Hortonworks Data Platform

HDP-2.4.2 Release Notes

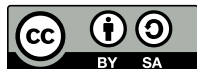## Hortonworks Data Platform: HDP-2.4.2 Release Notes
Copyright © 2012-2016 Hortonworks, Inc. Some rights reserved.

The Hortonworks Data Platform, powered by Apache Hadoop, is a massively scalable and 100% open source platform for storing, processing and analyzing large volumes of data. It is designed to deal with data from many sources and formats in a very quick, easy and cost-effective manner.

The Hortonworks Data Platform consists of the essential set of Apache Software Foundation projects that focus on the storage and processing of Big Data, along with operations, security, and governance for the resulting system. This includes Apache Hadoop – which includes MapReduce, Hadoop Distributed File System (HDFS), and Yet Another Resource Negotiator (YARN) – along with Ambari, Falcon, Flume, HBase, Hive, Kafka, Knox, Oozie, Phoenix, Pig, Ranger, Slider, Spark, Sqoop, Storm, Tez, and ZooKeeper. Hortonworks is the major contributor of code and patches to many of these projects. These projects have been integrated and tested as part of the Hortonworks Data Platform release process and installation and configuration tools have also been included.

Unlike other providers of platforms built using Apache Hadoop, Hortonworks contributes 100% of our code back to the Apache Software Foundation. The Hortonworks Data Platform is Apache-licensed and completely open source. We sell only expert technical support, training and partner-enablement services. All of our technology is, and will remain, free and open source.

Please visit the Hortonworks Data Platform page for more information on Hortonworks technology. For more information on Hortonworks services, please visit either the Support or Training page. Feel free to contact us directly to discuss your specific needs.

Table of Contents

List of Tables

1. HDP 2.4.2 Release Notes

This document provides you with the latest information about the HDP 2.4.2 release and its product documentation.

Component Versions

The official Apache versions of most HDP 2.4.2 components are unchanged from HDP 2.4.0.0, with the exception of Spark and Kafka. Spark is upgraded from 1.6.0 to 1.6.1; Kafka is upgraded from 0.9.0 to 0.9.0.1. See more details in the New Features section. All HDP components listed here are official Apache releases of the most recent stable versions available.

Hortonworks' philosophy is to provide patches only when absolutely necessary to assure the interoperability of the components. Unless you are explicitly directed by Hortonworks Support to take a patch update, each of the HDP components should remain at the following package version levels to ensure a certified and supported copy of HDP 2.4.2.

Official Apache versions for HDP 2.4.2.

• Apache Accumulo 1.7.0

• Apache Atlas 0.5.0

• Apache Calcite 1.2.0

• Apache DataFu 1.3.0

• Apache Falcon 0.6.1

• Apache Flume 1.5.2

• Apache Hadoop 2.7.1

• Apache HBase 1.1.2

• Apache Hive 1.2.1

• Apache Kafka 0.9.0.1

• Apache Knox 0.6.0

• Apache Mahout 0.9.0+

• Apache Oozie 4.2.0

• Apache Phoenix 4.4.0

• Apache Pig 0.15.0

• Apache Ranger 0.5.0

- Apache Slider 0.80.0

- Apache Solr 5.2.1

- Apache Spark 1.6.1

- Apache Sqoop 1.4.6

- Apache Storm 0.10.0

- Apache Tez 0.7.0

- Apache ZooKeeper 3.4.6

Additional component versions:

- Cascading 3.0.1

- Hue 2.6.1

1.1. New Features

This section highlights the new feature in HDP 2.4.2.

Table 1.1. New Features

Component	Feature
Kafka	General availability of Kafka 0.9.0.1
	Support for Mirror Maker with a Secure Cluster (Kerberos-enabled)
Spark	General availability of Spark 1.6.1
	ODBC/JDBC support for SparkSQL
	Support for Spark Streaming and Kafka in a Secure Cluster (Kerberos-enabled)
	Oozie Token support for Spark jobs

1.2. Unsupported Features

Some features exist within HDP 2.4.2, but Hortonworks does not currently support these specific capabilities.

- Technical Preview Features

- Community Features

1.2.1. Technical Preview Features

The following features are available within HDP 2.4.2, but are not ready for production deployment. We encourage you to explore these technical preview features in non-

production environments and provide feedback on your experiences through the
Hortonworks Community Forums.

Table 1.2. Technical Previews

Component	Feature
HBase and Phoenix	• Phoenix Query Server • Phoenix Query Server (PHOENIX-971) • Phoenix-Spark Integration • RPC Throttling • Support for `init.d` scripts
Hive	• Hive Streaming • ACID support
Slider	• Support for Docker-based application packaging (SLIDER-780)
Spark	• GraphX • SparkR • Spark-HBase connector
YARN	• Add support for network I/O isolation/scheduling for containers (YARN-2140) • NodeManager: add cgroup support for disk I/O isolation (YARN-2619)
Zeppelin	• Zeppelin 0.6.0 • Install & Manage Zeppelin with Ambari 2.2.2 • PySpark support in Zeppelin - improvements • Zeppelin support for SSL (one way) • Zeppelin support for form-based LDAP authentication

1.2.2. Community Features

The following features are developed and tested by the community, but are not officially
supported by Hortonworks. There are variety of reasons that these features are excluded,
including: insufficient reliability or incomplete test case coverage, declaration of non-
production readiness by the community at large, feature deviates from Hortonworks best
practices, and more. Do not use them in your production environments.

Table 1.3. Community Features

Component	Feature
Falcon	• Prism Server • User Recipes
HBase	• HBase Column Family Encryption: use HDFS data at rest encryption instead • Use of memcached as block cache is unsupported (HBASE-13170) • ZooKeeper-less region assignment

Component	Feature
	• Region size balancing (HBASE-13103)
HDFS	• Block-volume device choosing (HDFS-1804)
	• NameNode Federation (HDFS-1052)
	• viewFS (HADOOP-7257)
Kafka	• New Consumer API
Knox	• Storm REST APIs
Oozie	• Spark action (OOZIE-1983)
Slider	• Simplified Application Packaging
Spark	• Spark Standalone
	• Spark on Mesos
	• Jupyter/iPython Notebook
	• Oozie Spark action is not supported, but there is a tech note available for HDP customers
YARN	• Fair Scheduler
	• MapReduce Eclipse Plug-in
	• MapReduce Uber AM

1.3. Upgrading from HDP 2.4.x to HDP 2.4.2

- Before you begin

- Upgrade Procedure

- Optional: Spark Manual Upgrade Procedure

- Optional: Spark Manual Downgrade Procedure

HDP 2.4.2 is a feature-bearing maintenance release of HDP 2.4.x; it includes changes to HDP 2.4.x beyond a standard maintenance release. These Release Notes include abbreviated upgrade instructions; for full upgrade instructions, use the Non-Ambari Upgrade Guide.

- Keeping the same configuration files you used for HDP 2.4.x

- Keeping the same data and metadata in the same location you used for HDP 2.4.x

- Installing any new components (added for the first time in HDP 2.4.2) side-by-side with existing components

The following table summarizes HDP 2.2.x-to-2.4.2 upgrade options:

Cluster Management	Supporting Doc	
Cluster managed manually (HDP 2.1 and earlier)	Non-Ambari Upgrade Guide	
Cluster managed manually (HDP 2.3 and later)	These HDP Release Notes	
Cluster managed via Ambari 1.7.0	These HDP Release Notes	

Cluster Management	Supporting Doc	
Cluster managed via Ambari 2.0	Upgrading Ambari Guide	Ambari 2.0 supports rolling upgrade between HDP 2.2.x and HDP 2.4.2. **Note** Ambari does not support rolling upgrade between HDP 2.1 and 2.3 for Falcon. Use Configure and Validate Falcon to upgrade this component. When upgrading to HDP 2.4.2 using Ambari, Spark 1.6.0 will be automatically upgraded to 1.6.1. If you wish to return to using 1.6.0, use the Spark Manual Downgrade Procedure.
Cluster managed via Ambari 2.1	Upgrading Ambari Guide	Ambari 2.1 supports rolling upgrade between HDP 2.4.x and HDP 2.4.2. **Note** Ambari does not support rolling upgrade between HDP 2.1 and 2.3 for Falcon. Use Configure and Validate Falcon to upgrade this component. When upgrading to HDP 2.4.2 using Ambari, Spark 1.6.0 will be automatically upgraded to 1.6.1. If you wish to return to using 1.6.0, use the Spark Manual Downgrade Procedure.
Cluster managed via Ambari 2.2	Upgrading Ambari Guide	Ambari 2.2 supports rolling upgrade between HDP 2.4.x and HDP 2.4.2. **Note** Ambari does not support rolling upgrade between HDP 2.1 and 2.3 for Falcon. Use Configure and Validate Falcon to upgrade this component. When upgrading to HDP 2.4.2 using Ambari, Spark 1.6.0 will be automatically upgraded to 1.6.1. If you wish to return to using 1.6.0, use the Spark Manual Downgrade Procedure.

1.3.1. Before you begin

- Make sure you know what HDP components need to be upgraded at your installation.

- Think about whether you are going to upgrade using a local repository or a remote repository.

1.3.2. Upgrade Procedure

To upgrade your cluster from HDP 2.4.x to HDP 2.4.2:

1. **Download the appropriate HDP 2.4.2 hdp.repo file for your OS:**

Operating System	Repository Location
Debian 6	http://public-repo-1.hortonworks.com/HDP/debian6/2.x/updates/2.4.2.0/hdp.list
Debian 7	http://public-repo-1.hortonworks.com/HDP/debian7/2.x/updates/2.4.2.0/hdp.list
RHEL/CentOS/Oracle LINUX 6	http://public-repo-1.hortonworks.com/HDP/centos6/2.x/updates/2.4.2.0/hdp.repo
RHEL/CentOS/Oracle LINUX 7	http://public-repo-1.hortonworks.com/HDP/centos7/2.x/updates/2.4.2.0/hdp.repo
SLES 11 SP3/SP4	http://public-repo-1.hortonworks.com/HDP/suse11sp3/2.x/updates/2.4.2.0/hdp.repo
Ubuntu 12	http://public-repo-1.hortonworks.com/HDP/ubuntu12/2.x/updates/2.4.2.0/hdp.list
Ubuntu 14	http://public-repo-1.hortonworks.com/HDP/ubuntu14/2.x/updates/2.4.2.0/hdp.list

or

Download the HDP RPM single repository tarball. (For information on how to install the repositories, see the local repository instructions.)

