Discussion_643-2

Harpreet

Contents

In this article I would like to create a brief summary on how I feel about Hadoop and Spark.

Hadoop is an Apache.org project that is a software library and a framework that allows for distributed processing of large data sets (big data) across computer clusters using simple programming models. Hadoop can scale from single computer systems up to thousands of commodity systems that offer local storage and compute power. Whereas Apache Spark is a fast and general engine for large-scale data processing.

There are business applications where Hadoop outperforms the newcomer Spark, but Spark has its place in the big data space because of its speed and its ease of use. In our case, we had to calculate near real time data analytics, hence Spark was our first choice. MapReduce uses batch processing and was never build for speed. Spark is well known for its performance, but it's also somewhat well known for its ease of use in that it comes with user-friendly APIs for Scala (its native language), Java, Python, and Spark SQL. Spark SQL is very similar to SQL 92, so there's almost no learning curve required in order to use it. Both Hadoop and Spark are fault tolerant and highly scalable.

While working in a project, i was involved in comparing both Hadoop and Spark. Our requirements were quite similar to Spotify's requirements. Basically, we had a very huge consumer base close to around 40 billion and we have to run machine learning algorithm to determine fraudulent activities. In our project we decided to use Spark, but the truth is that Spark and MapReduce have a symbiotic relationship with each other. Hadoop provides features that Spark does not possess, such as a distributed file system and Spark provides real-time, in-memory processing for those data sets that require it.