 SP1 .	
Ubuntu/Debian	
Ubuntu	Trusty 14.04 - Long-Term Support (LTS) Precise 12.04 - Long-Term Support (LTS)
Debian	Wheezy 7.1, Wheezy 7.0

CDH and Cloudera Manager 5.2.x Supported Operating Systems

Operating System	Version
Red Hat Enterprise Linux (RHEL)-compatible	
RHEL (+ SELinux mode in available versions)	6.5, 6.4, 5.7
CentOS (+ SELinux mode in available versions)	6.5, 6.4, 5.7
Oracle Enterprise Linux (OEL) with Unbreakable Enterprise Kernel (UEK)	6.5 (UEK R2, UEK R3), 6.4 (UEK R2), 5.6 (UEK R2)
SLES	
SUSE Linux Enterprise Server (SLES)	11 with Service Pack 3, 11 with Service Pack 2
Hosts running Cloudera Manager Agents must use SUSE Linux Enterprise Software Development Kit 11 SP1 .	
Ubuntu/Debian	
Ubuntu	Trusty 14.04 - Long-Term Support (LTS) Precise 12.04 - Long-Term Support (LTS)

CDH 5 and Cloudera Manager 5 Requirements and Supported Versions

Operating System	Version
Debian	Wheezy 7.1, Wheezy 7.0

CDH and Cloudera Manager 5.1.x Supported Operating Systems

Operating System	Version
Red Hat Enterprise Linux (RHEL)-compatible	
RHEL (+ SELinux mode in available versions)	6.5, 6.4, 5.7
CentOS (+ SELinux mode in available versions)	6.5, 6.4, 5.7
Oracle Enterprise Linux (OEL) with Unbreakable Enterprise Kernel (UEK)	6.5 (UEK R2, UEK R3), 6.4 (UEK R2), 5.6 (UEK R2)
SLES	
SUSE Linux Enterprise Server (SLES)	11 with Service Pack 3, 11 with Service Pack 2
Hosts running Cloudera Manager Agents must use SUSE Linux Enterprise Software Development Kit 11 SP1 .	
Ubuntu/Debian	
Ubuntu	Precise 12.04 - Long-Term Support (LTS)
Debian	Wheezy 7.1, Wheezy 7.0

CDH and Cloudera Manager 5.0.x Supported Operating Systems

Operating System	Version
Red Hat Enterprise Linux (RHEL)-compatible	
RHEL (+ SELinux mode in available versions)	6.4, 6.2, 5.7
CentOS (+ SELinux mode in available versions)	6.4, 6.2, 5.7
Oracle Enterprise Linux (OEL) with Unbreakable Enterprise Kernel (UEK)	6.4, 5.6
SLES	
SUSE Linux Enterprise Server (SLES)	11 with Service Pack 3, 11 with Service Pack 2
Hosts running Cloudera Manager Agents must use SUSE Linux Enterprise Software Development Kit 11 SP1 .	
Ubuntu/Debian	
Ubuntu	Precise 12.04 - Long-Term Support (LTS)
Debian	Wheezy 7.1, Wheezy 7.0

Filesystem Requirements

Supported Filesystems

The Hadoop Distributed File System (HDFS) is designed to run on top of an underlying filesystem in an operating system. Cloudera recommends that you use either of the following filesystems tested on the [supported operating systems](#):

- **ext3:** This is the most tested underlying filesystem for HDFS.
- **ext4:** This scalable extension of ext3 is supported in more recent Linux releases.



Important: Cloudera does not support in-place upgrades from ext3 to ext4. Cloudera recommends that you format disks as ext4 before using them as data directories.

- **XFS:** This is the default filesystem in RHEL 7.

File Access Time

Linux filesystems keep metadata that record when each file was accessed. This means that even reads result in a write to the disk. To speed up file reads, Cloudera recommends that you disable this option, called `atime`, using the mount option in `/etc/fstab`:

```
/dev/sdb1 /data1 ext4 defaults,noatime 0
```

Apply the change without rebooting:

```
mount -o remount /data1
```

CDH and Cloudera Manager Supported Databases

Cloudera Manager requires several databases. The Cloudera Manager Server stores information about configured services, role assignments, configuration history, commands, users, and running processes in a database of its own. You must also specify a database for the Activity Monitor and Reports Manager roles.



Important: When you restart processes, the configuration for each of the services is redeployed using information saved in the Cloudera Manager database. If this information is not available, your cluster does not start or function correctly. You must schedule and maintain regular backups of the Cloudera Manager database to recover the cluster in the event of the loss of this database.

The database you use must be configured to support UTF8 character set encoding. The embedded PostgreSQL database installed when you follow [Installation Path A - Automated Installation by Cloudera Manager \(Non-Production Mode\)](#) automatically provides UTF8 encoding. If you install a custom database, you might need to enable UTF8 encoding. The commands for enabling UTF8 encoding are described in each database topic under [Cloudera Manager and Managed Service Datastores](#).

After installing a database, upgrade to the latest patch version and apply any other appropriate updates. Available updates may be specific to the operating system on which it is installed.

Cloudera supports the shipped version of MariaDB, MySQL and PostgreSQL for each supported Linux distribution.

Component	MariaDB	MySQL	SQLite	PostgreSQL	Oracle	Derby - see Note 5
Cloudera Manager	5.5, 10	5.6, 5.5, 5.1	–	9.4, 9.3, 9.2, 9.1. 8.4, 8.3, 8.1	12c, 11gR2	

Component	MariaDB	MySQL	SQLite	PostgreSQL	Oracle	Derby - see Note 5
Oozie	5.5, 10	5.6, 5.5, 5.1 See Note 3	–	9.4, 9.3, 9.2, 9.1. 8.4, 8.3, 8.1 See Note 3	12c, 11gR2	Default
Flume	–	–	–	–	–	Default (for the JDBC Channel only)
Hue	5.5, 10 See Note 6	5.6, 5.5, 5.1 See Note 3	Default	9.4, 9.3, 9.2, 9.1. 8.4, 8.3, 8.1 See Note 3	12c, 11gR2	–
Hive/Impala	5.5, 10 See Note 1	5.6, 5.5, 5.1 See Note 3	–	9.4, 9.3, 9.2, 9.1. 8.4, 8.3, 8.1 See Note 3	12c, 11gR2	Default
Sentry	5.5, 10 See Note 1	5.6, 5.5, 5.1 See Note 3	–	9.4, 9.3, 9.2, 9.1. 8.4, 8.3, 8.1 See Note 3	12c, 11gR2	–
Sqoop 1	5.5, 10	See Note 4	–	See Note 4	See Note 4	–
Sqoop 2	5.5, 10	See Note 9	–	–	–	Default

**Note:**

1. Cloudera supports the databases listed above provided they are supported by the underlying operating system on which they run.
2. MySQL 5.5 is supported on CDH 5.1. MySQL 5.6 is supported on CDH 5.1 and higher. The InnoDB storage engine must be enabled in the MySQL server.
3. Cloudera Manager installation fails if GTID-based replication is enabled in MySQL.
4. PostgreSQL 9.2 is supported on CDH 5.1 and higher. PostgreSQL 9.3 is supported on CDH 5.2 and higher. PostgreSQL 9.4 is supported on CDH 5.5 and higher.
5. For purposes of transferring data only, Sqoop 1 supports MySQL 5.0 and above, PostgreSQL 8.4 and above, Oracle 10.2 and above, Teradata 13.10 and above, and Netezza TwinFin 5.0 and above. The Sqoop metastore works only with HSQLDB (1.8.0 and higher 1.x versions; the metastore does not work with any HSQLDB 2.x versions).
6. Derby is supported as shown in the table, but not always recommended. See the pages for individual components in the [Cloudera Installation](#) guide for recommendations.
7. CDH 5 Hue requires the default MySQL version of the operating system on which it is being installed, which is usually MySQL 5.1, 5.5, or 5.6.
8. When installing a JDBC driver, only the `ojdbc6.jar` file is supported for both Oracle 11g R2 and Oracle 12c; the `ojdbc7.jar` file is not supported.
9. Sqoop 2 lacks some of the features of Sqoop 1. Cloudera recommends you use Sqoop 1. Use Sqoop 2 only if it contains all the features required for your use case.
10. MariaDB 10 is supported only on CDH 5.9 and higher.

CDH and Cloudera Manager Supported JDK Versions

For each supported major JDK version, CDH and Cloudera Manager are supported with any minor JDK update higher than the minimum required version (see table below). Cloudera reserves the right at any point to exclude a minor version from our support matrix, and, if so, list it under excluded versions. A reason to exclude a JDK might be that a major stability issue is discovered.

Cloudera Manager can install Oracle JDK 1.7.0_67 during installation and upgrade. If you prefer to install the JDK yourself, follow the instructions in [Java Development Kit Installation](#).



Note: Only 64 bit JDKs are supported.