Operating System	Tarball Location
Debian 6	http://public-repo-1.hortonworks.com/HDP/debian6/2.x/updates/2.4.2.0/HDP-2.4.2.0-debian6-deb.tar.gz
Debian 7	http://public-repo-1.hortonworks.com/HDP/debian7/2.x/updates/2.4.2.0/HDP-2.4.2.0-debian7-deb.tar.gz
RHEL/CentOS/Oracle LINUX 6	http://public-repo-1.hortonworks.com/HDP/centos6/2.x/updates/2.4.2.0/HDP-2.4.2.0-centos6-rpm.tar.gz
RHEL/CentOS/Oracle LINUX 7	http://public-repo-1.hortonworks.com/HDP/centos7/2.x/updates/2.4.2.0/HDP-2.4.2.0-centos7-rpm.tar.gz
SLES 11 SP3/SP4	http://public-repo-1.hortonworks.com/HDP/suse11sp3/2.x/updates/2.4.2.0/HDP-2.4.2.0-suse11sp3-rpm.tar.gz
Ubuntu 12	http://public-repo-1.hortonworks.com/HDP/ubuntu12/2.x/updates/2.4.2.0/HDP-2.4.2.0-ubuntu12-deb.tar.gz
Ubuntu 14	http://public-repo-1.hortonworks.com/HDP/ubuntu14/2.x/updates/2.4.2.0/HDP-2.4.2.0-ubuntu14-deb.tar.gz

2. **Run an update:**

```
apt-get update
```

3. **Install the HDP 2.4.2 bits:**

Operating System	Commands
RHEL/CentOS/Oracle LINUX	Install HDP 2.4.2 components on relevant nodes, according to the services that run on those hosts: yum install "hadoop_2.4.2_258*" "oozie_2.4.2_258*" "pig_2.4.2_258*" "sqoop_2.4.2_258*" "zookeeper_2.4.2_258*" "hbase_2.4.2_258*" "hive_2.4.2_258*" "tez_2.4.2_258*" "storm_2.4.2_258*" "falcon_2.4.2_258*" "flume_2.4.2_258*" "phoenix_2.4.2_258*" "accumulo_2.4.2_258*" "mahout_2.4.2_258*"
SLES	Install HDP 2.4.2 components on relevant nodes, according to the services that run on those hosts: zypper install "hadoop_2.4.2_258*" "oozie_2.4.2_258*" "pig_2.4.2_258*" "sqoop_2.4.2_258*" "zookeeper_2.4.2_258*" "hbase_2.4.2_258*" "hive_2.4.2_258*" "tez_2.4.2_258*" "storm_2.4.2_258*" "falcon_2.4.2_258*" "flume_2.4.2_258*" "phoenix_2.4.2_258*" "accumulo_2.4.2_258*" "mahout_2.4.2_258*"
Ubuntu/Debian	Install HDP 2.4.2 components on relevant nodes, according to the services that run on those hosts: apt-get install "hadoop_2.4.2_258*" "oozie_2.4.2_258*" "pig_2.4.2_258*" "sqoop_2.4.2_258*" "zookeeper_2.4.2_258*" "hbase_2.4.2_258*" "hive_2.4.2_258*" "tez_2.4.2_258*" "storm_2.4.2_258*" "falcon_2.4.2_258*" "flume_2.4.2_258*" "phoenix_2.4.2_258*" "accumulo_2.4.2_258*" "mahout_2.4.2_258*"

4. **Stop all HDP 2.4.x Services for your scenario:**

   • For non-Ambari managed clusters:

   a. Stop all HDP 2.4.x services using the Stopping HDP Services section of the *HDP Reference Guide*.

   • For Ambari 1.7.0-managed clusters:

   a. Open Ambari Web.

   b. Browse to **Services**.

   c. Use **Service Actions** to stop each service.

5. **For all services, switch the active version to HDP 2.4.2.**

   On each host in the cluster, use hdp-select to switch all services to the HDP 2.4.2 version:

   ```
   hdp-select set all 2.4.2.0-258
   ```

6. **Complete the Stack Upgrade for your scenario:**

   • For Ambari 1.7.0-managed clusters:

   Update the repository Base URLs to use the HDP 2.4.2 repositories for HDP and HDP-UTILS:

     a. Open Ambari Web.

     b. Browse to **Admin > Repositories**.

     c. Edit the Base URLs.

7. **Start all HDP 2.4.2 services, in the following order:**

     a. **ZooKeeper**

```
su - zookeeper export ZOOCFGDIR=/usr/hdp/current/zookeeper-
server/conf ; export ZOOCFG=zoo.cfg; source /usr/hdp/current/
zookeeper-server/conf/zookeeper-env.sh ; /usr/hdp/current/
zookeeper-server/bin/zkServer.sh start
```

     b. **(HA NameNode upgrade only) ZooKeeper Failover Controller Daemons**

```
/usr/hdp/current/hadoop-hdfs-namenode/../hadoop/sbin/hadoop-
daemon.sh start zkfc
```

     c. **(HA NameNode upgrade only) JournalNodes**

```
su - hdfs /usr/hdp/current/hadoop-hdfs-journalnode/../hadoop/
sbin/hadoop-daemon.sh start journalnode
```

     d. **HDFS NameNode(s)**

     Start the HDFS NameNode(s). Because there is no metadata schema update for this upgrade, start the NameNode(s) in normal mode:

```
su - hdfs /usr/hdp/current/hadoop-hdfs-namenode/../hadoop/
sbin/hadoop-daemon.sh start namenode
```

     e. **Remaining Services**

     Start the rest of the HDP services. On each host in the cluster, start the services that are relevant to that cluster. To identify the start commands for all services, see "Controlling HDP Services Manually" in the *HDP Reference Guide*.

You now have an upgraded cluster. Ensure that your workloads run correctly on this upgraded cluster.

1.3.3. Optional: Spark Manual Upgrade Procedure

(Optional) Upgrade Spark from 1.6.0 to 1.6.1. As root:

1. Stop Spark 1.6.0: `su - spark -c "/usr/hdp/current/spark-client/sbin/ stop-history-server.sh"`.

2. Remove Spark 1.6.0: `yum erase "spark*"`.

3. Add the node where you want Spark 1.6.1 History Server to run:

a. `su - root`

b. `wget -nv http://s3.amazonaws.com/dev.hortonworks.com/HDP/`
   `centos6/2.x/BUILDS/2.4.2.0-258/hdpbn.repo -O /etc/yum.repos.d/`
   `Spark141TP.repo`

c. `yum install spark_2_4_2_0_258-master -y`

d. **To use Python:** `yum install spark_2_4_2_0_258-python`

e. `conf-select create-conf-dir --package spark --stack-version`
   `2.4.2.0-258 --conf-version 0`

f. `cp /etc/spark/2.4.2.0-258/0/* /etc/spark/2.4.2.0-258/0/`

g. `conf-select set-conf-dir --package spark --stack-version`
   `2.4.2.0-258 --conf-version 0`

h. `hdp-select set spark-client 2.4.2.0-258`

i. `hdp-select set spark-historyserver 2.4.2.0-258`

4. Validate the Spark installation. As user spark, run SparkPI example:

a. `su - spark -c "cd /usr/hdp/current/spark-client"`

b. `./bin/spark-submit --class org.apache.spark.examples.SparkPi`
   `--master yarn-client --num-executors 3 --driver-memory`
   `512m --executor-memory 512m --executor-cores 1 lib/spark-`
   `examples*.jar 10`

5. Restart Spark on YARN in either yarn-cluster mode or yarn-client mode:

- yarn-cluster mode: `./usr/hdp/current/spark-client/bin/spark-submit`
  `--class path.to.your.Class --master yarn-cluster [options] <app`
  `jar> [app options]`

- yarn-client mode: `./usr/hdp/current/spark-client/bin/spark-shell --`
  `master yarn-client`

1.3.4. Optional: Spark Manual Downgrade Procedure

When upgrading to HDP 2.4.2 using Ambari, Spark 1.6.0 is automatically upgraded to 1.6.1. However, if you wish to return to using 1.6.0:

1. Remove Spark 1.6.1 from your HDP cluster using Ambari:

   curl -u admin:admin -H "X-Requested-By: ambari" -X DELETE

   http://<AMBARI_HOST>:8080/api/v1/clusters/<CLUSTER_NAME>/services/SPARK

2. Manually install Spark 1.6.0 with HDP 2.3.0 Installing HDP Manually: Installing and Configuring Apache Spark.

1.4. Behavioral Changes

Behavioral changes denote a marked change in behavior from the previously released version to this version of software. In HDP 2.4.2, behavioral changes affect the following Hadoop components.

Table 1.4. Behavioral Changes

Hortonworks Bug ID	Component	Apache JIRA	Summary	Details
BUG-52381	Spark		Spark Kafka streaming jars dependency	For developers using Spark Streaming with Kafka on a Kerberos-enabled cluster, use the HDP spark-streaming-kafka jar file and associated jar files. For more information, see Using Spark Streaming with Kafka on a Kerberos-Enabled Cluster in the Spark Guide.
BUG-55056	YARN	YARN-4414	Nodemanager connection errors are retried at multiple levels.	**Scenario:** During NodeManager restart, all YARN clients connect to NodeManager need to retry the connection to NodeManager. We formerly had a two layer retry for YARN Client connection with NodeManager: one was at the RPC layer, and the other was at the NodeManager proxy layer. This caused the log of retry activities to expand logarithmically. After this fix, we consolidated the RPC layer retry and NodeManager proxy retry; the retry log no longer has duplicate entries.

1.5. Apache Patch Information

The following sections list patches in each HDP 2.4.2 component beyond what was fixed in the base version of the Apache component.

Please see the Fixed Issues section for customer-reported issues.

- Hadoop

- Accumulo

- Atlas

- Calcite

- Falcon

- Flume

- HBase

- Hive

- Kafka

- Knox

- Mahout

- Oozie

- Phoenix

- Pig

- Ranger

- Slider

- Spark

- Sqoop

- Storm

- Tez

- Zookeeper

1.5.1. Hadoop

HDP 2.4.2 provides the following Apache patches:

- HADOOP-10365: BufferedOutputStream in FileUtil#unpackEntries() should be closed in finally block.

- HADOOP-11212: NetUtils.wrapException to handle SocketException explicitly.

- HADOOP-12100: ImmutableFsPermission should not override applyUmask since that method doesn't modify the FsPermission.

- HADOOP-12103: Small refactoring of DelegationTokenAuthenticationFilter to allow code sharing.

- HADOOP-12107: Long running apps may have a huge number of StatisticsData instances under FileSystem.

- HADOOP-12161: Add getStoragePolicy API to the FileSystem interface.

- HADOOP-12191: Bzip2Factory is not thread safe.

- HADOOP-12213: Interrupted exception can occur when Client#stop is called.

- HADOOP-12348: MetricsSystemImpl creates MetricsSourceAdapter with wrong time unit parameter.

