Apache Hive

Sinziana Gafitanu

Apache Hive ...

... is a data warehouse software that facilitates reading, writing, and managing large datasets residing in distributed storage and queried using SQL syntax.

Source: https://cwiki.apache.org/confluence/display/Hive/Home

Hive Features

- Tools to enable easy access to data via SQL, thus enabling data warehousing tasks such as extract/transform/load (ETL), reporting, and data analysis.
- Access to files stored either directly in HDFS or HBase
- Query execution via Map Reduce (default), Tez or Spark
- Procedural language with HPL-SQL
- Sub-second query retrieval via Hive LLAP, Apache YARN and Apache Slider.

Source: https://cwiki.apache.org/confluence/display/Hive/Home

A little bit of history

- Initially developed by Facebook
- Open Source project part of Apache Foundation
- Current contributors include Dropbox, Cloudera, LinkedIn, Microsoft, Yahoo!, Intel
- A fork of Hive is included in Amazon Elastic MapReduce

Major Components of Hive (I)

- Metastore
  - Stores metadata about each table such as the schma and location.
  - Works hand in hand with the driver to keep track of the distributed data and with the backup server for replication.
  - Data is stored in the RDBMS format.
- Driver
  - The Controller
  - Starts the execution of the HiveQL statements and monitors the lifecycle and execution.
  - Acts as a collection point of data or query result after the reduce operation.

Major Components of Hive (II)

- Compiler
  - Converts the query to an execution plan (the tasks and steps needed to be performed on the execution engine).
  - First converts the query to an Abstract Syntax Tree (AST).
  - Second it converts the AST to Directed Acyclic Graph
  - From DAG it builds the operators to run on the execution engine.
- Optimizer
  - Performs operations on the DAG to optimize (splits and joins the data)
- Executor:
  - Takes the DAG and interacts with the Hadoop job tracker to schedule tasks

HiveQL

Hive SQL Datatypes	Hive SQL Semantics
INT	SELECT, LOAD INSERT from query
TINYINT/SMALLINT/BIGINT	Expressions in WHERE and HAVING
BOOLEAN	GROUP BY, ORDER BY, SORT BY
FLOAT	Sub-queries in FROM clause
DOUBLE	GROUP BY, ORDER BY
STRING	CLUSTER BY, DISTRIBUTE BY
TIMESTAMP	ROLLUP and CUBE
BINARY	UNION
ARRAY, MAP, STRUCT, UNION	LEFT, RIGHT and FULL INNER/OUTER JOIN
DECIMAL	CROSS JOIN, LEFT SEMI JOIN
CHAR	Windowing functions (OVER, RANK, etc)
CARCHAR	INTERSECT, EXCEPT, UNION, DISTINCT
DATE	Sub-queries in WHERE (IN, NOT IN, EXISTS/ NOT EXISTS)
	Sub-queries in HAVING

Source: http://hortonworks.com/wp-content/uploads/2016/05/Hortonworks.CheatSheet.SQLtoHive.pdf

Hive Example


```
DROP TABLE IF EXISTS docs;
CREATE TABLE docs (line STRING);
LOAD DATA INPATH 'input_file' OVERWRITE INTO TABLE docs;
CREATE TABLE word_counts AS
SELECT word, count(1) AS count FROM
(SELECT explode(split(line, '\s')) AS word FROM docs) temp
GROUP BY word
ORDER BY word;
```

Running in CLI


```
hive  -e 'select a.col from tab1 a'
```

```
hive  -e 'select a.col from tab1 a'
 -hiveconf hive.root.logger=DEBUG,console
```

```
hive  -f script.sql
```

Source: http://hortonworks.com/wp-content/uploads/2016/05/Hortonworks.CheatSheet.SQLtoHive.pdf


HIVE