Pivotal

PRODUCT DOCUMENTATION

# Pivotal HD Enterprise

Version 1.1.1

## Release Notes

Rev: A02

# Table of Contents

**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.

# 1 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, here: http://hadoop.apache.org/

# 2 PHD Components

Pivotal HD Enterprise 1.1.1 includes the following open source Apache stack and additional Pivotal components as listed below:

## 2.1 Core Apache Stack

| Component | Description |
|---|---|
| Hadoop (MR2) | HDFS: A Hadoop distributed file system (HDFS).<br><br>YARN: Next-generation Hadoop data-processing framework. |
| Hadoop (MR1) | HDFS: A Hadoop distributed file system (HDFS).<br><br>MapReduce: A system for parallel processing of large data sets.<br><br>⚠ PHD 1.x is based upon Hadoop 2.0.5-alpha, if you don't want to install a YARN-based/Hadoop 2.0.5-alpha-base cluster, we have included the MR1 files from our PHD 1.1 release. The MR1 stack must be installed manually and cannot be mixed with components from a YARN-based cluster. |
| 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. |
| 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. |

# 2.2 Pivotal Components

| Component | Description |
|---|---|
| Pivotal Command Center | A command line and web-based tool for installing, managing and monitoring your Pivotal HD cluster. |
| Pivotal DataLoader | High-speed data ingest tool for your Pivotal HD cluster. |
| USS | Unified Storage System, a framework that provides HDFS protocol layer on top of external file systems. |
| Spring Data | Spring for Apache Hadoop provides support for writing Apache Hadoop applications that benefit from the features of Spring, Spring Batch, and Spring Integration. |
| Pivotal ADS - 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 ADS - PXF | Extensibility layer to provide support for external data formats such as HBase and Hive. |
| Pivotal CF Service Broker | A Cloud Foundry Service Broker installable on non-CF PHD clusters. This broker, when installed and configured, allows Cloud Foundry applications to bind to and use non-CF PHD clusters. |
| Pivotal Real Time Services (PRTS) | Pivotal HD 1.1 includes support for GemFire XD (GFXD) Beta 2, an offering of PRTS. |

# 3 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 1.1 Installation and User Guide.

JDK 1.6 is optional but has not been tested with PHD 1.1.x

# 4 What's New

This release includes new and improved features, bug fixes and performance and functionality improvements:

- The CLI (`icm_client`) now supports `-restart`. This stops, then restarts a cluster. For complete syntax see the *Pivotal HD Enterprise Installation and User Guide*.
- You can now setup an automatic failover mechanism via the CLI.
- This release includes a Cloud Foundry Service Broker installable on non-CF PHD clusters. This broker, when installed and configured, allows Cloud Foundry applications to bind to and use non-CF PHD clusters.
- This release of PHD supports an upgrade from the previous release.
- Pig has been upgraded to 0.12
- Piggy Bank (a place for Pig users to share functions) is supported in Pig 0.12.
- Hive Server 2 now supports security

# 5 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 1.1 Installation and User Guide*.

Pivotal HD Enterprise 1.1 is made up of the following tar files:

> ⚠ We provide both rpm and non-rpm tar files for those customers who are unable to perform RPM installs.

- Pivotal HD Enterprise: `PHD-1.1.1.0-82.tar.gz`, `PHD-1.1.1.0-bin-35.tar.gz`
- Pivotal HD Tools (DataLoader and USS): `PHDTools-1.1.1.0-108.tar.gz`, `PHDTools-1.1.1.0-bin-108.tar.gz`
- Pivotal Command Center: `PCC-2.1.1-73.x86_64.tar.gz`
- Pivotal Advanced Database Services (HAWQ, PXF): `PADS-1.1.4-34.tar.gz`, `PADS-1.1.4-bin-34.tar.gz` (Optional additional purchase)
- Pivotal HD Enterprise - MapReduce component from Apache 1.x: `PHDMR1-1.1.0.0-83.tar.gz`, `PHDMR1-1.1.0.0-bin-24.tar.gz` (Optional)

Pivotal HD 1.x supports YARN (MR2) resource manager by default. For those customers who don't want to deploy a YARN-based cluster, we provide MR1 files from our PHD 1.1 release as optional manually-installable software, instructions for which can be found in the *Pivotal HD 1.1 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.

- Pivotal-CF Service Broker.`PHD-CF-BROKER-1.0.0.0-1.tar.gz`.
- Pivotal GemFire XD (PRTS) is available via our gopivotal.com download page or by contacting your account manager.

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: See the *Pivotal HD 1.1 Stack and Tool Reference Guide* for manual installation information.
- Pivotal DataLoader: See the *Pivotal DataLoader 2.0 User Guide* for details.
- Spring Data: Use TAR to install `spring-data-hadoop-1.0.2.RELEASE`. Once you expand the file, find the `spring-data-hadoop-reference.pdf`. The installation instructions are in this pdf file.

- MRv1: MapReduce version 1. See the *Pivotal HD Stack and Tool Reference Guide* for more details about installation and use.

