Gary Coltrane

Big Data Summary 3/6/17

Hive - A Pedabyte 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

The main idea behind Hive's warehouse at Facebook

- Facebook developed Hive in order to process its large data tasks
- Hive is an open source Map Reduce infrastructure developed from Hadoop
- Hive is scalable and provides a SQL-like language called HiveQl, which is easy to learn and use.

How Hive is implemented

- HiveQL
  - Commands and syntax are similar to SQL, hence easy to learn
  - It supports primitive data types and complex data types such as structs, lists, and maps
  - Hive structures its database in the relational model, which makes room for nesting data types within each other.
- Hive uses the SerDe Java interface
  - SerDe serializes Java objects to the hdfs and can also deserialize Hive objects to the JVM
- Hive's architecture contains several components that serve as its building blocks.
  - Metastore: The system catalog of Hive, which is the metadata about each local database
  - HiveServer: CLI and a UI can connect to the a client
  - Query Compiler: Processes HiveQL through an Abstract Syntax Tree (AST)
- Facebook uses Hive to process all of its big data
  - Facebook's warehouse currently contains 700TB of data
  - Hive's infrastructure allows processing services to engineers and analysts for fraction of the cost than a more traditional warehouse infrastructure

Analysis of Hive

- Hive provides a simple source for any SQL developer to process big data
  - HiveQL's language syntax and structure is similar to SQL
  - Complex data benefit users who have a dabase system that require a level of layers.
  - Using Hive in an application stack is great for any developer.
    - Programmers can connect to HiveQL services in order to create REST applications

- HiveQL's complex data type allows users to store multiple contents in one single field

Comparison Paper: Main Ideas

- Choosing between MR (Map Reduce) and a parallel SQLdatabase systems (DBMS)
- MR is well suited for small development environments, while DBMS fit for large projects
- DBMS and MR have various differences and similarities in terms of its architecture, structure, and optimization.

Implementation of Comparison paper

- DBMS-X and Vertica represented DBMS and Hadoop represented Map Reduce
  - In the BenchMark Task, each system's performance was heavily measured with large amounts of data
  - Loading process was quick in Hadoop
  - Overall performance was better in both DBMS-X and Vertica
  - Specifically DBMS-X had higher load times than both Vertica and Hadoop
  - Hadoop had better performance in its Grep task results

Analysis of Comparison Paper

- MR is best suited for small applications and applications in development phase
  - MR is Open Source and still relatively new, therefore there is still some time and room for its improvements
  - MR is free for a lot of hacks and tweaks, which can be best fit for small scale applications that don’t require large chunks of heavyweight data.
  - MR is highly efficient with checking for faults
  - MR is simply easy to set up and use
- DBMS is more suit for large amounts of data
  - Developers are provided guaranteed comfort when it comes to DBMS performances
  - DBMS require less processes to run difficult tasks

Comparison of the two papers

- Hadoop is relatively supported in the Facebook paper since Hadoop has successfully processed Facebook’s large chunks of data.
- In the comparison paper, the flaws of Hadoop is shown, specifically for dealing with large structures of data.
- However, both articles tend to pinpoint how Hadoop is relatively easy to use and manipulate

Main ideas of Stonebreaker talk

- The belief of relational databases being “One Size Fits All” is extinct
- In the future, there will be many database engines with different capabilities
- Data Scientists will eventually become the next analysts
  - Data Scientists will use more mathematical advancements to measure data

Advantages and disadvantages

- Advantages
  - Hive is a quick non-DBMS solution that anybody can learn.
  - Hive has helped scale Facebook's data and process it more efficiently than its prior system
  - Hive is open source which allows developers to add on amazing features.
- Disadvantages
  - Hive is still relatively new
  - Since it's built on top of Hadoop, it's performance is still low when compared to DBMS.