- HADOOP-12374: Description of HDFS expunge command is confusing.

- HADOOP-12426: Add Entry point for Kerberos health check.

- HADOOP-12464: Interrupted client may try to fail-over and retry Interrupted client may try to fail-over and retry.

- HADOOP-12482: Race condition in JMX cache update.

- HADOOP-12589: Fix intermittent test failure of TestCopyPreserveFlag.

- HADOOP-12609: Fix intermittent failure of TestDecayRpcScheduler.

- HADOOP-12699: TestKMS#testKMSProvider intermittently fails during 'test rollover draining'.

- HADOOP-12706: TestLocalFsFCStatistics#testStatisticsThreadLocalDataCleanUp times out occasionally.

- HADOOP-12752: Improve diagnostics/use of envvar/sysprop credential propagation.

- HADOOP-12787: KMS SPNEGO sequence does not work with WEBHDFS.

- HADOOP-12795: KMS does not log detailed stack trace for unexpected errors.

- HADOOP-12825: Log slow name resolutions.

- HADOOP-12829: StatisticsDataReferenceCleaner swallows interrupt exceptions.

- HADOOP-12851: S3AFileSystem Uptake of ProviderUtils.excludeIncompatibleCredentialProviders.

- HADOOP-12903: IPC Server should allow suppressing exception logging by type, not log 'server too busy' messages.

- HADOOP-12958: PhantomReference for filesystem statistics can trigger OOM.

- HADOOP-13026: Should not wrap IOExceptions into a AuthenticationException in KerberosAuthenticator.

- HDFS-10199: Unit tests TestCopyFiles, TestDistCh, TestLogalyzer under org.apache.hadoop.tools are failing.

- HDFS-10270: TestJMXGet:testNameNode() fails.

- HDFS-10281: TestPendingCorruptDnMessages fails intermittently.

- HDFS-10283: o.a.h.hdfs.server.namenode.TestFSImageWithSnapshot#testSaveLoadImageWithAppending fails intermittently.

- HDFS-6101: TestReplaceDatanodeOnFailure fails occasionally.

- HDFS-8113: Add check for null BlockCollection pointers in BlockInfoContiguous structures.

- HDFS-8337: Accessing HttpFS via WebHDFS doesn't work from a jar with Kerberos.

- HDFS-8647: Abstract BlockManager's rack policy into BlockPlacementPolicy.

- HDFS-8659: Block scanner INFO message is spamming logs.

- HDFS-8676: Delayed rolling upgrade finalization can cause heartbeat expiration.

- HDFS-8772: Fix TestStandbyIsHot#testDatanodeRestarts which occasionally fails.

- HDFS-8806: Inconsistent metrics: number of missing blocks with replication factor 1 not properly cleared.

- HDFS-8815: DFS getStoragePolicy implementation using single RPC call.

- HDFS-8891: HDFS concat should keep srcs order.

- HDFS-9072: Fix random failures in TestJMXGet.

- HDFS-9130: Use GenericTestUtils#setLogLevel to the logging level.

- HDFS-9221: HdfsServerConstants#ReplicaState#getState should avoid calling values() since it creates a temporary array.

- HDFS-9239: DataNode Lifeline Protocol: an alternative protocol for reporting DataNode liveness.

- HDFS-9289: Make DataStreamer#block thread safe and verify genStamp in commitBlock.

- HDFS-9290: DFSClient#callAppend() is not backward compatible for slightly older NameNodes.

- HDFS-9313: Possible NullPointerException in BlockManager if no excess replica can be chosen.

- HDFS-9314: Improve BlockPlacementPolicyDefault's picking of excess replicas.

- HDFS-9347: Invariant assumption in TestQuorumJournalManager.shutdown() is wrong.

- HDFS-9383: TestByteArrayManager#testByteArrayManager fails.

- HDFS-9402: Switch DataNode.LOG to use slf4j.

- HDFS-9431: DistributedFileSystem#concat fails if the target path is relative.

- HDFS-9434: Recommission a datanode with 500k blocks may pause NN for 30 seconds for printing info log messages.

- HDFS-9445: Datanode may deadlock while handling a bad volume.

- HDFS-9478: Reason for failing ipc.FairCallQueue construction should be thrown.

- HDFS-9534: Add CLI command to clear storage policy from a path.

- HDFS-9557: Reduce object allocation in PB conversion.

- HDFS-9572: Prevent DataNode log spam if a client connects on the data transfer port but sends no data.

- HDFS-9574: Reduce client failures during datanode restart.

- HDFS-9600: Do not check replication if the block is under construction.

- HDFS-9625: Set replication for empty file failed when set storage policy.

- HDFS-9655: NN should start JVM pause monitor before loading fsimage.

- HDFS-9661: Deadlock in DN.FsDatasetImpl between moveBlockAcrossStorage and createRbw.

- HDFS-9710: DN can be configured to send block receipt IBRs in batches.

- HDFS-9724: Degraded performance in WebHDFS listing as it does not reuse ObjectMapper.

- HDFS-9726: Refactor IBR code to a new class.

- HDFS-9740: Use a reasonable limit in DFSTestUtil.waitForMetric().

- HDFS-9743: Fix TestLazyPersistFiles#testFallbackToDiskFull in branch-2.7.

- HDFS-9752: Permanent write failures may happen to slow writers during datanode rolling upgrades.

- HDFS-9768: Reuse ObjectMapper instance in HDFS to improve the performance.

- HDFS-9790: HDFS Balancer should exit with a proper message if upgrade is not finalized.

- HDFS-9839: Reduce verbosity of processReport logging.

- HDFS-9851: NameNode throws NPE when setPermission is called on a path that does not exist.

- HDFS-9854: Log cipher suite negotiation more verbosely.

- HDFS-9906: Remove spammy log spew when a datanode is restarted.

- HDFS-9941: Do not log StandbyException on NN, other minor logging fixes.

- MAPREDUCE-6436: JobHistory cache issue.

- MAPREDUCE-6460: TestRMContainerAllocator.testAttemptNotFoundCausesRMCommunicatorException fails.

- MAPREDUCE-6492: AsyncDispatcher exit with NPE on TaskAttemptImpl#sendJHStartEventForAssignedFailTask.

- 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-6670: TestJobListCache#testEviction sometimes fails on Windows with timeout.

- MAPREDUCE-6680: JHS UserLogDir scan algorithm sometime could skip directory with update in CloudFS.

- YARN-2046: Out of band heartbeats are sent only on container kill and possibly too early.

- YARN-2871: TestRMRestart#testRMRestartGetApplicationList sometimes fails in trunk.

- YARN-3102: Decommissioned Nodes not listed in Web UI.

- YARN-3695: ServerProxy (NMProxy, etc.) shouldn't retry forever for non network exception.

- YARN-3769: Consider user limit when calculating total pending resource for preemption policy in Capacity Scheduler.

- YARN-4155: TestLogAggregationService.testLogAggregationServiceWithInterval failing.

- YARN-4365: FileSystemNodeLabelStore should check for root dir existence on startup.

- YARN-4414: Nodemanager connection errors are retried at multiple levels.

- YARN-4422: Generic AHS sometimes doesn't show started, node, or logs on App page.

- YARN-4428: Redirect RM page to AHS page when AHS turned on and RM page is not available.

- YARN-4439: Clarify NMContainerStatus#toString method.

- YARN-4546: ResourceManager crash due to scheduling opportunity overflow.

- YARN-4598: Invalid event: RESOURCE_FAILED at CONTAINER_CLEANEDUP_AFTER_KILL.

- YARN-4610: Reservations continue looking for one app causes other apps to starve.

- YARN-4623: TestSystemMetricsPublisher#testPublishAppAttemptMetricsForUnmanagedAM fails with NPE on branch-2.7.

- YARN-4633: Fix random test failure in TestRMRestart#testRMRestartAfterPreemption.

- YARN-4680: Fix TimerTasks leak in Application Timeline Server (ATS) v1.5 Writer.

- YARN-4696: TimelineClient to add flush operation for deterministic writes (including testing) and Changes to EntityGroupFSTimelineStore for testability.

- YARN-4709: NMWebServices produces incorrect JSON for containers.

- YARN-4723: NodesListManager$UnknownNodeId ClassCastException.

- YARN-4737: Add CSRF filter support in YARN.

- YARN-4769: Add support for CSRF header in the dump capacity scheduler logs and kill app buttons in RM web UI.

- YARN-4785: Inconsistent value type of the "type" field for LeafQueueInfo in response of RM REST API.

- YARN-4814: ATS 1.5 timelineclient impl call flush after every event write.

- YARN-4815: ATS 1.5 timelineclinet impl try to create attempt directory for every event call.

- YARN-4817: TimelineClient ATSv1.5 logging is very noisy.

- YARN-4916: TestNMProxy.tesNMProxyRPCRetry fails.

- YARN-4928: Some yarn.server.timeline.* tests fail on Windows attempting to use a test root path containing a colon.

- YARN-4954: TestYarnClient.testReservationAPIs fails on machines with less than 4 GB available memory.

- YARN-4955: Add retry for SocketTimeoutException in TimelineClient.

- YARN-4965: Distributed shell AM failed due to ClientHandlerException thrown by jersey.

- YARN-4968: Fix two scheduler related UTs in YARN.

HDP 2.4.0 provided the following Apache patches:

- HADOOP-10406: TestIPC.testIpcWithReaderQueuing may fail.

- HADOOP-12551: Introduce FileNotFoundException for WASB FileSystem API.

- HADOOP-12608: Fix exception message in WASB when connecting with anonymous credential.

- HADOOP-12678: Handle empty rename pending metadata file during atomic rename in redo path.

- HDFS-8729: Fix TestFileTruncate#testTruncateWithDataNodesRestartImmediately which occasionally failed.

- HDFS-9358: TestNodeCount#testNodeCount timed out.

- HDFS-9406: FSImage may get corrupted after deleting snapshot.

- HDFS-9672: o.a.h.hdfs.TestLeaseRecovery2 fails intermittently.

- MAPREDUCE-6566: Add retry support to mapreduce CLI tool.

- MAPREDUCE-6618: YarnClientProtocolProvider leaking the YarnClient thread.

- MAPREDUCE-6621: Memory Link in JobClient#submitJobInternal().

- YARN-3480: Remove attempts that are beyond max-attempt limit from state store.

- YARN-4309: Add container launch related debug information to container logs when a container fails.

- YARN-4497: RM might fail to restart when recovering apps whose attempts are missing.

- YARN-4565: Sometimes when sizeBasedWeight FairOrderingPolicy is enabled, under stress appears that cluster is virtually in deadlock.

- YARN-4584: RM startup failure when AM attempts greater than max-attempts.