# 6 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 Pivotal HD Manager (ICM client). Instructions for upgrading components using Pivotal HD Manager are provided in the *Pivotal HD Enterprise Installation and Administrator Guide*.
- Instructions for manually upgrading Pivotal HAWQ are provided in the *Pivotal HAWQ Release Notes*.

# 7 Additions to Apache

## 7.1 Apache Patches

The following patches were applied to PHD 1.1:

- HADOOP-7206: Added Snappy compression
- HADOOP-8515: Upgrade to Jetty 7, including upgrades/build process changes to Hive, Hbase, Pig, Sqoop
- HDFS-3848: A Bug in recoverLeaseInternal method of FSNameSystem class
- HBASE-5665: Repeated split causes HRegionServer failures and breaks table
- HIVE-4619: Hive 0.11.0 is not working with pre-cdh3u6 and hadoop-0.23
- YARN-854: App submission fails on secure deploy

## 7.2 Pivotal Apache Modifications

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

- HAWQ - Pivotal’s HDFS truncate capability for HAWQ
- HVE Elastic Resource Extension - Enables scalable, on-demand, sharing of resources in a virtual environment.
- Vaidya - A performance diagnostic tool for MapReduce jobs.
- Snappy - A compression/decompressions library.

# 8 Resolved Issues

This section lists issues that have been resolved in Pivotal HD since release 1.1. A work-around is provided where applicable.

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

| Issue | Description |
|---|---|
| HD-8381 | When upgrading a cluster with approximately 400GB or more data on each node, the puppet process could consume most of the CPU, then hang, causing the upgrade to fail. |
| HD-7010 | Hive server does not need to be on the namenode . |
| HD-7346 | HBase Access Control (grant, revoke, alter, user_permission) were not supported in previous releases. |
| HD-7377 | `pig -useHCatalog` failed to launch mapreduce jobs (binary package). |
| HD-5679 | HDFS operations used the local PHD's block size as opposed to the block size configured in the filesystem where the file resides. |

# 9 Known Issues

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



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

| Component | Issue | Description |
|---|---|---|
| General | HD-8493 | Prepare host command fails if the root password contains certain special characters, for example: $ |
| General | HD-6658 | Webhcat server reports HTTP ERROR while running the REST operation after starting/restarting service (MR1 clusters). |
| General | HD-7296 | Some `icm_client` commands are not supported on FIPS mode-enabled clusters.<br><br>The following ICM commands work on FIPS mode:<br><br>`icm_client list`, `start`, `stop`, `preparehosts`, `scanhosts`, `import`, `fetch-template`, `fetch-configuration`<br><br>The following ICM commands DO NOT work on FIPS mode:<br><br>`deploy`, `uninstall`, `reconfigure`, `add-slaves`, `remove-slaves`<br><br>**Workaround**: Disable FIPS, run the required commands, then renable FIPS. |
| General | N/A | Pivotal Command Center hostnames can only contain lower case letters. |
| General | HD-2209 | After uninstalling a cluster, some of the following RPMs may be left behind:<br><br>• `bigtop-jsvc.x86_64`<br>• `bigtop-utils.noarch`<br>• `zookeeper.noarch`<br>• `zookeeper-server.noarch` |
| General | HD-2477 | Job History web URL is not correct: The resource manager Dashboard UI history link directs to the job history server. As it uses a short hostname instead of a fully qualified domain name as the job history server hostname, it will fail from outside the domain browser. |
| General | N/A | Pivotal CC CLI currently does not support downgrading the Hadoop version on the entire cluster. |

| Component | Issue | Description |
|---|---|---|
| General | N/A | Pivotal DataLoader, Mahout, Sqoop, Flume, and Spring Hadoop cannot be installed via the Pivotal HD CLI. See the *Pivotal HD Enterprise Stack and Tool Reference Guide* for details on installing these components manually. |
| General | HD-6149 | Configuration changes are not implemented following an upgrade or reconfiguration:<br><br>**Workaround**: Following an upgrade or reconfiguration, perform the following:<br><br>1. Fetch the new templates that come with the upgraded software by running `icm_client fetch-template`.<br>2. Retrieve the existing configuration from database using `icm_client fetch-configuration`.<br>3. Sync the new configurations (`hdfs/hadoop-env`) from the template directory to the existing cluster configuration directory.<br>4. Upgrade or reconfigure service by specifying the cluster configuration directory with updated contents. |
| General | N/A | Installation: The `preparehosts --hostfile` command creates the gpadmin user on the cluster nodes. Do NOT create this user manually. If gpadmin user already exists on the cluster nodes, the installation will fail. Delete existing gpadmin users by running:<br><br>`pkill -KILL -u gpadmin`<br><br>`userdel -r gpadmin` |
| General | HD-2273 | Use of file paths in the configuration XML files with a format similar to `file:///path/to/file` will not work due to a puppet handoff issue and error out. |
| General | N/A | The Apache Hadoop 2.0.5-alpha stack may not be reliable for mission-critical applications. |
| General | HD-2339 | Security warning when short-circuit read is not allowed. MapReduce jobs continue correctly by reading through HDFS. |