Minimum Required Version(s)	Excluded Version(s)	Comments
JDK1.8_31	JDK1.8_40, JDK1.8_45	N/A
JDK1.7_55	N/A	<ul style="list-style-type: none"> Using JDK 1.7.0_80 with CDH 5.1 and CDH 5.2 causes Kerberos authentication failures with HDFS clients.

Cloudera Manager Supported Browsers

Hue

Hue works with the two most recent [LTS](#) (long term support) or [ESR](#) (extended support release) browsers. Cookies and JavaScript must be on.

- **Chrome:** [Version history](#)
- **Firefox:** [Version history](#)
- **Internet Explorer:** [Version history](#)
- **Safari** (Mac only): [Version history](#)

Hue can display in older versions and even other browsers, but you might not have access to all of its features.

Cloudera Manager

The Cloudera Manager Admin Console, which you use to install, configure, manage, and monitor services, supports the latest version of the following browsers:

- Mozilla Firefox
- Google Chrome
- Internet Explorer
- Safari.

Supported Network Protocols

CDH requires IPv4. IPv6 is not supported.

See also [Configuring Network Names](#).

Multihoming Support

Multihoming CDH or Cloudera Manager is not supported outside specifically certified Cloudera partner appliances. Cloudera finds that current Hadoop architectures combined with modern network infrastructures and security practices remove the need for multihoming. Multihoming, however, is beneficial internally in appliance form factors to take advantage of high-bandwidth InfiniBand interconnects.

Although some subareas of the product may work with unsupported custom multihoming configurations, there are known issues with multihoming. In addition, unknown issues may arise because multihoming is not covered by our test matrix outside the Cloudera-certified partner appliances.

CDH and Cloudera Manager Supported Transport Layer Security Versions

The following components are supported by the indicated versions of Transport Layer Security (TLS):

Table 4: Components Supported by TLS

Component	Role	Name	Port	Version
Cloudera Manager	Cloudera Manager Server		7182	TLS 1.2
Cloudera Manager	Cloudera Manager Server		7183	TLS 1.2
Flume			9099	TLS 1.2
Flume		Avro Source/Sink		TLS 1.2
Flume		Flume HTTP Source/Sink		TLS 1.2
HBase	Master	HBase Master Web UI Port	60010	TLS 1.2
HDFS	NameNode	Secure NameNode Web UI Port	50470	TLS 1.2
HDFS	Secondary NameNode	Secure Secondary NameNode Web UI Port	50495	TLS 1.2
HDFS	HttpFS	REST Port	14000	TLS 1.1, TLS 1.2
Hive	HiveServer2	HiveServer2 Port	10000	TLS 1.2
Hue	Hue Server	Hue HTTP Port	8888	TLS 1.2
Impala	Impala Daemon	Impala Daemon Beeswax Port	21000	TLS 1.2
Impala	Impala Daemon	Impala Daemon HiveServer2 Port	21050	TLS 1.2
Impala	Impala Daemon	Impala Daemon Backend Port	22000	TLS 1.2
Impala	Impala StateStore	StateStore Service Port	24000	TLS 1.2
Impala	Impala Daemon	Impala Daemon HTTP Server Port	25000	TLS 1.2
Impala	Impala StateStore	StateStore HTTP Server Port	25010	TLS 1.2
Impala	Impala Catalog Server	Catalog Server HTTP Server Port	25020	TLS 1.2
Impala	Impala Catalog Server	Catalog Server Service Port	26000	TLS 1.2

Component	Role	Name	Port	Version
Oozie	Oozie Server	Oozie HTTPS Port	11443	TLS 1.1, TLS 1.2
Solr	Solr Server	Solr HTTP Port	8983	TLS 1.1, TLS 1.2
Solr	Solr Server	Solr HTTPS Port	8985	TLS 1.1, TLS 1.2
Spark	History Server		18080	TLS 1.2
YARN	ResourceManager	ResourceManager Web Application HTTP Port	8090	TLS 1.2
YARN	JobHistory Server	MRv1 JobHistory Web Application HTTP Port	19890	TLS 1.2

Cloudera Manager Resource Requirements

Cloudera Manager requires the following resources:

- **Disk Space**
 - **Cloudera Manager Server**
 - 5 GB on the partition hosting `/var`.
 - 500 MB on the partition hosting `/usr`.
 - For parcels, the space required depends on the number of parcels you download to the Cloudera Manager Server and distribute to Agent hosts. You can download multiple parcels of the same product, of different versions and different builds. If you are managing multiple clusters, only one parcel of a product/version/build/distribution is downloaded on the Cloudera Manager Server—not one per cluster. In the local parcel repository on the Cloudera Manager Server, the approximate sizes of the various parcels are as follows:
 - CDH 5 (which includes Impala and Search) - 1.5 GB per parcel (packed), 2 GB per parcel (unpacked)
 - Impala - 200 MB per parcel
 - Cloudera Search - 400 MB per parcel
 - **Cloudera Management Service** - The Host Monitor and Service Monitor databases are stored on the partition hosting `/var`. Ensure that you have at least 20 GB available on this partition.
 - **Agents** - On Agent hosts, each unpacked parcel requires about three times the space of the downloaded parcel on the Cloudera Manager Server. By default, unpacked parcels are located in `/opt/cloudera/parcels`.
- **RAM** - 4 GB is recommended for most cases and is required when using Oracle databases. 2 GB might be sufficient for non-Oracle deployments with fewer than 100 hosts. However, to run the Cloudera Manager Server on a machine with 2 GB of RAM, you must tune down its maximum heap size (by modifying `-Xmx` in `/etc/default/cloudera-scm-server`). Otherwise the kernel might kill the Server for consuming too much RAM.
- **Python** - Cloudera Manager requires Python 2.4 or higher (but is not compatible with Python 3.0 or higher). Hue in CDH 5 and package installs of CDH 5 require Python 2.6 or 2.7. All supported operating systems include Python version 2.4 or higher. Cloudera Manager is compatible with Python 2.4 through the latest version of Python 2.x. Cloudera Manager does not support Python 3.0 and higher.
- **Perl** - Cloudera Manager requires [perl](#).

CDH and Cloudera Manager Networking and Security Requirements

The hosts in a Cloudera Manager deployment must satisfy the following networking and security requirements:

- CDH requires IPv4. IPv6 is not supported and must be disabled.

CDH 5 and Cloudera Manager 5 Requirements and Supported Versions

See also

- Multihoming CDH or Cloudera Manager is not supported outside specifically certified Cloudera partner appliances. Cloudera finds that current Hadoop architectures combined with modern network infrastructures and security practices remove the need for multihoming. Multihoming, however, is beneficial internally in appliance form factors to take advantage of high-bandwidth InfiniBand interconnects.
- Although some subareas of the product might work with unsupported custom multihoming configurations, there are known issues with multihoming. In addition, unknown issues can arise because multihoming is not covered by the test matrix outside the Cloudera-certified partner appliances.
- Cluster hosts must have a working network name resolution system and correctly formatted `/etc/hosts` file. All cluster hosts must have properly configured forward and reverse host resolution through DNS. The `/etc/hosts` files must:
 - Contain consistent information about hostnames and IP addresses across all hosts
 - Not contain uppercase hostnames
 - Not contain duplicate IP addresses

Cluster hosts must not use aliases, either in `/etc/hosts` or in configuring DNS. A properly formatted `/etc/hosts` file should be similar to the following example:

```
127.0.0.1 localhost.localdomain localhost
192.168.1.1 cluster-01.example.com cluster-01
192.168.1.2 cluster-02.example.com cluster-02
192.168.1.3 cluster-03.example.com cluster-03
```

- In most cases, the Cloudera Manager Server must have SSH access to the cluster hosts when you run the installation or upgrade wizard. You must log in using a root account or an account that has password-less sudo permission. For authentication during the installation and upgrade procedures, you must either enter the password or upload a public and private key pair for the root or sudo user account. If you want to use a public and private key pair, the public key must be installed on the cluster hosts before you use Cloudera Manager.

Cloudera Manager uses SSH only during the initial install or upgrade. Once the cluster is set up, you can disable root SSH access or change the root password. Cloudera Manager does not save SSH credentials, and all credential information is discarded when the installation is complete.

- If [single user mode](#) is not enabled, the Cloudera Manager Agent runs as root so that it can make sure the required directories are created and that processes and files are owned by the appropriate user (for example, the `hdfs` and `mapred` users).
- No blocking is done by Security-Enhanced Linux (SELinux).



Note: Cloudera Enterprise is supported on platforms with Security-Enhanced Linux (SELinux) enabled. However, Cloudera does not support use of SELinux with Cloudera Navigator. Cloudera is not responsible for policy support nor policy enforcement. If you experience issues with SELinux, contact your OS provider.

