Hive – A Petabyte Scale Data Warehouse Using Hadoop

Colin May

Hive

- Problem
  - Map-reducing programming model requires custom programs for data analysis which are hard to maintain and reuse
- Solution through Hive
  - Open source data warehouse solution built on Hadoop
  - Supports queries expressed in SQL-like declarative language-*HiveQL*
  - Results compiled into map-reduce jobs that are executed using Hadoop

Hive

- Structures data into well-understood database concepts like tables, columns, rows, and partitions
- Query language very similar to SQL and therefore can be easily understood by SQL users
- Provides flexibility to incorporate data into table without transforming data
  - Saves substantial time with large data sets

Hive-Analysis

- Hadoop lacks expressiveness in data analysis
- Hive provides SQL like functionality to the unstructured world of Hadoop
  - Maintaining extensibility and flexibility of Hadoop
  - Allows for table, column, and partition organization
  - Provides query capacity for simple and complex analysis
  - Metadata
- Gives opportunity to use universal SQL query language with Hadoop giving analysis capacities to the storage and processing of large data sets

Comparison Paper

- Description and comparison of:
  - Parallel SQL database management systems (DBMS)
  - MapReduce
- Evaluate both kinds of systems in terms of performance and development complexity using tests and experiments
  - Benchmark (collection of tasks) run on open source version of MR and two parallel DBMSs

Comparison Paper-Implementation

- Approaches to large-scale data analysis are considered with the choices of MR and parallel databases and their trade-offs
- Organized by aspects of comparison to contrast approaches
  - Overview of each system
  - Architectural Elements
  - Performance Benchmarks
  - Discussion of comparison of test results
- Benchmark consisting of tasks and tests using:
  - Open source version of MapReduce-*Hadoop*
  - Two parallel SQL DBMS-*Vertica and DBMS-X*

Comparison Paper-Analysis

- SQL DBMSs were significantly faster and required less code to implement each task, but took longer to tune and load the data
  - Parallel DBMSs: advantage in executing a variety of data intensive analysis benchmarks
  - MR: advantage in set up and use as an application
  - MR: advantage in extensibility and tolerance

Hive and Comparison Paper

- Hive is a SQL open source extension to Hadoop, a MapReduce tool
- Parallel DBMS were found to have advantage in executing data analysis tasks
- Hive is an attempt to take the two types of approaches to big data and create a model that combines the benefits of cluster computing of MapReduce and the parallel database system with a SQL query language

Stonebraker Talk

- Pushed and developed relational database management system to be be universal-failed
- DB2, Oracle, and SQL systems were becoming obsolete
- Column stores in data storage is faster and more efficient than row stores
- Regressions and analysis with clustering as data organization based on arrays
- Dilemma between progress of database systems and maintaining market share of current database systems

Hive vs Comparison and Stonebraker

- Comparison Paper analyzed advantages and disadvantages to different approaches to large-scale data analysis
- Hive was designed to fulfill current efforts to find optimal balance between the MR and parallel DBMS
- Is able to take the advantages of big data approaches and allows a MR program Hadoop incorporate a parallel DBMS query language
- Doesn’t embrace column stores for query application