Hiroki Hayashi
Database Management
Professor: Alan Labouseur
Dec. 14, 2014

Big Data Paper Summary

Bibliography

Hive – A Petabyte Scale Data Warehouse Using Hadoop

- 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

- Andrew Pavlo, Erik Paulson, Alexander Rasin, Daniel J. Abadi, David J. DeWitt, Samuel Madden,Michael Stonebraker

Main Idea of First Paper

- The Facebook Data Infrastructure Team writes about their open source warehousing program, Hive.
- Hive was built on top of Hadoop, which is a open source map-reducing implementation
  - End users had troubles with Hadoop, where it required them to spend extensive amounts of time writing programs for even the simplest analysis.
  - Hive is a solution for this in terms of helping end users
  - It is expressed in the language “HiveQL”, and uses well-understood database concepts such as tables, rows, and columns.
- The whole idea of Hive is to create an easier and more productive version of Hadoop, and is still being developed by the Facebook team

Implementation of Idea

- Data Model and Type System
  - Data stored in tables (with rows and columns)
  - Columns have associated types
    - Primitive Type (Int, Float, String)
    - Complex Type (Arrays, Lists, Structs)
- Query Language
  - The language they use is a subset of SQL with some things they added
    - This allows users who know SQL to use Hive easier and faster

Idea & Implementation Analysis

- First of all, I like how one can understand most of what this article is saying just by reading the first part (I did read the whole article). It gave me a better idea of what the paper was talking about since the whole idea was in my head.
- It is great how Facebook, as many other companies, are making sure that the users get the best experience possible. Not only is Hive giving users the chance to better understand Hadoop by making it easier, but they are also making it more productive so that the users will spend less time writing programs.
- However, by making it “simpler” or “easier to use”, it says that Facebook’s declarative language has taken out some things from the SQL language. Although I understand that makes it easier, I believe that all SQL functions are there for a reason, therefore by taking them out, it would perhaps limit the work you can do with the queries.

Paper 1 vs. Paper 2

- The whole idea behind the second paper is to compare MapReduce and Database Management Systems' approach to large scale data analysis
  - MR, having only two functions "Map and Reduce", is similar to Hive in terms of trying to make the programming model "simple". By letting the programmers code their programs on the model, it gives users a wider range of things to do.
  - DBMS are more geared towards the performance, such as loading times, or task execution. Vertica, which is a type of parallel database system, seemed to excel in performance, as this paper compared it to DBMS-X and Hadoop.
- Hive focuses on these aspects as well; making the language simpler so that the users can have an easier time coding, which ultimately lead to faster processing times.

Advantages & Disadvantages

- Advantages
  - Hadoop was compared to DBMS-X and Vertica in the second paper, and it was shown that it loads fairly quickly. As Hive is built on top of Hadoop, it should have the same or faster performance.
  - The amount of work that needs to go into Hive's program is much less than that of MapReduce, which makes Hive simpler than the two-function model.
- Disadvantages
  - Although Hadoop had loaded quickly, it was not fast in executing this many times, according to studies in the second paper.
  - MR had loaded slower on Hadoop, and that it started executions a lot later than normal. As Hive runs on Hadoop, this may be a disadvantage in terms of performance speed