Hive – A Petabyte Scale Data Warehouse Using Hadoop Vs. A Comparison of Approaches to Large-Scale Data Analysis

Facebook Data Infrastructure Team. Hive-A Petabyte Scale Data Warehouse Using Hadoop. Facebook, Website. 8 December 2014.

Pavlo, Andrew, Erik Paulson, Alexander Rasin, Daniel Abadi, David DeWitt, Samuel Madden, Michael Stonebraker. A comparison of Approaches to Large-Scale Data Analysis. Website, 8 December 2014.

Kyle Berkoski

12/8/2014

Hive: Summary

- Designed to work on Hadoop due to map-reduce programming being very low level and hard to maintain
- Hive is a warehousing solution that uses a language and queries similar to SQL.
- Able to run jobs that previously took days in a matter of hours
- Has a system catalog that contains schemas, statisticsc, and query compilations
- Mainly used by FaceBook

Hive: Implementation

- Open Source, easy to use, similar to SQL which leads to users being able to adapt easily
- 15TB of data added to FaceBook daily
- Hive runs on pre-existing Hadoop
- HiveQL is compiled into map-reduce jobs executed by Hadoop
- Hadoop was hard to use and time consuming so users were very happy with Hive
- Open-Source

Hive: Analysis

- Does not support inserts into an existing table, however hasn’t been a problem
- Tables stored in directories, Partition stored in subdirectories, bucket is a file in a (sub) directory
- Supports Serialization/Deserialization java interface
- Supports many different file types

Hive Vs. Database Management System

- Hive: Files that are accessed by a program

  DBMS: A computerized record keeping system
- Hive: Does not require files to adhere to schema definitions

  DBMS: Requires data to fit into relational paradigm of rows and columns
- Hive: Since the model is so simple, does not provide built-in indexes

  DBMS: Uses hash (B-tree) indexes to access to data
- Hive: Provides a more sophisticated failure model

  DBMS: If a singe node fails the entire query must be restarted

Advantages and Disadvantages

- Advantages
  - Has a failure model
  - Only consists of two functions (Map and Reduce)
  - Reads from files, so data processing is quicker
- Disadvantages
  - Not very flexible
  - Pre-existing structures must be built into the MR programs
  - Sends large amount of data from the node instead of the other way around