Pivotal™

Rev: A02
Updated: July 15, 2013

Welcome to Pivotal HD Enterprise

Pivotal HD Enterprise is an enterprise-capable, commercially supported distribution of Apache Hadoop 2.0 packages targeted to traditional Hadoop deployments.

The Pivotal HD Enterprise product enables you to take advantage of big data analytics without the overhead and complexity of a project built from scratch. Pivotal HD Enterprise is Apache Hadoop that allows users to write distributed processing applications for large data sets across a cluster of commodity servers using a simple programming model. This framework automatically parallelizes Map Reduce jobs to handle data at scale, thereby eliminating the need for developers to write scalable and parallel algorithms.

For more information, visit the `Apache Hadoop home` page.

About Pivotal, Inc.

Greenplum is currently transitioning to a new corporate identity (Pivotal, Inc.). We estimate that this transition will be completed in 2013. During this transition, there will be some legacy instances of our former corporate identity (Greenplum) appearing in our products and documentation. If you have any questions or concerns, please do not hesitate to contact us through our web site: `http://gopivotal.com/about-pivotal/support`.

About Pivotal HD Enterprise 1.0.1

Pivotal HD Enterprise 1.0.1 is an open source Apache stack and includes the following components:

- **Hadoop 2.0.2-alpha:** HDFS and MapReduce
- **Yarn**: Next-generation Hadoop data-processing framework.
- **MapReduce V1**: An alternative to Yarn.
- **Pig 0.10.1**: Procedural language that abstracts lower level MapReduce.
- **Hive 0.9.1**: Data warehouse infrastructure built on top of Hadoop.
- **HBase 0.94.2**: Database for random real time read/write access.
- **Mahout 0.7**: Scalable machine learning and data mining library.
- **Zookeeper 3.4.5**: Hadoop centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services.

- **Flume 1.3.1**: A tool used for collecting and aggregating data from multiple sources to a centralized data store.
- **Sqoop 1.4.2**: A tool for transferring bulk data between Apache Hadoop and structured datastores.
- **HVE** (Hadoop Virtualization Extension): Helps Hadoop to be truly aware of underlying vSphere virtualization. HVE allows Hadoop clusters implemented on virtualized infrastructure full awareness of the topology on which they are running, thus enabling the elastic, reliable operation and high performance of these deployments. To learn how to install, configure, and manage HVE, see the *Pivotal HD Enterprise 1.0 Installation and User Guide*.
- **Vaidya**: A performance diagnostic tool for MapReduce jobs.
- **Snappy**: Snappy a compression/decompression library.
- **HVE Elastic Resource Extension**: Enables scalable, on-demand, sharing of resources in a virtual enviroment.

Pivotal components in this release:

- **Pivotal Data Loader 2.0.1**: High-speed data ingest tool for your Pivotal HD cluster
- **USS Beta**: Unified Storage System, a framework that provides HDFS protocol layer on top of external file systems.

Other components included in this release:

- **Spring Hadoop 1.0.1.RC1**: Integrates with the Spring Framework to create and run Hadoop MapReduce, Hive, and Pig jobs as well as work with HDFS and HBase.

Refer to the *Pivotal HD Enterprise 1.0 Installation and User Guide f*or further details related to these components and other related points to this software release.

About Pivotal Advanced Database Services and HAWQ

- Pivotal Advanced Database Services (ADS) powered by HAWQ, is an optional add-on to Pivotal HD that extends Pivotal Hadoop (HD) Enterprise, adding rich, proven parallel SQL processing facilities.
- HAWQ is a parallel SQL query engine that combines the merits of the Greenplum Database Massively Parallel Processing (MPP) relational database engine and the Hadoop parallel processing framework.

Requirements

- Java: The Oracle JDK 1.6 is required to be installed prior to a cluster installation. Instructions for checking for, and downloading the Oracle JDK are included in the installation process described in the *Pivotal HD Enterprise 1.0 Installation and User Guide*.

Supported Platforms

- Red Hat Enterprise Linux (RHEL): 6.1 - 64 bit; 6.2 - 64 bit
- CentOs: 6.1 - 64 bit; 6.2 - 64 bit

Installation Notes

For a brief summary of the contents of this release and Getting Started instructions, refer to the *readme.txt* file.

Pivotal Command Center (PCC) provides command line tools and a Web-based user interface for installing and upgrading, monitoring, and management of Pivotal HD, as such, it needs to be installed first. To install Pivotal Command Center, refer to the *Pivotal Command Center 2.0 Installation and User Guide*. Once PCC is successfully installed and configured, proceed with installation of Pivotal HD by following the instructions in the *Pivotal HD Enterprise 1.0 Installation and User Guide*.

