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

Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Ning Zhang, Suresh Antony, Hao Liu , Raghotham Murthy, Andrew Pavlo, Erik Paulson, Alexander Rasin, Daniel J. Abadi, David J. DeWitt, Samuel Madden and Michael Stonebraker

Ronald Dartey

12th December, 2014

# Main Idea

- Facebook's processing infrastructure was built around a data warehouse and this was based on a Relational database management system.
- The amount of data Facebook had was growing rapidly.
- Some daily processing jobs at Facebook were taking more than a day to process.
- The situation gets worse as the amount of data Facebook had increased.
- Facebook resorted to Hadoop/Hive to solve this scaling problem.

# Implementation Of Idea

- Storing a table in a directory in a Hadoop file system
- Storing a partition of a table in a sub directory within a table's directory
- Storing a bucket in a file within the partition's or tables directory.
- Pruning data
- Using Metastore as a way to store system catalog and meta data
- Hash based partial aggregation to reduce the amount of time spent in sorting and merging data.

# Analyzing Idea

- This idea of Hadoop/hive is necessary because it is fast will also meet future demands as the amount of data for a growing business increases rapidly.
- The idea also appears to be user friendly as Hadoop/hive reduces its complexity for end users.

# Comparison to the ideas and implementations presented in the comparison paper

- The first paper shows how MapReduce in is helpful in processing meta data and how fast processing the data is.
- The second paper shows how Parallel DBMSs is faster compared to MapReduce

# Advantages and Disadvantages of the Main Idea

- **Advantages**
- Open source and can even be made better than its current functionality
- It's scalable
- **Disadvantages**
- Hadoop/hive is not as fast as parallel data base management systems.