# Andrew Lohr
# 5/9/2014

## HIVE

Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Ning Zhang, Suresh Antony, Hao Liu and Raghotham Murthy

*Facebook Data Infrastructure Team*

## A COMPARISON OF APPROACHES TO LARGE-SCALE DATA ANALYSIS


Andrew Pavlo
Brown University
pavlo@cs.brown.edu

Erik Paulson
University of Wisconsin
epaulson@cs.wisc.edu

Alexander Rasin
Brown University
alexr@cs.brown.edu

Daniel J. Abadi
Yale University
dna@cs.yale.edu

David J. DeWitt
Microsoft Inc.
dewitt@microsoft.com

Samuel Madden
M.I.T. CSAIL
madden@csail.mit.edu

Michael Stonebraker
M.I.T. CSAIL
stonebraker@csail.mit.edu

# Hive – Main Idea

- Facebook used to use RDBMS
  - Was okay with 15TB, now they have data up to 700TB!
- Switched to Hadoop (open-source)
  - Great performance enhancement with Big data
  - End-users had trouble coding programs to get analytics
    - Used Map-Reduce, and lacked expressiveness of most languages (ex SQL)
- Made Hive off of Hadoop to incorporate the lost expressiveness.

# Implementation of Hive

- Structures data into well understood concepts
  - Tables, rows, columns, partitions
  - Stores data in Hadoop Cluster
- Supports most primitive types and some complex types
  - Signed Integer types, Floating point numbers, Strings
  - Associative Arrays, lists, structs.
- Hive Query Language (HQL)
  - Extension of SQL + more.
- Read based, not appropriate for fast responses or transactions.
  - Long sequential scans.
- Does not support inserting into an existing table/partition
  - All inserts overwrite

```
INSERT OVERWRITE TABLE t1
SELECT * FROM t2;
```

# Analysis

- Great way to add to Hadoop.
- I love that complex data types are added to the data types.
  - Being able to store arrays, lists, and structs can make life a little more easier.
- Since Hive is similar to SQL syntax and theory, it makes it easier for people to start with it right out of the gate.
- I do not like that there is no support for row level inserts, updates, and deletes.

# Comparison

- Hive with Map Reduce VS SQL DBMS
  - SQL DBMS faster and require less code
  - Map Reduce is faster with tuning and loading the data
  - Map reduce is forced to start a query with a scan of an entire input file.
- SQL DBMS provides indexing for faster lookup
  - Map Reduce framework is so simple it does not provide this and the programmer must implement indexing themselves.
- Hive was must easier to set up than the SQL DBMS (Vertica, DBMS-X) and much more cost effective.

# Advantages / Disadvantages

- Advantages
  - Hive is simple to set up
  - Does an awesome job minimizing the amount of data lost on a hardware failure.
- Disadvantages
  - Hive Map Reduce Benchmark tests were slower
  - Hive has “schema later” paradigm means it will parse records at run time compare to at load time.
    - Makes compression less valuable