BIG DATA SUMMARY: HIVE A PETABYTE SCALE DATA WAREHOUSE USING HADOOP

ASHISH THUSOO, JOYDEEP SEN SARMA, NAMIT JAIN, ZHENG SHAO, PRASAD CHAKKA, NING ZHANG, SURESH ANTONY, HAO LIU, AND RAGHOTHAM MURTHY

A COMPARISON OF APPROACHES TO LARGE-SCALE DATA ANALYSIS

BY ANDREW PAVLO, ERIK PAULSON, ALEXANDER RASIN, DANIEL J. ABADI, DAVID J. DEWITT, SAMUELE MADDEN, AND MICHAEL STONEBRAKER

ONE SIZE FITS ALL- AN IDEA WHOSE TIME HAS COME AND GONE (2005)

MICHAEL STONEBRAKER

BY ANDREW ARRIGO 3/3/17

HIVE: MAIN IDEA

- Facebook’s Data Infrastructure Team realized:
  - RDBMS are too slow
  - Solutions are too expensive and complicated for users
- What is Hive?
  - An open-source data map-reducing implementation
- Facebook’s Data Infrastructure team’s focus is to implement Hive to run on top of Hadoop to assist in ad-hoc analysis
- The goal is that implementing Hive on Facebook will allow for an easier user experience for ad-hoc analysis

HIVE: IDEA IMPLEMENTATION

- Hive’s storage of data:
  - Tables: rows and columns
- Runs *HiveQL*- an SQL-like declarative language
- HiveQL in comparison to SQL:
  - Inserts take precedence
- Primitive Supported Data Types:
  - Integers, Floating point numbers, String
- Natively Supports:
  - Associative arrays, Lists, Strings
- Will allow for ease of use to users with knowledge of SQL

HIVE: IDEA AND IMPLEMENTATION ANALYSIS

- Hive allows for advantages and disadvantages
- Advantages:
  - Simplistic functionality (similar to SQL)
  - Speed, power and scalability
  - Open Source
- Disadvantages:
  - Places limitations
- Overall is good for a company like Facebook allotting minimal space with quick accessibility

A COMPARISON OF APPROACHES TO LARGE-SCALE DATA ANALYSIS: MAIN IDEAS

- A Comparison of Approaches to Large-Scale Data Analysis
- MapReduce is simple as it has two functions:
  - Map and Reduce
- Map Reduce is similar to Hive in that they’re both:
  - Simple and Easy to use
- DBMS are oriented towards faster speeds
- Parallel DBMS such as Vertica and DBMS-X in comparison to Hadoop are:
  - Faster
  - Functionality across more nodes
  - More energy efficient
- Hadoop has its benefits, but also its disadvantages when working with Big Data sets

A COMPARISON OF APPROACHES TO LARGE-SCALE DATA ANALYSIS: IDEAS IMPLEMENTATION

- Testing is performed to show Hadoop lacks in several crucial categories applying to Big Data:
  - Load Times
  - Join Task Results
  - Aggregation Task Results
  - Selection Task Results
  - Grep Task Results
- These results were performed to show that Parallel Database Management Systems prove faster in almost every category tested and that they are the best for Big Data

A COMPARISON OF APPROACHES TO LARGE-SCALE DATA ANALYSIS: IDEA AND IMPLEMENTATION ANALYSIS

- Hadoop in comparison to RDBMS does not compete very well
- Although MapReduce, Hive, and Hadoop are meant to be simpler and easier solutions to given problems, parallel systems outperform almost every time

COMPARISON OF IDEAS AND IMPLEMENTATION OF BOTH PAPERS

- The Hive paper went more into the logistics of Hive and Hadoop with examples of how to use their software
- The comparison paper was more oriented towards showing the difference in speeds amongst multiple types of systems.
- I think both papers were important to focus on when faced with understanding what way to go about storage of Big Data
- I believe the comparison paper is more important than the Hive article because it brings many factors into pay rather than solely Hive and Hadoop

MAIN IDEAS OF THE STONEBRAKER TALK

- Stonebraker Believes “One size fits none”
- Markets are shying away from data warehouses and using column stores instead
- New software and technology uses such little memory that physical data row storage isn’t necessary
    - OLTP for example is used for transaction processing that doesn’t require much memory

ADVANTAGES AND DISADVANTAGES OF HIVE'S MAIN IDEA IN CONTEXT TO COMPARISON PAPER AND STONEBRAKER TALK

- The Hive paper focuses on the benefits of using Hive and Hadoop
- Advantages:
  - Will provide solutions for Facebook specifically
  - Provides ease of use
- Disadvantages:
  - Basic with low functionality in comparison to other programs
- The comparison paper focused on providing graphic representations of data implementations
- The comparison paper also provides enough information to show Hive and Hadoop are not the best solution to handle Big Data
- The Stonebrakers talk focused on Traditional Data Models and how they are inefficient for what will be the future of data