

# Pivotal HD Enterprise

Version 2.0.1

Rev: A01 – April 30, 2014

This document provides information related to the Pivotal HD Enterprise 2.0.1 release. It includes the following topics:

- Welcome to Pivotal HD Enterprise
- PHD Components
  - Core Apache Stack
  - Pivotal and Other Components
- Requirements
- What's New
- Installation Notes
- Upgrade Notes
- Additions to Apache
  - Apache Patches
  - Pivotal Apache Modification
- Resolved Issues
- Known Issues
- Versioning and Compatibility
  - Pivotal and Other Components
  - Apache
- Pivotal HD Enterprise Documentation

# Copyright

---

Copyright © 2014 Pivotal Software, Inc. All Rights reserved.

Pivotal Software, Inc. believes the information in this publication is accurate as of its publication date. The information is subject to change without notice. THE INFORMATION IN THIS PUBLICATION IS PROVIDED "AS IS." Pivotal Software, Inc. ("Pivotal") MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND WITH RESPECT TO THE INFORMATION IN THIS PUBLICATION, AND SPECIFICALLY DISCLAIMS IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Use, copying, and distribution of any Pivotal software described in this publication requires an applicable software license.

All trademarks used herein are the property of Pivotal or their respective owners.

## **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, Pivotal will provide a copy of the source code that is required to be made available in accordance with the applicable open source license. Pivotal may charge reasonable shipping and handling charges for such distribution.

## **About Pivotal Software, Inc.**

Greenplum transitioned to a new corporate identity (Pivotal, Inc.) in 2013. As a result of 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>.

## Welcome to Pivotal HD Enterprise

---

Pivotal HD Enterprise is an enterprise-capable, commercially supported distribution of Apache Hadoop 2.2 packages targeted to traditional Hadoop deployments.

Pivotal HD Enterprise 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 about Apache Hadoop, see the Apache Hadoop home page: <http://hadoop.apache.org/>

## PHD Components

---

Pivotal HD Enterprise 2.0.1 includes the following open source Apache stack and other components:



For specific version numbers for all components, see [Versioning and Compatibility](#) later in this document.

## Core Apache Stack

---

Component	Description
Hadoop	HDFS: A Hadoop distributed file system (HDFS). YARN: Next-generation Hadoop data-processing framework.
Pig	Procedural language that abstracts lower level MapReduce.
Hive	Data warehouse infrastructure built on top of Hadoop.
HCatalog	HCatalog is a table and storage management layer for Hadoop that enables users with different data processing tools – Pig, MapReduce, and Hive – to more easily read and write data on the grid.
HBase	Database for random real time read/write access.
Mahout	Scalable machine learning and data mining library.
Zookeeper	Hadoop centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services.
Flume	A tool used for collecting and aggregating data from multiple sources to a centralized data store.
Sqoop	A tool for transferring bulk data between Apache Hadoop and structured datastores.

Component	Description
Oozie	A workflow scheduler system to manage Apache Hadoop jobs. Oozie Workflow jobs are Directed Acyclical Graphs (DAGs) of actions. Oozie Coordinator jobs are recurrent Oozie Workflow jobs triggered by time (frequency) and data availability.


## Pivotal and Other Components

Component	Description
Pivotal Command Center	A command line and web-based tool for installing, managing and monitoring your Pivotal HD cluster.
Pivotal HAWQ	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.
Pivotal HAWQ - PXF	Extensibility layer to provide support for external data formats such as HBase and Hive.
Pivotal Real Time Services (PRTS)	Pivotal HD 2.0 includes support for GemFireXD (GFXD) 1.0, an offering of PRTS.
* Hamster	Developed by Pivotal, Hamster (beta) is a framework which enable users running MPI programs on Apache Hadoop YARN platform.  (OpenMPI is a A High Performance Message Passing Library)
* GraphLab	GraphLab is a powerful new system for designing and implementing parallel algorithms in machine learning. It is a graph-based, high performance, distributed computation framework written in C++ that makes use of MPI and has its own programming model.  Note that this service is in Beta, so only community support is provided. Interested customers should contact their account managers if they plan to use GraphLab in production.

\* New in PHD 2.0

## Requirements

- Java: The Oracle JDK 1.7 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 2.0 Installation and User Guide.