- No blocking by iptables or firewalls; port 7180 must be open because it is used to access Cloudera Manager after installation. Cloudera Manager communicates using specific ports, which must be open.
- For RHEL and CentOS, the `/etc/sysconfig/network` file on each host must contain the hostname you have just set (or verified) for that host.
- Cloudera Manager and CDH use several user accounts and groups to complete their tasks. The set of user accounts and groups varies according to the components you choose to install. Do not delete these accounts or groups and do not modify their permissions and rights. Ensure that no existing systems prevent these accounts and groups from functioning. For example, if you have scripts that delete user accounts not in a whitelist, add these accounts to the list of permitted accounts. Cloudera Manager, CDH, and managed services create and use the following accounts and groups:

Table 5: Users and Groups

Component (Version)	Unix User ID	Groups	Notes
Cloudera Manager (all versions)	cloudera-scm	cloudera-scm	<p>Cloudera Manager processes such as the Cloudera Manager Server and the monitoring roles run as this user.</p> <p>The Cloudera Manager keytab file must be named <code>cmf.keytab</code> since that name is hard-coded in Cloudera Manager.</p> <div style="border: 1px solid black; padding: 5px; margin-top: 10px;">  Note: Applicable to clusters managed by Cloudera Manager only. </div>
Apache Accumulo (Accumulo 1.4.3 and higher)	accumulo	accumulo	Accumulo processes run as this user.
Apache Avro			No special users.
Apache Flume (CDH 4, CDH 5)	flume	flume	The sink that writes to HDFS as this user must have write privileges.
Apache HBase (CDH 4, CDH 5)	hbase	hbase	The Master and the RegionServer processes run as this user.
HDFS (CDH 4, CDH 5)	hdfs	hdfs, hadoop	The NameNode and DataNodes run as this user, and the HDFS root directory as well as the directories used for edit logs should be owned by it.
Apache Hive (CDH 4, CDH 5)	hive	hive	<p>The HiveServer2 process and the Hive Metastore processes run as this user.</p> <p>A user must be defined for Hive access to its Metastore DB (for example, MySQL or Postgres) but it can be any identifier and does not correspond to a Unix uid. This is <code>javax.jdo.option.ConnectionUserName</code> in <code>hive-site.xml</code>.</p>
Apache HCatalog (CDH 4.2 and higher, CDH 5)	hive	hive	The WebHCat service (for REST access to Hive functionality) runs as the <code>hive</code> user.
HttpFS (CDH 4, CDH 5)	httpfs	httpfs	The HttpFS service runs as this user. See HttpFS Security Configuration for instructions on how to generate the merged <code>httpfs-http.keytab</code> file.
Hue (CDH 4, CDH 5)	hue	hue	Hue services run as this user.
Hue Load Balancer (Cloudera Manager 5.5 and higher)	apache	apache	The Hue Load balancer has a dependency on the <code>apache2</code> package that uses the <code>apache</code> user name. Cloudera Manager does not run processes using this user ID.

CDH 5 and Cloudera Manager 5 Requirements and Supported Versions

Component (Version)	Unix User ID	Groups	Notes
Cloudera Impala (CDH 4.1 and higher, CDH 5)	impala	impala, hive	Impala services run as this user.
Apache Kafka (Cloudera Distribution of Kafka 1.2.0)	kafka	kafka	Kafka services run as this user.
Java KeyStore KMS (CDH 5.2.1 and higher)	kms	kms	The Java KeyStore KMS service runs as this user.
Key Trustee KMS (CDH 5.3 and higher)	kms	kms	The Key Trustee KMS service runs as this user.
Key Trustee Server (CDH 5.4 and higher)	keytrustee	keytrustee	The Key Trustee Server service runs as this user.
Kudu	kudu	kudu	Kudu services run as this user.
Llama (CDH 5)	llama	llama	Llama runs as this user.
Apache Mahout			No special users.
MapReduce (CDH 4, CDH 5)	mapred	mapred, hadoop	Without Kerberos, the JobTracker and tasks run as this user. The LinuxTaskController binary is owned by this user for Kerberos.
Apache Oozie (CDH 4, CDH 5)	oozie	oozie	The Oozie service runs as this user.
Parquet			No special users.
Apache Pig			No special users.
Cloudera Search (CDH 4.3 and higher, CDH 5)	solr	solr	The Solr processes run as this user.
Apache Spark (CDH 5)	spark	spark	The Spark History Server process runs as this user.
Apache Sentry (CDH 5.1 and higher)	sentry	sentry	The Sentry service runs as this user.
Apache Sqoop (CDH 4, CDH 5)	sqoop	sqoop	This user is only for the Sqoop1 Metastore, a configuration option that is not recommended.
Apache Sqoop2 (CDH 4.2 and higher, CDH 5)	sqoop2	sqoop, sqoop2	The Sqoop2 service runs as this user.
Apache Whirr			No special users.
YARN (CDH 4, CDH 5)	yarn	yarn, hadoop	Without Kerberos, all YARN services and applications run as this user. The LinuxContainerExecutor binary is owned by this user for Kerberos.

Component (Version)	Unix User ID	Groups	Notes
Apache ZooKeeper (CDH 4, CDH 5)	zookeeper	zookeeper	The ZooKeeper processes run as this user. It is not configurable.

Product Compatibility Matrix for Apache Accumulo

This matrix contains compatibility information across versions of Apache Accumulo, and CDH and Cloudera Manager. For detailed information on each release, see [Apache Accumulo documentation](#).

Product	Lowest supported Cloudera Manager Version	Lowest supported CDH Version	Lowest supported Impala Version	Lowest supported Search Version	Integrated into CDH
Accumulo 1.7.2	Cloudera Manager 5.0.0	CDH 5.5.0	Not Supported	Not Supported	No
Accumulo 1.6.0	Cloudera Manager 5.0.0	CDH 4.6.0 - 4.x.x, CDH 5.1.0	Not Supported	Not Supported	No
Accumulo 1.4.4	Cloudera Manager 5.0.0	CDH 4.5.0 (Not for use with CDH 5)	Not Supported	Not Supported	No
Accumulo 1.4.3	Cloudera Manager 5.0.0	CDH 4.3.0 (Not for use with CDH 5)	Not Supported	Not Supported	No

Product Compatibility Matrix for Impala

This matrix contains compatibility information across versions of Impala and CDH/Cloudera Manager. For detailed information on each release, see [the Impala documentation](#).

Product	Supported Cloudera Manager Versions	Supported CDH Versions	Integrated into CDH
Impala 2.7.x for CDH 5.9.x	Cloudera Manager 5.0.0 - 5.x.x	CDH 5.9.x	CDH 5.9.x
Impala 2.6.x for CDH 5.8.x	Cloudera Manager 5.0.0 - 5.x.x	CDH 5.8.x	CDH 5.8.x
Impala 2.5.x for CDH 5.7.x	Cloudera Manager 5.0.0 - 5.x.x	CDH 5.7.x	CDH 5.7.x
Impala 2.4.x for CDH 5.6.x	Cloudera Manager 5.0.0 - 5.x.x	CDH 5.6.x	CDH 5.6.x
Impala 2.3.x for CDH 5.5.x	Cloudera Manager 5.0.0 - 5.x.x	CDH 5.5.x	CDH 5.5.x
Impala 2.2.x for CDH 5.4.x	Cloudera Manager 5.0.0 - 5.x.x	CDH 5.4.x	CDH 5.4.x
Impala 2.2.0	Cloudera Manager 5.0.0 - 5.x.x	CDH 5.4.0	CDH 5.4.0

CDH 5 and Cloudera Manager 5 Requirements and Supported Versions

Product	Supported Cloudera Manager Versions	Supported CDH Versions	Integrated into CDH
Impala 2.1.8	Cloudera Manager 5.0.0 - 5.x.x	CDH 5.3.10	CDH 5.3.10
Impala 2.1.7	Cloudera Manager 5.0.0 - 5.x.x	CDH 5.3.9	CDH 5.3.9
Impala 2.1.6	Cloudera Manager 5.0.0 - 5.x.x	CDH 5.3.8	CDH 5.3.8
Impala 2.1.5	Cloudera Manager 5.0.0 - 5.x.x	CDH 5.3.6	CDH 5.3.6
Impala 2.1.4	Cloudera Manager 5.0.0 - 5.x.x	CDH 5.3.4 and CDH 5.3.5	CDH 5.3.4 and CDH 5.3.5
Impala 2.1.3	Cloudera Manager 5.0.0 - 5.x.x	CDH 5.3.3	CDH 5.3.3
Impala 2.1.2	Cloudera Manager 5.0.0 - 5.x.x	CDH 5.3.2	CDH 5.3.2
Impala 2.1.1	Cloudera Manager 5.0.0 - 5.x.x	CDH 5.3.1	CDH 5.3.1
Impala 2.1.0	Cloudera Manager 5.0.0 - 5.x.x	CDH 5.3.0	CDH 5.3.0
Impala 2.0.5	Cloudera Manager 5.0.0 - 5.x.x	CDH 5.2.6	CDH 5.2.6
Impala 2.0.4	Cloudera Manager 5.0.0 - 5.x.x	CDH 5.2.5	CDH 5.2.5
Impala 2.0.3	Cloudera Manager 5.0.0 - 5.x.x	CDH 5.2.4	CDH 5.2.4
Impala 2.0.2	Cloudera Manager 5.0.0 - 5.x.x	CDH 5.2.3	CDH 5.2.3
Impala 2.0.1	Cloudera Manager 5.0.0 - 5.x.x	CDH 5.2.1	CDH 5.2.1
Impala 2.0.0	Cloudera Manager 5.0.0 - 5.x.x	CDH 5.2.0	CDH 5.2.0
Impala 1.4.4	Cloudera Manager 5.0.0 - 5.x.x	CDH 5.1.5	CDH 5.1.5
Impala 1.4.3	Cloudera Manager 5.0.0 - 5.x.x	CDH 5.1.4	CDH 5.1.4
Impala 1.4.2	Cloudera Manager 5.0.0 - 5.x.x	CDH 5.1.3	CDH 5.1.3
Impala 1.4.1	Cloudera Manager 5.0.0 - 5.x.x	CDH 5.1.2	CDH 5.1.2

