

Big Data Papers

CHRIS KORINISKIE

MAY 8, 2015

- A COMPARISON OF APPROACHES TO LARGE-SCALE DATA ANALYSIS
ANDREW PAVLO, ERIK PAULSON, ALEXANDER RASIN, DANIEL J.
ABADI, DAVID J DEWITT, SAMUEL MADDEN, MICHAEL STONEBRAKER
- MICHAEL STONEBRAKER ON HIS 10-YEAR MOST INFLUENTIAL PAPER
AWARD AT ICDE 2015
- HIVE – A PETABYTE SCALE DATA WAREHOUSE USING HADOOP
ASHISH THUSOOO, JOYDEEP SEN SARMA, NAMIT JAIN, ZHENG
SHAO, PRASAD CHAKKA, NING ZHANG, SURESH ANTONY, HAO
LIU, RAGHOTHAM MURTHY

Main Ideas on Hive

- ▶ Open-source project
- ▶ Petabyte scale data processing system
- ▶ Map and reduce implementation
- ▶ HiveQL: uses similar syntax to SQL
- ▶ Runs on top of Hadoop (HDFS)
- ▶ Scalable analysis on large data sets
- ▶ Data organized into tables and partitions stored in HDFS
- ▶ Serialization/Deserialization of different data formats

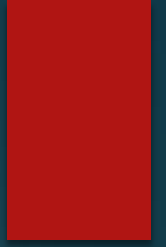
How it was implemented

- ▶ Need for large data set analytics mainly for Facebook.
- ▶ Facebook went from 15tb of data to several hundreds of terabytes of data to process in a few years.
- ▶ HiveQL statements are compiled into a MapReduce and then sent to Hadoop for execution.
- ▶ Data Storage: Stored in tables, partitions and buckets

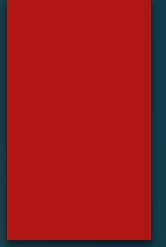
Analysis

- ▶ The HiveQL is similar to SQL but you can create and add your own code since it is open source which is pretty interesting.
- ▶ Querying of information is much faster using Hive
- ▶ The metastore (tables, partitions, schemas, columns and their types, table locations) are all stored in a regular RDBMS for low latency.
- ▶ Hive is continually growing and being used by more companies such as Amazon and Netflix for large data.

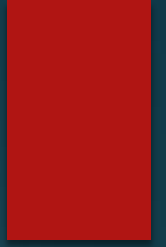
Main ideas of A Comparison of Approaches to Large-Scale Data Analysis



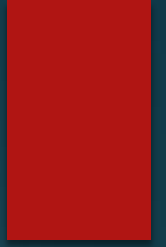
Implementation



Analysis



Comparison on both papers



Main ideas of Stonebraker talk

