Hive – A Petabyte Scale Data Warehouse Using Hadoop

Matthew Velasquez

March 6, 2017

CMPT 308 – Database Systems

Main Ideas of Hive

- Open-source data warehousing solution built on top of Hadoop
- Supports queries similar to SQL called HiveQL
- Capable of running jobs on Hadoop/Hive cluster for a large variety of applications

Implementation of Hive

- Used by Facebook for data processing needs
- Since most of the workload on Facebook is on adhoc queries, it is allowed on the Hadoop cluster
- Used for summarization jobs as well as machine learning algorithms
- Provides data processing services to engineers and analysts at a smaller cost than a main common warehousing infrastructure

Analysis of Hive

- The system has a lot of data that is allowed to be stored since the warehouse currently has over 700TB of data
- HiveQL can only accept certain subsets of SQL as valid queries
- User friendly
- Jobs that would require a long period of time to finish running can now be completed in less time with the use of Hadoop and Hive

Main Ideas of the Comparison Paper

- MapReduce takes less time in loading data and tuning the execution of parallel DBMSs
- Comparison of DBMSs on time consumption of Large-Scale Data tasks
- Discussion of system architectures that are necessary for processing large amounts of data

Implementation of Comparison Paper

- MapReduce can be enabled into a distributed processing framework
- Allows familiar concepts and logic that is used in SQL
- MapReduce can run jobs faster containing minor data processing tasks

Analysis of Comparison Paper

- MapReduce uses a pull method to transfer data
- There is more flexibility in database systems
- Runs jobs faster that contain minor data processing tasks

Comparison between both Articles

- Hive can run jobs that contain large data in less time along with the use of Hadoop where MapReduce is faster at running jobs with small data
- There is more data that is able to be stored in Hive than in MapReduce
- Hive is more compatible with many services that require different types of functions

Main Ideas of Stonebraker Talk

- Data Warehouse have or soon will have column stores because they are faster
- There is starting to be more data scientists in the field
- SAP would be competing against Oracle in the markets in the future
- New ideas are being implemented and it is a good time to become a DBMS researcher

Advantages and Disadvantages of Hive

- Hive provides a faster time consumption in running jobs that contain large amounts of data
- It was not made for running small data processing tasks
- Not capable of conforming to all systems because it was made for applications such as Facebook containing large amounts of data
- Since SQL code will eventually be converted to arrays according to the Stonebraker talk, Hive will have lost its importance