Big Data Paper Summary

Nicholas DePaul 10/30/2017

Paper Titles and Bibliographic Data

- 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 J. Abadi, David J. DeWitt, Samuel Madden and Michael Stonebraker

- 10-Year Most Influential Paper Award at ICDE 2015

  By Michael Stonebraker

Hive – A Petabyte Scale Data Warehouse Using Hadoop – Main Idea

- Hadoop is a very popular open-source map-reduce system which is being used by many large companies such as Facebook, Yahoo and many more.

  -Stores and processes very large data on hardware.

  -Makes developers write custom programs which can be hard to manage.

- Hive is and open-source data warehousing solution which is built on top of Hadoop.

  -Supports queries that is expressed in a language that is similar to SQL.

Implementation of Hive

- Since Hive is built on top of Hadoop it makes it easy to use since they are both open source. That's why large companies use it because its free. ☺
- Execution Engine, Query Compiler, Server, Driver, Metastore, and Client are in Hive to allow you to interface with Hadoop.

Hive Idea and Implementation Analysis

- I really like that Hive is a thing because it has given Hadoop more features since it has been implemented into Hadoop.
- If you are a computer science major or have any prior knowledge with using SQL then it makes it easier to convert to Hive.

-But as with everything there are similarities between Hive and SQL but not everything is the same. You will still have to learn some things to be able to fully grasp Hive.

A Comparison of Approaches to Large-Scale Data Main Idea

- Was comparing the MapReduce approach which allows it to perform a large-scale data analysis with the RDBMS approach.
- SQL DBMS are extremely quicker and does not require as much code for it to execute a task. But all good things come with a downside. It took longer to load the data.

Comparison Paper Implementation

- MapReduce
  - Process key and value data pairs.
  - Input data is stored in a file system which is then sent to every node.
- DBMS
  - Tables partitioned in a cluster.
  - System optimizes and then will translate SQL.
  - Execution is then divided between the nodes.

Comparison Paper Analysis

- DBMS has been better then the MapReduced Model after it took all of the tests.
- Even displayed graphs to show us that MapReduce did worse.
- DBMS although outperforms MapReduce it is limited to SQL. MapReduce does not have a limitation like that.

Comparison of Both Papers

- They made Hive just so they could make Hadoop's MapReduce better.
  - Since it is open source it allowed large companies such as Facebook and Yahoo to use it.
- In the Comparison Paper the DBMS was the slowest but it required less code. I would rather less code and suffer the load times.

Stonebraker Talk Main Idea

- One Size Does Not Fit All
- Moving away from Relational DBMSs
- Wants to change: MVRAM, big main memory, processor diversity, LLVM and vectorization

Main Idea of the Chosen Paper in the Context of the Comparison Paper and the Stonebraker Talk

- If you are working with a lot of data then the MapReduce system may be better to use.

  -If you are working on Facebook or Yahoo then you have a ton of data to work with. It might be more frustrating to use a traditional database.

- There are many limitations with DBMS as I have stated before.
- Since MapReduce is still kind of new but not really. It needs more updates so it can get faster and more fluid like how Hive is. I feel that eventually it will get better and more people will be willing to use MapReduce.