- YARN-4625: ApplicationSubmissionContext and ApplicationSubmissionContextInfo more consistent.

1.5.2. Accumulo

HDP 2.4.2 provides Accumulo 1.7.0 and the following Apache patches:

- ACCUMULO-4080 TabletServers should be less aggressively "monitoring RO filesystems".

- ACCUMULO-4185 Use SYNC durability and expect the update to be durable in ProxyDurabilityIT.

HDP 2.4.0 provided Accumulo 1.7.0 and the following Apache patches:

- ACCUMULO-4135 Add impersonation configuration keys which don't put the principal in the key.

1.5.3. Atlas

HDP 2.4.2 provides Atlas 0.5.0 with the patches specified below. No new additional Apache patches have been included in this release.

HDP 2.4.0 provided Atlas 0.5.0 and the following Apache patches:

- ATLAS-448: Hive IllegalArgumentException with Atlas hook enabled on SHOW TRANSACTIONS AND SHOW COMPACTIONS.

1.5.4. Calcite

HDP 2.4.2 provides Calcite 1.2.0 and the following Apache patches:

- CALCITE-1128: Support addBatch()/executeBatch() in remote driver.

- CALCITE-1180: Support clearBatch() in remote driver.

HDP 2.4.0 provided Calcite 1.2.0 and the following Apache patch:

- CALCITE-984: NPE in ErrorResponse construction.

1.5.5. Falcon

HDP 2.4.2 provides Falcon 0.6.1 with no additional Apache patches included.

HDP 2.4.0 provided Falcon 0.6.1 with no additional Apache patches included.

1.5.6. Flume

HDP 2.4.2 provides Flume 1.5.2 and the following Apache patch:

- FLUME-2889 Fixes to DateTime computations.

HDP 2.4.0 provided Flume 1.5.2 and the following Apache patch:

- FLUME-2865 Upgrade thrift version (0.9.2).

1.5.7. HBase

HDP 2.4.2 provides HBase 1.1.2 and the following Apache patches:

- HBASE-14283: Reverse scan doesn't work with HFile inline index/bloom blocks.

- HBASE-14307: Incorrect use of positional read API in HFileBlock.

- HBASE-14812: Fix ResultBoundedCompletionService deadlock.

- HBASE-14883: TestSplitTransactionOnCluster#testFailedSplit flakey.

- HBASE-15220: Change two logs in SimpleRegionNormalizer to INFO level.

- HBASE-15221: Reload the cache on re-tried puts in HTableMultiplexer and adds a close() method to HTableMultiplexer.

- HBASE-15232: Handle region location cache management in AsyncProcess for multi()'s.

- HBASE-15621: Suppress Hbase SnapshotHFile cleaner error messages when a snaphot is going on.

HDP 2.4.0 provided HBase 1.1.2 and the following Apache patches:

- HBASE-14107: Administrative Task: Provide an API to List all procedures.

- HBASE-14108: Administrative Task: provide an API to abort a procedure.

- HBASE-14432: Enforce ACL on procedure admin tasks.

- HBASE-14468: Compaction improvements: FIFO compaction policy.

- HBASE-14471: Thrift - HTTP Error 413 full HEAD if using kerberos authentication.

- HBASE-14487: Shell command to list all procedures.

- HBASE-14488: Shell command to abort a procedure.

- HBASE-14575: Relax region read lock for compactions.

- HBASE-14761: Deletes with and without visibility expression do not delete the matching mutation.

- HBASE-14866: VerifyReplication should use peer configuration in peer connection.

- HBASE-14928: Start row should be set for query through HBase REST gateway involving globbing option.

- HBASE-14963: Remove Guava dependency from HBase client code.

- HBASE-14987: Compaction marker whose region name doesn't match current region's needs to be handled.

- HBASE-15014: Fix filterCellByStore in WALsplitter is awful for performance.

- HBASE-15022: replication_admin.rb throws undefined method 'getZooKeeperClusterKey' for ZKUtil.

- HBASE-15035: bulkloading hfiles with tags that require splits do not preserve tags.

1.5.8. Hive

HDP 2.4.2 provides Hive 1.2.1 and the following Apache patches:

- HIVE-11097: HiveInputFormat uses String.startsWith to compare splitPath and PathToAliases.

- HIVE-11333: ColumnPruner prunes columns of UnionOperator that should be kept.

- HIVE-11470: NPE in DynamicPartFileRecordWriterContainer on null part-keys..

- HIVE-11816: Upgrade groovy to 2.4.4.

- HIVE-11841: KeyValuesInputMerger creates huge logs.

- HIVE-12022: NPE in SARG with timestamp cast.

- HIVE-12064: prevent transactional=false.

- HIVE-12165: wrong result when hive.optimize.sampling.orderby=true with some aggregate functions.

- HIVE-12189: The list in pushdownPreds of ppd.ExprWalkerInfo should not be allowed to grow very large.

- HIVE-12352: CompactionTxnHandler.markCleaned() may delete too much.

- HIVE-12684: NPE in stats annotation when all values in decimal column are NULLs.

- HIVE-12742: NULL table comparison within CASE does not work as previous hive versions.

- HIVE-12749: Constant propagate returns string values in incorrect format.

- HIVE-12784: Group by SemanticException: Invalid column reference.

- HIVE-12879: RowResolver of Semijoin not updated in CalcitePlanner.

- HIVE-12947: SMB join in tez has ClassCastException when container reuse is on.

- HIVE-13016: ORC FileDump recovery utility fails in Windows.

- HIVE-13092: Vectorized java.lang.ClassCastException: org.apache.hadoop.hive.serde2.typeinfo.ListTypeInfo cannot be cast to org.apache.hadoop.hive.serde2.typeinfo.PrimitiveTypeInfo.

- HIVE-13129: CliService leaks HMS connection.

- HIVE-13144: HS2 can leak ZK ACL objects when curator retries to create the persistent ephemeral node.

- HIVE-13151: Clean up UGI objects in FileSystem cache for transactions.

- HIVE-13174: Remove Vectorizer noise in logs.

- HIVE-13232: Aggressively drop compression buffers in ORC OutStreams.

- HIVE-13233: Use min and max values to estimate better stats for comparison operators.

- HIVE-13263: Vectorization: Unable to vectorize regexp_extract/regexp_replace " Udf: GenericUDFBridge, is not supported".

- HIVE-13285: Orc concatenation may drop old files from moving to final path.

- HIVE-13298: nested join support causes undecipherable errors in SemanticAnalyzer.

- HIVE-13325: Excessive logging when ORC PPD fails type conversions.

HDP 2.4.0 provided Hive 1.2.1 and the following Apache patches:

- HIVE-11141: Improve RuleRegExp when the Expression node stack gets huge.

- HIVE-11291: Avoid allocation storm while doing rule matching on operator/expression trees.

- HIVE-11310: Avoid expensive AST tree conversion to String for expressions in WHERE clause.

- HIVE-11311: Avoid dumping AST tree String in Explain unless necessary.

- HIVE-11328: Avoid String representation of expression nodes in ConstantPropagateProcFactory unless necessary.

- HIVE-11330: Add early termination for recursion in StatsRulesProcFactory.evaluateExpression.

- HIVE-11397: Parse Hive OR clauses as they are written into the AST.

- HIVE-11405: Add early termination for recursion in StatsRulesProcFactory.evaluateExpression for OR expression.

- HIVE-11406: Vectorization: StringExpr::compare() == 0 is bad for performance.

- HIVE-11981: ORC Schema Evolution Issues (Vectorized, ACID, and Non-Vectorized).

- HIVE-12625: Backport to branch-1 HIVE-11981 ORC Schema Evolution Issues (Vectorized, ACID, and Non-Vectorized).

- HIVE-12660: HS2 memory leak with .hiverc file use.

- HIVE-12706: Incorrect output from from_utc_timestamp()/to_utc_timestamp when local timezone has DST.

- HIVE-12728: Apply DDL restrictions for ORC schema evolution.

- HIVE-12766: TezTask does not close DagClient after execution.

- HIVE-12799: Always use Schema Evolution for ACID.

1.5.9. Kafka

HDP 2.4.2 provides Kafka 0.9.0.1 and the following Apache patches:

- KAFKA-2803: Add hard bounce system test for Kafka Connect.

- KAFKA-2812: Improve consumer integration tests.

- KAFKA-2862: Fix MirrorMaker's message.handler.args description.

- KAFKA-2872: Unite sink nodes with parent nodes in addSink.

- KAFKA-2877: Handle request timeout in sync group.

- KAFKA-2878: Guard against OutOfMemory in Kafka broker.

- KAFKA-2879: Make MiniKDC test service slightly more generic.

- KAFKA-2880: Consumer should handle disconnect/timeout for metadata requests.

- KAFKA-2881: Improve Consumer Configs and API Documentation.

- KAFKA-2882: Add constructor cache for Snappy and LZ4 Output/Input streams in Compressor.java.

- KAFKA-2892: Consumer Docs Use Wrong Method.

- KAFKA-2899: Improve logging when unexpected exceptions thrown in reading local log.

- KAFKA-2906: Fix Connect javadocs, restrict only to api subproject, and clean up javadoc warnings.

- KAFKA-2913: Missing partition check when removing groups from cache.

- KAFKA-2942: Inadvertent auto-commit when pre-fetching can cause message loss.

- KAFKA-2950: Fix performance regression in the producer.

HDP 2.4.0 provided Kafka 0.9.0 with no additional Apache patches included.

1.5.10. Knox

HDP 2.4.2 provides Knox 0.6.0 and the following Apache patches:

- KNOX-677: Upgrade to latest Groovy.

- KNOX-695: Expose configuration of HttpClient's connection and socket timeout settings.

HDP 2.4.0 provided Knox 0.6.0 and the following Apache patch:

- KNOX-647: Rename LDAP artifacts from test to demo.

1.5.11. Mahout

In HDP-2.3.x and 2.4.x, instead of shipping a specific Apache release of Mahout, we synchronized to a particular revision point on Apache Mahout trunk. This revision point is after the 0.9.0 release, but before the 0.10.0 release. This provides a large number of bug fixes and functional enhancements over the 0.9.0 release, but provides a stable release of the Mahout functionality before the complete conversion to new Spark-based Mahout in 0.10.0.

The revision point chosen for Mahout in HDP 2.3.x and 2.4.x is from the "mahout-0.10.x" branch of Apache Mahout, as of 19 December 2014, revision 0f037cb03e77c096 in GitHub.

In addition, we have provided the following patches:

- MAHOUT-1493 Port Naive Bayes to Scala DSL.

- MAHOUT-1589 mahout.cmd has duplicated content.

1.5.12. Oozie

HDP 2.4.2 provides Oozie 4.2.0 with no additional Apache patches included.

