HIVE – A PETABYTE SCALE DATA WAREHOUSE USING HADOOP

Chris Belmonte

May 9, 2014

Thusoo, Ashish, Joydeep Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Ning Zhang, Suresh Anthony, Hao Liu, and Raghotham Murthy. "Hive – A Petabyte Scale Data Warehouse Using Hadoop." (n.d.): 996-1005. Web.

Pavlo, Andrew, Erik Paulson, Alexander Rasin, Daniel J. Abadi, David J. DeWitt, Samuel Madden, Michael Stonebreaker. “A Comparison to Large-Scale Data Analysis.” (n.d.): 165-178. Web.

MAIN IDEA

- Data sets that are collected and analyzed are growing faster causing for traditional solutions to become more expensive
- Hadoop is an “open-source map-reduce implementation” that is currently used by huge companies such as Yahoo and Facebook
- Hive, built upon Hadoop, is an open-source data warehousing solution
- HiveQL is a SQL-like declarative language that is compiled into map-reduce jobs and are executed by Hadoop
- Hive also has a system catalog called Metastore which contains schemas
- Facebook has a Hive warehouse with over tens of thousands of tables and over 700TB of data

HOW IT'S IMPLEMENTED

- Hive uses these building blocks: Metastore, Driver, Query compiler, execution, HiveServer, user-inputted command line interface, and interfaces such as SerDe and ObjectInspector
- Metastore is a system catalog for Hive
- The query compiler processes queries by using data stored in Metastore to create an execution plan
- Tasks are executed in the order of their dependencies
  - A task can only execute if all prerequisite tasks have been executed
- A new user, reportedly, is able to use the system after about an hour which means that training a new employee to work with Hive takes essentially no time at all

ANALYSIS

- Hive is an open sourced project backed by an already popular application called Hadoop
- This is an extremely powerful program because if it can handle Facebook 700TB of data, then it can handle just about anything.
- If an employee can be trained in just an hour, then that means it takes essentially no time for a new employee to be trained to use it
- The database system itself seems like it is pretty similar to a relational database model
- With plans to even further optimize Hive, I think this will be one of the best options for companies especially with companies like Facebook and Yahoo using it

IDEAS OF SECOND PAPER

Advantages to Parallel

- Requires data to fit into the relational paradigm
- Support for user-defined functions, stored-procedures, and user-defined aggregates in SQL
- Strives to balance computational workloads while minimizing the amount data transmitted over the network connecting the nodes of the cluster

Disadvantages to Parallel

- If breaking relational database paradigm is something you wish to do, then you can't
- Less sophisticated failure model. MapReduce is much better at handling node failures

HIVE IN LIGHT OF THE COMPARISON PAPER

- Hadoop is extremely fast with data loading
  - The amount of seconds taken slightly increases when the amount of nodes are increased
- Hadoop gets outperformed in task execution by Vertica and DBMS-X
- Hadoop gets generally outperformed with queries and gets seriously outperformed in joins
- Hive is still attractive even after these tests because it can handle all of Facebook's data
- Hive is also extremely fast with loading data and this can be seen as more important than queries in some cases
- With Facebook being the giant it is today, it has the ability to change the industry with what it decides to use and that is why I think Hive will always be at least a decent option but more so a good option.