‘Hive - A Petabyte Scale Data Warehouse Using Hadoop’

By Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Ning Zhang, Suresh Antony, Hao Liu and Raghotham Murthy

‘A Comparison of Approaches to Large-Scale Data Analysis’

By Andrew Pavlo, Erik Paulson, Alexander Rasin, Daniel Abadi, David DeWitt, Samuel Madden, and Michael Stonebraker

●●●

10/19/16

Presentation by Ray Mattingly

The Main Idea

The paper that I chose to read focuses on a few key parts of the Facebook big data infrastructure.

1. The relational database simply took too long to analyze the TBs of data that Facebook was routinely collecting.
2. Hadoop, an open source project, made it possible to analyze in hours what formerly took days when using relational databases.
3. Branching off from Hadoop, the project Hive allows analysis of big data at extremely fast rates and with a more similar query structure to that of a relational database.

Implementation

Hive and Hadoop are now implemented extensively within the Facebook data processing infrastructure.

Statistics listed tell us that, each day, 7500 jobs are executed with Facebook's data and that more than 75TB of compressed data is submitted to processing.

By creating the open source project Hive to accompany Hadoop, the Facebook team has allowed for a more familiar query language with which to operate. This language, HiveQL, allows for quick creation of queries, while Hadoop allows for quick execution of said queries. All in all, the cluster allows for far more efficient data analysis.

Analysis of the Idea and the Implementation

Just based upon the information in the chosen paper, the idea and implementation seem to be a great idea. Not only does the Hadoop implementation allow for faster analysis of information, but the Hive addition allows for simple querying to be done by those familiar with SQL already.

The comparison paper does, however, bring up some evidence to suggest that there may be more effective alternatives already out there, that have been out there “for decades”.

The Main Idea - Comparison Paper

The main idea of this comparison paper is that people may be jumping to conclusions in assuming that Hadoop is the best way of analysing big data.

This paper suggests that there may be a better data handling option for different necessities, and that it is certainly jumping the gun to, at this point in time, call Hadoop the superior data modeling.

Tests are done to show that Vertica and DBMS-X do have instances where in which they show superiority to Hadoop.

Implementation - Comparison Paper

The comparison paper conducted many different experiments to contrast the efficacy of different big data models.

The major comparisons were done between Vertica, DBMS-X, and Hadoop. The tests compared:

1. Load Times
2. Task Execution
3. Aggregation Tasks
4. Join Tasks
5. And others metrics as compression, failure models, and system configuration

Analysis of Ideas & Implementation - Comparison Paper

This paper seemed to take a fairly objective and inductive approach toward its comparisons. The conclusions were created based upon information derived from data that was created through experimentation, thus my analysis of the ideas and implementation is that they are valid.

One cannot really question the validity of the idea either way, as questioning the efficacy of one model over another, then testing said question, is always a valid and worthwhile idea or activity when done correctly.

Through the experimentation, this paper concludes that there are many instances where in Vertica or DBMS-X are, indeed, more efficient than Hadoop. The results of this paper suggest that there may be much more behind finding the best data model for one's own usage than one would expect at first glance.

Comparison of Two Paper Ideas and Implementations

I believe that the first paper certainly lists some convincing ways in which the Hive and Hadoop cluster has bettered the data analysis at Facebook. The second paper does a good job of not really contradicting that Hadoop can show improvements over standard practices, while also suggesting that that does not mean that Hadoop is the end-all-be-all best option for data modeling.

The second paper does a good job of suggesting that the Hive and Hadoop clusters may not be the best setup for all instances of data modeling and analysis. In that, I believe both papers were accurate despite conveying countering ideas. I believe the second paper, accurately, suggests that the first paper may be narrow minded in its search of the best data modeling system.

The Main Idea - Stonebraker Talk

The main idea of the Stonebraker Talk is that the standard row based relational database models no longer fit today's data needs. His claim is that 10 years ago one size fit some, and today one size fits none.

"The jury is out" seems to be a common phrase relating to the best options for data warehouses, complex analytics, streaming markets, or graph analytics, but the speaker seems fairly certain that row-stores will see an incredible drop to negligible market share among new databases.

Two Papers and Talk - Overall Analysis

The second paper and the talk both have a way of supporting and discrediting the first paper.

The support comes in in that both the second paper and the talk suggest that the traditional way of modeling a relational database may not be the best solution for big data handling. This coincides well with the idea that the Hive and Hadoop cluster is the best option for handling big data at this time.

The discrediting of the first paper's findings comes in the general vagueness of both the second paper and the talk. Neither of the two find any sort of evidence that suggests there is any single best option for handling big data at this time, which suggests that diving all of one's resources into Hive and Hadoop without proper recognition of other options may be short sighted.