HDP 2.4.0 provided Oozie 4.2.0 with no additional Apache patches included.

1.5.13. Phoenix

HDP 2.4.2 provides Phoenix 4.4.0-HBase-1.1 and the following Apache patches:

- PHOENIX-2601: Query result is incorrect when both index hint and limit are used.

- PHOENIX-2817: Phoenix-Spark plugin doesn't work in secured environment.

HDP 2.4.0 provided Phoenix 4.4.0-HBase-1.1 and the following Apache patch:

- PHOENIX-2608: Incompatibility between Jackson1 version shipped with Phoenix, YARN.

1.5.14. Pig

HDP 2.4.2 provides Pig 0.15.0 and the following Apache patches:

- PIG-4690: Union with self replicate join will fail in Tez.

- PIG-4760: TezDAGStats.convertToHadoopCounters is not used, but impose MR counter limit.

- PIG-4790: Join after union fail due to UnionOptimizer.

- PIG-4814: AvroStorage does not take namenode HA as part of schema file URL.

- PIG-4816: Read a null scalar causing a Tez failure.

HDP 2.4.0 provided Pig 0.15.0 with no additional Apache patches included.

1.5.15. Ranger

HDP 2.4.2 provides Ranger 0.5.0 and the following Apache patches:

- RANGER-218: LDAP Groups incorrectly labelled internal.

- RANGER-746: Addressing suggestions from Review - Add wildcard, multiple CN & SAN support when validating plugins' SSL certs.

- RANGER-771: Fix 4+ Log entries upon login in X_AUTH_SESS.

- RANGER-777: Remove the maven profile that was added for building Kafka plugin.

- RANGER-789: Fix incorrect policy list paging for non-admin users.

- RANGER-794: Ranger policy engine performance measurement.

- RANGER-798: Handle different timezone issue while saving audit logs to Solr.

- RANGER-799: Ranger UI fixes - partial search not working on Policy listing page.

- RANGER-804: Delete groups associated with User causes Exception in UserSync.

- RANGER-809: Audit framework need to cache the getHostName() values to reuse for successive calls.

- RANGER-821: Ranger shutdown hook should not only do its processing asynchronously but also terminate itself if it runs more than a configurable amount of time.

- RANGER-825: groupId in ranger's child poms should be consistent and follow convention.

- RANGER-831: policy version incremented twice when resources are updated.

- RANGER-833: In Ranger UI add support for usernames containing a plus + symbol.

- RANGER-834: Correct the excludes flag's treatment when resource value denotes everything.

- RANGER-836: Optimize policy retrieval.

- RANGER-844: Optimize policy retrieval for non-admin users.

- RANGER-848: Policy Listing page fix: users column is empty for non-admin users.

- RANGER-857: Unify (and update) Tomcat versions.

- RANGER-863: Make parameters like maxHttpHeaderSize configurable for EmbeddedServer.

- RANGER-882: Scrub received policies before policy engine uses it to guard against inadvertent data corruption: remove null policy resource values.

- RANGER-889: Policy engine API to find list of users/groups having access to a resource.

- RANGER-894: Fixing few issues in the ldap tool when user search base or group search base is configured as part of input.

HDP 2.4.0 provided Ranger 0.5.0 and the following Apache patches:

- RANGER-173: Utility scripts to create HDFS audit folders and policies.

- RANGER-725: Add the right .gitignore file to the newly added projects so that directory listing is clean after a build.

- RANGER-767: Refactor UserGroupSink implementation and consolidate performance improvements.

- RANGER-772: Hive plugin Update Ranger authorizer to mimic changes made by Hive standard authorizer for the case when IMPORT can end up creating a table.

- RANGER-773: Fix newly found Coverity scan issues for Ranger KMS.

- RANGER-778: Fix user update issue.

- RANGER-809: Audit framework need to cache the getHostName() values to reuse for successive calls.

1.5.16. Slider

HDP 2.4.2 provides Slider 0.80.0 and the following Apache patches:

- SLIDER-1079 Cache metainfo object in AgentClientProvider.

- SLIDER-1088 Slider client getting UnknownHost exception trying to RPC to AM in different subdomain.

- SLIDER-657 Introduce –force switch for slider destroy command.

HDP 2.4.0 provided Slider 0.80.0 with no additional Apache patches included.

1.5.17. Spark

HDP 2.4.2 provides Spark 1.6.1 and the following Apache patches:

- SPARK-10582: Using dynamic-executor-allocation, if AM failed, the new AM will be started. But the new AM does not allocate executors to driver.

- SPARK-11137: Make StreamingContext.stop() exception-safe.

- SPARK-11182: HDFS Delegation Token will be expired when calling "UserGroupInformation.getCurrentUser.addCredentials" in HA mode.

- SPARK-11314: Add service API and test service for Yarn Cluster schedulers.

- SPARK-11627: Spark Streaming backpressure mechanism has no initial rate limit, receivers receive data at the maximum speed, it might cause OOM exception.

- SPARK-11969: SQL UI does not work with PySpark.

- SPARK-12001: StreamingContext cannot be completely stopped if the stop() is interrupted.

- SPARK-12009: Avoid re-allocate yarn container while driver want to stop all executors.

- SPARK-12142: Can't request executor when container allocator is not ready.

- SPARK-12241: Improve failure reporting in Yarn client obtainTokenForHBase().

- SPARK-12353: Wrong output for countByValue and countByValueAndWindow.

- SPARK-12513: SocketReceiver hang in Netcat example.

- SPARK-12523: Support long-running of the Spark on HBase and hive metastore.

- SPARK-12920: Fix high CPU usage in Spark thrift server with concurrent users.

- SPARK-12925: Improve HiveInspectors.unwrap for StringObjectInspector.getPrimitiveWritableObject.

- SPARK-12948: OrcRelation uses HadoopRDD which can broadcast conf objects frequently.

- SPARK-12998: Enable OrcRelation when connecting via Spark thrift server.

- SPARK-13021: Fail fast when custom RDD's violate RDD.partition's API contract.

- SPARK-13117: WebUI should use the local IP not 0.0.0.0.

- SPARK-13308: ManagedBuffers passed to OneToOneStreamManager need to be freed in non error cases.

- SPARK-13360: Pyspark related environment variable is not propagated to driver in yarn-cluster mode.

- SPARK-13468: Fix a corner case where the page UI should show DAG but it doesn't show.

- SPARK-13478: Use real user when fetching delegation tokens.

- SPARK-13599: Groovy-all ends up in spark-assembly if Hive profile set.

- SPARK-13642: Properly handle signal kill in ApplicationMaster.

- SPARK-13885: Fix attempt ID regression for Spark running on Yarn.

- SPARK-14062: Fix log4j and upload metrics.properties automatically with distributed cache.

- SPARK-6847: Stack overflow on updateStateByKey which followed by a stream with checkpoint set.

HDP 2.4.0 provided Spark 1.6.0 and the following Apache patches:

- SPARK-11315: Add YARN extension service to publish Spark events to YARN timeline service.

- SPARK-11323: Add History Service Provider to service application histories from YARN timeline server.

- SPARK-12417: Support to have ORC bloom filters during write code path.

- SPARK-12898: Consider having dummyCallSite for HiveTableScan.

1.5.18. Sqoop

HDP 2.4.2 provides Sqoop 1.4.6 with no additional Apache patches included

HDP 2.4.0 provided Sqoop 1.4.6 with no additional Apache patches included.

1.5.19. Storm

HDP 2.4.2 provides Storm 0.10.0-beta and the following Apache patches:

- STORM-1001:Undefined STORM_EXT_CLASSPATH adds '::' to classpath of workers.

- STORM-1005:Supervisor do not get running workers after restart.

- STORM-1024:log4j changes leaving ${sys:storm.log.dir} under STORM_HOME dir.

- STORM-1027:Use overflow buffer for emitting metrics.

- STORM-1037:Do not remove STORM-code in supervisor until kill job.

- STORM-1044:Setting dop to zero does not raise an error.

- STORM-1096:Fix some issues with impersonation on the UI.

- STORM-1108:Fix NPE in simulated time.

- STORM-1121:Deprecate test only configuration nimbus.reassign.

- STORM-139:hashCode does not work for byte[].

- STORM-1481:Avoid Math.abs(Integer) get a negative value.

- STORM-1482:Add missing 'break' for RedisStoreBolt.

- STORM-1521:When using Kerberos login from keytab with multiple bolts/executors ticket is not renewed.

- STORM-584:Fix logging for LoggingMetricsConsumer metrics.log file.

- STORM-793:Made change to logviewer.clj in order to remove the invalid http 500 response.

- STORM-810:PartitionManager in STORM-kafka should commit latest offset before close.

- STORM-837:HdfsState ignores commits.

- STORM-866:Use storm.log.dir instead of storm.home in log4j2 config.

- STORM-966:ConfigValidation.DoubleValidator doesn't really validate whether the type of the object is a double.

- STORM-977:Incorrect signal (-9) when as-user is true.

- STORM-992:A bug in the timer.clj might cause unexpected delay to schedule new event.

HDP 2.4.0 provided Storm 0.10.0-beta and the following Apache patches:

- STORM-1476: Filter -c options from args and add them as part of storm.options.

- STORM-422: Allow more arguments to be passed to storm jar.

- STORM-745: Fix storm.cmd to evaluate 'shift' correctly with 'storm jar'.

1.5.20. Tez

HDP 2.4.2 provides Tez 0.7.0 and the following Apache patches:

- TEZ-1961: Remove misleading exception "No running dag" from AM logs.

- TEZ-2863: Container, node, and logs not available in UI for tasks that fail to launch.

- TEZ-2963: RecoveryService#handleSummaryEvent exception with HDFS transparent encryption & Kerberos authentication.

- TEZ-3066: TaskAttemptFinishedEvent ConcurrentModificationException in recovery or history logging services.

- TEZ-3101: Tez UI: Task attempt log link doesn't have the correct protocol.

- **TEZ-3117**: Deadlock in Edge and Vertex code.

- **TEZ-3123**: Containers can get re-used even with conflicting local resources.

- **TEZ-3128**: Avoid stopping containers on the AM shutdown thread.

- **TEZ-3137**: Tez task failed with illegal state exception.

- **TEZ-3156**: Tez client keeps trying to talk to RM even if RM does not know about the application.

- **TEZ-3177**: Non-DAG events should use the session domain or no domain if the data does not need protection.

- **TEZ-3189**: Pre-warm dags should not be counted in submitted dags count by DAGAppMaster.

HDP 2.4.0 provided Tez 0.7.0 and the following Apache patches:

- **TEZ-2307**: Possible wrong error message when submitting new DAG.