Product	Supported Cloudera Manager Versions	Supported CDH Versions	Integrated into CDH
Impala 1.4.0	Cloudera Manager 5.0.0 - 5.x.x; Recommended: Cloudera Manager 5.1.0	CDH 5.1.0	CDH 5.1.0
Impala 1.3.3	Cloudera Manager 5.0.0 - 5.x.x	CDH 5.0.5	CDH 5.0.5
Impala 1.3.2	Cloudera Manager 5.0.0 - 5.x.x	CDH 5.0.4	CDH 5.0.4
Impala 1.3.1	Cloudera Manager 5.0.0 - 5.x.x	CDH 5.0.1 - 5.0.x	CDH 5.0.1
Impala 1.3.0	Cloudera Manager 5.0.0 - 5.x.x	CDH 5.0.0	CDH 5.0.0

Product Compatibility Matrix for Cloudera Distribution of Apache Kafka

Cloudera Distribution of Apache Kafka is currently distributed as a package and in a parcel that is independent of the CDH parcel. The parcel integrates with Cloudera Manager using a Custom Service Descriptor (CSD).

For the latest documentation, see [Kafka Documentation](#).

Product	Feature	Lowest Supported Cloudera Manager Version	Supported CDH Versions	Integrated into CDH
Kafka 2.0.x	Enhanced security	Cloudera Manager 5.5.3	CDH 5.4.x and higher	No
Kafka 1.4.x	Distributed both as package and parcel	Cloudera Manager 5.2.x	CDH 5.4.x and higher	No
Kafka 1.3.x	Includes Kafka Monitoring	Cloudera Manager 5.2.x	CDH 5.4.x and higher	No
Kafka 1.2.x		Cloudera Manager 5.2.x	CDH 5.4.x and higher	No

Product Compatibility Matrix for Cloudera Navigator

This matrix contains compatibility information across versions of Cloudera Navigator, Cloudera Manager, and CDH. For detailed information on each release, see [Cloudera Navigator documentation](#).



Note: Cloudera Navigator requires HiveServer2 for complete governance Hive queries. Cloudera Navigator does not capture audit events for queries that are run on HiveServer1/Hive CLI, and lineage is not captured for certain types of operations that are run on HiveServer1.

[HiveServer1 and Hive CLI are deprecated in CDH 5](#). If you use Cloudera Navigator to capture auditing, lineage, and metadata for Hive operations, upgrade to HiveServer2 if you have not done so already.

CDH 5 and Cloudera Manager 5 Requirements and Supported Versions

Product	Feature	Lowest supported Cloudera Manager Version	Lowest supported CDH Version	Lowest supported Impala Version	Lowest supported Search Version
Cloudera Navigator 2.8.x	Auditing, Metadata, Analytics, and Security	5.9.0	<ul style="list-style-type: none"> • Audit Component <ul style="list-style-type: none"> – HDFS, HBase - 4.0.0 – Hue - 4.4.0, 5.5.0 for LDAP user operations – HiveServer2 - 4.2.0, 4.4.0 for operations denied due to lack of privileges. – Sentry - 5.1.0 • Metadata Component <ul style="list-style-type: none"> – HDFS, HiveServer2, MapReduce, Oozie, Sqoop 1 - 4.4.0 – Pig - 4.6.0 – YARN - 5.0.0 – Impala - 5.4.0 – Spark (available but unsupported) - 5.4.0 	Impala 1.2.1 with CDH 4.4.0	CDH 5.4.0
Cloudera Navigator 2.7.x	Auditing, Metadata, Analytics, and Security	5.8.0	<ul style="list-style-type: none"> • Audit Component <ul style="list-style-type: none"> – HDFS, HBase - 4.0.0 – Hue - 4.4.0, 5.5.0 for LDAP user operations – Hive - 4.2.0, 4.4.0 for operations 	Impala 1.2.1 with CDH 4.4.0	CDH 5.4.0

Product	Feature	Lowest supported Cloudera Manager Version	Lowest supported CDH Version	Lowest supported Impala Version	Lowest supported Search Version
			<p>denied due to lack of privileges.</p> <ul style="list-style-type: none"> – Sentry - 5.1.0 • Metadata Component <ul style="list-style-type: none"> – HDFS, Hive, MapReduce, Oozie, Sqoop 1 - 4.4.0 – Pig - 4.6.0 – YARN - 5.0.0 – Impala - 5.4.0 – Spark (available but unsupported) - 5.4.0 		
Cloudera Navigator 2.6.x	Auditing, Metadata, Analytics, and Security	5.7.0	<ul style="list-style-type: none"> • Audit Component <ul style="list-style-type: none"> – HDFS, HBase - 4.0.0 – Hue - 4.4.0, 5.5.0 for LDAP user operations – Hive - 4.2.0, 4.4.0 for operations denied due to lack of privileges. – Sentry - 5.1.0 • Metadata Component <ul style="list-style-type: none"> – HDFS, Hive, MapReduce, Oozie, Sqoop 1 - 4.4.0 – Pig - 4.6.0 	Impala 1.2.1 with CDH 4.4.0	CDH 5.4.0

CDH 5 and Cloudera Manager 5 Requirements and Supported Versions

Product	Feature	Lowest supported Cloudera Manager Version	Lowest supported CDH Version	Lowest supported Impala Version	Lowest supported Search Version
			<ul style="list-style-type: none"> – YARN - 5.0.0 – Impala - 5.4.0 – Spark (available but unsupported)- 5.4.0 		
Cloudera Navigator 2.5.x	Auditing, Metadata, Analytics, and Security	5.6.0	<ul style="list-style-type: none"> • Audit Component <ul style="list-style-type: none"> – HDFS, HBase - 4.0.0 – Hue - 4.2.0 – Hive - 4.2.0, 4.4.0 for operations denied due to lack of privileges. – Sentry - 5.1.0 • Metadata Component <ul style="list-style-type: none"> – HDFS, Hive, Impala, MapReduce, Oozie, Sqoop 1 - 4.4.0 – Pig - 4.6.0 – YARN - 5.0.0 – Impala and Spark (available but unsupported)- 5.4.0 	Impala 1.2.1 with CDH 4.4.0	CDH 5.4.0
Cloudera Navigator 2.4.x	Auditing, Metadata, Analytics, and Security	5.5.0	<ul style="list-style-type: none"> • Audit Component <ul style="list-style-type: none"> – HDFS, HBase - 4.0.0 – Hue - 4.2.0 	Impala 1.2.1 with CDH 4.4.0	CDH 5.4.0

Product	Feature	Lowest supported Cloudera Manager Version	Lowest supported CDH Version	Lowest supported Impala Version	Lowest supported Search Version
			<ul style="list-style-type: none"> - Hive - 4.2.0, 4.4.0 for operations denied due to lack of privileges. - Sentry - 5.1.0 • Metadata Component <ul style="list-style-type: none"> - HDFS, Hive, Impala, MapReduce, Oozie, Sqoop 1 - 4.4.0 - Pig - 4.6.0 - YARN - 5.0.0 - Impala and Spark (available but unsupported)- 5.4.0 		
Cloudera Navigator 2.3.x	Auditing, Metadata, and Security	5.4.0	<ul style="list-style-type: none"> • Audit Component <ul style="list-style-type: none"> - HDFS, HBase - 4.0.0 - Hue - 4.2.0 - Hive - 4.2.0, 4.4.0 for operations denied due to lack of privileges. - Sentry - 5.1.0 • Metadata Component <ul style="list-style-type: none"> - HDFS, Hive, Impala, MapReduce, Oozie, Sqoop 1 - 4.4.0 	Impala 1.2.1 with CDH 4.4.0	CDH 5.4.0

