CISC 525-90-2016 - Assignment 6
SQL-like languages - Hive, Pig and Drill

Nelson Corrocher

August 2016

1 Overview

Hive, Pig and Drill all have a similar goal: Provide an interface for a data scientist to access Big Data and perform analytics in a familiar manner without having to learn MapReduce programming. However, each of these tools implements that goal slightly differently, and is focused on different use cases and approaches.

2 Hive

Hive is a declarative language (one that tells the hardware what has to be done, contrary to imperative lanaguages, that tells the computer how to do it) invented at Facebook as means to query large datasets in Hadoop, developed to feel similar SQL. Despite being more limited than regular SQL by having little support for inner queries and limitation in update or delete statements, it still permits all types of joins and Group functions. It also provide User Defined Functions(UDFs) which can be written in Java or any other language supported by Hive.

It is typically used by end users and business analysts who wouldn't know how to develop in the MapReduce framework but would regularly use SQL for accessing and extracting data prior to analysis, in the same way they would in regular RDBMs. It supports only structured data.

3 Pig

Pig is a scripting and procedural (kind of, it is called data-flow oriented) language developed by Yahoo. It's main goal is to work as an abstraction layer over the complexity of MapReduce programing (its compiler translates its language, Pig Latin, into sequences of MapReduce programs). Being similar to a scripting language, it can do everything that Hive can and many things it can't. However, it normally involves more steps and lines of code (more complex) to do simpler tasks that Hive would with its SQL-like language. It can have UDFs like Hive. One key difference is that it can store data in variables during the execution of code. It supports both structure and unstructured data.

Pig, being more like a imperative computer language than Hive, is often used by developers, system administrators and engineers who are more familiar with scripting language the regular users and data analysts.

4 Drill

Drill is an open source, schema-less low-latency query engine for big data that delivers secure and interactive SQL analytics. With the ability to discover schemas on-the-fly, Drill is a pioneer in delivering self-service data exploration capabilities on data stored in multiple formats in files or NoSQL databases. Drill is fully ANSI SQL compliant and integrates seamlessly with visualization tools. It doesn't depend on Hadoop components (although it benefits from it). Since it is not a hadoop MapReduce programming abstraction(it has its own SQL engine), it works faster than Hive or Pig. It also supports semi-structured data like JSON.

Drill, being a ANSI-SQL compliant language, works well for any user familiar with SQL specialist or even end-users analysts experienced with databases, since they can query data directly without the need of pre-define DB schemas.

References

[1] What is main differences between hive vs pig vs sql? - Quora. (2016). Quora.com. Retrieved 5 August 2016, from https://www.quora.com/What-is-main-differences-between-hive-vs-pig-vs-sql

[2] Apache Drill — MapR. (2016). Mapr.com. Retrieved 5 August 2016, from https://www.mapr.com/products/apache-drill

Total words (excluding title and references): 481 words