HIVE- A PETABYTE SCALE DATA WAREHOUSE USING HADOOP

Facebook Data Infrastructure Team

A COMPARASION OF APPROACHES TO LARGE-SCALE DATA ANALYSIS

By different research groups

PAYAL SHARMA

07/05/2014

Main Idea of First Paper

Because of the increasing data set it is difficult to maintain and render data at fast pace. In order to achieve speed in large data access, Big companies are using Hadoop. But the drawback of using Hadoop is that it is completely customized and cannot be used anywhere else, making it impossible to understand or trouble shoot by person other than developer. Hive concept was introduced on the top of Hadoop to makes things less complicated and easy to manage. Hive manage the data using language called HiveQL. It manages the data by creating tables, setting up mapping and arrays. Hive also include Metastore which contains schema and statistics which make data retrieval much easier.

Idea Implementation

- Use Hadoop's MAP-REDUCE, on which Hive is built , to store large data sets.
- The MAP task indicates how the input columns can be transformed using a user program into output columns.
- The REDUCE task specifies the user program to invoke on the output columns of the sub query.
- Usage of Hive's serializers and deserializers that helps in incorporating data into tables without transforming data sets and compresses data extensively thus reducing the size of data to a great extent.

My Analysis of that Idea and its Implementation

- Facebook uses Hive and Hadoop extensively thus storing large amount of compressing data that is growing every day. It consists of large number of tables that contain 700TB of data which is a good example in real time.
- This idea can also be implemented in the areas where we create our social profiles or customized accounts. Netflix can be another example where we customize our account choosing movies which we want to watch. Every time we login, it comes up with the page consisting of movies that we chose.

Comparison of Ideas & Implementation in Comparison Paper

The comparison paper states that the basic control flow of MAP-REDUCE lies in parallel SQL Database Management Systems. The performance and development complexity of MAP-REDUCE has been compared with two other parallel SQL DBMS (Vertica and DBMS-X) on a cluster of 100 nodes. The performance of parallel SQL DBMS is far better than MAP-REDUCE in certain areas. So, it says to implement the features of both kind of architectures. It suggests to use parallel SQL DBMS features like less power usage, index execution, schema and structure creation and high level compression. Similarly the features like easy installation , extensibility and coding language of MAP-REDUCE should be used.

Advantages & Disadvantages of Main Idea of First Paper

Advantages	Disadvantages
The programming model used in MAP-REDUCE is java which is easier to understand when compared to SQL.	MAP-REDUCE does not support schemas and indexing. Its compression level is less when compared to SQL.
It is fault tolerant i.e. if a unit of work fails, only that is restarted by MR scheduler.	It is not flexible enough. It does not contain any additional tools like parallel SQL DBMS.
Data loading is easier using the command line utility or by creating a custom data loading program.	MR model does not have inherent property to join two statements which is implemented in three phases.
Its installation, configuration and usage is easier when compared to parallel SQL DBMS.	It takes time before all nodes of MAP-REDUCE get started and running fully.