CMPT 308

Big Data Paper

Hive - A Petabyte Scale Data Warehouse Using Hadoop

Resources:

-Database Systems:The complete book

-URL : http://labouseur.com

By- Natnael Mengistu

Date - Feb 27, 2017

Hive - is a data warehouse infrastructure built on Hadoop which facilitates data analysis, query and summarization.

- Hadoop is an open opensource framework used for storage and processing of applications and big data using the Mapreduce programming model.
- Hive Query Language(HiveQL) - is an SQL format language , however does not support Insert and uses a different join format.
- Hive structures data in tables and support all the primitive data types. It also supports complex data types as maps, lists and structs.
- Hive being an open source , is still being actively worked on by Facebook and other contributors to handle more complicated operations.

Implementation -

- Hive structures data into the well-understood database concepts like tables, columns, rows, and partitions.
  It supports all the major primitive and complex datatypes.
  Example
  Primitive - integers , floats , strings , doubles
  Complex - maps , lists . structs
- The building components of Hive are Metastore, Drive ,Query compiler, Execution engine, Hive server, client components and Interfaces
  Hadoop includes Mapreduce and hdfs
- Table metadata associates the data in a table to hdfs directories.
- Primary data units and their mapping in the hdfs namespace are tables, partitions and buckets

* Hive can take an implementation of the SerDe java interface provided by the user and associate it to a table or partition.
  - The default SerDe implementation in Hive is called the Lazy SerDe.

Analysis -

- Hive projects structure onto a big data and queries using HiveQL to perform Map and reduce tasks on large datasets.
- Hive users traditional database instances to store data instead of distributed file system.

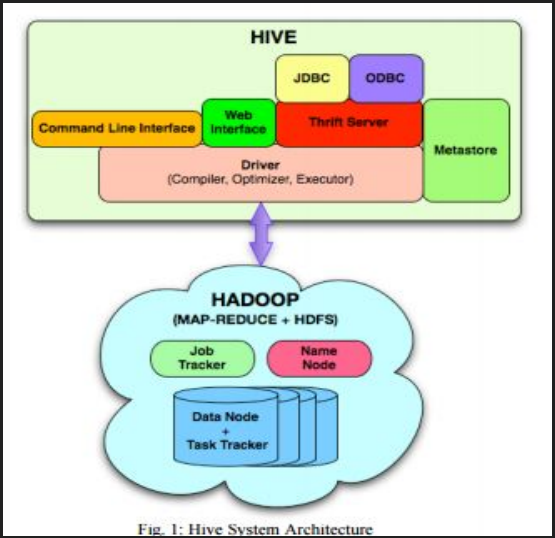

Fig. 1: Hive System Architecture

- HiveQL currently only support a subset of SQL as valid queries. This and other performance improvements are still being worked on

Mapreduce -

- Consists of two functions, Map and Reduce
- Is simple and easy to interact
- MR programmer is free to structure their data in any manner

Parallel DBMS -

- Capable of running in different clusters of shared nothing nodes.
- Parallel execution is possible because almost all tables are partitioned over the nodes and using optimizer that translates SQL command.
- Require data to fit into the relational paradigm.

Implementation:

- Hadoop system is the most popular open source implementations of the MR framework.

  - It provides implementation of google distributed file system.

  - Uses central job tracker and a Master daemons to coordinate node activities.

- Vertica and DBMS-X are examples of Parallel DBMS.

  - In vertica all data is stored as a column

  - DBMS-X is a parallel DBMS from a major relational database vendor which stores data in row format.

Analysis -

- Although MapReduce is simple to use , Parallel systems execute most tasks better than MapReduce.

  They show significant performance advantage that Hadoop and DBMS-X.

- MR frame works provide a more sophisticated failure system than parallel DBMSs.

- While the original MR framework is written in c++ , Hadoop is completely Java

- Parallel DBMSs construct a complete query plan to all processing nodes while MR use control messages to synchronize processing.

Hive vs. Approaches to Large-scale data analysis

- Hive loads data quicker than DBMS-X and Vertica because it is built on Hadoop.

  - It is a great fit for facebook because there will be a large amount of data to store everyday.

- Parallel DBMS allows optional compression of stored data while Hive use both Block and Record level compression on Input data.

- Executions are a lot slower in Hive than parallel DBMS

- Hadoop is much easier to setup and use than parallel databases.

The main ideas of the Stonebraker talk

- From 1970 - 2000 RDBMS grew and was going to be an universal answer , But they could not embed it on streaming applications.

- C - Store : column store
- Paper - talks about text

2015 - One size does not fit all

- Data warehouse Market : have column stores
- OLTP Market : Moving toward main memory deployment
- NoSQL Market : 100 or so vendors with a potpouri of data Models
- Complex Analytics : Regression , eigen factors , svd , data clustering
- Streaming Market : can add streaming to OLTP engine
- Graph analytic market :

- The ‘main tent’ Conflict

Advantages -

- Suitable for Facebook's data processing
- Simple to use
- Compared to the other DBMS, It is easy to load

Disadvantages -

- Although it's easy to load data, execution is slow.
- Because Hive is built on Hadoop , It shares most of the drawbacks.