**Sara hacheche 5672156**

**1/**

|  |  |  |
| --- | --- | --- |
| ////////// | Hadoop | Spark |
| **Architecture** | Supports Hive for SQL-like querying | Supports Spark SQL for querying structured data |
| **Performance** | Limited performance tuning at MapReduce level | Optimized DAGs for efficient step-level optimizations |
| **Costs** | Hadoop experts more widely available | Spark experts are rarer and costlier |
| **Faults and tolerance** | Highly fault tolerant is using HDFS | RDD Lineage Iis using reconstruction |
| **Security** | Supports fine-grained security controls with Apache Sentry | Sparse security model with shared secret authentication |

**3/**

|  |  |  |
| --- | --- | --- |
| Tool | feauture |  |
| HDFS | is an open source distributed file system designed to store and manage large volumes of data on server clusters. It is part of the Hadoop ecosystem |  |
| MAPREDUCE | is the processing engine of Apache Hadoop It allows to easily calculate huge amounts of data by applying mapping and reduction steps in order to find a solution to the given problem |  |
| PIG | is an open-source technology that provides a high-level mechanism for parallel programming of MapReduce jobs to run on Hadoop clusters Pig allows developers to create query execution routines to analyze large, distributed datasets |  |
| HIVE | is a Data Warehouse software. It allows you to easily and quickly perform SQL queries to efficiently extract data from Apache Hadoop. |  |
| SPARK | is a unified, lightning-fast analytics engine for large-scale data processing. It enables large-scale analysis Fast, flexible and offers many advantages for developers |  |
| YARN | is a unified, lightning-fast analytics engine for large-scale data processing. It enables large-scale analysis Fast, flexible and offers many advantages for developers |  |
| SQOOP | Sqoop is a tool for transfer data between hadoop and relational databases . |  |
| NOSQL | A query language that allows storing unstructured data, widely used with Oracle Database, Oracle Wallet and Hadoop. |  |
| MAHOUT | A machine learning library developed by Apache, implemented on Apache Hadoop and using BigData's MapReduce paradigm to generate learning models |  |
| AVRO | A tool to quickly represent complex data structures generated by Hadoo's MapReduce |  |
| GIS TOOLS | Tools used to manage geographical information, in particular allowing the manipulation of geographical coordinates instead of character strings |  |
| FLUME | A data transfer tool used to move large amounts of activity-generated log data into a database, enabling analysis |  |
| CLOUDS | Cloud platforms use Hadoop to process large datasets efficiently, using temporary machines to perform the calculations |  |
| IMPALA | Cloudera software for massively parallel processing of SQL queries, running on Apache Hadoop |  |