 PHD 2.x. has been tested with JDK 1.7 (u15)

## What's New

---

In addition to bug fixes and performance and functionality improvements, this release includes the following new and improved features:

- Apache
  - Hadoop has been upgraded to 2.2.0
  - HBase has been upgraded to 0.96.0
  - Hive/HCatalog has been upgraded to 0.12.0
  - Oozie has been upgraded to 4.0.0
  - Flume has been upgraded to 1.4.0
  - HDFS NFS Gateway Support has been added
- Pivotal HAWQ has been upgraded to 1.2.0.1.
- Pivotal's Hamster 1.0 (beta) has been added to this distribution.
- GraphLab 2.2 (beta) has been added to this distribution.
- Binary distribution files: Pivotal is no longer including binary archives of the product with this release. If you happen to depend upon the binary archives as we delivered with previous releases, contact your sales/field representative for assistance.
- MapReduce 1 (MR1): With improved MapReduce compatibility in Hadoop 2.2's YARN, Pivotal is no longer including the Hadoop 1.x MapReduce files with the rest of a Hadoop 2.x based stack.
- DataLoader: Pivotal is no longer shipping the DataLoader tool. For data ingestion, consider SpringXD, Flume, Sqoop or standard HDFS functionality such as HDFS put scripts.
- USS: USS Beta is no longer being shipped with PHD. In a future release it will be subsumed by PXF to support HAWQ queries on remote clusters/file systems
- Spring Data: Spring Data is no longer being shipped with PHD.
- Security Scripts: We provide instructions for manually enabling Kerberos authentication in the *PHD 2.0 Stack and Tools Reference Guide*. We also can provide scripts to automate this process. To obtain these scripts and instructions how to use them, contact either your PHD Account Manager, or open up a service request with support at <https://support.emc.com/> and ask for the PHD Secure Install Tools.
- GemFireXD: GemFireXD has been upgraded to 1.0.

## 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 a command line tool (CLI) and a Web-based user interface for installing and upgrading, monitoring, and management of Pivotal HD, as such, it must be installed first. To install Pivotal Command Center and the other Pivotal HD components via the CLI, follow the instructions in the *Pivotal HD Enterprise 2.0 Installation and User Guide*.

Pivotal HD Enterprise 2.0.1 is made up of the following tar files:

- Pivotal HD Enterprise: `PHD-2.0.1.0-148.tar.gz`
- Pivotal Command Center: `PCC-2.2.1-150.tar.gz`
- Pivotal HAWQ, PXF (Pivotal ADS): `PADS-1.2.0.1-8119.tar.gz`
- Pivotal GemFireXD (PRTS): `PRTS-1.0.0-14.tar.gz`.

These files are available from [Pivotal Network](#) or by contacting your account manager.

Binaries for previous releases are available from [EMC's Download Center](#).

Pivotal Command Center's CLI does not currently support the installation of the following Pivotal HD components, which have to be installed manually.

- Flume, Sqoop, Oozie, Hcatlog, Hamster and GraphLab: See the latest *Pivotal HD 2.0 Stack and Tool Reference Guide* for manual installation information.

## Upgrade Notes

---

- If you are upgrading to a new version of Pivotal HD, make sure you are also upgrading to compatible versions of Pivotal Command Center and Pivotal ADS (optional). See [Versioning and Compatibility](#) for more information)
- We recommend that you always back up your data before performing any upgrades.
- We recommend you upgrade Pivotal HD via the PCC command line interface (CLI) (ICM client). Instructions for upgrading components using the CLI, see the *PHD Installation and Administrator Guide*.
- You cannot upgrade High Availability enabled clusters or secure clusters. Before upgrading, revert clusters to non-secure and non-HA enabled. See the *PHD Installation and Administrator Guide* for details.
- Instructions for manually upgrading Pivotal HAWQ are provided in the *Pivotal HAWQ Installation and Upgrade Guide*.