Pivotal HD Enterprise 1.0.1 is made up of the following tar files:

- `PHD-1.0.1.0-19.tar.gz`
- `PHDTools-1.0.1-18.tar.gz`
- `PCC-2.0.1.84.121.163.x86_64.tar.gz`
- `PADS-1.1.0-8.tar.gz` (Optional additional purchase)
- `PHDMR1-1.0.1.0-45.tar.gz` (Optional - an alternative to Yarn)

  Pivotal HD 1.0.1 supports YARN (MR2) resource manager by default. For those customers who don't want to deploy a YARN based cluster, we provide MR1 as optional manually-installable software, instructions for which can be found in the *Pivotal HD 1.0 Stack and Tool Reference Guide*. Note that since MR1 needs to be installed manually, you won't be able to use Pivotal Command Center for monitoring and management of the cluster.

We have also provided the following non-rpm tar files for those customers who are unable to perform RPM installs:

- `PHD-1.0.1.0-bin-3.tar.gz`
- `PHDTools-1.0.1-bin-18.tar.gz`
- `PHDMR1-1.0.1.0-bin-3.tar.gz`
- `PADS-1.1.0.8.tar.gz`

Pivotal Command Center 2.0 does **not** currently support the installation of the following Pivotal HD components.

- **Flume, Sqoop, HVE**: See the *Pivotal HD 1.0 Stack and Tool Reference Guide* for manual installation information.
- **Pivotal DataLoader**: See the *Greenplum DataLoader 1.0 User Guide* for details.

- **Spring_data_hadoop**: Use TAR to install. Once you expand the file, find the `spring-data-hadoop-1.0.1.RC14-docs.zip`. The installation instructions are in the zip archive.
- **MRv1**: MapReduce version 1. See the *Pivotal HD Stack and Tool Reference Guide* for more details about installation and use.

You can download Pivotal HD Enterprise 1.0.1 from the EMC Download Center.

Upgrade Notes

We recommend backing up your data before performing any upgrades.

Instructions for upgrading components using Pivotal HD Manager are provided in the *Pivotal HD Enterprise 1.0 Installation and Administrator Guide*.

Instructions for manually upgrading Pivotal HAWQ are provided in the *Pivotal HAWQ Release Notes*.

Apache Patches

Pivotal HD 1.0.1 includes the following patches that were not included in the 1.0 release.

Component	Patch
Hadoop	HADOOP-6762: Exception while doing RPC I/O closes channel
	HADOOP-7206: Added Snappy compression
	HADOOP-8515: Upgrade to Jetty 7, including upgrades/build process changes to Hive, Hbase, Pig, Sqoop
	HADOOP-8843: Old trash directories are never deleted on upgrade from 1.x
	HADOOP-8906: Paths with multiple globs are unreliable
	HADOOP-8962: RawLocalFileSystem.listStatus fails when a child filename contains a colon
	HADOOP-8986: Server$Call object is never released after it is sent
	HADOOP-9041: FileSystem initialization can go into infinite loop
	HADOOP-9093: Move all the Exception in PathExceptions to o.a.h.fs package
	HADOOP-9105: FsShell -moveFromLocal erroneously fails
	HADOOP-9215: When using cmake-2.6, libhadoop.so doesn't get created (only libhadoop.so.1.0.0)
	HADOOP-9255: Relnotes.py missing last jira
	HADOOP-9260: Hadoop version may be not correct when starting name node or data node
	HADOOP-9289: FsShell rm -f fails for non-matching globs
	HADOOP-9299: Kerberos name resolution is kicking in even when kerberos is not configured
	HADOOP-9444: $var shell substitution in properties are not expanded in hadoop-policy.xml
HDFS	HDFS-3598: WebHDFS support for file concat
	HDFS-4104: dfs -test -d prints inappropriate error on nonexistent directory
	HDFS-4456: Add concat to HttpFS and WebHDFS REST API docs

Component	Patch
Yarn	YARN-127: Move RMAdmin tool to the client package
	YARN-179: Bunch of test failures on trunk
	YARN-189: Deadlock in RM - AMResponse object
	YARN-212: NM state machine ignores an APPLICATION_CONTAINER_FINISHED event when it shouldn't
	YARN-217: Yarn rmadmin commands fail in secure cluster
	YARN-320: RM should always be able to renew its own tokens
	YARN-325: RM CapacityScheduler can deadlock when getQueueInfo() is called and a container is completing
	YARN-354: WebAppProxyServer exits immediately after startup
	YARN-355: RM app submission jams under load
	YARN-429: Capacity-scheduler config missing from yarn-test artifact
MapReduce	MAPREDUCE-4730: AM crashes due to OOM while serving up map task completion events
	MAPREDUCE-4740: Only .jars can be added to the Distributed Cache classpath
	MAPREDUCE-4748: Invalid event: T_ATTEMPT_SUCCEEDED at SUCCEEDED
	MAPREDUCE-4782: NLineInputFormat skips first line of last InputSplit
	MAPREDUCE-4819: AM can rerun job after reporting final job status to the client
	MAPREDUCE-4894: Renewal / cancellation of JobHistory tokens
	MAPREDUCE-4921: JobClient should acquire HS token with RM principal

Earlier Patches

Component	Patch
MapReduce	MAPREDUCE-2454: Allow external sorter plugin for MR
HDFS	HDFS-3077: Quorum-based protocol for reading and writing edit logs
Hadoop	HADOOP-8816: HTTP Error 413 full HEAD if using kerberos authentication
Hive	HIVE-2800: NPE in "create index" without comment clause in external metastore
N/A	Pivotal's HDFS truncate capability for HAWQ

Resolved Issues in Pivotal HD Enterprise 1.0.1

This section lists issues that have been resolved in Pivotal HD since release 1.0. A work-around is provided where applicable.

