Created by Rahul Pupreja, linkedin.com/in/rahulpupreja

Differences between Hive and Spark!! Today's Big Data Dose for you!!

Hive is specific in nature and only work with selected resources whereas Spark is flexible & smart in nature and seeks for best resources out there.

Hive is a data Warehouse platform with capabilities for managing massive data volumes. The datasets are usually present in Hadoop Distributed File Systems and other databases integrated with the platform. Hive is built on top of Hadoop and provides the measures to read, write, and manage the data. HQL is the query language in use with Hive to perform querying and analytics activities. The operations in Hive are slower than Apache Spark in terms of memory and disk processing as Hive runs on top of Hadoop.

Spark, on the other hand, is an analytics framework to process high-volume datasets. The tool offers a interface with easy usage by offering APIs in numerous languages, such as Scala ,Python, etc. Spark also offers hassle-free integration with other high-level tools. Spark SQL, for instance, enables structured data processing with SQL. Spark also comes with faster operational and computational speed. Intermediate operations occur in Spark within the memory, thereby bringing down the number of reading/write operations.park does not come with its own File Management System. It has to rely on different FMS like Hadoop, Amazon S3 etc.

Read/Write operations: – The number of read/write operations in Hive are greater than in Apache Spark. This is because Spark performs its intermediate operations in memory itself.

Memory Consumption: – Spark is highly expensive in terms of memory than Hive due to its in-memory processing.

Functionalities: – Apache Hive is used for managing the large scale data sets using HiveQL. It does not support any other functionalities. Apache Spark provides multiple libraries for different tasks like graph processing, machine learning algorithms, stream processing etc.

Conclusion

Spark and Hive are essential tools for big data and analytics. Apache Hive provides functionalities like extraction and analysis of data using SQL-like queries. Apache Spark is a great alternative for big data analytics and high speed performance.
It also supports multiple programming languages and provides different libraries for performing various tasks. Both the tools have their pros and cons which are listed above. It depends on the objectives of the organizations whether to select Hive or Spark.
As Spark is highly memory expensive, it will increase the hardware costs for performing the analysis. Hive is going to be temporally expensive if the data sets are huge to analyse. As both the tools are open source, it will depend upon the skillsets of the developers to make the most of it.