Big Data Summary

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 BY ANDREW PAVLO, ERIK PAULSON, ALEXANDER RASIN, DANIEL ABADI, DAVID DEWITT, SAMUEL MADDEN, AND MICHAEL STONEBRAKER

MICHAEL STONBRAKER'S ICDE 2015 TALK ABOUT HIS "10 YEAR TEST OF TIME" PAPER AWARD

BY TADD BINDAS 10/19/16

The Main Ideas

I chose the paper " *Hive – A Petabyte Scale Data Warehouse Using Hadoop*"

In this paper the authors write about how with the increasing size of data sets, traditional warehouse solutions are becoming increasingly expensive. The authors' focus is on Facebook's ad hoc analysis and what software would be best paired with Hadoop to provide the best ad hoc analysis. HiveQL is an open-source, SQL-like declarative language is dubbed as the solution to this issue.

The idea is that Hive will solve all of the needs of Facebook and work perfectly with Facebook and Hadoop.

The Implementation

Currently, Hadoop and Hive are used extensively at Facebook.

Facebook warehouses store around 75TB of compressed data and submit over 7500 jobs to the cluster each day.

Hive, in combination with Hadoop, enables this kind of workload because of the simplicity to which ad hoc analysis can be done

Analysis of the idea and The implementation

Hive has a lot of upside to its ad hoc analysis campaign.

- It is fast
- Most of its functions are very simple

There are some cons however

- There is some unpredictability with Hive's ad hoc jobs.
- It is a work in progress and continues to be tweaked and tuned.

Main Ideas of the Comparison Paper

The comparison paper, "*A Comparison of Approaches to Large-Scale Data Analysis*" writes about Hadoop being a new computing model that is drawing lots of enthusiasm from coders.

Hadoop is then put up against Parallel DBMSs (Vertica and DBMS-X) and Parallel Database Systems were found to have a significant advantage over Hadoop in a variety of benchmark assessments.

The idea is that Hadoop may not be the best tool for Big Data.

Comparison Idea Implementation

To prove the paper's central thesis, that Hadoop is not the best Big Data software, a series of tests are performed to determine the better Big Data software.

The following tests are performed to prove Hadoop may not be the best Big Data Software.

- Load Times
- Aggregation Task
- Selection Task
- Join Task

Analysis of the Comparison Paper

The paper highlights the pros and cons of both Hadoop and Parallel DBMSs objectively and through many comparisons. In the variety testing, Hadoop came in second to both Vertica and DBMS-X. In section three, many of Hadoop's features are highlighted such as Hadoop having freedom to structure their data which makes Hadoop the better option.

This paper does a great job at showcasing the strengths and weaknesses of both Hadoop and Parallel DMBMs.

Comparison on the Two Papers

- I personally liked the second paper, the comparison paper, more because it had more facts about different types of Big Data software.
- The paper on Hive was very interesting because it talked about how Hadoop, in combination with Hive, is used in the world of today with Facebook ad hoc analysis.

Both papers do a very good job describing each software and their pros and cons. The first paper could be improved if it included examples of what Hive is better than.

The second paper could be improved by including references to Hive instead of just Hadoop.

Main Ideas of the Stonebraker Talk

- The paper was about “One size does not fit all” and now Stonebraker believes “One size doesn’t fit anything”
- Stonebraker talks about how many types of data modeling software do not have a place in the modern markets and are being phased out by newer software, specifically row-stores.
- His prediction for the future is that column-stores will be one of the biggest things in the future.

Advantages and Disadvantages of the Main Idea of the Chosen Paper in Context of Talk

The comparison paper and the talk both talk about how the traditional model for data will not be the best model for the future of data. The talk discusses the different markets that the traditional data-model does not work for. The comparison paper breaks down the different ways of data modeling to show that the traditional model is not the best way to query and sort through data.

The first paper talks about Hive and Hadoop, however the talk does not specifically indicate which software is better than Hive and Hadoop. From what I can conclude Hive and Hadoop is currently still one of the top open-source programs on the market.