CDH 5 and Cloudera Manager 5 Requirements and Supported Versions

Product	Feature	Lowest supported Cloudera Manager Version	Lowest supported CDH Version	Lowest supported Impala Version	Lowest supported Search Version
			<ul style="list-style-type: none"> – Pig - 4.6.0 – YARN - 5.0.0 – Impala and Spark (available but unsupported)- 5.4.0 		
Cloudera Navigator 2.2.x	Auditing, Metadata, and Security	5.3.0	<ul style="list-style-type: none"> • Audit Component <ul style="list-style-type: none"> – HDFS, HBase - 4.0.0 – Hue - 4.2.0 – Hive - 4.2.0, 4.4.0 for operations denied due to lack of privileges. – Sentry - 5.1.0 • Metadata Component <ul style="list-style-type: none"> – HDFS, Hive, Oozie, MapReduce, Sqoop 1 - 4.4.0 – Pig - 4.6.0 – YARN - 5.0.0 	Impala 1.2.1 with CDH 4.4.0	Not Supported
Cloudera Navigator 2.1.x	Auditing, Metadata, and Security	5.2.0	<ul style="list-style-type: none"> • Audit Component <ul style="list-style-type: none"> – HDFS, HBase - 4.0.0 – Hue - 4.2.0 – Hive - 4.2.0, 4.4.0 for operations denied due to lack of privileges. – Sentry - 5.1.0 	Impala 1.2.1 with CDH 4.4.0	Not Supported

Product	Feature	Lowest supported Cloudera Manager Version	Lowest supported CDH Version	Lowest supported Impala Version	Lowest supported Search Version
			<ul style="list-style-type: none"> • Metadata Component <ul style="list-style-type: none"> – HDFS, Hive, Oozie, MapReduce, Sqoop 1 - 4.4.0 – Pig - 4.6.0 – YARN - 5.0.0 		
Cloudera Navigator 2.0.1	Auditing, Metadata, and Security	5.1.2	<ul style="list-style-type: none"> • Audit Component <ul style="list-style-type: none"> – HDFS, HBase - 4.0.0 – Hue - 4.2.0 – Hive - 4.2.0, 4.4.0 for operations denied due to lack of privileges. – Sentry - 5.1.0 • Metadata Component <ul style="list-style-type: none"> – HDFS, Hive, Oozie, MapReduce, Sqoop 1 - 4.4.0 – Pig - 4.6.0 – YARN - 5.0.0 	Impala 1.2.1 with CDH 4.4.0	Not Supported
Cloudera Navigator 2.0.0	Auditing, Metadata, and Security	5.1.0	<ul style="list-style-type: none"> • Audit Component <ul style="list-style-type: none"> – HDFS, HBase - 4.0.0 – Hue - 4.2.0 – Hive - 4.2.0, 4.4.0 for operations denied due to lack of privileges. 	Impala 1.2.1 with CDH 4.4.0	Not Supported

CDH 5 and Cloudera Manager 5 Requirements and Supported Versions

Product	Feature	Lowest supported Cloudera Manager Version	Lowest supported CDH Version	Lowest supported Impala Version	Lowest supported Search Version
			<ul style="list-style-type: none"> - Sentry - 5.1.0 • Metadata Component <ul style="list-style-type: none"> - HDFS, Hive, Oozie, MapReduce, Sqoop 1 - 4.4.0 - Pig - 4.6.0 - YARN - 5.0.0 		
Cloudera Navigator 1.2.x	Auditing	5.0.0	<ul style="list-style-type: none"> • HDFS, HBase - 4.0.0 • Hue - 4.2.0 • Hive - 4.2.0, 4.4.0 for operations denied due to lack of privileges. 	Impala 1.2.1 with CDH 4.4.0	Not Supported
	Metadata (2.0 beta 2)	5.0.0	<ul style="list-style-type: none"> • HDFS, Hive, Oozie, MapReduce, Sqoop 1 - 4.4.0 • Pig - 4.6.0 	Not Supported	Not Supported

Cloudera Navigator Supported Databases

Cloudera Navigator supports the following databases:

- Maria DB 5.5, 10
- MySQL 5.1, 5.5, and 5.6
- Oracle 11gR2 and 12c

Cloudera Navigator Supported Browsers

The Cloudera Navigator UI supports the following browsers:

- Mozilla Firefox 24 and higher
- Google Chrome 36 and higher
- Internet Explorer 11
- Safari 5 and higher

Cloudera Navigator Supported CDH and Managed Service Versions

This section describes the CDH and managed service versions supported by the Cloudera Navigator auditing and metadata features.

Cloudera Navigator Auditing

This section describes the audited operations and service versions supported by Cloudera Navigator auditing.

Component	Operations (For details, see Cloudera Navigator Auditing).	Minimum Supported Service Version
HDFS	<ul style="list-style-type: none"> Operations that access or modify a file's or directory's data or metadata Operations denied due to lack of privileges 	CDH 4.0.0
HBase	<ul style="list-style-type: none"> In CDH versions less than 4.2.0, for grant and revoke operations, the operation in log events is <code>ADMIN</code> In simple authentication mode, if the HBase Secure RPC Engine property is <code>false</code> (the default), the username in log events is <code>UNKNOWN</code>. To see a meaningful username: <ol style="list-style-type: none"> Click the HBase service. Click the Configuration tab. Select Service-wide > Security. Set the HBase Secure RPC Engine property to <code>true</code>. Save the change and restart the service. 	CDH 4.0.0
Hive	<ul style="list-style-type: none"> Operations (except grant, revoke, and metadata access only) sent to HiveServer2 Operations denied due to lack of privileges <p>Limitations:</p> <ul style="list-style-type: none"> Actions taken against Hive using the Hive CLI are <i>not</i> audited. Therefore if you have enabled auditing you should disable the Hive CLI to prevent actions against Hive that are not audited. In simple authentication mode, the username in log events is the username passed in the HiveServer2 connect command. If you do not pass a username in the connect command, the username in log events is anonymous. 	CDH 4.2.0, CDH 4.4.0 for operations denied due to lack of privileges.
Hue	<ul style="list-style-type: none"> Operations (except grant, revoke, and metadata access only) sent through the Beeswax Server 	CDH 4.4.0
	<ul style="list-style-type: none"> User operations such as log in, log out, add and remove user, add and remove LDAP group, add and remove user from LDAP group 	CDH 5.5.0
Impala	<ul style="list-style-type: none"> Queries denied due to lack of privileges Queries that pass analysis 	Impala 1.2.1 with CDH 4.4.0
Navigator Metadata Server	<ul style="list-style-type: none"> Viewing and changing audit reports Viewing and changing authorization configurations Viewing and changing metadata Viewing and changing policies Viewing and changing saved searches 	Cloudera Navigator 2.3
Sentry	<ul style="list-style-type: none"> Operations sent to the HiveServer2 and Hive Metastore Server roles and Impala service Adding and deleting roles, assigning roles to groups and removing roles from groups, creating and deleting privileges, granting and revoking privileges Operations denied due to lack of privileges 	CDH 5.1.0

Component	Operations (For details, see Cloudera Navigator Auditing).	Minimum Supported Service Version
	You do not directly configure the Sentry service for auditing. Instead, when you configure the Hive and Impala services for auditing, grant, revoke, and metadata operations appear in the Hive or Impala service audit logs.	
Solr	<ul style="list-style-type: none"> Index creation and deletion Schema and configuration file modification Index, service, document tag access 	CDH 5.4.0

Cloudera Navigator Metadata

This section describes the CDH and managed service versions supported by the Cloudera Navigator metadata feature.

Component	Minimum Supported Version
HDFS. However, federated HDFS is <i>not supported</i> .	CDH 4.4.0
Hive	CDH 4.4.0
Impala	CDH 5.4.0
MapReduce	CDH 4.4.0
Oozie. Supported actions:	CDH 4.4.0
<ul style="list-style-type: none"> 2.4 - map-reduce, pig, hive, hive2, sqoop 2.3 and lower - map-reduce, pig, hive, sqoop 	
Pig	CDH 4.6.0
Spark	CDH 5.4.0
<div style="border: 1px solid #ccc; padding: 10px;"> Important: Spark metadata and lineage is not supported or recommended for production use. By default it is disabled. To try this feature, use it in a test environment until Cloudera resolves currently existing issues and limitations to make it ready for production use. </div>	
Sqoop 1. All Cloudera connectors are supported.	CDH 4.4.0
YARN	CDH 5.0.0

Product Compatibility Matrix for Cloudera Navigator Encryption

Cloudera Navigator encryption comprises several components.

