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

Facebook Data Infrastructure Team

Dan Swezey
Alan Labouseur
CMPT 308
Big Data Paper Summary
10 October 2014

HIVE: BETTER PERFORMANCE EVERYWHERE

- Hive is an open-source data warehousing solution build on top of Hadoop (an open-source map-reduce implementation that can handle very large amounts of data on the average hardware).
- Increase infrastructure to support and process immense amounts of data.
- Make the data easier to manage for end users and to save time doing even the simplest of queries.
- Structure data so it is familiar to the majority of users (e.g. tables, columns, rows, and partitions). Also supports major primitive types (e.g. integers, floats, doubles, and structs). Can also allow users to create their own types of functions.
- "Hive provides the flexibility to incorporate that data into a table without having to transform the data, which can save substantial amount of time for large data sets."

HIVE: IMPLEMENTATION

- Hive uses a system catalog named, Metastore.
- Metastore is useful when it comes to exploring data and compiling and optimizing queries. This is because it contains schemas and statics.
- The information in Metastore can be "queried or modified using a thrift interface and as a result it can be called from clients in different programming languages."
- The information is stored on a RDBMS (relational database management system). It is not stored in hdfs due to the need for very low latency.
- Data Storage:

  - Tables – A table is stored in a directory in hdfs.

  -Partitions – A partition of the table is stored in a subdirectory within a table's directory.

  -Buckets – A bucket is stored in a file within the partition's or table's directory depending on whether the table is a partitioned table or not.

- Tasks are executed in order of their dependencies if and only if all of its prerequisites have already been executed.

WHAT DO I THINK?

- The infrastructure of Hive compared to other open-source warehousing seems most ideal.
- The main reason for this is because it is setup to be user friendly. It uses very similar syntax to SQL, which most users already know.
- Being able to rely on its simplicity allows for better processing and organization. It also means a substantial amount of time is saved when dealing with large amounts of data sets.
- The use of the system catalog, Metastore, creates very low latency which is a must. It also stores all the info about many things such as the tables, partitions, schemas, columns, and table locations.
- How data is stored in different ways (e.g. tables, partitions, and buckets) allows for well organized and accessible data that users can easily specify where and how that data is stored.

HIVE IN COMPARISON TO LARGE-SCALE DATA ANALYSIS

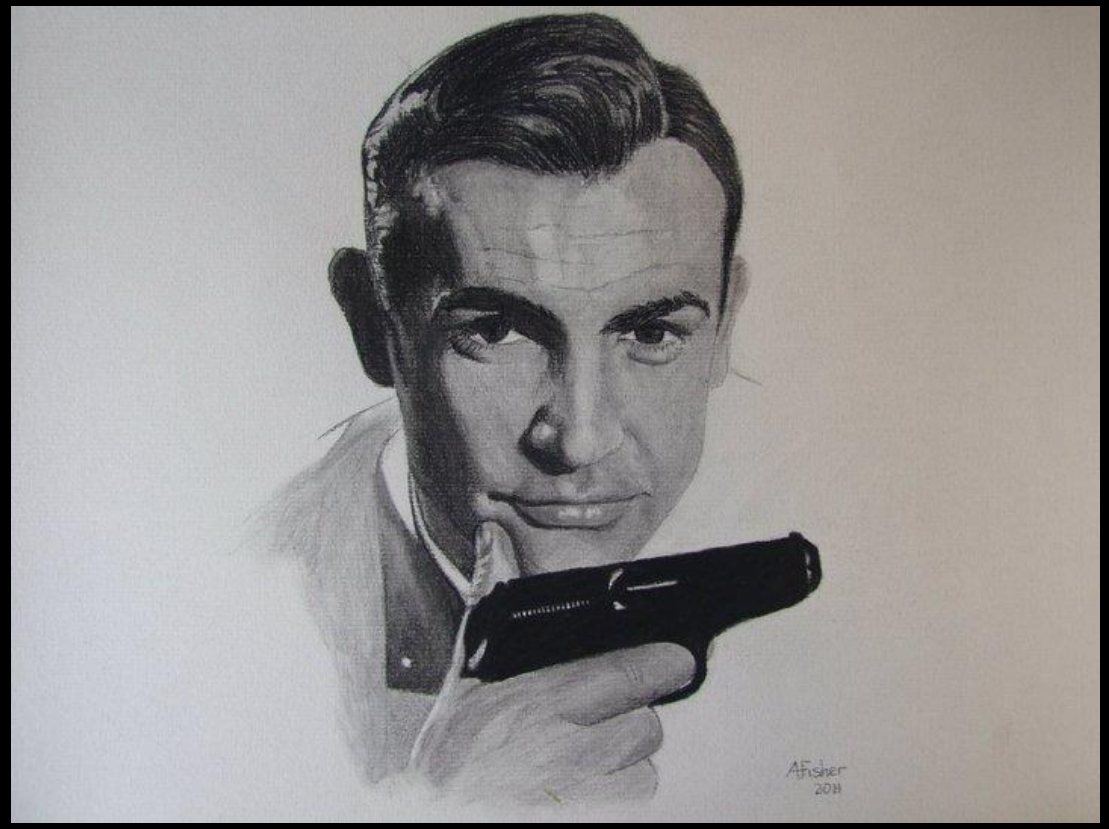
- MapReduce (MR) model is based of its simplicity. It only uses a MAP and REDUCE function to process key/value data pairs
- In Parallel DBMSs, programmers don't have to worry about the underlying storage details (indexing and join strategies).
- MR model is slower than the Parallel DBMS.
- MR is used primarily for "environments with a small number of programmers and limited application domain." This might not be ideal for "long=term and larger-sized projects."
- Parallel DBMSs need its data to fit into the relational paradigm of rows and columns whereas the MR model does not. This means that MR programmers are "free to structure their data in any manner or even to have no structure at all."

ADVANTAGES + DISADVANTAGES OF HIVE

- Advantages:
  - Hive works more efficiently with larger amounts of data saving precious time.
  - Focuses mainly on user familiarity and its system catalog, Metastore.
  - Exploring data and compiling and optimizing queries due to its statics and schemas.
  - Learning curve is slight.

- Disadvantages:
  - "Hive currently does not support inserting into an existing table or data partition and all inserts overwrite the existing data."
  - MR model can structure data in ways programmer sees best fit.

THANKS FOR THE TIME!



All quotes are from either the, "Hive – A Petabyte Scale Data Warehouse Using Hadoop" article or the, "A Comparison of Approaches to Large-Scale Data Analysis" article.