- **TEZ-2886**: Ability to merge AM credentials with DAG credentials.

- **TEZ-2898**: Tez tools: swimlanes.py is broken.

- **TEZ-2900**: Ignore V_INPUT_DATA_INFORMATION when vertex is in Failed/Killed/Error.

- **TEZ-3017**: HistoryACLManager does not have a close method for cleanup.

- **TEZ-3025**: InputInitializer creation should use the DAG GUI.

- **TEZ-3032**: Incorrect start time in different events for DAG history events.

- **TEZ-3037**: History URL should be set regardless of which history logging service is enabled.

- **TEZ-3126**: Log reason for not reducing parallelism.

1.5.21. ZooKeeper

HDP 2.4.2 provides ZooKeeper 3.4.6 with no additional Apache patches included.

HDP 2.4.0 provided ZooKeeper 3.4.6 with no additional Apache patches included.

1.6. Common Vulnerabilities and Exposures

No CVE fixes apply to HDP 2.4.2.

1.7. Third-party Licenses

Global: Apache 2.0

Component	Subcomponents	License
Accumulo	JCommander	JCommander
Falcon	cern.colt* , cern.jet*, cern.clhep	CERN
Knox	ApacheDS, Groovy	ANTLR
Knox	SL4J	MIT
Knox	Jetty and Jerico	EPL
Knox	ApacheDS	Bouncy Castle
Oozie	JDOM Oro	
Phoenix		EPL
Storm	Logback	EPL

1.8. Fixed Issues

Fixed issues represents selected issues that were previously logged via Hortonworks Support, but are now addressed in the current release. These issues may have been reported in previous versions within the Known Issues section; meaning they were reported by customers or identified by Hortonworks Quality Engineering team.

Potential Data Loss

Hortonworks Bug ID	Apache JIRA	Component	Summary
BUG-52963	HDFS-9289	HDFS	Make DataStreamer#block thread safe and verify genStamp in commitBlock.

Security

Hortonworks Bug ID	Apache JIRA	Component	Summary
BUG-48092	RANGER-746	Ranger	Ranger Admin: For validating plugins SSL certs - Add wildcard, multiple CN & SAN support.
BUG-51298	HBASE-15145	HBase	Default hbase-env.sh breaks add peer command that writes to /hbase-secure/ replication znode for kerberized clusters.
BUG-52193	RANGER-798	Ranger	Ranger "Access > Audit" not showing anything with Source "Solr" due to Time filter Issue (GMT).
BUG-52636	STORM-1521	Storm	Kerberos login from keytab with multiple bolts/ executors ticket is not renewed for HBase Bolt.
BUG-53108	HIVE-13120	Hive	Hiveserver2 doAs=true, local fetch task on orc table permission checks on wrong user.
BUG-53766	STORM-1535	Storm	Storm-HDFS - When using keytab, Kerberos ticket is not renewed with multiple bolts/executors.

Incorrect Results

Hortonworks Bug ID	Apache JIRA	Component	Summary
BUG-49949		Hive	Query with duplicate columns in GROUP BY clause fails.
BUG-51192	HIVE-12905	Hive	Hive LEFT JOIN with empty data set produces empty data set.
BUG-52706	HIVE-12742	Hive	NULL table comparison within CASE does not work as previous hive versions.
BUG-53777	PIG-4690	Pig	Pig on TEZ creates wrong result with replicated join.
BUG-53918	PHOENIX-2601	Phoenix	Query result is incorrect when both index hint and limit are used.
BUG-53927	HIVE-13330	Hive	Vectorization returns NULL for empty values for varchar/string data type.
BUG-54958	HIVE-13361, HIVE-13362	Hive, Hive2	Orc concatenation cannot enforce the compression buffer size.

Stability

Hortonworks Bug ID	Apache JIRA	Component	Summary
BUG-51303	HIVE-11470	HCatalog, Hive	NPE in DynamicPartFileRecordWriterContainer on null part-keys.
BUG-51808	HIVE-11097	Hive	MR mode query fails if one table path string starts with another's.
BUG-51829	YARN-4820	YARN	ResourceManager JMX Redirects in HA Mode Drops the Query Parameters.
BUG-52479	PHOENIX-1968, PHOENIX-2036, PHOENIX-2088, PHOENIX-2112, PHOENIX-2196, PHOENIX-2287, PHOENIX-2288, PHOENIX-2328, PHOENIX-2426, PHOENIX-2469, PHOENIX-2503, PHOENIX-2567, PHOENIX-2568, PHOENIX-2599, PHOENIX-2638	Phoenix	Phoenix-Spark integration inoperative in HDP 2.3.4.
BUG-52732	HIVE-12947	Hive	SMB mapjoin query runtime error "FileSinkOperator cannot be cast to org.apache.hadoop.hive.ql.exec.DummyStoreOpe
BUG-52957	PHOENIX-2295	Phoenix	NullPointerException when parsing a query with hints in UNION ALL.

Hortonworks Bug ID	Apache JIRA	Component	Summary
BUG-53260	HIVE-13092	Hive	Vectorized java.lang.ClassCastException: org.apache.hadoop.hive.serde2.typeinfo.ListTypeI cannot be cast to org.apache.hadoop.hive.serde2.typeinfo.PrimitiveT
BUG-53482	PIG-4814	Pig	AvroStorage does not take namenode HA as part of schema file URL.
BUG-53499	PIG-4790	Pig	Cannot use values iterator on the previous K-V pair after moveToNext has been invoked to move to the next K-V.
BUG-53671	HIVE-10308	HDP / Stack, Hive	Vectorization execution throws java.lang.IllegalArgumentException: Unsupported complex type: MAP.
BUG-54106	HIVE-13263	Hive	Vectorization: Unable to vectorize regexp_extract " Udf: GenericUDFBridge, is not supported".
BUG-54199	PHOENIX-2616	Phoenix	Indexes over immutable tables not marked as immutable.
BUG-54214	HIVE-13144	Hive, Hive2	HS2 can leak ZK ACL objects when curator retries to create the persistent ephemeral node.
BUG-54650	KNOX-695	Knox	Intermittent java.net.ConnectException: Connection timed out between Knox server and Hiveserver2.
BUG-54713	STORM-1689	Storm	Large request header causes 413 ERROR when accessing Storm Logviewer.
BUG-54822	HIVE-13151	Hive, Hive2	Hive metastore memory leak of fs cache with ACID compaction on.

Query Failure

Hortonworks Bug ID	Apache JIRA	Component	Summary
BUG-54388, BUG-51701, BUG-53494	MAPREDUCE-6635	MapReduce, Tez	java.lang.IndexOutOfBoundsException running select count(*) query on table with skip.header.line.count.

Upgrade

Hortonworks Bug ID	Apache JIRA	Component	Summary
BUG-53878		Kafka	Running bin/kafka-console-consumer.sh outputs "No brokers found in ZK". Script modified to address this issue.

Hortonworks Bug ID	Apache JIRA	Component	Summary
BUG-54155	RANGER-882	Ranger	Ranger policy engine initialization failure caused namenode startup to fail.

Usability

Hortonworks Bug ID	Apache JIRA	Component	Summary
BUG-47662	RANGER-218	Ranger	UserSync - group sync is marked as INTERNAL even if the group is synced from LDAP.
BUG-48573	OOZIE-2322	Oozie	Oozie Web UI doesn't work with Kerberos in Internet Explorer 10 or 11 and curl.
BUG-52257	FALCON-1315	Falcon	Falcon UI Hive mirroring : Unable to provide security credentials.
BUG-52260	SQOOP-2737	Sqoop	Cannot import table from Oracle with column with spaces in name.
BUG-52702	SQOOP-2779	Sqoop	Sqoop metastore doesn't seem to recognize –schema option.
BUG-52951	FALCON-1647	Falcon	Unable to create feed : FilePermission error under cluster staging directory.
BUG-54681	RANGER-889	Ranger	Create new method for partners to use for integration.
BUG-55216	FALCON-748	Falcon	Falcon webUI return 413 error when SSL is enabled.

Performance

Hortonworks Bug ID	Apache JIRA	Component	Summary
BUG-52708	RANGER-809	Ranger	Remove reference to getLocalHost() call from Audit code to avoid performance impact.
BUG-52729	RANGER-821	Ranger	NameNode shutdown hangs due to Ranger HDFS plugin.
BUG-53028	RANGER-794, RANGER-836, RANGER-844	Ranger	Policy download optimizations.

Other

Hortonworks Bug ID	Apache JIRA	Component	Summary
BUG-51738	HBASE-15147	HBase	Backport HBASE-15147: Shell should use Admin.listTableNames() instead of Admin.listTables().
BUG-53095	HADOOP-12825	Hadoop Common	Log slow name resolutions.
BUG-53871	HBASE-15128	HBase	Disable region splits and merges switch in master.
BUG-54764	MAPREDUCE-6436	MapReduce	JobHistoryServer cache issues.

1.9. Known Issues

Hortonworks Bug ID	Apache JIRA	Component	Summary
BUG-14789		Flume	Receiving Error when using bzip compression.
BUG-17827		Sqoop	Sqoop gails to create tables with reserved keywords.
BUG-22452		Flume	Got java.lang.NegativeArraySizeException when filechannel.capacity is set to more than 268434426.
BUG-27069	FLUME-2118	Flume	Occasional multi-hour pauses in file channel replay.
BUG-30022	HIVE-7693	Hive, Hive2	Invalid column ref error in order by when using column alias in select clause and using having.
BUG-31377		Falcon	Deleting an entity in Falcon will tell "falcon/test(cluster) removed successfully" even though the entity does not exist.
BUG-33453		Flume	Multiple sinks definitions result in Flume treating the last in the file as the definition.
BUG-33577	FLUME-2119	Flume	When using spool directory for Flume agent, the flume-ng process stops when a file with the same name exist in the spool dir with completed suffix.
BUG-33598	FALCON-1480	Falcon	Gather data transfer detail of replication job from HiveDR.
BUG-35149	YARN-4392	YARN	ATS logs show start timestamp later than end timestamp.
BUG-35792		Hive	Query on view results fails with table not found error, view created with subquery alias (CTE).
BUG-37146		Falcon	Falcon recipe does not support NN HA endpoint.
BUG-39283		Falcon, YARN	Falcon should support RM HA for remote cluster.
BUG-39988	HIVE-11110	Hive	CBO: Default partition filter is from MetaStore query causing TPC-DS to regress by 3x.
BUG-40923		Hive	HIVE_METASTORE_HADOOP_OPTS memory configs - Xmx overridden by HADOOP_OPTS or HADOOP_CLIENT_OPTS.
BUG-41362	SQOOP-1361	Sqoop	Sqoop import with --as-avrofile does not work when