Although the version numbers differ, Cloudera Navigator encryption component releases are generally coordinated with Cloudera Enterprise. See the following table for information on which component versions correspond with a given Cloudera Enterprise release:

Table 6: Cloudera Navigator Encryption Components Release Versions

Cloudera Enterprise Release	Cloudera Navigator Key Trustee Server Version	Key Trustee KMS Version	Cloudera Navigator Key HSM Version	Cloudera Navigator Encrypt Version
5.9.x	5.9.x	5.9.x	1.8.x	<i>None</i>
5.8.x	5.8.x	5.8.x	1.7.x	3.10.x
5.7.x	5.7.x	5.7.x	1.6.x	3.9.x
5.6.x	<i>None</i>	<i>None</i>	<i>None</i>	<i>None</i>
5.5.x	5.5.x	5.5.x	1.5.x	3.8.x
5.4.x	5.4.x	5.4.x	1.4.x	3.7.x
5.3.x	3.8.x	5.3.x	1.3.x	3.6.x

See below for the individual compatibility matrices for each component:

Cloudera Navigator Key Trustee Server

Because of a change in the ports used by Key Trustee Server, Navigator Encrypt versions lower than 3.7 and Key Trustee KMS versions lower than 5.4 are not supported in Key Trustee Server 5.4 and higher.

Table 7: Cloudera Navigator Key Trustee Server Compatibility Matrix

Cloudera Navigator Key Trustee Server Version	Supported Operating Systems	Lowest Supported Cloudera Manager Version	Lowest Supported Cloudera Navigator Key HSM Versions	Supported Key Trustee KMS Versions	Supported Cloudera Navigator Encrypt Versions
5.9.x	<ul style="list-style-type: none"> • RHEL and CentOS: 6.4, 6.5, 6.6, 6.7, 6.8, 7.1, 7.2 	5.9.x	1.3.x	5.4.x, 5.5.x, 5.7.x, 5.8.x, 5.9.x	3.7.x, 3.8.x, 3.9.x, 3.10.x
5.8.x	<ul style="list-style-type: none"> • RHEL and CentOS: 6.4, 6.5, 6.6, 6.7, 7.1, 7.2 	5.8.x	1.3.x	5.4.x, 5.5.x, 5.7.x, 5.8.x	3.7.x, 3.8.x, 3.9.x, 3.10.x
5.7.x	<ul style="list-style-type: none"> • RHEL and CentOS: 6.4, 6.5, 6.6, 6.7, 7.1, 7.2 	5.7.x	1.3.x	5.4.x, 5.5.x, 5.7.x	3.7.x, 3.8.x, 3.9.x, 3.10.x
5.5.x	<ul style="list-style-type: none"> • RHEL and CentOS: 6.4, 6.5, 6.6, 6.7, 7.1 	5.5.x	1.3.x	5.4.x, 5.5.x, 5.7.x	3.7.x, 3.8.x, 3.9.x, 3.10.x
5.4.x	<ul style="list-style-type: none"> • RHEL and CentOS: 6.4, 6.5, 6.6 	5.4.x	1.3.x	5.4.x, 5.5.x, 5.7.x	3.7.x, 3.8.x, 3.9.x, 3.10.x
3.8.x	<ul style="list-style-type: none"> • RHEL and CentOS: 6.4, 6.5 	Not supported	1.3.x	5.3.x	3.6.x

Key Trustee KMS

Table 8: Key Trustee KMS Compatibility Matrix

Key Trustee KMS Version	Supported Operating Systems	Supported Key Trustee Server Versions	Lowest Supported Cloudera Manager Version	Supported CDH Versions
5.9.x	<ul style="list-style-type: none"> RHEL and CentOS: 5.7, 5.10, 6.4, 6.5, 6.6, 6.7, 6.8, 7.1, 7.2 Oracle Enterprise Linux: 5.7, 5.10, 5.11, 6.4, 6.5, 6.6, 6.7, 6.8, 7.1, 7.2 SLES: 11 SP2, 11 SP3, 11 SP4, 12 SP1 Debian: 7.1, 7.8, 8.2, 8.4 Ubuntu: 12.04, 14.04 	5.4.x, 5.5.x, 5.7.x, 5.8.x, 5.9.x	5.9.x	5.3.x, 5.4.x, 5.5.x, 5.7.x, 5.8.x, 5.9.x
5.8.x	<ul style="list-style-type: none"> RHEL and CentOS: 5.7, 5.10, 6.4, 6.5, 6.6, 6.7, 7.1, 7.2 Oracle Enterprise Linux: 5.7, 5.10, 5.11, 6.4, 6.5, 6.6, 6.7, 7.1, 7.2 SLES: 11 SP2, 11 SP3, 11 SP4 Debian: 7.1, 7.8, 8.2 Ubuntu: 12.04, 14.04 	5.4.x, 5.5.x, 5.7.x, 5.8.x	5.8.x	5.3.x, 5.4.x, 5.5.x, 5.7.x, 5.8.x
5.7.x	<ul style="list-style-type: none"> RHEL and CentOS: 5.7, 5.10, 6.4, 6.5, 6.6, 6.7, 7.1, 7.2 Oracle Enterprise Linux: 5.7, 5.10, 5.11, 6.4, 6.5, 6.6, 6.7, 7.1, 7.2 SLES: 11 SP2, 11 SP3, 11 SP4 Debian: 7.1, 7.8 Ubuntu: 12.04, 14.04 	5.4.x, 5.5.x, 5.7.x	5.7.x	5.3.x, 5.4.x, 5.5.x, 5.7.x
5.5.x	<ul style="list-style-type: none"> RHEL and CentOS: 5.7, 	5.4.x, 5.5.x	5.5.x	5.3.x, 5.4.x, 5.5.x

Key Trustee KMS Version	Supported Operating Systems	Supported Key Trustee Server Versions	Lowest Supported Cloudera Manager Version	Supported CDH Versions
	<p>5.10, 6.4, 6.5, 6.6, 6.7, 7.1</p> <ul style="list-style-type: none"> Oracle Enterprise Linux: 5.7, 5.10, 6.4, 6.5, 6.6, 6.7, 7.1 SLES: 11 SP2, 11 SP3 Debian: 7.1 Ubuntu: 12.04, 14.04 			
5.4.x	<ul style="list-style-type: none"> RHEL and CentOS: 5.7, 5.10, 6.4, 6.5, 6.6 Oracle Enterprise Linux: 5.7, 5.10, 6.4, 6.5, 6.6 SLES: 11 SP2, 11 SP3 Debian: 7.1 Ubuntu: 12.04, 14.04 	5.4.x	5.4.x	5.3.x, 5.4.x
5.3.x	<ul style="list-style-type: none"> RHEL and CentOS: 5.7, 6.4, 6.5 Oracle Enterprise Linux: 5.7, 6.4, 6.5 SLES: 11 SP2, 11 SP3 Debian: 7.1 Ubuntu: 12.04, 14.04 	3.8.x	5.3.x	5.3.x

Cloudera Navigator Key HSM

Cloudera Navigator Key HSM must be installed on the same host as Key Trustee Server. Although Key HSM is compatible across all versions of Key Trustee Server, Cloudera strongly recommends also upgrading Key HSM after you upgrade Key Trustee Server.

Table 9: Cloudera Navigator Key HSM Compatibility Matrix

Cloudera Navigator Key HSM Version	Supported Operating Systems	Lowest Supported Key Trustee Server Version
1.8.x	<ul style="list-style-type: none"> RHEL and CentOS: 6.4, 6.5, 6.6, 6.7, 6.8, 7.1, 7.2 	3.8.x
1.7.x	<ul style="list-style-type: none"> RHEL and CentOS: 6.4, 6.5, 6.6, 6.7, 7.1, 7.2 	3.8.x

Cloudera Navigator Key HSM Version	Supported Operating Systems	Lowest Supported Key Trustee Server Version
1.6.x	<ul style="list-style-type: none"> RHEL and CentOS: 6.4, 6.5, 6.6, 6.7, 7.1, 7.2 	3.8.x
1.5.x	<ul style="list-style-type: none"> RHEL and CentOS: 6.4, 6.5, 6.6, 6.7, 7.1 	3.8.x
1.4.x	<ul style="list-style-type: none"> RHEL and CentOS: 6.4, 6.5, 6.6 	3.8.x
1.3.x	<ul style="list-style-type: none"> RHEL and CentOS: 6.4, 6.5 	3.8.x

Cloudera Navigator Encrypt

Table 10: Cloudera Navigator Encrypt Compatibility Matrix

Cloudera Navigator Encrypt Version	Supported Operating Systems	Supported Key Trustee Server Versions
3.10.x	<ul style="list-style-type: none"> RHEL and CentOS: 5.7, 5.10, 6.4, 6.5, 6.6, 6.7, 7.1, 7.2 Oracle Enterprise Linux: 6.4, 6.5, 6.6, 6.7, 7.1 SLES: 11 SP2, 11 SP3, 11 SP4 Debian: 7.1, 7.8 Ubuntu: 12.04, 14.04, 14.04.3 	5.4.x, 5.5.x, 5.7.x, 5.8.x
3.9.x	<ul style="list-style-type: none"> RHEL and CentOS: 5.7, 5.10, 6.4, 6.5, 6.6, 6.7, 7.1, 7.2 Oracle Enterprise Linux: 6.4, 6.5, 6.6, 6.7, 7.1 SLES: 11 SP2, 11 SP3, 11 SP4 Debian: 7.1, 7.8 Ubuntu: 12.04, 14.04, 14.04.3 	5.4.x, 5.5.x, 5.7.x
3.8.x	<ul style="list-style-type: none"> RHEL and CentOS: 5.7, 5.10, 6.4, 6.5, 6.6, 6.7, 7.1 Oracle Enterprise Linux: 6.4, 6.5, 6.6, 6.7, 7.1 SLES: 11 SP2, 11 SP3 Debian: 7.1 Ubuntu: 12.04, 14.04 	5.4.x, 5.5.x, 5.7.x
3.7.x	<ul style="list-style-type: none"> RHEL and CentOS: 5.7, 5.10, 6.4, 6.5, 6.6 Oracle Enterprise Linux: 6.4, 6.5, 6.6 SLES: 11 SP2, 11 SP3 Debian: 7.1 Ubuntu: 12.04, 14.04 	5.4.x, 5.5.x, 5.7.x
3.6.x	<ul style="list-style-type: none"> RHEL and CentOS: 5.7, 6.4, 6.5 Oracle Enterprise Linux: 6.4, 6.5 SLES: 11 SP2, 11 SP3 	3.8.x

