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
- Hadoop is written in Java

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

- Parallel DBMS and MapReduce large scale data analysis
- DBMS seperates schema from the application and stores in a catalog that may be used for queries.
- Modern DBMSs use indexing to reduce scope of a search.
- Hadoop, DBMS-X, and Vertica database systems.
- DMBS more structured data with data constraints
- SQL queries fast and easy

Implementation

- Both are used but MapReduce has become the standard for large scale databases.
- Parallel RDBMS's have been around since the 1980's
- They tested three systems Haddop, DBMs-X, and Vertica(Stonebraker) on performance to compare one another
- All results pointed that Hadoop was extremely outperformed

Analysis

- Map Reduce is simple because it only has two functions, essentially read and output (map and reduce).
- MapReduce has higher fault-tolerance/flexibility
- Parallel DBMSs can be faster

Comparison on both papers

Hive Paper

- Large data summarizations, queries, and analysis
- Metadata in Relational databases
- SQL like with HiveQL but not the same thing….converted into MapReduce

Comparison Paper

- This paper basically just talks about how DBMSs are better than MapReduce.
- Schema support
- Indexes
- User result statements

Main ideas of Stonebraker talk

- The video doesn't like to play too easily
- One size does not fit all
- New Ideas: NVRAM(no more flash), main memory storage, faster processing and networks
- Row stores vs column stores(column stores common now)
- Market has no standards with many data models and architectures with a ton of new ideas being worked on at all times
- Data warehousing

Advantages and Disadvantages of Hive paper in comparison

Advantages

- SQL like queries makes it easy
- MapReduce
- Hive uses indexing as well
- Only need two functions
- Works on top of Hadoop and is open source.
- Petabyte size scalable analysis

Disadvantages

- Parallel DBMS implementations (vertica, DBMS-X) are faster than MapReduce
  - Use an index on the pageRank column and store the rankings table already sorted by pageRank
- SQL is slow
- Cannot join two different data sets