Big Data assessment: -

1) **Difference between MapReduce and Spark?**
   - Spark's data processing model includes a variety of operations such as Map, Reduce, filter, join, and others, whereas MapReduce is limited to two phases of data processing, namely Map and Reduce.

   - Unlike MapReduce, which has a single built-in execution engine, Spark offers multiple execution engines such as Spark Standalone, Apache Mesos, and Hadoop YARN

2) **Difference between Flume and Sqoop?**
   - While Sqoop is primarily designed for ingesting data from relational databases, Flume is capable of ingesting data from a diverse range of sources such as log files, social media platforms, web servers, and sensors.

   - Flume has the ability to handle a wide range of data formats such as CSV, JSON, XML, and plain text, whereas Sqoop is limited to importing and exporting structured data formats like CSV and relational database tables only.

3) **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

Source	Full Name	Address	Phone number	Email id	Skills	Experience	Projects Worked
Indeed	ABC	XYZ	234	Abc@gmail.com	Python, Java, SQL	1 years	Customer Churn Prediction Model, Fraud Detection Model
Hirect	DFG	qwe	467	dfg@gmail.com	Java, JavaScript, React	0 years	Restaurant Ordering System, Online Marketplace
Naukri	ABC	Xyz	234	Abc@gmail.com	Python, Java, SQL	1 years	Customer Churn Prediction Model, Fraud Detection Model
LinkedIn	HIJ	Hjk	890	hij@gmail.com	C++, Python, R	5 years	Inventory Management System, CRM System
indeed	KLM	uio	789	klm@gmail.com	Python, SQL, AWS	3 years	Migration to AWS, Implementation of Kubernetes

A)-> First one to extract the important data.
If you are looking to hire someone for a job, the important columns in the candidate's profile could be:

- Full Name: This is important to identify the candidate.

- Phone Number: This is necessary for the recruiter or the hiring manager to contact the candidate.

- Email Address: This is essential for communication with the candidate and sharing important details about the job.

- Skills: This column should list the candidate's skills relevant to the job.

- Experience: This should include the candidate's work experience, particularly as it relates to the job you are hiring for.
- Projects Worked: This should include a list of projects the candidate has worked on in the past, particularly those that demonstrate their skills and experience in the field you are hiring for.

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

Full Name	Phone number	Email id	Skills	Experience	Projects Worked
ABC	234	Abc@gmail.com	Python, Java, SQL	1 years	Customer Churn Prediction Model, Fraud Detection Model
DFG	467	dfg@gmail.com	Java, JavaScript, React	0 years	Restaurant Ordering System, Online Marketplace
ABC	234	Abc@gmail.com	Python, Java, SQL	1 years	Customer Churn Prediction Model, Fraud Detection Model
HIJ	890	hij@gmail.com	C++, Python, R	5 years	Inventory Management System, CRM System
KLM	789	klm@gmail.com	Python, SQL, AWS	3 years	Migration to AWS, Implementation of Kubernetes

B)-> second one what transformation you perform.

If you want to filter the candidates' profiles for hiring purposes, you may consider the following steps:

1. Removing candidates with more than 3 years of experience: This is a criterion you have set to hire freshers, so removing profiles with more than 3 years of experience would ensure that you only consider candidates who meet this requirement.

2. Filtering candidates based on relevant skillset: You may want to remove candidates who do not have the necessary skills for the job you are hiring for. You can filter

candidates based on the skills listed in their profile and remove any profiles that do not match the required skillset. This will help you narrow down your selection to only those candidates who are the best fit for the job.

3. In the dataset, the columns containing more than one value are being split.

Full Name	Phone number	Email id	Skills1	Skills 2	Skill3	Experience	Project 1	Project 2
ABC	234	Abc@gmail.com	Python	SQL	Java	1	Customer Churn Prediction Model,	Fraud De

c)-> last one Entity Relationship model

Table 1: Employee Information

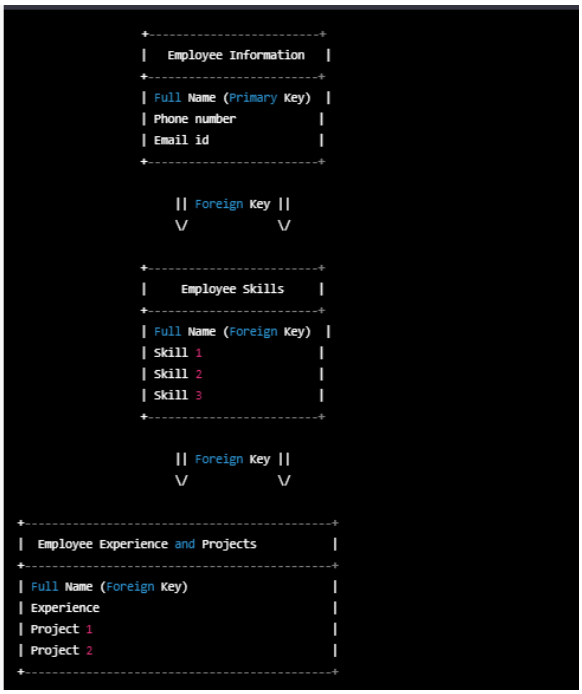
- Full Name
- Phone number
- Email id

Table 2: Employee Skills

- Full Name (Foreign key referencing Employee Information table)
- Skill 1
- Skill 2
- Skill 3

Table 3: Employee Experience and Projects

- Full Name (Foreign key referencing Employee Information table)
- Experience
- Project 1
- Project 2



4.

The following technologies can be used for the two phases of data processing:

- Data Extraction: Flume, Kafka, and Sqoop are commonly used tools for extracting data from various sources such as databases, social media platforms, web servers, and sensors.

- Data Transformation: MapReduce and Spark are popular tools used for processing and transforming the extracted data. These tools can perform operations such as filtering, aggregating, and joining data to convert it into a structured format suitable for analysis and storage.

- Data warehouses: Data warehouses such as Snowflake and Amazon Redshift can be used to store large volumes of data extracted from job sites. These platforms provide scalable storage and processing capabilities and can integrate with a range of data processing tools

- Cloud-based data platforms: Cloud-based data platforms such as Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure provide a range of services for data extraction and processing. These services include web scraping tools, API integration, ETL tools, and data storage options.