Hortonworks Data Platform Win

Data Integration Services with HDP

(May 21, 2013)

docs.hortonworks.com

Hortonworks Data Platform Win 1.1.0 (GA) : Data Integration Services with HDP

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 Hadoop projects including MapReduce, Hadoop Distributed File System (HDFS), HCatalog, Pig, Hive, HBase, Zookeeper and Ambari. 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

1. Using Apache Hive 1
    1.1. Hive Documentation 1
    1.2. Using Microsoft Excel to Query Hive 2
    1.3. Resources 3
2. Using Apache Pig 4
3. Using HDP for Metadata Services (HCatalog) 5
4. Using Templeton (WebHCat) 6
5. Using HDP for Workflow and Scheduling (Oozie) 7
6. Using Apache Sqoop 8
    6.1. Apache Sqoop 8
        6.1.1. Sqoop Connectors 8
        6.1.2. Sqoop Import Table Commands 8

1. Using Apache Hive

Hortonworks Data Platform deploys Apache Hive for your Hadoop cluster.

Hive is a data warehouse infrastructure built on top of Hadoop. It provides tools to enable easy data ETL, a mechanism to put structures on the data, and the capability for querying and analysis of large data sets stored in Hadoop files.

Hive defines a simple SQL-like query language, called QL, that enables users familiar with SQL to query the data. At the same time, this language also allows programmers who are familiar with the MapReduce framework to be able to plug in their custom mappers and reducers to perform more sophisticated analysis that may not be supported by the built-in capabilities of the language.

1.1. Hive Documentation

Documentation for Hive release 0.9.0 can be found in wiki docs and javadocs.

1. The Hive wiki is organized in four major sections:
   - General Information about Hive
     - Getting Started
     - Presentations and Papers about Hive
     - Hive Mailing Lists
   - User Documentation
     - Hive Tutorial
     - HiveQL Language Manual
     - Hive Operators and Functions
     - Hive Web Interface
     - Hive Client
     - Avro SerDe
   - Administrator Documentation
     - Installing Hive
     - Configuring Hive
     - Setting Up the Metastore
     - Setting Up Hive Web Interface
     - Setting Up Hive Server

- Hive on Amazon Web Services
- Hive on Amazon Elastic MapReduce

- Resources for Contributors
  - Hive Developer FAQ
  - How to Contribute
  - Hive Developer Guide
  - Plugin Developer Kit
  - Unit Test Parallel Execution
  - Hive Architecture Overview
  - Hive Design Docs
  - Full-Text Search over All Hive Resources
  - Project Bylaws

2. Javadocs describe the Hive API.

3. Hive indexing was added in version 0.7.0; documentation and examples can be found here:
   - Indexes – design document (lists the indexing JIRAs with current status, starting with HIVE-417)
   - Create/Drop Index – HiveQL language manual
   - Bitmap indexes – added in Hive version 0.8.0 (JIRA HIVE-1803)
   - Indexed Hive – overview and examples by Prafulla Tekawade and Nikhil Deshpande, October 2010
   - Tutorial: SQL-like join and index with MapReduce using Hadoop and Hive – blog by Ashish Garg, April 2012

1.2. Using Microsoft Excel to Query Hive

Use the following instructions to query data from Hive through Microsoft Excel:

1. On the client machine where you will run Microsoft Excel:
   a. Download the Windows 64-bit Hortonworks ODBC driver from here.
   b. Execute the MSI and follow the instructions to install the ODBC driver.
   c. Set up an ODBC DSN using the following instructions:

i. For 64-bit ODBC driver: Open the **64-bit ODBC Administrator** pane.

ii. Navigate to the **System DSN** tab.

iii. Click **Add**, select the **Hortonworks Hive** driver and click **Finish**.

iv. Configure the driver using the following instructions:

A. Under **Host**, provide the hostname of the cluster node that runs the **Apache Hadoop hiveserver2** service.

B. Under **Port**, enter the port of the Hive Server 2 service. By default, the Hive Server port is set to `10001`.

v. Choose **Username** for authentication, and enter 'Hadoop' as the user name.

vi. Click **Ok**.

2. You can now use this ODBC DSN from Excel, to see tables in Hive and pull data from these tables.

1.3. Resources

Hive JIRAs

Issue tracking for Hive bugs and improvements can be found here: Hive JIRAs.

2. Using Apache Pig

Hortonworks Data Platform deploys Apache Pig for your Hadoop cluster.

Apache Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The structure of Pig programs is amenable to substantial parallelization, which enables them to handle very large data sets.

At the present time, Pig's infrastructure layer consists of a compiler that produces sequences of Map-Reduce programs. Pig's language layer currently consists of a textual language called Pig Latin, which is easy to use, optimized, and extensible.

