# BIG DATA PAPER SUMMARY

Using “Hive” and “A Comparison of Approaches to Large-Scale Data Analysis”

Talia Rossi

5/9/2014

# Hadoop

- Large data sets need to be collected and analyzed in an inexpensive and convenient way
- Hadoop is an implementation that can store large data sets on hardware
  - Is an open-source map-reduce implementation
  - Used by many large companies like Facebook and Yahoo
  - Is low level and is made up of custom programs that are difficult to maintain and reuse
  - The solution to this: Hive

# Hive

- Hive is an open source data warehouse that is built on top of and serves as a solution to Hadoop
  - Used as SQL like language (HiveQL)
    - Supports all major primitive types: integers, floats, doubles, and strings
    - Processed in three steps: Parse, Type Checking, Semantic Analysis, and Optimization
  - Structures data into database concepts (tables, columns, rows, etc.)
  - Uses a system catalog called Metastore
  - Uses a SerDe Java interface to easily interpret and query custom data formats

# HiveQL Implementation Steps

1. Statements are submitted using thrift, odbc, or jdbc interfaces
2. The query undergoes the parse, type check, and systematic analysis phase after being passed from the driver to the compiler
3. A logical plan is created by the compiler
4. The logical plan is optimized
5. A DAG of map-reduce tasks and hdfs tasks is created from the optimized plan
6. Using Hadoop, the tasks are finally executed in order of dependencies by using the executed engine

# Implementation of Hadoop and Hive

- Hadoop and Hive are used for data processing by many large companies such as Yahoo and Facebook.
- Facebook implements a Hadoop cluster that monitors the continuous growth of the Facebook network and data.
  - This can only be done on such a large scale because of Hive
- Hive also runs various jobs daily such as summarization jobs and advanced machine learning algorithms
- Hive and Hadoop allow data processing services to be performed at a much cheaper cost than other data warehousing operations

# My Analysis

- I think Hive sounds like a great addition to Hadoop. Together these two programs seem to make analysis of large data sets much more cost effective and simpler. The fact that this can be done with Terabytes of data, I think, is a great feat.
- However, there seem to still be some issues regarding the performance of reporting queries. In the section of the article referring to Facebook's use of Hadoop and Hive it was stated that they had to separate Hadoop clusters to keep up performance. This is obviously something that needs to be worked on. The fact, however, that Hive is open source, I think, will allow it to be improved greatly in time.
- All in all I think these are great ideas for database warehouses and within time will become even better.

# Comparison

- MapReduce is a relatively new concept as opposed to DBMSs, which have been used for a much longer time period.
- MapReduce uses a relational database model and allows data to be in any format while DBMSs require data to specifically fit into a relational paradigm of rows and columns.
- MapReduce programmers must distribute data manually while DBMSs use a parallel query optimizer to balance the amount of data being transmitted.
- MapReduce has had much more flexibility in terms of data expressiveness than DBMSs. However, over time DBMSs have greatly improved although they are still not up to par with MapReduce.
- MapReduce tends to load data quicker than DBMSs, however, once the data is attained DBMSs tend to have an overall better performance, resulting in fewer errors.
- MapReduce seems to be easier and quicker for beginners to learn rather than DBMSs. However, MapReduce tends to require a lot of maintenance, which can be frustrating for developers.

# Advantages and Disadvantages

- Map-reduce Advantages
  - Simple while still being able to distribute relatively complex formats
  - Data can be in any format
  - Handles node failures well
- Map-reduce Disadvantages
  - Data sharing between programmers can cause issues
  - Data transfer causes performance issues
  - Some data formats result in slower load and execution times

# Bibliography

- Pavlo, Andrew and et al. “A Comparison of Approaches to Large Scale Data Analysis.” *Labouseur*. Web. 7 May 2014.
- Thusoo, Ashish and et al. “Hive – A Petabyte Scale Data Warehouse Using Hadoop.” *Labouseur*. Web. 7 May 2014.