Pivotal™

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 1.1

Please refer to the following sections for more information about this release.

- Components
- What's New
- Requirements
- Installation Notes
- Patches
- Resolved Issues
- Known Issues
- Versioning and Compatibility
- Pivotal HD Enterprise Documentation

Components

Pivotal HD Enterprise 1.1 includes the following open source Apache stack and additional Pivotal components as listed below:

Table A.1 PHD Components

CORE APACHE STACK		
Hadoop (MR2)	HDFS	A Hadoop distributed file system (HDFS).
	Yarn	Next-generation Hadoop data-processing framework.
Hadoop (MR1)	HDFS	A Hadoop distributed file system (HDFS).
	MapReduce	A system for parallel processing of large data sets.
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.
OPTIONAL PIVOTAL COMPONENTS		
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 Beta		Unified Storage System, a framework that provides HDFS protocol layer on top of external file systems.
Sping 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.
	PXF	Extensibility layer to provide support for external data formats such as HBase and Hive.

What's New

Pivotal HD 1.1 includes new and improved features, bug fixes and performance and functionality improvements:

Features:

- **GemFire XD Beta**. This release of Pivotal HD Enterprise integrates with GemFireXD Beta, available separately.

- **High Availability (HA)**. HA can now be enabled across a cluster.
  Notes:
  NameNode High availability only supports manual failover in this release.

  If NameNode HA is configured in the cluster, then HAWQ will not be able to take advantage of HA capability when the primary NameNode fails. We recommend that if you want to use HAWQ, don't configure HA for HDFS. In the case of the primary NameNode failing, HAWQ will not be able to failover to the secondary NameNode. Manual re-direction to secondary NameNode is not currently supported in this release. Therefore we highly recommend that HA should be disabled when used with HAWQ.

Deployment:

- Simplification of deployment via CLI.

- You can add and remove services using `icm_client reconfigure` command. Refer to the *Pivotal HD Enterprise 1.1 Installation and User Guide* for further details.

- You can define `JAVA_HOME` in the `ClusterConfig.xml` configuration file.

Apache Stack:

- Rebase Hadoop to 2.0.5-alpha

- Oozie is included in PHD 1.1 RPM and Binary Distro

- Hive has been updated to 0.11.0. See Versioning and Compatibility for more Apache component version information.

- Hcatalog included with Hive 0.11.0

- HBase has been updated to 0.94.8. See Versioning and Compatibility for more Apache component version information.

- Hcatalog has been enabled in Pig

USS:

- USS upgraded to 0.5.0

- Compatible with secured Pivotal HD cluster

- Supports streaming from NFS, Isilon, and PHD (secure/insecure)

- Supports CDH4 (Cloudera) access

- New USS shell

DataLoader:

- Push stream performance and scalability enhancements

- Integration with Spring XD

- Simplified Start/Stop procedures

- Metrics collection enhancement

- Elimination of standalone mode

- Enhanced pseudo-distributed mode

Security:

- To address security concerns, sudo actions available to the gpadmin user have been significantly restricted. To implement these restrictions on an existing cluster, perform the following:

  a. Install the latest version of Pivotal Command Center

  b. Locate and delete the `/etc/sudoers.d/gpadmin` file from all cluster nodes.

  c. Run the following command:

  `icm_client preparehosts`

  d. Secure the clusters, as described in the *Security* chapter of the *Pivotal HD Stack and Tool Reference Guide*.

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*.

- Note that JDK 1.6 is optional but has not been tested with PHD 1.1

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 1.0 Installation and User Guide*.

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

Note that we provide both rpm and non-rpm tar files for those customers who are unable to perform RPM installs.

- **Pivotal HD Enterprise**: `PHD-1.1.0.0-76.tar.gz`, `PHD-1.1.0.0-bin-25.tar.gz`

- **Pivotal HD Tools (DataLoader and USS)**: `PHDTools-1.1.0.0-97.tar.gz`, `PHDTools-1.1.0.0-bin-25.tar.gz`

- **Pivotal Command Center**: `PCC-2.1.0-460.x86_64.tar.gz`

- **Pivotal Advanced Database Services (HAWQ, PXF)**: `PADS-1.1.3-31.tar.gz, PADS-1.1.3-bin-31.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 1.0.3 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.

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.

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.

- You can only upgrade Pivotal HD via Pivotal HD Manager (ICM client). Instructions for upgrading components using Pivotal HD Manager are provided in the *Pivotal HD Enterprise 1.1 Installation and Administrator Guide*.

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

Patches

Pivotal HD 1.1 includes the following patches. These patches were applied to Pivotal HD, which is based on Apache 2.0.5-alpha base.

Apache Patches

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

Pivotal Patches

Feature	Description
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.

Resolved Issues

This section lists issues that have been resolved in Pivotal HD since release 1.0.1 A work-around is provided where applicable.