Hortonworks Bug ID	Apache JIRA	Component	Summary
			DB table column names have '$' character in them.
BUG-42190		Ambari, Storm	Ambari start Storm UI reports failure, although UI starts.
BUG-44798	FALCON-1437	Falcon	In Falcon UI, perform changes for Notification in recipes.
BUG-46095	HDFS-8831, HDFS-9799	HDFS	Trash support for TDE.
BUG-46235	FLUME-2799	Flume	Kafka Source - Message Offset and Partition add to headers.
BUG-46316	SQOOP-2471	Sqoop	Import of arrays and struct datatypes not supported in Sqoop Hcatalog.
BUG-46771	HIVE-11716	Hive	Reading ACID table from non-acid session should raise an error.
BUG-46789	HIVE-13151	Hive, Hive2	Hive metastore memory leak of fs cache with ACID compaction on.
BUG-46862		Hive, Hive2	Hive metastore memory leak with ACID compaction on - SQLConnection leak observed as well.
BUG-47069	FALCON-1102	Falcon	Capture data transfer details of feed replication and HDFS DR recipe.
BUG-47441		Ambari, Hive	Ambari Rolling Upgrade-Upgraded from HDP 2.2.4.2 to HDP 2.3.2 changes the hiveserver 2 port from 10000 to 10010 after upgrade is complete.
BUG-47635		Sqoop	Sqoop scripts can exhibit unexpected behavior with Accumulo client installed.
BUG-47948	HIVE-11388, HIVE-13344	Hive	Allow ACID Compactor components to run in multiple metastores.
BUG-48990	HIVE-12837	Hive	MicroStrategy query 117 fails with out of memory error.
BUG-49531	HIVE-11740	HCatalog, Hive	NPE in DynamicPartFileRecordWriterContainer on null part-keys.
BUG-49726	HIVE-10632	Hive	Hive compactor processing partitions of table that does not exist.
BUG-50390		Sqoop	Sqoop import from Teradata does not support CHAR data type as column type for hcat table.
BUG-50767		Falcon	Can't edit Falcon process specifications on Falcon Web UI once process is saved.

Hortonworks Bug ID	Apache JIRA	Component	Summary
BUG-51516	AMBARI-15500	Ambari, Atlas	Atlas alert present when WE is enabled.
BUG-51596	HIVE-12439	Hive	Hive Compactor cleaner thread fails to clean aborted txns due to ORA-01795 limit.
BUG-51723	HADOOP-12444	Hadoop Common	Consider implementing lazy seek in S3AInputStream.
BUG-51896	HIVE-13013	Hive	Further improve concurrency in TxnHandler.
BUG-51986	HIVE-12996	Hive	Temp tables shouldn't be stored in metastore tables for ACID.
BUG-52058		Ambari, Spark	Spark History Server heap size is not exposed (History Server crashed with OOM).
BUG-52313	HIVE-13392	Hive	Hive compactor job failing because multiple attempts using same tmp dir on hdfs.
BUG-52573	HIVE-13051	Hive, Hive2	Compactor failing with timeout to getTable.
BUG-52844	RANGER-843	Ranger	Policy download optimizations. These optimizations improve performance of Ranger while handling a large number of authorization policies – like tens of thousands of policies. The optimizations include: • Addition of appropriate database indexes
BUG-53072	KAFKA-2854	Kafka	The full Kerberos principal must be passed through for incoming requests to Kafka.
BUG-53267	HIVE-12022	Hive	NPE: hive.ql.io.sarg.SearchArgumentImpl$PredicateLeafImpl.
BUG-53270		Hive	Query with duplicate columns in GROUP BY clause fails.
BUG-53536	OOZIE-2185	Oozie	Make Oozie CLI source conf/oozie-env.sh.
BUG-53563	HIVE-13187	Hive, Hive2	Hiveserver2 can suppress OOM errors in some cases.
BUG-53654	HDFS-8791	HDFS	Block ID-based DN storage layout can be very slow for datanode on ext4. HDFS-8791 introduces a new datanode layout format. This layout is identical to the previous block ID based layout except it has a smaller 32x32 sub-directory structure in each data storage. On startup, the datanode will

Hortonworks Bug ID	Apache JIRA	Component	Summary
			automatically upgrade its storages to this new layout. Currently, datanode layout changes support rolling upgrades, but downgrading is not supported between datanode layout changes and a rollback would be required.
BUG-53656	HDFS-8999	HDFS	Allow a file to be closed with COMMITTED but not yet COMPLETE blocks.
BUG-53757		Hive	Stats are not properly accounted for when multiple COUNT(DISTINCT) are in the query.
BUG-53789	HIVE-13201	Hive	Compaction shouldn't be allowed on non-ACID table.
BUG-53830	HIVE-13189	Hive	Consider using Joda DateTimeFormatter instead of SimpleDateFormat in GenericUDFDateAdd.
BUG-53938	HIVE-13043	Hive	Backport HIVE-13043: Reload function has no impact to function registry.
BUG-54518		Hive, Hive2	Hybrid mapjoin allocates memory the same for multi broadcast.
BUG-54732	AMBARI-15431	Ambari, Atlas	Atlas Server start failed after enabling security.
BUG-55089	YARN-4863	YARN	Mahout Client check failed during Kerberos setup.
BUG-55105	HIVE-13369	Hive	AcidUtils.getAcidState() is not paying attention toValidTxnList when choosing the "best" base file.
BUG-55223		Kafka	Mirror maker command is failing When running the Kafka MirrorMaker tool: **Issue**If you specify `--white-list=".*"`, MirrorMaker tries to fetch data from the system-level topic `__consumer-offsets` and produce the data to the target cluster. This can result in the following error: Producer cannot send requests to `__consumer-offsets` **Workaround:** add `--whitelist="topic1,topic2"` instead of `.*`
BUG-55879		HBase	#1

Hortonworks Bug ID	Apache JIRA	Component	Summary
			Issue: Given one table name, creating, pre-splitting, deleting and then recreating with the same name and split point in quick succession can result in the HBase region locator cache returning invalid region locations. This has been observed in tests around the HBase replication feature.
			Error Message: On the source HBase cluster: `2016-04-19 13:07:28,355 WARN [main-EventThread.replicationSource,TestPee regionserver.HBaseInterClusterReplica Can't replicate because of an error on the remote cluster: org.apache.hadoop.hbase.ipc.RemoteWit`
			Workaround: When re-creating a table which has already been configured for replication, restart the sink (destination) cluster after the table was recreated to ensure that invalid region locations are not cached.
			#2
			Issue: It has been observed that, occasionally, when configuring HBase to use an HDFS encryption zone (aka TDE - transparent data encryption), some WAL updates were lost in the HBase replication process. This results in not all updates to be propagated from the source cluster to the sink cluster.
			Error Message: No error message, but fewer entries in the sink cluster's table than the source cluster's table when TDE is enabled would be the sign that there was a problem.
			Workaround: None.
BUG-56664		Hive	Hive TPC-DS query4 and microstrategy fails with NPE in custom partition edge

1.10. Documentation Errata

The following section contains late additions or corrections to the product documentation.

- Flume: Kafka Sink

- Hive Sink

- Configuring Pig Scripts to Use HCatalog in Oozie Workflows

- Configuring a Sqoop Action to Use Tez to Load Data into a Hive Table

1.10.1. Flume: Kafka Sink

This is a Flume Sink implementation that can publish data to a Kafka topic. One of the objectives is to integrate Flume with Kafka so that pull-based processing systems can process the data coming through various Flume sources. This currently supports Kafka 0.8.x series of releases.

Property Name	Default	Description
type	-	Must be set to org.apache.flume.sink.kafka.KafkaSink.
brokerList	-	List of brokers Kafka-Sink will connect to, to get the list of topic partitions. This can be a partial list of brokers, but we recommend at least two for HA. The format is a comma separated list of hostname:port.
topic	default-flume-topic	The topic in Kafka to which the messages will be published. If this parameter is configured, messages will be published to this topic. If the event header contains a "topic" field, the event will be published to that topic overriding the topic configured here.
batchSize	100	How many messages to process in one batch. Larger batches improve throughput while adding latency.
requiredAcks	1	How many replicas must acknowledge a message before it is considered successfully written. Accepted values are 0 (Never wait for acknowledgement), 1 (wait for leader only), -1 (wait for all replicas) Set this to -1 to avoid data loss in some cases of leader failure.
Other Kafka Producer Properties	-	These properties are used to configure the Kafka Producer. Any producer property supported by Kafka can be used. The only requirement is to prepend the property name with the prefix "Kafka.". For example: kafka.producer.type.

Note: Kafka Sink uses the topic and key properties from the FlumeEvent headers to send events to Kafka. If the topic exists in the headers, the event will be sent to that specific

topic, overriding the topic configured for the Sink. If key exists in the headers, the key will used by Kafka to partition the data between the topic partitions. Events with same key will be sent to the same partition. If the key is null, events will be sent to random partitions.

An example configuration of a Kafka sink is given below. Properties starting with the prefix Kafka (the last 3 properties) are used when instantiating the Kafka producer. The properties that are passed when creating the Kafka producer are not limited to the properties given in this example. It is also possible include your custom properties here and access them inside the preprocessor through the Flume Context object passed in as a method argument.

```
a1.sinks.k1.type = org.apache.flume.sink.kafka.KafkaSink a1.sinks.k1.topic =
 mytopic
a1.sinks.k1.brokerList = localhost:9092
a1.sinks.k1.requiredAcks = 1
a1.sinks.k1.batchSize = 20
a1.sinks.k1.channel = c1
```

1.10.2. Hive Sink

This sink streams events containing delimited text or JSON data directly into a Hive table or partition. Events are written using Hive transactions. As soon as a set of events are committed to Hive, they become immediately visible to Hive queries. Partitions to which flume will stream to can either be pre-created or, optionally, Flume can create them if they are missing. Fields from incoming event data are mapped to corresponding columns in the Hive table.