Table 1 All Resolved Issues in Pivotal HD Enterprise 1.0.1

Component	Issue	Description
DataLoader	HD-1947	Web page formats incorrectly when Chinese characters are shown.
Data Loader	HD-1897	A file cannot be transferred through the command line without a specification XML file.
Stack issues	HD-1704	Uninstallation of Hadoop, HBase, Hive, Zookeeper or Pig will remove all the configuration files under /etc/gphd

Known Issues in Pivotal HD Enterprise 1.0.1

This section lists the known issues in Pivotal HD Enterprise. A work-around is provided where applicable.

Table 2 All Known Issues in Pivotal HD Enterprise 1.0.1

Component	Issue	Description
General	N/A	The Apache Hadoop 2.0.2-alpha stack may not be reliable for mission-critical applications.
	HD-2339	Security warning when short-circuit read is not allowed. MapReduce jobs continue correctly by reading through HDFS.
	HD-2909	nmon does not monitor when there are multiple clusters. Workaround: After the second cluster install perform the following from the Admin node: Copy `/etc/nmon/conf/nmon-site.xml` to all the cluster hosts (same location) `massh hostfile verbose 'sudo service nmon restart'` (hostfile must contain all the existing cluster hosts)
	N/A	Client hosts cannot be used to enter hbase or hive shell (unless they are configured as one of the hbase or hive roles as well)
	N/A	If Hive support for HAWQ is required then the Hive server needs to be collocated with namenode. This restriction is due to a known bug which will be fixed in the future releases.
	HD-5246	Secure PHD HBase configuration. There are unresolved issues with Java DNS lookup failures when trying to set up the secure Zookeeper for HBase to use secured HDFS.
	HD-5048	RPM upgrade from PHD1.0 to PHD1.0.1 is not supported.
	HD-5283	Yarn nodemanager does not decommission after adding it into yarn exclude host list.
	HD-5110	Vaidya Report is not available in MR1.

Table 2 All Known Issues in Pivotal HD Enterprise 1.0.1

Component	Issue	Description
USS	HD-1705	`Mount Point not defined` error when run command `-copyFromLocal` to FTP through USS.
	HD-1772	`Seek not supported` error when running wordcount example using a FTP input
Data Loader	N/A	Streaming job configuration is only supported through the command-line interface.
	HD-2922	Data Loader may fail to bind to port 0.0.0.0:12320 due to a previous installation. User encounters error while adding datastores or creating jobs. Scheduler log will show the port bind error message. **Workaround**: Manually kill the Data Loader process listening to the port (find the PID through netstat -anop|grep 12320) and restart the Data Loader scheduler.
	HD-5319	HDFS2 DataStore only supports Apache Hadoop 2.0.2-alpha, Pivotal HD 1.0 and 1.0.1.
	HD-5310	You must mount NFS on both master and slave machines, and have same mount directory.
	HD-5097	LocalFS data stores do no work with default Yarn schedulers. In distributed mode, for localfs data stores, you must replace the YARN fairscheduler with DataLoader's modified scheduler. For Hadoop 2.x clusters, you must also restart the cluster.
	HD-5390	Localfs job hangs when Hadoop NodeManager Host name is not FQHN. Workaround: Reset the hostname of NodeManager to FQHN, and restart NodeManager.

Pivotal HD Enterprise Documentation

The following Pivotal HD Enterprise and related documentation is available in PDF format on our website at `www.gopivotal.com`.

Additionally, you can still access product documentation from EMC's `Support Zone`:

Table 3 Pivotal HD Enterprise 1.0.1 and related Documentation

Title	Revision
Pivotal HD Enterprise 1.0 Installation and Administrator Guide	A05
Pivotal HD Enterprise 1.0.1 Release Notes (this document)	A02
Pivotal Command Center 2.0 Installation and User Guide	A03
Pivotal HD DataLoader 2.0 Installation and User Guide	A02
Pivotal HD Stack and Tool Reference Guide	A03
HAWQ 1.1.0.1 Installation Guide	A06
Pivotal ADS 1.1 Administrator Guide	A03

Use of Open Source

This product may be distributed with open source code, licensed to you in accordance with the applicable open source license. If you would like a copy of any such source code, EMC will provide a copy of the source code that is required to be made available in accordance with the applicable open source license. EMC may charge reasonable shipping and handling charges for such distribution. Please direct requests in writing to EMC Legal, 176 South St., Hopkinton, MA 01748, ATTN: Open Source Program Office.