## Additions to Apache

---

### Apache Patches

---

The following patches were applied to PHD 2.0.1:

Apache Issue	Description
HDFS-5728	[Diskfull] Block recovery will fail if the metafile does not have crc for all chunks of the block
HDFS-5438	Flaws in block report processing can cause data loss
HDFS-3848	A Bug in recoverLeaseInternal method of FSNameSystem class
HDFS-5526	Datanode cannot roll back to previous layout version
BIGTOP-802	Add rollback option to DataNode service script
HIVE-4388	Upgrade HBase to 0.96
PIG-3512	Reducer estimator is broken by PIG-3497
Flume-1618	Make Flume NG build and tests work with Hadoop 2.0 & HBase 0.96
Flume-2172	Update protocol buffer from 2.4.1 to 2.5.0

### Pivotal Apache Modification

---

The following changes were applied to PHD's Apache stack:

Apache Component	Pivotal Issue	Description
HADOOP	HD-6508	Added HVE topology support for Hadoop 2.2.0
HADOOP	HD-8088	Added HVE elasticity support for Hadoop 2.2.0
HADOOP	HD-6938	Added Rack Awareness Support for Hadoop 2.2.0
HADOOP	HD-7669	Added PXF functionality support for Hadoop 2.2.0
HADOOP	HD-6936	Added HAWQ support for Hadoop 2.2.0
HADOOP	HD-7780	Added Vaidya support for Hadoop 2.2.0
HADOOP	HD-6937	Added Jetty 7 support for Hadoop 2.2.0
HBASE	HD-7642	Added Jetty 7 support for HBase 0.96.0
HADOOP	HD-7833	nfs3.server.port in core-site.xml doesn't take effect

## Resolved Issues

This section lists issues that have been resolved in Pivotal HD 2.0.1.



For resolved issues relating to Pivotal Command Center's UI functionality, see the corresponding PCC Release Notes.

Issue	Description
HD-9527	Nodemanager failed to launch more than 4 tasks in parallel in secure mode.
HD-9238	Especially under high volume usage, Yarn can throw text file busy exceptions.
HD-7105	Error messages and product documentation gave the incorrect location of the ScanCluster.xxx.log files.
HD-8005	Polling logs could fill up the log files causing it difficult to debug the actual problems reported in error logs. Polling and gphdmgr-webservices.log logs are now separated.
HD-8353	PCC installation failed if postgresql-server was running and pre-initialized.
HD-6062	Installation can fail due to rpm permission issues.
HD-6063	Null values in the cluster configuration could cause a memory leak. Null values are now dealt with properly; If they are a mandatory field, validation should fail; if optional, ignored.
HD-7867	PHD installation failed due to a puppet synchronization error.
HD-8944	Some job details were not being collected properly due to the inconsistency of history server and polling service. A new API now collects the specified job details from Hadoop and updates the database.
HD-9059	Misleading error message was thrown after <code>icm_client import</code> failure due to missing rpms.
HD-9095	Unable to secure a HA cluster because HDFS was started in the wrong order, before Zookeeper. Zookeeper is now automatically started first.
HD-9184	Cluster service could not be started after PHD stack upgrade.
HD-9355	Typo in <code>hadoop-env.sh</code> file caused following error to be thrown: <code>ERROR Could not find value for key log4j.appender.DRFAAUDIT</code>
HD-2339	Security warning when short-circuit read is not allowed. MapReduce jobs continue correctly by reading through HDFS.
HD-7856	When using QJM, running BootstrapStandby while the existing NN was active could result in an exception.
HD-7010	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.
N/A	Single node installations are now supported but are only recommended for demonstration/POC purposes.
HD-8730	Start/Reconfigure or upgrade actions failed after upgrading from PCC 2.1.0 to a later version.
HD-9062	MapReduce jobs failed after upgrading from PHD 1.1.0 to 1.1.1 due to NodeManager failing to cleanup local directories.



Issue	Description
HD-8712	Zookeeper automatic failover fails as both NameNodes being in Standby mode.