Property Name	Default	Description
channel	–	
type	–	The component type name, needs to be hive.
hive.metastore	–	Hive metastore URI (e.g. thrift://a.b.com:9083).
hive.database	–	Hive database name .
hive.table	–	Hive table name.
hive.partition	–	Comma separated list of partition values identifying the partition to write to. May contain escape sequences. E.g.: If the table is partitioned by (continent: string, country :string, time : string) then 'Asia,India,2014-02-26-01-21' will indicate continent=Asia,country=India,time=2014-02-26-01-21.
hive.txnsPerBatchAsk	100	Hive grants a batch of transactions instead of single transactions to streaming clients like Flume. This setting configures the number of desired transactions per Transaction Batch. Data from all transactions in a single batch end up in a single file. Flume will write a maximum of batchSize events in each transaction in the batch. This setting in conjunction with batchSize provides control

Property Name	Default	Description
		over the size of each file. Note that eventually Hive will transparently compact these files into larger files.
heartBeatInterval	240	(In seconds) Interval between consecutive heartbeats sent to Hive to keep unused transactions from expiring. Set this value to 0 to disable heartbeats.
autoCreatePartitions	true	Flume will automatically create the necessary Hive partitions to stream to.
batchSize	15000	Max number of events written to Hive in a single Hive transaction.
maxOpenConnections	500	Allow only this number of open connections. If this number is exceeded, the least recently used connection is closed.
callTimeout	10000	(In milliseconds) Timeout for Hive & HDFS I/O operations, such as openTxn, write, commit, abort.
serializer	–	Serializer is responsible for parsing out field from the event and mapping them to columns in the hive table. Choice of serializer depends upon the format of the data in the event. Supported serializers: DELIMITED and JSON.
roundUnit	minute	The unit of the round down value - second, minute or hour.
roundValue	1	Rounded down to the highest multiple of this (in the unit configured using hive.roundUnit), less than current time.
timeZone	Local	Name of the timezone that should be used for resolving the escape sequences in partition, e.g. Time America/Los_Angeles.
useLocalTimeStamp	false	Use the local time (instead of the timestamp from the event header) while replacing the escape sequences.

Following serializers are provided for Hive sink:

- **JSON:** Handles UTF8 encoded Json (strict syntax) events and requires no configuration. Object names in the JSON are mapped directly to columns with the same name in the Hive table. Internally uses org.apache.hive.hcatalog.data.JsonSerDe but is independent of the Serde of the Hive table. This serializer requires HCatalog to be installed.

- **DELIMITED:** Handles simple delimited textual events. Internally uses LazySimpleSerde but is independent of the Serde of the Hive table.

Property Name	Default	Description
serializer.delimiter	,	(Type: string) The field delimiter in the incoming data. To use special characters, surround them with double quotes like "\t".
serializer.fieldnames	–	The mapping from input fields to columns in hive table. Specified as a comma separated list (no spaces) of

Property Name	Default	Description
		hive table columns names, identifying the input fields in order of their occurrence. To skip fields leave the column name unspecified. E.g.. 'time,,IP,message' indicates the 1st, 3rd and 4th fields in input map to time, IP and message columns in the hive table.
serializer.serdeSeparator	Ctrl-A	(Type: character) Customizes the separator used by underlying serde. There can be a gain in efficiency if the fields in serializer.fieldnames are in same order as table columns, the serializer.delimiter is same as the serializer.serdeSeparator and number of fields in serializer.fieldnames is less than or equal to number of table columns, as the fields in incoming event body do not need to be reordered to match order of table columns. Use single quotes for special characters like '\t'. Ensure input fields do not contain this character. **Note:** If serializer.delimiter is a single character, preferably set this to the same character.

The following are the escape sequences supported:

Alias	Description
%{host}	Substitute value of event header named "host". Arbitrary header names are supported.
%t	Unix time in milliseconds .
%a	Locale's short weekday name (Mon, Tue, ...)
%A	Locale's full weekday name (Monday, Tuesday, ...)
%b	Locale's short month name (Jan, Feb, ...)
%B	Locale's long month name (January, February, ...)
%c	Locale's date and time (Thu Mar 3 23:05:25 2005)
%d	Day of month (01)
%D	Date; same as %m/%d/%y
%H	Hour (00..23)
%I	Hour (01..12)
%j	Day of year (001..366)
%k	Hour (0..23)
%m	Month (01..12)
%M	Minute (00..59)
%p	Locale's equivalent of am or pm
%s	Seconds since 1970-01-01 00:00:00 UTC
%S	Second (00..59) %y last two digits of year (00..99)
%Y	Year (2015)
%z	+hhmm numeric timezone (for example, -0400)

Example Hive table:

```
create table weblogs ( id int , msg string )

partitioned by (continent string, country string, time string)
clustered by (id) into 5 buckets
stored as orc;
```

Example for agent named a1:

```
a1.channels = c1

a1.channels.c1.type = memory

a1.sinks = k1

a1.sinks.k1.type = hive

a1.sinks.k1.channel = c1

a1.sinks.k1.hive.metastore = thrift://127.0.0.1:9083
a1.sinks.k1.hive.database = logsdb

a1.sinks.k1.hive.table = weblogs
a1.sinks.k1.hive.partition = asia,%{country},%y-%m-%d-%H-%M
a1.sinks.k1.useLocalTimeStamp = false

a1.sinks.k1.round = true

a1.sinks.k1.roundValue = 10

a1.sinks.k1.roundUnit = minute

a1.sinks.k1.serializer = DELIMITED
a1.sinks.k1.serializer.delimiter = "\t"
a1.sinks.k1.serializer.serdeSeparator = '\t'
a1.sinks.k1.serializer.fieldnames =id,,msg
```

Note: For all of the time related escape sequences, a header with the key "timestamp" must exist among the headers of the event (unless useLocalTimeStampis set to true). One way to add this automatically is to use the TimestampInterceptor.

The above configuration will round down the timestamp to the last 10th minute. For example, an event with timestamp header set to 11:54:34 AM, June 12, 2012 and 'country' header set to 'india' will evaluate to the partition (continent='asia',country='india',time='2012-06-12-11-50'. The serializer is configured to accept tab separated input containing three fields and to skip the second field.

1.10.3. Configuring Pig Scripts to Use HCatalog in Oozie Workflows

To access HCatalog with a Pig action in an Oozie workflow, you need to modify configuration information to point to the Hive metastore URIs.

There are two methods for providing this configuration information. Which method you use depends upon how often your Pig scripts access the HCatalog.

1.10.3.1. Configuring Individual Pig Actions to Access HCatalog

If only a few individual Pig actions access HCatalog, do the following:

1. Identify the URI (host and port) for the Thrift metastore server.

   a. In Ambari, click **Hive > Configs > Advanced**.

   b. Make note of the URI in the **hive.metastore.uris** field in the General section.

      This information is also stored in the `hive.default.xml` file.

2. Add the following two properties to the <configuration> elements in each Pig action.

   > **Note**
   >
   > Replace `[host:port(default:9083)]` in the example below with the host and port for the Thrift metastore server.

```
<configuration>
    <property>
        <name>hive.metastore.uris</name>
        <value>thrift://[host:port(default:9083)]</value>
        <description>A comma separated list of metastore uris the client can
 use to contact the
        metastore server.</description>
    </property>
    <property>
        <name>oozie.action.sharelib.for.pig</name>
        <value>pig,hive,hcatalog</value>
        <description>A comma separated list of libraries to be used by the
 Pig action.</description>
    </property>
</configuration>
```

1.10.3.2. Configuring All Pig Actions to Access HCatalog

If all of your Pig actions access HCatalog, do the following:

1. Add the following line to the `job.properties` files, located in your working directory:

```
oozie.action.sharelib.for.pig=pig,hive,hcatalog
<!-- A comma separated list of libraries to be used by the Pig action.-->
```

2. Identify the URI (host and port) for the Thrift metastore server.

   a. In Ambari, click **Hive > Configs > Advanced**.

   b. Make note of the URI in the **hive.metastore.uris** field in the General section.

      This information is also stored in the `hive.default.xml` file.

3. Add the following property to the <configuration> elements in each Pig action.

> **Note**
>
> Replace *[host:port(default:9083)]* in the example below with the host and port for the Thrift metastore server.

```
<configuration>
    <property>
        <name>hive.metastore.uris</name>
        <value>thrift://[host:port(default:9083)]</value>
        <description>A comma separated list of metastore uris the client can
 use to contact the
        metastore server.</description>
    </property>
    </configuration>
```

1.10.4. Configuring a Sqoop Action to Use Tez to Load Data into a Hive Table

You can use the Tez execution engine to load data into a Hive table using the `--hive-import` option,

In the code example in each step, replace the sample text in [square brackets] with the appropriate information for your configuration.

1. Create a workflow directory.

   ```
   hdfs dfs -mkdir -p [/user/dummy/app]
   ```

2. Create a `lib` directory in the workflow directory.

   ```
   hdfs dfs -mkdir -p [/user/dummy/app/lib]
   ```

3. Copy the database JDBC driver jar file to the `lib` directory.

   ```
   hadoop fs -copyFromLocal [/usr/share/java/mysql-connector-java.jar]
           [/user/dummy/app/lib]
   ```

4. Copy the `hive-site.xml` and `tez-site.xml` files to a location accessible by the workflow. For example:

   ```
   hadoop fs -copyFromLocal [/etc/oozie/conf/action-conf/hive/hive-site.xml /
   user/dummy/app]
   hadoop fs -copyFromLocal [/etc/oozie/conf/action-conf/hive/tez-site.xml /
   user/dummy/app]
   ```

5. In the Sqoop action of the workflow, do the following:

   • Add `hive-site` and `tez-site` resources in the `<file>` element of the Sqoop action in the workflow.

   ```
   <file>/user/dummy/app/hive-site.xml#hive-site.xml</file>
   <file>/user/dummy/app/tez-site.xml#tez-site.xml</file>
   ```

   • Include the `--hive-import` option in the `<command>` element.

   ```
   <command>import --connect [jdbc:mysql://db_host:port/database] --
   username [user]
   ```

```
   --password [pwd] --driver c[om.mysql.jdbc.Driver] --table [table_name]
--hive-import -m 1 </command>
```

6. Add the following into the `job.properties` file.

```
oozie.use.system.libpath=true
oozie.action.sharelib.for.sqoop=sqoop,hive
```

More information regarding the Sqoop parameters can be found in
the Apache documentation at https://sqoop.apache.org/docs/1.4.6/
SqoopUserGuide.html#_importing_data_into_hive

Example Workflow Action

Replace all sample text in [square brackets] in the example below with the appropriate
workflow name, URI, paths, file names, etc. for your configuration.

```
<action name="sqoop-node">
      <sqoop xmlns="uri:oozie:sqoop-action:0.2">
            <job-tracker>${jobTracker}</job-tracker>
            <name-node>${nameNode}</name-node>
            <configuration>
                  <property>
                        <name>mapred.job.queue.name</name>
                        <value>${queueName}</value>
                  </property>
            </configuration>
            <command>import --connect [jdbc:mysql://db_host:port/database] --
username [user]
--password [pwd] --driver [com.mysql.jdbc.Driver] --table [table_name] --hive-
import -m 1</command>
            <file>[/user/dummy/app/hive-site.xml#hive-site.xml]</file>
            <file>[/user/dummy/app/tez-site.xml#tez-site.xml]</file>
      </sqoop>
      <ok to="end"/>
      <error to="killJob"/>
</action>
```