

Big Data Paper Summary

Hive – *A Petabyte Scale Data Warehouse Using Hadoop*
(Ashish [Thusoo](#), Sen Sarma, Jain, Shao, Chakka, Zhang, Antony,
Liu, and Murthy)

A Comparison of Approaches to Large-Scale Data Analysis (Pavlo,
Paulson, Rasin, Abadi, DeWitt, Madden, and Stonebraker)
*Michael stone breaker on his 10-Year Most Influential Paper Award at
ICDE 2015*


By: Joe Tursi
Professor Labouseur
October 30 2017



Hive – A *Petabyte Scale Data Warehouse Using Hadoop*

Main Ideas:

- RDBMS used too much data and proved to be inefficient with processing large scale data taking days to complete
- The Facebook Data Infrastructure team utilized Hadoop, which is an open source project, which processed data on a petabyte scale. This reduced the data processing, but it was still taking too long
 - End users had trouble with Hadoop writing programs for simple analysis
- Hive, an open source data warehousing solution built on top of Hadoop, was made to help end users
 - Uses the Hadoop file system as storage



Hive – *A Petabyte Scale Data Warehouse Using Hadoop Implementation*


Data Model

- Data is stored in tables, columns, rows, and partitions
 - Supports primitive (Int,Float,String)and complex (Arrays, Lists)data types
- Hive query language is similar to SQL with some limitations
 - Inserts aren't possible, write over the data
 - More user friendly with background in SQL
- File Formats allows the user to specify how Hadoop files are stored in a file



Hive – *A Petabyte Scale Data Warehouse Using Hadoop Analysis / Implementation*


- Facebook is focusing on users needs and making data processing infrastructure more efficient
 - Facebook data keeps increasing
 - Retained data storage types making it user friendly
 - Data will be handled more effectively not wasting space or time
- HiveQL keeps the traditional constructs of SQL but is also unique constructs of its own
 - That being said it does put a limit on what you are able to do not having full capabilities of SQL



A Comparison of Approaches to Large-Scale *Data Analysis*


Main Ideas:

- Compares the data analysis of: Parallel Database Management Systems(DBMS) and MapReduce(MR)
 - MR has two function, Map and Reduce, to process key value and data pairs
 - Cluster Computing: harnessing large numbers of low end processors working in parallel to solve computing problem
- DBMS out performs MR in terms of task execution and loading times. MapReduce is easier to carry out, the program model is simple.




A Comparison of Approaches to Large-Scale *Data Analysis Implementation*

- Both systems tested on HTML document processing
 - Selection, Loading, and Join
 - Aggregation
 - Execution time
 - DBMS performed better
- You can manipulate MapReduce easier, but the system requires more maintenance
- DBMS provides built in indexing, while MR relies on user



A Comparison of Approaches to Large-Scale *Data Analysis* *Analysis of Implementation*

- DBMS and MapReduce have both advantages over each other
 - MR is open source and easier to implement, DBMS was able to execute tasks more efficiently
- DBMS saves developers from focusing on creating indexes and joins
- Process to load data into and tune the execution of DBMSs took longer than MR system, performance of the DBMSs was drastically better
- MR programs in Hadoop primarily written in Java, which most programmers are familiar with OOP, instead of SQL. (SQL is easier/better)




Comparison of the ideas and implementations of the two papers

- Hive uses a relational parallel DBMS
- RDBMS took too long to process data, some features were implemented into Hive
- Hives solves all the problems MR had
- Approaches to Large-scale data tested the execution of both systems, Hive utilized RDBMS concepts



Stonebraker Talk

- Stonebraker says RDBMS - one size fits none
- C-Stores(database management system based on column oriented DBMS) are more efficient
 - Column stores are going to replace row stores
- The traditional DBMS architecture is no longer applicable to the database market
 - The future promises many opportunities in term of data



Advantages and Disadvantages of Hive in context of comparison paper & Stonebraker Talk

Advantages

1. Users can add functions/types in Hadoop
2. Easy interface to learn/use
 - a. MapReduce model is simple, easy for users to navigate

Disadvantage

1. RDBMS are not one size fits all
 - a. No longer application to database market
2. Less efficient compared to other databases
 - a. Hive cannot be used with unstructured databases