Hive A Petabyte Scale Data Warehouse Using Hadoop
And
A Comparison of Approaches to Large-Scale Data Analysis

Joseph Gust
3/7/2017

The main idea of this paper was the creation of HIVE, an open-source data warehousing solution built on top of Hadoop, which is an open-source implementation of map-reduce. HiveQL a higher level language to be used with HIVE. HiveQL borrows heavily from SQL, but it’s meant to have more extensibility and flexibility.

Slide 3 - your analysis of that idea and its implementation

The comparison paper attempts to determine the pros and cons of using different map reduce methods vs. parallel DBMSs. It takes in factors such as the speed for initially loading the data and speed it takes to complete various tasks.

Slide 5 - how those ideas are implemented

As we can see from the comparison paper, the Hadoop(map-reduce) was much faster at loading the data than the parallel DBMS counterparts. On the other hand Hadoop was much slower than the parallel DBMSs when it came to task execution. When they used a select statement, Hadoop took significantly longer than the DBMSs. When asked to perform an aggregate function, Hadoop was left in the dust once again. For performing a join statement, Hadoop was vastly slower than the other two, but for the UDF Aggregation task Hadoop was faster than DBMS-X and much slower than Vertica. So despite Hadoop's excellent speed for initially loading the data, it's much slower than the other two when it comes to task execution.

Slide 7 - comparison of the ideas and implementations of the two papers

The Stonebraker talk was primarily about how Relational Databases are losing their effectiveness, and that while it used to be “one size fits all” they’ve realised it’s actually not going to work moving forward.

Slide 9 - advantages and disadvantages of the main idea of the chosen paper in the context of the comparison paper and the Stonebraker talk