## Known Issues

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



For known issues relating to Pivotal Command Center's UI functionality, see the corresponding *PCC Release Notes*

Component	Issue	Description
HDFS	HD-9912	Users may encounter a rare, but known issue with Apache: JIRA HDFS-5557 - <i>Write pipeline recovery for the last packet in the block may cause rejection of valid replicas</i> . Customers are advised to be aware of this rare HDFS anomaly and should feel free to contact Pivotal Support if they see symptoms similar to the ones mentioned in HDFS-5557. This bug only impacts the Hadoop 2.2 release, and will be resolved once Pivotal HD moves to a Hadoop release later than 2.2.
HDFS Upgrade	HD-7339	If Hbase master is installed all alone without any other Hadoop roles, the upgrade from an older version of stack to newer version fails.  <b>Workaround:</b> Stop Hbase, manually run <code>yum install hadoop-hdfs-&lt;new version&gt;</code> on the Hbase master node and restart Hbase.
Install/Upgrade	CC-3494	Install/Upgrades: If the RHEL ssl certificate for subscription-manager plugin has expired, yum will fail.  <b>Workaround:</b> If <code>yum -list</code> reports an error on any of the cluster nodes then check the yum configurations files, <code>/etc/yum.repos.d/</code> and make sure all remote repositories are DNS resolvable.
Hive	HD-9676	If Hive is part of your PHD deployment; <code>java-1.5.0-gcj-1.5.0.0-29.1.el6.x86_64</code> will be installed on your hive server node overriding the <code>/usr/bin/java</code> to point to <code>java-1.5.0-gcj</code> instead of the Java7 you have provided as part of your base PHD deployment. This can cause issues with services that don't explicitly set the Java path (GemfireXD for example).  <b>Workaround:</b>  Following PHD deployment, set <code>/usr/bin/java</code> as follows:  Either: <pre>ln -f -s /usr/java/default/bin/java /usr/bin/java</pre> or <pre>ln -f -s /usr/java/latest/bin/java /usr/bin/java</pre>

Component	Issue	Description
HD-8960		<p>In environments using LDAP authentication, Namenode can fail when running queries due to JVM crash.</p> <p>Workaround:</p> <p>Add the following parameter to core-site.xml:</p> <pre>&lt;property&gt;   &lt;name&gt;hadoop.security.group.mapping&lt;/name&gt;   &lt;value&gt;org.apache.hadoop.security.ShellBasedUnixGroupsMapping&lt;/value&gt; &lt;/property&gt;</pre>
Sqoop	HD-9429	SQOOP: Import to HBase does not work on a secure cluster.
HBase	HD-9399	Writing to an HBase table throws an IllegalArgumentException, causing job submission to fail.
HBase	HD-9090	Zookeeper exception appears in HBase shell running on a security phd cluster by non-root user
HDFS	HD-9153	Namenode WebUI browse file system fails with secure HDFS.
HDFS NFS	HD-8084	Attempts to copy a large file to a HDFS NFS mounted directory can cause the system to hang.
General	HD-8493	Prepare host command fails if the root password contains certain special characters, for example: \$
General	HD-7296	<p>Some <code>icm_client</code> commands are not supported on FIPS mode-enabled clusters.</p> <p>The following ICM commands work on FIPS mode:</p> <pre>icm_client list, start, stop, preparehosts, scanhosts, import, fetch-template fetch-configuration</pre> <p>The following ICM commands DO NOT work on FIPS mode:</p> <pre>deploy, uninstall, reconfigure, add-slaves, remove-slaves</pre> <p><b>Workaround:</b> Disable FIPS, run the required commands, then renable FIPS.</p>

Component	Issue	Description
General	HD-8926	<p>Single-node installations fail due to a Postgres configuration issue that blocks the ip connection needed to create hive metastore to create schema.</p> <p><b>Workaround:</b></p> <p>After installing PCC, but before cluster deployment, do the following:</p> <ol style="list-style-type: none"> <li>If they do not already exist, add the following lines to <code>/var/lib/pgsql/data/postgresql.conf</code>: <pre>listen_addresses = '*' standard_conforming_strings = off</pre> <p>Notes:</p> <p>If these two lines did not exist they would automatically be added after the deployment operation.</p> <p>Another, <code>listen_address = 'localhost'</code> also exists but is commented out.</p></li> <li>Check that the following lines exist as shown below in <code>/var/lib/pgsql/data/pg_hba.conf</code> <pre># "local" is for Unix domain socket connections only local all all trust  # IPv4 local connections: host all all 0.0.0.0 0.0.0.0 trust  # IPv6 local connections: host all all ::1/128 trust</pre> </li> <li>For your changes to take effect, restart postgres: <pre>sudo /etc/init.d/postgresql restart</pre> </li> </ol>
General	CC-3501	Pivotal Command Center hostnames can only contain lower case letters.
General	HD-2209	<p>After uninstalling a cluster, some of the following RPMs may be left behind:</p> <ul style="list-style-type: none"> <li><code>bigtop-jsvc.x86_64</code></li> <li><code>bigtop-utils.noarch</code></li> <li><code>zookeeper.noarch</code></li> <li><code>zookeeper-server.noarch</code></li> </ul>
General	HD-9126	PHD CLI currently does not support downgrading the Hadoop version on the entire cluster.
General	HD-9105	<p>Sqoop cannot be installed via the PHD CLI.</p> <p>See the <i>Pivotal HD Enterprise Stack and Tool Reference Guide</i> for details on installing Sqoop manually.</p>