Pig Documentation

See the Pig documentation including:

- Getting Started
- Pig Latin Basics
- Pig API Docs

Pig Wiki Docs

See the Pig wiki for additional documentation.

Pig JIRAs

Issue tracking for Pig bugs and improvements can be found here: Pig JIRAs.

Pig Mailing Lists

Information about the Pig mailing lists and their archives can be found here: Apache Pig Mailing Lists.

3. Using HDP for Metadata Services (HCatalog)

Hortonworks Data Platform deploys Apache HCatalog to manage the metadata services for your Hadoop cluster.

Apache HCatalog is a table and storage management service for data created using Apache Hadoop. This includes:

- Providing a shared schema and data type mechanism.
- Providing a table abstraction so that users need not be concerned with where or how their data is stored.
- Providing interoperability across data processing tools such as Pig, MapReduce, and Hive.

Start the HCatalog CLI with the command '`<hadoop-install-dir>\hcatalog-0.4.1\bin\hcat.cmd`'

For more details, see the following resources:

- HCatalog Overview
- Installation from Tarball
- Load and Store Interfaces
- Input and Output Interfaces
- Command Line Interface
- Storage Formats
- Dynamic Partitioning
- Notification
- API Documentation

For more details on the Apache HCatalog project, use the following resources:

- HCatalog Wiki
- HCatalog Mailing Lists

4. Using Templeton (WebHCat)

Templeton provides a REST-like web API for HCatalog and related Hadoop components.

Note

The name *Templeton* is being replaced by *WebHCat*, so both terms may be used interchangeably.

For more details, see the following resources:

Overview

- Introduction
- URL Format
- Security
- WebHDFS and Code Push
- Error Codes and Responses
- Log Files

Installation

Configuration

Reference

- DDL

5. Using HDP for Workflow and Scheduling (Oozie)

Hortonworks Data Platform deploys Apache Oozie for your Hadoop cluster.

Oozie is a server-based workflow engine specialized in running workflow jobs with actions that execute Hadoop jobs, such as MapReduce, Pig, Hive, Sqoop, HDFS operations, and sub-workflows. Oozie supports coordinator jobs, which are sequences of workflow jobs that are created at a given frequency and start when all of the required input data is available. A command-line client and a browser interface allow you to manage and administer Oozie jobs locally or remotely.

For additional Oozie documentation, use the following resources:

- Quick Start Guide
- Developer Documentation
  - Oozie Workflow Overview
  - Running the Examples
  - Workflow Functional Specification
  - Coordinator Functional Specification
  - Bundle Functional Specification
  - EL Expression Language Quick Reference
  - Command Line Tool
  - Workflow Rerun
  - Kerberos Authentication
  - Email Action
  - Writing a Custom Action Executor
  - Oozie Client Javadocs
  - Oozie Core Javadocs
  - Oozie Web Services API
- Administrator Documentation
  - Oozie Installation and Configuration
  - Oozie Monitoring
  - Command Line Tool

6. Using Apache Sqoop

6.1. Apache Sqoop

Hortonworks Data Platform deploys Apache Sqoop for your Hadoop cluster.

Sqoop is a tool designed to transfer data between Hadoop and relational databases. You can use Sqoop to import data from a relational database management system (RDBMS) such as MySQL or Oracle into the Hadoop Distributed File System (HDFS), transform the data in Hadoop MapReduce, and then export the data back into an RDBMS. Sqoop automates most of this process, relying on the database to describe the schema for the data to be imported. Sqoop uses MapReduce to import and export the data, which provides parallel operation as well as fault tolerance.

For additional information see the Sqoop documentation, which includes:

- User Guide
  - Basic Usage
  - Sqoop Tools
  - Troubleshooting
- Developer's Guide
- API Documentation

6.1.1. Sqoop Connectors

Sqoop uses a connector based architecture which supports plugins that provide connectivity to new external systems. Using specialized connectors, Sqoop can connect with external systems that have optimized import and export facilities, or do not support native JDBC. Connectors are plugin components based on Sqoop's extension framework and can be added to any existing Sqoop installation.

A Sqoop connector for SQL Server is available from Microsoft:

- **SQL Server R2 connector:** This connector and its documentation can be downloaded from here.

6.1.2. Sqoop Import Table Commands

When connecting to an Oracle database, the Sqoop **import** command requires case-sensitive table names and usernames (typically uppercase). Otherwise the import fails with error message "Attempted to generate class with no columns!"

Prior to the resolution of SQOOP-741, **import-all-tables** would fail for an Oracle database. See the JIRA for more information.

The **import-all-tables** command has additional restrictions. See Chapter 8 in the Sqoop User Guide.