Hive – A Petabyte Scale Data Warehouse Using Hadoop

Facebook Data Infrastructure Team. Hive – A Petabyte Scale Data Warehouse Using Hadoop. Facebook. Web. 6 December 2014.

A Comparison of Approaches to Large-Scale Data Analysis

Pavlo, Andrew, Erik Paulson, Alexander Rasin, Daniel Abadi, David DeWitt, Samuel Madden, Michael Stonebraker. A comparison of Approaches to Large-Scale Data Analysis. Web. 7 December 2014.

Kevin Skocypec
December 8, 2014

Hive

-A Petabyte Scale Data Warehouse Using Hadoop-

- Due to the rapid increase in size of data sets, the typical process for warehouse solutions is expensive and impractical
- Hadoop is currently a popular map-reduce implementation, and it is open-source, but requires lots of maintenance and time spent by programmers
- *Hive* is another open-source data warehouse implementation using Hadoop
- Uses *HiveQL*, which is similar to any other SQL language
- Then is compiled into map-reduce jobs, which are then run in Hadoop
- Hive can use an implementation of SerDe (Serialization/Desrialization) java interface when the user associates the provided one to a table
- Hive contains: Metastore, a system catalog; Driver, which manages the lifestyle; Query Compiler, the component that compiles the *HiveQL*; Execution Engine, the part that runs tasks created by the compiler; HiveServer, that has a thrift interface and a JDBC/ODBC server and helps integrate Hive amongst other applications;

Implementation

- Facebook uses, with 5TB (15 after replication) of compressed data added daily
- Runs on Hadoop, so it is not fully distinctive
- Hive can take implementation of SerDe java interface
- Due to Hadoop's questionable efficiency, Hive was well-accepted after implementation
- Uses language, *HiveQL*, which is similar to any other SQL language (parts are identical)
- Open-Source project that was easily adapted to and is a work in progress

Analysis

- Much more efficient than previous processes
- Lack of inserting into preexisting tables seems to be a problem, but has apparently not caused one yet
- Great support amongst DataStorage, file formats, and SerDe
- Nice idea to move towards more productive methods

Comparison

- Hive compiles files using *HiveQL* that are then accessed later, while MapReduce uses DBMS as a database management system
- Hive contains easy join functionality, while MR does not
- MapReduce works on both structured and unstructured data, while Hive only works on structured data
- MR is better for business logic than Hive
- Hive contains more sound fail-protection, while MR must restart with a single node failure

Advantages and Disadvantages

(Hive)

- Advantages:
  - Failure model incorporated
  - Easy join functionality
  - Compiles and reads from files
- Disadvantages
  - Very Strict
  - Only Structured Data
  - Not for business logic
  - Inability to insert into preexisting tables, but contains other methods to compensate