Component	Issue	Description
General	HD-9106	<p>Flume cannot be installed via the PHD CLI.</p> <p>See the <i>Pivotal HD Enterprise Stack and Tool Reference Guide</i> for details on installing Flume manually.</p>
General	HD-6149	<p>Reconfiguration changes are not implemented following an upgrade or reconfiguration:</p> <p><b>Workaround:</b> Following an upgrade or reconfiguration, perform the following:</p> <ol style="list-style-type: none"> <li>1. Fetch the new templates that come with the upgraded software by running <code>icm_client fetch-template</code>.</li> <li>2. Retrieve the existing configuration from database using <code>icm_client fetch-configuration</code>.</li> <li>3. Sync the new configurations (<code>hdfs/hadoop-env</code>) from the template directory to the existing cluster configuration directory.</li> <li>4. Upgrade or reconfigure service by specifying the cluster configuration directory with updated contents.</li> </ol>
General	HD-2909	<p>nmon does not monitor when there are multiple clusters.</p> <p><b>Workaround:</b> After the second cluster install perform the following from the Admin node:</p> <p>Copy <code>/etc/nmon/conf/nmon-site.xml</code> to all the cluster hosts (same location)</p> <pre>massh hostfile verbose 'sudo service nmon restart'</pre> <p>(hostfile must contain all the existing cluster hosts)</p>

## Versioning and Compatibility

### Pivotal and Other Components

Product	Version	OS/Browser
Pivotal HD See the table below for Apache Stack component versioning information.	2.0.1	RedHat 64-bit: 6.2, 6.4 CentOS 64-bit: 6.2, 6.4
Pivotal Command Center	2.2.1	RedHat 64-bit: 6.2, 6.4 CentOS 64-bit: 6.2, 6.4 Firefox 21, 22 Chrome Version 28.0.1500.95 IE 9, 10
HAWQ *	1.2.0.1	RedHat 64-bit: 6.2, 6.4 CentOS 64-bit: 6.2, 6.4

Product	Version	OS/Browser
PXF *	2.2	RedHat 64-bit: 6.2, 6.4 CentOS 64-bit: 6.2, 6.4
Hamster	1.0	N/A
GraphLab	2.2	N/A

\* Distributed with Pivotal ADS 1.2

## Apache

---

Component	Version
Hadoop HDFS	2.2.0
Hadoop YARN	2.2.0
Pig	0.12.0
Hive	0.12.0
HBase	0.96.0
Mahout	0.7
Zookeeper	3.4.5
Flume	1.4.0
Sqoop	1.4.2
HCatalog	0.12.0
Oozie	4.0.0

## Pivotal HD Enterprise Documentation

---

The following Pivotal HD Enterprise and related documentation is available in HTML and PDF format on our website at [docs.gopivotal.com/pivotalhd/](https://docs.gopivotal.com/pivotalhd/). Documentation for previous releases is available from EMC's [Support Zone](#).

Title	Revision
Pivotal HD Enterprise 2.0 Installation and Administrator Guide	A02
Pivotal HD Enterprise 2.0.1 Release Notes (this document)	A01
Pivotal Command Center 2.2 User Guide	A02
Pivotal HD 2.0 Stack and Tool Reference Guide	A02
Pivotal HAWQ 1.2 Administrator Guide	A02
Pivotal HAWQ 1.2 Installation Guide	A02

Title	Revision
Pivotal Extension Framework 2.2 Installation and User Guide	A02