**Apache Spark**

Apache Spark is an open-source cluster computing framework for big data workloads, such as batch applications, streaming, iterative algorithms, and interactive queries. It is suitable to process real-time data and larger data in network. When compared to MapReduce, Spark performs 100 times faster for a few applications with in-memory primitives and up to 10 times when accessing. The components of a Spark project are as following. Spark Core and Resilient Distributed Dataset (RDDs) are the foundations that provide basic input/output and distributed tasks. RDDs are a collection of data partitioned across machine logically, and they can contain any type of Python, Java, or Scala objects. They can be created by parallelizing an existing collection or referencing external datasets; example of these transformations are reduce, join, filter, and map. The *lazy* RDD transformation operations are optimized for performance by DAGScheduler, which transforms using stages. Spark SQL resides on top of Spark Core, and it introduces SchemaRDD, which is a new data abstraction and supports semi-structured and structured data. Spark Streaming utilizes the fast scheduling by Spark Core, ingests data in small batches, and performs RDD transformation on them. Another component in Spark is the Machine Learning Library (MLlib), which is a distributed framework and runs 9 times faster than the Hadoop disk-based version of Apache Mahout. Last, but not the least, is GraphX, which is a distributed graph-processing framework that provides an API with optimized runtime for the Pregel abstraction; Pregel is a system for large-scale graph processing. Like RDDs, DataFrames are immutable distributions of data organized into named columns that can filter, group, or compute aggregates, and are designed to process large data sets easier. The DataFrame API is accessible in Python, Java, Scale, and R. Spark SQL is a structured data processing module that runs both SQL queries and DataFrame API. Catalyst Optimizer is the core of the Spark SQL. As it eases the optimization techniques and extends the optimizer, it helps to run queries faster than their RDD counterparts. Since a Dataset is a distributed collection of data, a DataFrame is an organized Dataset and faster. In conclusion, when comparing Hadoop Ecosystem and Apache Spark, every type of data can be processed using Spark. The performances are as following: batch processing, structured data analysis, machine learning analysis, interactive SQL analysis, and real-time analysis. The advantages are: speed, as it supports stream processing and interactive queries; combination, as it covers and combines different workloads and processing types for easy tools management; and Hadoop support, as it can create distributed datasets from any file in the Hadoop Distributed File System (HDFS) or other supported storage systems.