**Process**

**HDFS –** Distributed Storage System using Nodes and in batches

**MapReduce -** HDFS service that reduces the data and manages data distribution

**YARN –** Resource Negotiator in HDFS to process the data blocks

MapReduce runs first, then YARN

**PySpark** – Python Interface for Apache Spark

Difference between **HDFS** and **Spark**

|  |  |
| --- | --- |
| **HDFS** | **Spark** |
| Memory on Disk | RAM (In - House Memory) |
| MapReduce process data blocks Sequentially | Cluster based |
| Batchwise | Batchwise, Real Time, Graph Process |
| Code Complexity High | User Friendly |
| Written on Java. Supports Python, R and C++ | Written on Scala, Supports Python, R and Java |
| Stores in Data Nodes, Batches | Stores in Clusters |
| Came first to connect bunch of computers | Spark came alter to enhance mapreduce and uses in-memory |
| Hadoop can used if there is huge amount of data and spark can be used on top of it | If only few Giga Bytes of Memory than we can uses Spark only |
| Hadoop uses **Mahout** (now old school) for processing data and building Models. **Samsara** (Written on Scala) based for algorithms that uses in-memory | Spark has built in M.L & Algorithms and M.L Pipelines |
|  | Spark is 2x faster than MapReduce |
| Hadoop MapReduce depends on External Scheduler (Example ZooKeeper) | Spark has built in Scheduler |

|  |  |  |
| --- | --- | --- |
| **1.** | Hadoop is an open source framework which uses a MapReduce algorithm | Spark is lightning fast cluster computing technology, which extends the MapReduce model to efficiently use with more type of computations. |
| **2.** | Hadoop’s MapReduce model reads and writes from a disk, thus slow down the processing speed | Spark reduces the number of read/write cycles to disk and store intermediate data in-memory, hence faster-processing speed. |
| **3.** | Hadoop is designed to handle batch processing efficiently | Spark is designed to handle real-time data efficiently. |
| **4.** | Hadoop is a high latency computing framework, which does not have an interactive mode | Spark is a low latency computing and can process data interactively. |
| **5.** | With Hadoop MapReduce, a developer can only process data in batch mode only | Spark can process real-time data, from real time events like twitter, facebook |
| **6.** | Hadoop is a cheaper option available while comparing it in terms of cost | Spark requires a lot of RAM to run in-memory, thus increasing the cluster and hence cost. |
| 7. | The PageRank algorithm is used in Hadoop. | Graph computation library called GraphX is used by Spark. |