The “Big Data Paper”

Michael Hercules Sirico - Due 3/7/2017

Paper Titles and Bibliographic Data

A Comparison of Approaches to Large-Scale Data Analysis,
By Pavlo, Paulson, Rasin, Abadi, DeWitt, Madden, and Stonebraker

Hive - A Petabyte Scale Data Warehouse Using Hadoop
By Thusoo, Sarma, Jain, Shao, Chakka, Zhang, Antony, Liu, and Murthy

Speech on 10-Year Most Influential Paper Award at ICDE 2015
By Michael Stonebraker

Main Idea of Hive

- Hadoop is an open-source ‘MapReduce’ system used by large companies; Yahoo, Facebook
  - Used at the very high level; Many Terabytes of Data
  - Very Low Level, a lot of hard to maintain, low level custom programs
- Hive is an open-source “data warehousing solution” that is built on top of the already existing Hadoop
  - Supports Queries and functions in a special language similar to SQL
  - Hive would increase the ease of use of Hadoop, and, when properly optimized, would improve the performance of Hadoop.

How Hive is Implemented

- Hive is Built on top of Hadoop, it’s essentially a High-Level way to interface with Hadoop
  - Ease of doing so due to both being open source
- Hive has a few Major Building blocks that allow it to interface with The Hadoop MapReduce:
  - Metastore, Driver, Query Compiler, Execution Engine, Server, and Client
  - These Different components are the bridge between the high level Hive and the Lower Level Hadoop

My Analysis of the Hive Idea/Implementation

- I feel hive is a beneficial thing, as it has helped Hadoop become more accessible
- Those who are only familiar with Relational / SQL Databases will be able to transition more easily.
  - New Skills must still be learned as Hive is not a complete SQL
- Since Hadoop is excellent at handling Enormous Amounts of Data, Hive paves the way for easier use of that data
  - The fact that Facebook has used Hive to enhance their own Services and cut costs, and has a lot of Faith in Hive shows potential

Main idea of comparison paper

- Comparison of MapReduce Model to Parallel DBMSs
  - Observed that DBMS performance was far superior to MapReduce results
- Hadoop (From previous paper) implements a MapReduce Model
- DBMS-X, a parallel SQL DBMS, and Vertica database (Column Based Rather than Row Based)
- MapReduce ultimately, while superior at minimizing loss, struggles with performance penalties when compared to Relational Databases

Analysis of Comparison Ideas/Implementation

- Five Tests were taken; on Data Loading, Selection, Aggregation, Joining, and UDF Aggregation
  - In General, The DBMSs outperformed Hadoop, the MapReduced Model, in all categories, with the exception of UDF aggregation
  - Provided Graphs show how in Hadoop really under performed
- The paper touches on this, and in some regard I feel it is correct- The DBMSs have been able to mature over 25 years
  - Hadoop and MapReduction have not been as refined as the DBMSs

Comparison of Ideas and Implementations of Both Papers

- Each System has Pros and Cons, Positives and Negatives
- Hive was Created to Improve Hadoop's MapReduce System
  - It borrows some ideas from DBMSs for ease of use
  - It allows for huge companies like Facebook to process Massive Data
- The Comparison Paper Analyzed Performance Between DBMS-x, Vertica, and Hadoop
  - Hadoop was the slowest, but is also the newest and least Refined
- Perhaps Further Improvements to MapReduce Systems, such as Hive will allow MapReduce Systems like Hadoop to rival DBMS Performance

Main ideas of Stonebraker Talk

- From the 70s until 2005, The goal was a “One size Fits All” Relational Database that could be used for anything
  - In 2005, Streaming database, as well as Column-storage led to problems
  - Inability to use “Complex Analytics” very slow to simulate in SQL
- Movement away from Traditional Relational DBMSs
  - Data Warehouses, OLTP (streaming), and other replacements in the Future
- He feels that there is a lot more than just Databases for the future
  - Different needs require different systems, Relational Databases Don’t work for every situation
  - “Great opportunity for New Ideas”

Hive: In Context of Comparison Paper and The Talk

- I think that for very large data needs, A MapReduction system, Like Hadoop, may be more useful than a traditional Database
  - Facebook with Many Many Terabytes of Data
  - Stonebraker mentioning DBMS don’t fit all applications
- Looking at Hadoop from a DBMS and SQL perspective can be helpful
  - Don’t limit it to a SQL perspective - remember Limitations of DBMS
- Hadoop and MapReduction are still relatively New
  - Still under performs the Traditional DBMS shown in the paper study
  - With maturity and new advancements like Hive, It will improve in time
  - Important not to limit options with just a database; look at new ideas.