Cloudera Navigator Encrypt Version	Supported Operating Systems	Supported Key Trustee Server Versions
	<ul style="list-style-type: none"> • Debian: 7.1 • Ubuntu: 12.04, 14.04 	

Product Compatibility Matrix for Apache Sentry

Sentry enables role-based, fine-grained authorization for HiveServer2 and provides classic database-style authorization for Hive, Cloudera Impala and Cloudera Search. You can use either the Sentry service (introduced in Cloudera Manager 5.1.0 and CDH 5.1.0) or the policy file approach to secure your data.



Note: It's possible for a single cluster to use both the Sentry service (for Hive and Impala) and Sentry policy files (for Solr).

Product	Feature	Lowest supported Cloudera Manager Version	Lowest supported CDH Version	Lowest supported Impala Version	Lowest supported Search Version	Integrated into CDH
Apache Sentry	Sentry Service	Cloudera Manager 5.1.0	CDH 5.1.0	Impala 1.4.0 for CDH 5	Search for CDH 5.8.0	Yes
	Policy File	Cloudera Manager 4.7.0	CDH 4.3.0	Impala 1.2.1	Search 1.1.0	Yes; Starting CDH 4.4.0

Support for Other Filesystems

Filesystems	Lowest Release Supported
Amazon S3 (s3a)	Apache Sentry for CDH 5.8
Amazon RDS	Apache Sentry for CDH 5.9

Product Compatibility Matrix for Apache Spark

Spark is a fast, general engine for large-scale data processing. For installation and configuration instructions, see . To see new features introduced with each release, refer to the [CDH 5 Release Notes](#) on page 14.

Since Spark has been integrated into the CDH package, its compatibility with Cloudera Manager and CDH depends on the CDH 5.x.x release it is shipped with.

Product	Lowest Supported Cloudera Manager Version	Lowest Supported CDH Version	Lowest supported Impala Version	Lowest supported Search Version	Integrated into CDH
Apache Spark 1.6.x	Cloudera Manager 5.7.x	CDH 5.7.x	Not Supported	Not Supported	CDH 5.7.0
Apache Spark 1.5.x	Cloudera Manager 5.5.x and Cloudera Manager 5.6.x	CDH 5.5.x and CDH 5.6.x	Not Supported	Not Supported	CDH 5.5.0 and CDH 5.6.0

CDH 5 and Cloudera Manager 5 Requirements and Supported Versions

Product	Lowest Supported Cloudera Manager Version	Lowest Supported CDH Version	Lowest supported Impala Version	Lowest supported Search Version	Integrated into CDH
Apache Spark 1.3.x	Cloudera Manager 5.4.x	CDH 5.4.x	Not Supported	Not Supported	CDH 5.4.0
Apache Spark 1.2.x	Cloudera Manager 5.3.x	CDH 5.3.x	Not Supported	Not Supported	CDH 5.3.0
Apache Spark 1.1.x	Cloudera Manager 5.2.x	CDH 5.2.x	Not Supported	Not Supported	CDH 5.2.0
Apache Spark 1.0.x	Cloudera Manager 5.1.x	CDH 5.1.x	Not Supported	Not Supported	CDH 5.1.0
Apache Spark 0.9.x	Cloudera Manager 5.0.x	CDH 5.0.x	Not Supported	Not Supported	CDH 5.0.0

Product Compatibility Matrix for EMC DSSD D5

This matrix contains compatibility information for clusters that use the EMC DSSD D5 storage appliance as the storage for DataNodes.

Cloudera Manager Version	DSSD Hadoop Plugin Version	CDH Version	Operating System Support for DataNodes
5.9	1.3	5.9	<ul style="list-style-type: none"> • RHEL 6.6 • RHEL 7.1 • RHEL 7.2
5.8	1.2	5.8	<ul style="list-style-type: none"> • RHEL 6.6 • RHEL 7.1 • RHEL 7.2

Product Compatibility for EMC Isilon

CDH 5.4.4 and higher is compatible with Isilon OneFS version 7.2.0.3 and higher. Cloudera recommends Isilon OneFS version 7.2.1.1. See [Using CDH with Isilon Storage](#).

For compatibility with Cloudera Manager Replication and Snapshot features, see [Product Compatibility Matrix for Backup and Disaster Recovery](#) on page 537.

Table 11:

CDH Version	Isilon OneFS Version
5.8	8.0
5.4.4 and higher	7.2.0.3 and higher Cloudera recommends 7.2.1.1

Product Compatibility Matrix for Backup and Disaster Recovery

This matrix contains compatibility information across features of Cloudera Manager Backup and Disaster Recovery and CDH and Cloudera Manager.

Product	Feature	Lowest supported Cloudera Manager Version	Lowest supported CDH Version	Lowest supported Impala Version	Lowest supported Search Version	Integrated into CDH
Backup & Disaster Recovery	Replication	Cloudera Manager 4.5.0	CDH 4.0.1	Impala 1.0.x	Not Supported	No
	Snapshots	Cloudera Manager 5.0.0	CDH 5.0.0	Not Supported	Not Supported	No
	Snapshots from Isilon storage	Not Supported	Not Supported	Not Supported	Not Supported	n/a

The tables below list the supported and unsupported replication scenarios:

Supported Replication Scenarios

In Cloudera Manager 5, replication is supported between CDH 5 or CDH 4 clusters. The following tables describe support for HDFS and Hive replication.

Service	Source			Destination		
	Cloudera Manager Version	CDH Version	Comment	Cloudera Manager Version	CDH Version	Comment
HDFS, Hive	4	4		5	4	
HDFS, Hive	4	4.4 or higher		5	5	
HDFS, Hive	4 or 5	5	SSL enabled on Hadoop services	4 or 5	5	SSL enabled on Hadoop services
HDFS, Hive	4 or 5	5	SSL enabled on Hadoop services	4 or 5	5	SSL <i>not</i> enabled on Hadoop services
HDFS, Hive	4 or 5	5.1	SSL enabled on Hadoop services and YARN	4 or 5	4 or 5	
HDFS, Hive	5	4		4.7.3 or higher	4	
HDFS, Hive	5	4		5	4	
HDFS, Hive	5	5		5	5	
HDFS, Hive	5	5		5	4.4 or higher	

CDH 5 and Cloudera Manager 5 Requirements and Supported Versions

Service	Source			Destination		
	Cloudera Manager Version	CDH Version	Comment	Cloudera Manager Version	CDH Version	Comment
HDFS, Hive	5.7	5.7, with Isilon storage	See Supported Replication Scenarios for Clusters using Isilon Storage on page 540.	5.7	5.7, with Isilon storage	See Supported Replication Scenarios for Clusters using Isilon Storage on page 540.
HDFS, Hive	5.7	5.7		5.7	5.7, with Isilon storage	
HDFS, Hive	5.7	5.7, with Isilon storage		5.7	5.7	
HDFS, Hive	5.8	5.8, with or without Isilon Storage	<ul style="list-style-type: none"> Kerberos can be enabled See Supported Replication Scenarios for Clusters using Isilon Storage on page 540. 	5.8	5.8, with or without Isilon Storage	<ul style="list-style-type: none"> Kerberos can be enabled See Supported Replication Scenarios for Clusters using Isilon Storage on page 540.

Unsupported Replication Scenarios



Note: If you are using Isilon storage for CDH, see [Supported Replication Scenarios for Clusters using Isilon Storage](#) on page 540.

Service	Source			Destination		
	Cloudera Manager Version	CDH Version	Comment	Cloudera Manager Version	CDH Version	Comment
Any	4 or 5	4 or 5	Kerberos enabled.	4 or 5	4 or 5	Kerberos not enabled
Any	4 or 5	4 or 5	Kerberos not enabled.	4 or 5	4 or 5	Kerberos enabled
HDFS, Hive	4 or 5	4	Where the replicated data includes a directory that contains a large number of files or subdirectories (several hundred thousand entries), causing out-of-memory errors. To work around this issue, follow this procedure .	4 or 5	5	
Hive	4 or 5	4	Replicate HDFS Files is disabled.	4 or 5	4 or 5	Over-the-wire encryption is enabled.
Hive	4 or 5	4	Replication can fail if the NameNode fails over during replication.	4 or 5	5, with high availability enabled	Replication can fail if the NameNode fails over during replication.

