Apache Spark: A Unified Engine for Big Data Processing

Apache spark is an open source unified analytics engine that processes large scale data. The paper gives us an overview of how spark has enabled new applications to combine a simple programming model capture tasks like capture streaming and interactive workload.

The spark programming model depends on the Resilient Distributed Datasets(RDDs). RDDs are collections of objects segmented over a cluster. These clusters can be run and manipulated parallelly. The spark application can belong to a wide range of fields like finance, scientific data processing which can combine the libraries for graphs, SQL and machine learning.

Benefits of spark as discussed in paper are as follows

1. It is makes the application development easier as they an unified API
2. There is no need to write the data to a storage to pass it to another engine since spark enables multiple functions to be run on the same data in memory.
3. Spark enables machine learning, graphs and SQL that were not available in the previous machines.

In the paper the spark’s design and storage patterns are compared with Google’s MapReduce.

Spark is resilient, it stores the replica of the files it stores and manipulates. It is most commonly used with the cluster file system like HDFS, S3 and Cassandra. The paper also discusses why one would choose spark over MapReduce.

Finally the paper concludes by highlighting the ongoing work in the Apache Spark project, to which both industries and research are contributing.