Hive, Pig and Drill

Bhavneet Soni

**Hive, Pig and Drill**

Power of DFS in analyzing big data lies with the massive parallel processing it brings. This is made possible by the MapReduce that divides up the analysis such that it can be executed at the node the data is being hosted. Since the data is distributed over wide number of machines, traditional databases that could be queried using SQL, are not enough. NoSQL databases such as MongoDb, Hbase etc. provided a great alternative and benefit to deal with the issues of storing data over DFS and yet being queried easily. To query a database over the big data we need quite an overhead of converting the query into a MapReduce job that could be executed on the nodes. Hadoop MapReduce being a complicated tool hence there was need for an interface that would take in queries and convert them to MapReduce jobs to get the results back. Some of these interfaces are Hive, Pig and Drill. While all of these are open source and managed by Apache org they were developed by different teams. Pig was developed by Yahoo while Drill was developed by Google. Each of these interfaces provide different advantages and have their own drawbacks also.

Hive was developed by Facebook and was targeted for people comfortable and experienced in using SQL. Hive uses an SQL abstraction layer known as Hive Query Language (HQL) that implements schema on read strategy. Hive is inherently a server side too and has implicit DFS direct access. Hive provides for smooth transition from SQL based applications. While its remarkably similar to SQL users still have to learn something new. Hive provides for built-in User defined functions to manipulate data these can be further extended to handle customized requests. Hive can only handle structured data and It supports text files, sequence files, Record Columnar files (RC) and ORC files. It is best suited for doing daily analysis and generating daily reports. Hive is widely used by companies like Netflix and FINRA.

Pig was developed by Yahoo back in 2006, and is a data-flow (procedural) language. Pig uses Pig Latin as the main querying language and there is no need to import java classes. Pig can be implemented to both structured and semi structured data, however it does not support partitions. Pig is best suited for queries with complex joins and filters and is primarily used by programmers and data researchers. It resides on the client side and needs an explicit DFS access. Pig comes with a command line interface(CLI) and do not have a web interface. Like Hive, Pig also supports UDFs.

Drill is the first schema free SQL engine, based on Google Dremel system, that can be scaled to up to 10000 servers. It supports NoSQL database and can join data from multiple data stores in a single query. It supports all types of file type Hive does with an addition of JSON and text. Drill has a data store aware optimizer that can restructure the query plan for efficient analysis.

References

*Apache Drill*. (n.d.). Retrieved from mapr.com: https://www.mapr.com/products/product-overview/apache-drill

*Apache Hive*. (2015). Retrieved from wikipedia: https://en.wikipedia.org/wiki/Apache\_Hive

*Difference between Pig and Hive-The Two Key Components of Hadoop Ecosystem*. (2016, Dec 15). Retrieved from dezyre.com: https://www.dezyre.com/article/difference-between-pig-and-hive-the-two-key-components-of-hadoop-ecosystem/79

Manlove, D. (2016, Feb 26). *SQL ON HADOOP THE DIFFERENCES AND MAKING THE RIGHT CHOICE*. Retrieved from agildata.com: http://www.agildata.com/sql-on-hadoop-the-differences-and-making-the-right-choice/