**Big Data assessment: -**

1. Difference between MapReduce and Spark?

* MapReduce processes data in two phases - Map and Reduce, whereas Spark has data processing model that allows Map, Reduce, and other operations like filter, join etc.
* MapReduce has a built-in execution engine, whereas Spark has multiple execution engines like Spark Standalone, Apache Mesos, and Hadoop YARN.

1. Difference between Flume and Sqoop?

* Flume can ingest data from a wide range of sources like log files, social media, web servers, and sensors, whereas Sqoop is mainly used for ingesting data from relational databases.
* Flume can handle various data formats like CSV, JSON, XML, and plain text, while Sqoop can only import and export data in structured formats like CSV and relational database tables.

1. You have database of 3 employment websites. All resumes are in same template. Your task is to make 3 sheets.

Source| Full Name  | Address | Phone number | Email id | Skills | Experience | Projects Worked

A)-> First one to extract the important data.

So, for the job important columns are

Full Name | Phone number| Email id| Skills| Experience| Projects Worked

B)-> second one what transformation you perform.

-Removing the duplicates.

-Because it the hiring of the fresher will remove the 3+ year experience.

-Removing the row which doesn’t have the relevant skillset.

c)-> last one Entity Relationship model

|  |
| --- |
| table 1 |
| Email |
| Skills |
| Experience |

|  |
| --- |
| Table 2 |
| Full name |
| Phone number |
| Email |

One to many

One to one

|  |
| --- |
| Table 3 |
| Email |
| Project work |

Many to one

* Data Extraction

Flume,Kafka,Sqoop

* Data Tranfromation

Mapreduce,Spark