BIG DATA PAPER SUMMARY

Hive – A Petabyte Scale Data Warehouse Using Hadoop

A Comparison of Approaches to Large-Scale Data Analysis

Michael Stonebraker's ICDE 2015 talk about his "10 Year Test of Time" paper award

By Mark Rajovic 3/3/17

MAIN IDEAS - HIVE

- Hive is an open-source data warehousing solution built on top of Hadoop. It supports queries expressed with a declarative language similar to SQL; called HiveQL.
- The idea behind this article is that due to increasing sizes of data sets, traditional warehousing solutions are no longer cost effective.
- The Authors provide the example of the ad hoc analysis present in Facebook and proposes ideas to the best Hadoop software to perform optimal analysis. Hive is offered as the solution to on top of Hadoop to allow Facebook and its ad hoc analysis to function smoothly.

IMPLEMENTATION - HIVE

- In terms of modern day implementation and application, Facebook currently employs a variety of the functionalities available through Hive and Hadoop.
- Facebook is constantly growing, therefore it continues to take on more and more data. Its data warehouses store approximately 75 TB of compressed data, as well as submits more than 7500 jobs to the cluster everyday.
- Facebook is able to manage this amount of data because the Ad Hoc analyses made possible and easier through the use of Hive, coupled on top of Hadoop.

ANALYSIS OF IDEAS AND IMPLEMENTATION - HIVE

- Hive provides a variety of benefits to an ad hoc analysis platform. Its simple and effective functions allow for quick and accurate results.
- The ability for Hive to handle increasing numbers of data sets proves how valuable of an asset it can be for any company, especially a rapidly expanding entity, such as Facebook.
- Upon further analysis of Hives effectiveness for ad hoc tasks, it is clear that Hive on Hadoop is still a work in process and by no means completely fine tuned. Because of this there is some inherent degree of unpredictability.

MAIN IDEAS - COMPARISON OF APPROACHES

- According to the Comparison article, MapReduce and Parallel DBMS are the two main approaches in large-scale data analysis.
- Both approaches are known for different reasons; MapReduce is more focused on the user end and is equip with more fault tolerant characteristics, making it more user friendly.
- On the other hand, the support tools available in Parallel DBMS are extremely useful and in turn, lead to better overall performance.
- Hadoop was compared to Parallel DBMS and its disadvantages were clearly highlighted. Demonstrating that Hadoop is not the ultimate tool for big data analysis.

IMPLEMENTATION – COMPARISON OF APPROACHES

- In terms of real world application, there are differences when analyzing the fault tolerances of the two approaches.
- For MapReduce, failure is avoided by skipping one node to another that works in order to run.
- On the other hand, in DBMS, files are saved on a network instead of being stored locally, this makes them accessible to more than one user at the same time.
- Through observation of both approaches, it becomes clear that for reasons such as load times, selection tasks, and join tasks, that Hadoop is not an optimal big data software.

ANALYSIS OF IDEAS AND IMPLEMENTATION – COMPARISON OF APPROACHES

- Through analysis of DBMS and MapReduce, which employs Hadoop. It's apparent that DBMS holds a clear advantage. This is mainly because time management is the top priority.
- Therefore whichever system is going to achieve the desired accurate result faster, needs to be viewed as the most valuable option.
- The time speed difference can most likely be attributed to the increased amount of code necessary in MapReduce rather than DBMS. That said, some of MapReduce's features are noticeable and useful, such as the flexibility of its structure and the ability to process data with minimal errors.

COMPARING THE TWO

- Following careful analysis of ideas and implementations of Hive as well as DBMS, its clear that there are advantages to both.
- However, if I had to pick, I would still agree that Hive is a more than suitable option to continue ad hoc analysis for Facebook.
- The comparison paper makes it clear that MapReduce and DBMS are intricate and amazing systems, which provide clear benefits in almost any situation. As mentioned in the previous slide, time management is a key component, but each approach should be looked at in a situational context as well. In the case of Hive, if it is something that continues to work in its current role of ad hoc analysis, there is no reason to completely change it!

MAIN IDEAS – STONEBRAKER'S TALK

- Stonebraker addresses the idea of proposing RDBMS as a possible "*one size fits all*" solution to everything. But after consultation and attempts at implementation, the concept was deemed clearly impossible.
- He also raises a point about the variety of database markets that were not at all productive. Also, that modern markets continue to be taken over by new software. Row-stores provides an example of this.
- Another topic brought up by Stonebraker is the concept of column stores. He believes that eventually, this is the most likely successor to the modern row format.

ADVANTAGES & DISADVANTAGES – HIVE AND STONEBRAKER

- In today's market Hive is still a top open-source program, capable of handling vast amounts of data. However in the future it is clear that Hive, like many other current facts in modern software and database systems will eventually become obsolete and replaced by more efficient forms of sorting and querying data.
- The advantage would be that it is currently still able to operate as one of the top software options available. It is able to maintain functionality even with the increasing demands of a company like Facebook.
- Unfortunately, there is a clear disadvantage with is inherent in Hive. It is highlighted in both the comparison article and Stonebraker's talk. He describes that with changing markets, and inability to adapt fast enough, new software will need to step into the place of Hive it is no longer efficient to deal with rapidly growing data sets quickly and accurately.