Paper 1: Hive – A Petabyte Scale Data Warehouse Using Hadoop
Source: http://www.labouseur.com/courses/db/papers/thusoo.icde2010.hive.pdf
Comparison Paper: A Comparison of Approaches to Large-Scale Data Analysis
Source:
http://www.labouseur.com/courses/db/papers/pavlo.sigmod09.comparison.pdf

Big Data Paper Summary

BY ZAIN QAYYUM 12/14/2014

Main idea of Hive – A Petabyte Scale Data Warehouse Using Hadoop

- In this paper, it is clear that in modern times, several hundred petabytes of data must be stored and analyzed rapidly, and Hadoop, an Apache open-source framework for distributed processing of large data, is capable of massive data processing but here's the catch: Hadoop uses a very low-level map-reduce programming model which means most users take days to write simple queries.
- Hive was created as a framework for running simple SQL-like (HiveQL) queries on top of Hadoop, which are then compiled into map-reduce scripts. This enables users to save a lot of time, and companies money since time is money.

Implementation

- Hive structures the data into popular database concepts such as tables columns rows etc.
- It is an extension of Hadoop, not merely a replacement
- All code is ultimately compiled and optimized at the driver level of Hive and then sent to Hadoop.
- System catalog stored In metastore

My Analysis

- As my company is a soon-to-be Hadoop user, it's great to see that something like Hive exists which should undoubtedly save us a lot of time and money by making queries a lot easier to write and execute, while still giving us the flexibility of writing map-reduce functions from scratch should we need to.
- What sucks though is that Hive cannot update an existing table, instead, the entire table must be overwritten with new data… which is ridiculous.

Comparison to Comparison Paper

- The comparison paper offered a performance comparison between Map Reduce and parallel SQL database management systems (DBMS.) Parallel SQL has been very popular in the past, and Map Reduce is being seen as a radically new way to analyze big data. The tested systems include Hadoop on the map reduce side, and DBMS-X for the parallel SQL.

Advantages / Disadvantages

- The comparison paper makes it seem as though the setup and query execution time is the main disadvantage of the map reduce system Hadoop, while admitting that the processing time is much better for Hadoop though. But the comparison paper did not make use of Hive in their testing procedure, which would essentially solve this very issue, mostly diminishing any advantage that the parallel SQL system had. But to get to the level that will allow Hive to truly replace all Hadoop queries, support for more functions must be added, such as updating an existing table.