| Component | Issue | Description |
|---|---|---|
| General | HD-2909 | nmon does not monitor when there are multiple clusters.<br><br>**Workaround**: After the second cluster install perform the following from the Admin node:<br><br>Copy `/etc/nmon/conf/nmon-site.xml` to all the cluster hosts (same location)<br><br>`massh hostfile verbose 'sudo service nmon restart'`<br><br>(hostfile must contain all the existing cluster hosts) |
| General | 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. |
| General | HD-5110 | Vaidya Report is not available in MR1. |
| General | HD-7856 | When QJM is used, running BootstrapStandby while the existing NN is active can result in an exception.<br><br>**Workaround**: Manually sync FSImnage and FSEdit to standby |
| USS | HD-6745 | MapReduce jobs fail with `org.apache.hadoop.util.Shell$ExitCodeException` when input is on the non-secure cluster and output on the secure cluster. |
| USS | HD-6759 | Map-reduce jobs fail with "`SIMPLE authentication is not enabled` Exception when input is on one secure hdfs cluster and output on another secure hdfs cluster. |
| USS | HD-1705 | Mount Point not defined error when run command `-copyFromLocal` to FTP through USS. |
| USS | HD-1772 | `Seek not supported` error when running wordcount example using a FTP input. |
| USS | HD-7140 | `hadoop fs -rm` operation fails when used against path on a remote filesystem.<br><br>**Workaround**: Always use with a `-skipTrash` option. |
| | HD-1625 | `hadoop fs -mv` operation is not supported when source and target locations are on different filesystems. |
| USS | HD-7367 | There are issues when using java FileSystem functions to access remote filesystem through USS. |
| DataLoader | N/A | Streaming job configuration is only supported through the command-line interface. |

| Component | Issue | Description |
|---|---|---|
| DataLoader | HD-5319 | HDFS2 DataStore only supports Pivotal HD 1.1 and Apache Hadoop 2.0.5-alpha, Pivotal HD 1.0 and 1.0.1. |
| DataLoader | HD-5310 | You must mount NFS on both master and slave machines, and have same mount directory. |
| DataLoader | HD-5097 | LocalFS data stores do no work with default Yarn schedulers.<br><br>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. |
| DataLoader | HD-5390 | Localfs job hangs when Hadoop NodeManager Host name is not FQHN.<br><br>**Workaround**: Reset the hostname of NodeManager to FQHN, and restart NodeManager. |

# 10 Versioning and Compatibility

## 10.1 Pivotal

| Product | Version | OS/Browser |
|---|---|---|
| Pivotal HD<br><br>See the table below for Apache Stack component versioning information. | 1.1.1 | RedHat 64-bit: 6.2, 6.4<br><br>CentOS 64-bit: 6.2, 6.4 |
| Pivotal Command Center | 2.1.1 | RedHat 64-bit: 6.2, 6.4<br><br>CentOS 64-bit: 6.2, 6.4<br><br>Firefox 21, 22<br><br>Chrome Version 28.0.1500.95<br><br>IE 9, 10 |
| Pivotal DataLoader | 2.0.4 | RedHat 64-bit: 6.2, 6.4<br><br>CentOS 64-bit: 6.2, 6.4 |
| Pivotal USS | 0.5 | N/A |
| HAWQ * | 1.1.4 | RedHat 64-bit: 6.2, 6.4<br><br>CentOS 64-bit: 6.2, 6.4 |
| PXF * | 2.1.1 | RedHat 64-bit: 6.2, 6.4<br><br>CentOS 64-bit: 6.2, 6.4 |

* Distributed with Pivotal ADS 1.1.4

## 10.2 Apache

| Component | Version |
|---|---|
| Hadoop (MR2) HDFS | 2.0.5-alpha |
| Hadoop (MR2) YARN | 2.0.5-alpha |
| Hadoop (MR1) MapReduce | 1.0.3 |

| Component | Version |
|---|---|
| Pig | 0.12 |
| Hive | 0.11.0 |
| HBase | 0.94.8 |
| Mahout | 0.7 |
| Zookeeper | 3.4.5 |
| Flume | 1.3.1 |
| Sqoop | 1.4.2 |
| HCatalog | N/A |
| Oozie | 3.3.2 |

# 11 Pivotal HD Enterprise Documentation

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

HTML versions of our documentation is available here: docs.gopivotal.com/pivotalhd/

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

| Title | Revision |
|---|---|
| Pivotal HD Enterprise 1.1 Installation and Administrator Guide | A03 |
| Pivotal HD Enterprise 1.1.1 Release Notes (this document) | A02 |
| Pivotal Command Center 2.1 User Guide | A02 |
| Pivotal HD DataLoader 2.0 Installation and User Guide | A06 |
| Pivotal HD 1.1 Stack and Tool Reference Guide | A02 |
| Pivotal HAWQ 1.1 Administrator Guide | A08 |
| Pivotal HAWQ 1.1 Installation Guide | A11 |
| Pivotal Extension Framework Installation and User Guide | A02 |
| Pivotal HD Service Broker for Pivotal CF v1.0.0.0 Release Notes | A01 |