# Comparison of Approaches to Large-Scale Data Analysis & Hive – A Petabyte Scale Data Warehouse Using Hadoop

JENNA FICULA

10/18/2016

# MAIN IDEA: HIVE

- Data processing infrastructure needs to scale along with data growth
- Map-reduce programming model requires developers to write custom programs - there must be a solution
- Hive is the an open-source data warehousing solution built on top of Hadoop
- Hive supports high level query language, HiveQL, similar to SQL which makes data analysis and compression more efficient

# IMPLEMENTATION: HIVE

- The architecture and capabilities of Hive are broken down into sections including
  - The data model – similar to traditional database (tables & rows)
  - Type system – integer, float, string, associative arrays, lists, structs
  - Query language – HiveQL - subset of SQL
  - Hive file storage – tables stored in a directory in hdfs, partitions, buckets
  - System architecture – metastore, driver, query compiler, execution engine, server
  - Usage statistics – Facebook uses hive for the simplicity of ad hoc analysis

# ANALYSIS: HIVE

- Hive is a work in progress
- This paper is effective in addressing hive as a solution for scaling infrastructure to big data
- It is descriptive but only mentions preliminary experiments in the conclusion
  - the paper would be better supported with more quantitative experimentation with Hive
- The paper was motivated by a specific business interest to improve efficiency of data analytics

# MAIN IDEA: COMPARISON PAPER

- Parallel database systems performed more favorable as compared to Hadoop MapReduce (MR) in executing a variety of data intensive analysis benchmarks.
- MapReduce model: a brute force solution that wastes vast amounts of energy
- Database management system (DBMS) has several advantages:
  - B-tree indices to speed the execution of selection operations
  - Novel storage mechanisms
  - Compression techniques
  - Sophisticated method for querying data

# IMPLEMENTATION: COMPARISON PAPER

- Defines DBMS and MapReduce
- Discusses differences between the two and architectural tradeoffs:
  - DBMS – uses high level query language, MR – uses low level algorithm
  - DBMS – uses rows/ columns, MR – does not adhere to a schema
  - DBMS – b tree index, MR – no built in indexes
- Compares parallel DBMS with MapReduce with 5 performance benchmarks:

(Grep task, Selection task, Aggregation task, Join task, UDF aggregation task)

# ANALYSIS: COMPARISON PAPER

- The paper successfully examines the pros and cons of both DBMS and MapReduce
- The experiments conducted in this research were thorough and demonstrated the differences between DBMS and MR
- Two DBMSs (Vertica & system from major relational vendor) were compared to Hadoop - increases validity of results as opposed to using just one DBMS
- The experiments conducted using 100 nodes are not as representative even though few data sets are in the petabyte range

# COMPARISON: BOTH PAPERS

| Compare | Contrast |
| --- | --- |
| • Since SQL is a main advantage in the effectiveness and popularity of RDBMS, Hive makes MapReduce a more competitive option by simplifying queries<br>• Both agree hive is needed to improve the expressiveness of query capabilities for more extensive data analytics<br>• Both papers agree that RDBMS is an outdated solution to data storage and needs replacement | • Hive is described whereas MapReduce & DBMS are compared through experimentation/ performance tasks in the second paper<br>• Hive paper focusses on the specific details of syntax and system logistics whereas the comparison paper is a overview of both systems<br>• Hive paper prompted by business interests, Comparison paper prompted by testing two data management systems |

# STONEBRAKER TALK

- 80s / 90s: one size fits all – RDBMS - traditional row stores were the solution to database needs
- 2000s: Traditional row stores are obsolete for all markets – not one size fits all
  - Complex Analytics
  - Streaming market
  - Graph Analytics
- Markets are innovating to implement a diversity of engines to store data focusing in column stores
- Old legacy vendors such as Oracle will lose market share while trying to adapt to new engines while maintaining their user interface

# ADVANTAGES AND DISADVANTAGES OF HIVE

| Advantages | Disadvantages |
| --- | --- |
| • Markets are moving away from the traditional relational db model row stores with SQL queries.<br>• Hive is an innovative solution to increase the popularity of MapReduce by exploring columnar storage and more intelligent data placement.<br>• Hive is unique from other innovations such as Pig because Hive provides a system catalog that persists metadata about tables within the system. | • Hive is new and many legacy vendors will have to compete with experimental open source database technologies.<br>• Relational Database Models are being phased out<br>• Hive has been tested and compared with a few other systems but could use further testing and experimentation prior to implementation.<br>• One size no longer fits all – Hive will have to compete with a variety of engines on the market |