ASSIGNMENT # 2

**1.** In your own words: (Please do not copy and paste)

**A)** What is *Hadoop Ecosystem*? (3 or maximum 4 lines)

The Hadoop Ecosystem is a set of tools that are built around the platform of larg datasets on computers using simple programming models, Hadoop also includes many other technologies, HDFS YARN By combining these different technologies, Hadoop allows companies to store, process and analyze large quantities of data

**B)** What is *HDFS*? (2 or maximum 3 lines)

Hadoop Distributed File System is a system made to store large amounts of data on machines. It is used as the main storage in the Hadoop ecosystem, and they run on server machines that work together to store and process data

**C)** What is the difference between Master Nodes and Slave Nodes? (4 lines)

|  |  |
| --- | --- |
| Master nodes | Slave nodes |
| It is responsible for managing | responsible for storing |
| It manages the resource allocation | Responsible for serving |
| It keeps track of the location of data | perform tasks |
| It coordinates the activities of the slave nodes | executing the tasks assigned to them by the master node |

**D)** What is the difference between Hadoop and HDFS? (4 lines)

|  |  |
| --- | --- |
| Hadoop | HDFS |
| It used for distributed storage | It is a distributed file system used by Hadoop |
| processing of large datasets | storing and managing large datasets |
| adoop consists of a lot of components, including the HDFS | HDFS is designed to handle large data sets |
| Hadoop is a software framework | HDFS is used by Hadoop to storage system for large datasets |

**2.** Go to Omnivox (Section Documents)

Make the reading *How to set properties for Secondary NameNode in Hadoop?* And answer the following questions using your own words:

**A)** What is the meaning of EditLog and FsImage?

1/ EditLog saves data modification transactions to the file system and is used for crash recovery.

2/Fslmage It represents the state of the filesystem as it should be after applying all changes recorded in the EditLog. The FsImage is used to quickly boot the system after a crash

**B)** Is the Secondary NameNode a back up of NameNode server?

The Secondary NameNode is not a back up for the NameNode server.

**C)** What is the main function of Secondary NameNode?

The main function of secondory NameNode is to create a new merge file Fslmage in merged periods, the NameNode picks up the files FsImage and EditLog

**D)** What is the property that defines the number of uncheckpointed transaction on the Name Node?

the name of property used for importing data is dfs.namenode.checkpoint.txns,

This property defines the number of transactions on the NameNode

**3.** Explain using your own words.

**A)** Why do we use the expression *Data ingestion*? (3 lines)

We use Data ingestion because it is an important step in data processing, especially in Big Data systems. Data ingestion is the collection and storage of data from different sources, Data ingestion can also include data quality operations, such as error detection and correction

**B)** What is the difference between *Data ingestion* and *ETL*? (4 lines)

|  |  |
| --- | --- |
| Data integration | ETL |
| process of collecting data | ETL is a data integration process that involves extracting data from various sources |
| Process storing data from various sources | ETL is more complex than data ingestion |
| Data ingestion is the first step in the data processing | ETL is focused on transforming |
| prepare the data for analysis | ETL is focused on integrating data from multiple sources |

**C)** Why *Data ingestion* and *ETL* are important in Big Data?

In big data the collection, storage, transformation and integration of data are dificult steps to ensure that data is available, accessible and useful for analysis and decision-making, Data collection and storage Preparation of data for analysis informed decision making because these last two are very important in big data

**D)** What is Sqoop and what are its features? (6 lines)

Sqoop is a big data for transferring data between Hadoop and relational database servers. Sqoop transfers data, Big Data Sqoop can also be used to transform data in Hadoop and export them We can also export data to relational databases using Sqoop.

The main features of Sqoop are provides connectivity to a wide variety of data sources, Sqoop is able to import data in large quantities Sqoop allows to perform data transformations during transfer Sqoop was designed to handle massive quantities of data .

**4.** Go to:

<https://www.projectpro.io/hadoop-tutorial/hadoop-sqoop-tutorial-data-aggregation>

**A)** Make the reading and explain:

1. What is the problem statement? (3 lines)

the problem is when the site is visited by many people the online application greatly increases the storage of data on the MYSQL and the processing of voluminous data of the data it receives every day is impossible to manage with RDMS

2. What is the proposed solution? (2 lines)

The solution that distributed systems can be used to process data, which involves bringing the data to the distributed systems and processing it in the system distribute, and with this method we will solve the problem

**5.** Practice: Intro to *HDFS* and *Sqoop*.

- Attach the screenshots of the practice in the assignment.





