Distributed Storage Layer

Hadoop distributed File System

- HDFS is the one, which makes it possible to store different types of large data sets (i.e. structured, unstructured and semi structured data).
- HDFS creates a level of abstraction over the resources, from where we can see the whole HDFS as a single unit.
- It helps us in storing our data across various nodes and maintaining the log file about the stored data (metadata).

Cluster Resource Management

Yet Another Resource Negotiator

- Consider YARN as the brain of your Hadoop Ecosystem. It performs all your processing activities by allocating resources and scheduling tasks.
- It has two major components, i.e. ResourceManager and NodeManager.

Processing Framework Layer

Map Reduce

- It is the core component of processing in a Hadoop Ecosystem as it provides the logic of processing. In other words, MapReduce is a software framework which helps in writing applications that processes large data sets using distributed and parallel algorithms inside Hadoop environment.

Apache Spark

- Apache Spark is a framework for real time data analytics in a distributed computing environment.
- The Spark is written in Scala and was originally developed at the University of California, Berkeley.

Apache Tez

- Apache Tez is a framework for creating a complex directed acyclic graph (DAG) of tasks for processing data. In some cases, it is used as an alternative to Hadoop MapReduce. For example, Pig and Hive workflows can run using Hadoop MapReduce or they can use Tez as an execution engine.

Application Program Interface (API)

Apache Pig

- PIG has two parts: Pig Latin, the language and the pig runtime, for the execution environment. You can better understand it as Java and JVM.
- It supports *pig latin* language, which has SQL like command structure.