Service	Source			Destination		
	Cloudera Manager Version	CDH Version	Comment	Cloudera Manager Version	CDH Version	Comment
Hive	4 or 5	4	The clusters use different Kerberos realms.	4 or 5	5	An older JDK is deployed. (Upgrade the CDH 4 cluster to use JDK 7 or JDK6u34 to work around this issue.)
Any	4 or 5	4	SSL enabled on Hadoop services.	4 or 5	4 or 5	
Hive	4 or 5	4.2 or higher	If the Hive schema contain views.	4 or 5	4	
HDFS	4 or 5	4, with high availability enabled	Replications fail if NameNode failover occurs during replication.	4 or 5	5, without high availability	Replications fail if NameNode failover occurs during replication.
HDFS	4 or 5	4 or 5	Over the wire encryption is enabled.	4 or 5	4	
HDFS	4 or 5	5	Clusters where there are URL-encoding characters such as % in file and directory names.	4 or 5	4	
Hive	4 or 5	4 or 5	Over the wire encryption is enabled and Replicate HDFS Files is enabled.	4 or 5	4	
Hive	4 or 5	4 or 5	From one cluster to the same cluster.	4 or 5	4 or 5	From one cluster to the same cluster.
HDFS, Hive	4 or 5	5	Where the replicated data includes a directory that contains a large number of files or subdirectories (several hundred thousand entries), causing out-of-memory errors. To work around this issue, follow this procedure .	4 or 5	4	
HDFS	4 or 5	5	The clusters use different Kerberos realms.	4 or 5	4	An older JDK is deployed. (Upgrade the CDH 4 cluster to use JDK 7 or JDK6u34 to work around this issue.)
Hive	4 or 5	5	Replicate HDFS Files is enabled and the clusters use different Kerberos realms.	4 or 5	4	An older JDK is deployed. (Upgrade the CDH 4 cluster to use JDK 7 or JDK6u34 to work around this issue.)
Any	4 or 5	5	SSL enabled on Hadoop services and YARN.	4 or 5	4 or 5	

CDH 5 and Cloudera Manager 5 Requirements and Supported Versions

Service	Source			Destination		
	Cloudera Manager Version	CDH Version	Comment	Cloudera Manager Version	CDH Version	Comment
Any	4 or 5	5	SSL enabled on Hadoop services.	4 or 5	4	
HDFS	4 or 5	5, with high availability enabled	Replications fail if NameNode failover occurs during replication.	4 or 5	4, without high availability	Replications fail if NameNode failover occurs during replication.
HDFS, Hive	5	5		4	4	
Hive	5.2	5.2 or lower	Replication of Impala UDFs is skipped.	4 or 5	4 or 5	

Supported Replication Scenarios for Clusters using Isilon Storage

Note the following when scheduling replication jobs for clusters that use Isilon storage:

- As of CDH 5.8 and higher, Replication is supported for clusters using Kerberos and Isilon storage on the source or destination cluster, or both. See [Configuring Replication with Kerberos and Isilon](#). Replication between clusters using Isilon storage and Kerberos is not supported in CDH 5.7.
- Make sure that the `hdfs` user is a superuser in the Isilon system. If you specify alternate users with the **Run As** option when creating replication schedules, those users must also be superusers.
- Cloudera recommends that you use the Isilon `rroot` user for replication jobs. (Specify `rroot` in the **Run As** field when creating replication schedules.)
- Select the **Skip checksum checks** property when creating replication schedules.
- Clusters that use Isilon storage do not support [snapshots](#). Snapshots are used to ensure data consistency during replications in scenarios where the source files are being modified. Therefore, when replicating from an Isilon cluster, Cloudera recommends that you do not replicate Hive tables or HDFS files that could be modified before the replication completes.

See [Using CDH with Isilon Storage](#).

Supported Configurations with Virtualization and Cloud Platforms

This section lists supported configurations for deploying Cloudera software on virtualization and cloud platforms, and provides links to reference architectures for these platforms.

Amazon Web Services

For information on deploying Cloudera software on a Amazon Web Services (AWS) cloud infrastructure, see the [Cloudera Enterprise Reference Architecture for AWS Deployments](#).

Google Cloud Platform

For information on deploying Cloudera software on a Google Cloud Platform infrastructure, see the [Cloudera Enterprise Reference Architecture for Google Cloud Platform Deployments](#).

Microsoft Azure

For information on deploying Cloudera software on a Microsoft Azure cloud infrastructure, see the [Cloudera Enterprise Reference Architecture for Azure Deployments](#).

VMware

For information on deploying Cloudera software on a VMware-based infrastructure, see the [Reference architecture for deploying on VMware](#).

Recommendation when deploying on VMware in the current release:

- Use the part of Hadoop Virtual Extensions that has been implemented in [HADOOP-8468](#). This will prevent data loss when a physical node that hosts two or more DataNodes goes down .

Deprecated Items

The section lists OSs, Java versions, databases, platforms, CDH components and subcomponents, and product functionality that have been deprecated.

Terminology

Deprecated

Feature, component, platform, or functionality that Cloudera is planning to remove in a future release. Cloudera supports items that are deprecated until they are removed, and the deprecation gives customers time to plan for removal.

Removed

Feature, component, platform, or functionality that has been removed from the product and is no longer supported.

OSs, Java Versions, Databases, and Platforms

Item	Related Information	Release in Which Item Is Deprecated	Release in Which Item Is Removed
Operating System <ul style="list-style-type: none"> Ubuntu 14.04 	CDH and Cloudera Manager Supported Operating Systems on page 505	5.8.0	6.0.0
Operating System <ul style="list-style-type: none"> RHEL 5, CentOS 5, Oracle Enterprise Linux 5 Ubuntu 10.04 (already EOL by Canonical) and 12.04 SLES 11 Debian 7 	CDH and Cloudera Manager Supported Operating Systems on page 505	August 2015	6.0.0
Java 7	CDH and Cloudera Manager Supported JDK Versions on page 513	5.0.0	6.0.0
Database <ul style="list-style-type: none"> MySQL 5.0 PostgreSQL 8.1 		August 2015	6.0.0
Database <ul style="list-style-type: none"> MySQL 5.1 PostgreSQL 8.14 	CDH and Cloudera Manager Supported Databases on page 511	5.9.0	6.0.0
Database <ul style="list-style-type: none"> Oracle 11g 		5.7.0	6.0.0
Filesystem Amazon S3 and S3n connectors. S3 and S3n are replaced by S3a.	Storing HBase Snapshots on Amazon S3 and Copying Data Between Two Clusters Using Distcp	5.5.0	5.7.0

Item	Related Information	Release in Which Item Is Deprecated	Release in Which Item Is Removed
Tarball CDH Tarball Distribution		5.9.0	6.0.0
Cloudera Manager Tarball Distribution		5.9.0	6.0.0

CDH Components, Subcomponents, and Product Functionality

Item	Related Information	Release in Which Item Is Deprecated	Release in Which Item Is Removed
Activity Monitor	Activity Monitor is only used and deployed by Cloudera Manager when the MapReduce service (MRv1) is deployed.	5.9.0	7.0.0
CapacityScheduler	Use FairScheduler	5.8.0	6.0.0
DataFu		5.9.0	TBD
Hive CLI	About Hive	5.0.0	6.0.0
HiveServer1	Starting HiveServer1 and the Hive Console	5.3.0	6.0.0
Llama		5.5.0	6.0.0
MRv1, MapReduce v1 APIs, MapReduce service	Managing YARN (MRv2) and MapReduce (MRv1)	5.0.0	6.0.0
Mahout	Mahout Installation	5.5.0	6.0.0
Management of Key Trustee Server without Cloudera Manager	Cloudera Navigator Key Trustee Server	5.9.0	6.0.0
MR Pipes		5.9.0	6.0.0
Navigator Encrypt Filesystem-Level Encryption Using eCryptfs	Filesystem-Level Encryption with eCryptfs	August 2015	6.0.0
Old NameNode UI		5.5.0	6.0.0
Oozie Hive action	Adding Schema to Oozie Using Cloudera Manager	5.7.0	6.0.0
Parquet libraries named com.twitter.*	Using Apache Parquet Data Files with CDH	5.8.0	6.0.0
Sentry policy files	Migrating from Sentry Policy Files to the Sentry Service	5.8.0	6.0.0
Spark Standalone	Managing Spark Standalone Using the Command Line	5.5.0	6.0.0
Whirr	Whirr Installation	5.5.0	6.0.0
YARN Capacity Scheduler	Configuring the YARN Scheduler	5.9.0	6.0.0