HACHECHE SARA

Assignment # 3

**1/ Answer 1 :**

**A/**MapReduce is a distributed programming model and processing system in the field of Big Data. it is used for processing and analyzing large quantitis of data on servers. and it allows performing transformation and operations on large amounts of data in a parallel and efficient way, and the great main advantage of MapReduce is its ability to do all the work in large quantities and quickly .

**B/**

1/ Splits are usually created based on a fixed size or a predefined amount of data , it can convert large file and data into small file and keeping the original data .

2/ mapping is the process of transforming raw data from a data source into a specific data model or a structure suitable for the analysis or processing of this data and this is the first step.

3/ Shyffling is the process of redistributing and organizing data during a distributed processing step, and its tasks to relevant from mapping similar data together and prepare it for the next step of processing.

4/ Reducing is the phase that comes after shuffling and is the final step of the MapReduce programming model this is the step where the data is gathered and combined to have a final result

**2/ Answer 2:**

A / 1.Processing: this is the process of transforming data

2.Writing to disk: he output generated by the mapper, known as intermediate output, is written to the local disk of each node

3.Copy: the intermediate output from the mappers is copied or transferred to the reducer node.

4.Merging and sorting: Once all the mappers have completed their tasks and their outputs have been transferred to the reducer nodes

5.Reducing: the reducer receives input data from all the mappers, which is then processed using anf the output generated by the reducer represents the final result

B/ MapReduce has various applications in machine learning, including building systems for learning from data, distributed searching, distributed sorting, query processing, .....

C / MapReduce is a powerful solution for different tasks in data science , it is able to process and analyze large and huge data in a distributed and parallel way .

**3 / Answer 3 :**

|  |  |
| --- | --- |
| Metadata | Metastore |
| **1** the Hive Metadata processor for records to be written to HDFS  **2** The processor also generates record header attributes  **3** it define the connection information for Hive  **4** italso specify the data format  **5** it define the location of the Hive and Hadoop  **6** Metadata is aprocessor includes a data and a metadata output stream  **7** s the metadata records downstream to the Hive Metastore | 1 In the Metastore, the database keeps all the information  2 it help to find data easily in the Metastore.  3 it contains all the information about tables and relations.  4 l Hive implementations need a metastore service, because all the data is stored inside .  5 It provides single process storage  6 when we use Derby, we cannot run instances of Hive CLI.  7 Metastore is the motor of data because it contains all the information |

4/ Answer 4 :

A/

|  |  |
| --- | --- |
| Hive | HDFS |
| - Hive is an ETL and data warehouse tool  -Hive is a database present in Hadoop ecosystem performs Data Description Language and Data Manipulation Language  - Hive provides different drivers for communication with a different type of applications | -HDFS manages the physical storage of data  -HDFS is a distributed file system designed to -store large amounts of data  -Hadoop Distributed File System is the primary storage system used by Hadoop. |

B/

|  |  |
| --- | --- |
| Hive | Mapreduce |
| - Hive is an ETL and data warehouse tool  -Hive is a database present in Hadoop ecosystem performs Data Description Language and Data Manipulation Language  - Hive provides different drivers for communication with a different type of applications | - MapReduce is a programming model and software framework  - MapReduce provides fine-grained control over the data processing  - MapReduce process large data set |

C/

|  |  |
| --- | --- |
| Hive | Sqoop |
| - Hive is an ETL and data warehouse tool  -Hive is a database present in Hadoop ecosystem performs Data Description Language and Data Manipulation Language  - Hive provides different drivers for communication with a different type of applications | - Sqoop is a tool designed to transfer data  - Sqoop is specialized in importing and exporting data  - Sqoop simplifies the data process between Hadoop and databases. |

D/

|  |  |
| --- | --- |
| Hive | Yarn |
| - Hive is an ETL and data warehouse tool  -Hive is a database present in Hadoop ecosystem performs Data Description Language and Data Manipulation Language  - Hive provides different drivers for communication with a different type of applications | - yarn is a framework responsible for managing resources in a Hadoop  - YARN abstracts the details of resource management  - YARN is responsible in scheduling of applications running on a Hadoop |

**5 / Answer 5 :**

1. Apache Impala : Apache Impala is an open-source interactive query engine designed to perform rapid analysis on large volumes of data. It is part of the Hadoop ecosystem.

2. Presto DB : is a distributed, open-source SQL query processing system designed to perform interactive, ad-hoc analysis on large amounts of data, and it has high performance.

3. Spark SQL : it allows to perform relational data analysis and to process structured data and its characteristics are Distributed and High Performance Processing

4. Shark : It is an open-source SQL query engine which is written in Scala , and he is not well known

5. BigSQL by IBM : It is not open source as it is provided by IBM and they provide SQL AND we can not use it in the query data .

6/ Answer 6:













