Denodo and Hadoop

Revision 20200811

Denodo Inc.: 525 University Avenue, Suite 31, Palo Alto, CA 94301. USA www.denodo.com

CONTENTS

1 GOAL

This document provides an overview on the various ways to connect Denodo with Hadoop.

2 OVERVIEW

Hadoop is an open source ecosystem of technologies and tools for processing large volumes of data on commodity hardware. Hadoop is built around the Hadoop Distributed File System (HDFS) and MapReduce methods to data processing but a range of tools have been layered on top of HDFS, for instance, HCatalog and Hive. Various other tools have been added by vendors (e.g. Impala) or are loosely associated with Hadoop (like Apache Spark).
Since Hadoop is not a vendor supported software nor follows a single well defined standard, Denodo supports the most common deployed technology standards for connecting to Hadoop.

3 TECHNOLOGIES COVERED IN THIS ARTICLE

3.1 HIVE

Hive is a SQL engine and data warehouse infrastructure designed to run SQL queries on HCatalog data through MapReduce jobs. Hive comes with a JDBC driver which Denodo can readily use to connect.

3.2 IMPALA

Cloudera Impala is a SQL engine provided with the Cloudera Hadoop distribution that provides fast interactive SQL queries directly on Hadoop data stored in HDFS or HBase. Impala provides a JDBC driver which Denodo can readily use to connect.

3.3 SQOOP

Sqoop is a connectivity tool for moving data from non-Hadoop data stores into Hadoop. Sqoop can access data prepared in Denodo through the standard Denodo JDBC driver for movement into Hadoop.

3.4 PIG

PIG is a simple scripting language and command line tool (commonly called `grunt`) for performing various common tasks in a Hadoop environment. PIG is commonly used for data preparation, import, export and maintenance tasks. PIG has no remote calling interface. Denodo will call PIG scripts through the Denodo-Connect SSH wrapper and consume the result of PIG scripts by other means.

3.5 DENODO DISTRIBUTED FILE SYSTEM CUSTOM WRAPPER AND HADOOP

HDFS is the core of Hadoop. It is a highly fault-tolerant distributed file system designed to run on commodity hardware. Denodo can read directly from HDFS files in a Hadoop cluster through the Denodo Distributed File System Custom Wrapper. The data in the following common file formats will be parsed into a relational format:

- Delimited text files

- Sequence files
- Map files
- Avro files
- Parquet files

3.6 HCATALOG

HCatalog is a metadata and table management system for Hadoop. It enables interoperability across data processing tools such as PIG, MapReduce, Streaming, and Hive. Denodo can access HCatalog through the HCatalog REST interface.

3.7 PIVOTAL HAWQ

HAWQ is a parallel/distributed SQL query engine built on top of Hadoop based on PostgreSQL 8.2.15. Denodo can connect to Pivotal HAWK using standard PostgreSQL JDBC drivers.

3.8 HADOOP MAPREDUCE JOBS

Map-Reduce scripts have no remote calling interface. Denodo will call MapReduce scripts through the Denodo Connects SSH and Distributed File System Custom Wrappers and will consume the results directly.

3.9 HBASE

HBase is a column-oriented NoSQL data storage environment designed to support large, sparsely populated tables in Hadoop. Denodo will connect to HBase through the Denodo HBase Custom Wrapper.

3.10 SPARKSQL

SparkSQL is a SQL engine built on top of Spark. It is largely Hive compatible, but faster and has shorter response times. Denodo will connect to SparkSQL through the Hive JDBC driver.

Note that SparkSQL and Hive may coexist on the same Hadoop cluster but will connect on different ports. When browsing Hive and SparkSQL metadata on the same Hadoop Cluster you will see the same tables. Although the **same** JDBC driver is used, Denodo needs to know which technology it connects to since SparkSQL is not 100% Hive compatible.

3.11 SPARK

A fast, parallel, general-purpose data processing and analytics engine with streaming capabilities. Spark is not SQL compatible and has no native remote interface. Spark is generally used to run complex analytic scripts on Hadoop clusters. Denodo can call Spark scripts through the Denodo Connect SSH Custom Wrapper and consume the resulting output, usually written to a Hive table, by other means.

3.12 SPARK AND SPARKSQL

Note that SparkSQL is not identical to Spark. SparkSQL makes use of the Spark framework, but has different capabilities.

- SparkSQL provides SQL capabilities and remote access, but does not allow access to Spark scripts.

- Spark scripts allow access to a wide range of analytical libraries written in Java and Spark can internally use SparkSQL to pre-process data.

4 ACCESSING HADOOP DATA AS A RELATIONAL DATA SOURCE

The following technologies/tools can be integrated as JDBC data sources into Denodo:

4.1 HIVE

- Fully supported standard product feature

4.2 IMPALA

- Fully supported standard product feature

4.3 SPARKSQL

- Fully supported standard product feature

4.4 HAWQ

- Standard Product feature - PostgreSQL compatible

5 OTHER RELATIONAL DATA SOURCES

5.1 HDFS

- Provided through the Denodo Distributed File System Custom Wrapper

5.2 HBASE

- Provided through the HBase Denodo Connect Custom Wrapper

5.3 MAP-REDUCE

- Provided through the Distributed File System and SSH Custom Wrappers.

6 ACCESSING DENODO FROM HADOOP

Denodo should be accessed through the standard Denodo JDBC driver

6.1 SQOOP

- Access to Denodo through the Denodo JDBC driver has been tested with Sqoop.

7 OTHER WAYS OF ACCESSING HADOOP

7.1 PIG

- Access through the SSH Denodo Connect wrapper

7.2 SPARK

- Access through the SSH Denodo Connect wrapper

7.3 HCATALOG

- Access through standard REST wrapper

8 CERTIFICATIONS

8.1 CLOUDERA

The Denodo Platform is certified with the following Cloudera 5 Hadoop features for both Kerberos-secured and unsecured environments:

- Apache Hive
- Cloudera Impala
- Hadoop MapReduce
- Apache HBase
- HDFS (Hadoop Distributed File System)

8.2 HORTONWORKS

The Denodo Platform is certified with the following Hortonworks Data Platform 2.1 Hadoop features for both Kerberos-secured and unsecured environments:

- Apache Hive
- Apache Avro
- Hadoop MapReduce
- Apache HBase
- HDFS (Hadoop Distributed File System)