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

Table 2 All Resolved Issues in Pivotal HD Enterprise 1.1

Component	Issue	Description
General	HD-2691	Upgraded to Jetty version 7.
General	HD-5283	Yarn nodemanager does not decommission after adding it into yarn exclude host list.
Installation	HD-2537	IP address are not supported. In previous releases, if you attempted to deploy using IP addresses the deployment failed but no error message was thrown. IP addresses are still not supported but now appropriate error messages are displayed for the user.
HBase	HD-5246	Secure PHD HBase configuration. There were unresolved issues with Java DNS lookup failures when trying to set up the secure Zookeeper for HBase to use secured HDFS.
Upgrade	HD-5048	RPM upgrade from PHD1.0 to PHD1.0.1 was not supported.
DataLoader	HD-2922	Data Loader could fail to bind to port 0.0.0.0:12320 due to a previous installation. User encountered error while adding datastores or creating jobs. Scheduler log showed the port bind error message.
DataLoader	HD-1897	A file could not be transferred through the command line without a specification XML file.
DataLoader	HD-1947	Web page formatted incorrectly when Chinese characters were shown.
General	HD-2436	Install "ruby" and "facter" from Admin's local repo onto all cluster nodes: You may have had to install/upgrade ruby and facter on all the cluster nodes.
General	HD-2386	A cluster node could not be on the same host as the Admin node.
Installation	HD-2532	HD Cluster had to be started with `-f` option. A HD cluster appeared to be successfully deployed, but could only be started using the `-f` option to force start it.
Stack	HD-1704	Uninstallation of Hadoop, HBase, Hive, Zookeeper or Pig removed all the configuration files under `/etc/gphd`.

Known Issues

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

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

Table 3   All Known Issues in Pivotal HD Enterprise 1.1

Component	Issue	Description
General	HD-6658	Webhcat server reports HTTP ERROR while running the REST operation after starting/restarting service (MR1 clusters).
	HD-7346	HBase Access Control (grant, revoke, alter, user_permission) are not supported in this release.
	HD-7296	Some icm_client commands are not supported on FIPS mode-enabled clusters. The following ICM commands work on FIPS mode: `icm_client list, start, stop, preparehosts, scanhosts, import, fetch-template, fetch-configuration` The following ICM commands DO NOT work on FIPS mode: `deploy, uninstall, reconfigure, add-slaves, remove-slaves` Workaround: Disabling FIPS, run the required commands, then renable FIPS.
	HD-7377	`pig -useHCatalog` fails to launch mapreduce jobs (binary package).
	N/A	Pivotal Command Center hostnames can only contain lower case letters.
	HD-2209	After uninstalling a cluster, some of the following RPMs may be left behind: ● `bigtop-jsvc.x86_64` ● `bigtop-utils.noarch` ● `zookeeper.noarch` ● `zookeeper-server.noarch`
	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.
	N/A	Pivotal CC CLI currently does not support upgrading or downgrading the Hadoop version on the entire cluster.
	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 further details on installing these components.
	HD-6149	Configuration changes are not implemented following an upgrade or reconfiguration: Workaround: Following an upgrade or reconfiguration, perform the following: **1.** Fetch the new templates that come with the upgraded software by running `icm_client fetch-template.` **2.** Retrieve the existing configuration from database using `icm_client fetch-configuration.` **3.** Sync the new configurations (`hdfs/hadoop-env`) from the template directory to the existing cluster configuration directory. Upgrade or reconfigure service by specifying the cluster configuration directory with updated contents.

Table 3 All Known Issues in Pivotal HD Enterprise 1.1

Component	Issue	Description

General (cont'd)	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:
		`pkill -KILL -u gpadmin`
		`userdel -r gpadmin`

	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.
	N/A	The Apache Hadoop 2.0.5-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	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-5110	Vaidya Report is not available in MR1.
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.
	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.
	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.
	HD-7140	`hadoop fs -rm` operation fails when used against path on a remote filesystem. 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
	HD-7367	There are issues when using java FileSystem functions to access remote filesystem through USS.
	HD-5679	HDFS operations uses the local PHD's block size as opposed to the block size configured in the filesystem where the file resides.

Table 3 All Known Issues in Pivotal HD Enterprise 1.1

Component	Issue	Description
Data Loader	N/A	Streaming job configuration is only supported through the command-line interface.
	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.
	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.

Versioning and Compatibility

The following versions of Pivotal products have been tested for interoperability/compatibility.

Table 4 Pivotal Support Matrix

Product		Version	OS/Browser
Pivotal HD See Table 5 below, for Apache Stack component versioning information.		1.1	RedHat 64-bit: 6.2, 6.4 CentOS 64-bit: 6.2, 6.4
Pivotal Command Center		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
Pivotal DataLoader		2.0.4	RedHat 64-bit: 6.2, 6.4 CentOS 64-bit: 6.2, 6.4
Pivotal USS		0.5	N/A
Pivotal ADS 1.1.3	HAWQ	1.1.3	RedHat 64-bit: 6.2, 6.4 CentOS 64-bit: 6.2, 6.4
	PXF	2.1.0	RedHat 64-bit: 6.2, 6.4 CentOS 64-bit: 6.2, 6.4

Pivotal 1.1 consists of the following Apache stack components:

Table 5 Apache Stack Components

Stack Component		Version
Hadoop (MR2)	HDFS	2.0.5-alpha
	Yarn	
Hadoop (MR1)	MapReduce	1.0.3
Pig		0.10.1
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

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 6 Pivotal HD Enterprise 1.1 and related Documentation

Title	Revision
Pivotal HD Enterprise 1.1 Installation and Administrator Guide	A01
Pivotal HD Enterprise 1.1 Release Notes (this document)	A01
Pivotal Command Center 2.1 User Guide	A01
Pivotal HD DataLoader 2.0 Installation and User Guide	A05
Pivotal HD 1.1 Stack and Tool Reference Guide	A01
HAWQ 1.1 Installation Guide	A09
Pivotal ADS 1.1 Administrator Guide	A06

Use of Open Source