BIG DATA ASSIGNMENT

1.Difference between MapReduce and Spark?

Feature	MapReduce	Spark
Processing model	Batch processing	Batch and real-time processing
Speed	Slower, reads data from disk after every operation	Faster, uses in-memory processing
Ease of use	More complex programming model	Simpler programming model and more intuitive API
Built-in modules	No built-in modules for SQL, streaming, ML, graph processing	Built-in modules for SQL, streaming, ML, graph processing, and more
Resource management	Limited resource management	Better resource management for complex workflows

2.Difference between Flume and Sqoop?

Feature	Flume	Sqoop
Use case	Transfer streaming data into Hadoop	Import/export data between Hadoop and relational databases
Data source	Collects data from various sources, such as log files and event streams	Works specifically with relational databases
Data transfer mode	Real-time or near-real-time data transfer	Batch processing and transferring large volumes of data
Data format	Designed for unstructured data, such as log files or event streams	Optimized for structured data in a database, such as tables or queries
Data processing	Can process data as it is being transferred, allowing for cleansing, filtering, and transformation	Does not have built-in data processing capabilities, relies on external tools or scripts

3.For below use case

- You have database of 3 employment websites. All resumes are in same template.
- Your task is to make 3 sheets. First one to extract the important data, second one what transformation you perform, last one Entity Relationship model.

1.Data Extraction:

To extract important data from resumes, the following information can be considered important:

- Candidate Name
- Contact Information
- Education Qualifications
- Work Experience
- Skills and Achievements

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

Full Name	Phone number	Skills	Experience	Projects Worked
John Smith	555-123-4567	Python, Java, SQL	Software Engineer, 3 years experience	Inventory Management System, CRM System
Sarah Lee	555-987-6543	HTML, CSS, JavaScript	Front-end Developer, 2 years experience	E-commerce Website, Online Learning Portal
Michael Johnson	555-555-1212	C++, Python, R	Data Scientist, 4 years experience	Customer Churn Prediction Model, Fraud Detection Model
Emily Wang	555-111-2222	Java, JavaScript, React	Full-stack Developer, 1 year experience	Restaurant Ordering System, Online Marketplace
Kevin Kim	555-222-3333	Python, SQL, AWS	Cloud Architect, 5 years experience	Migration to AWS, Implementation of Kubernetes

2.Data Transformation:

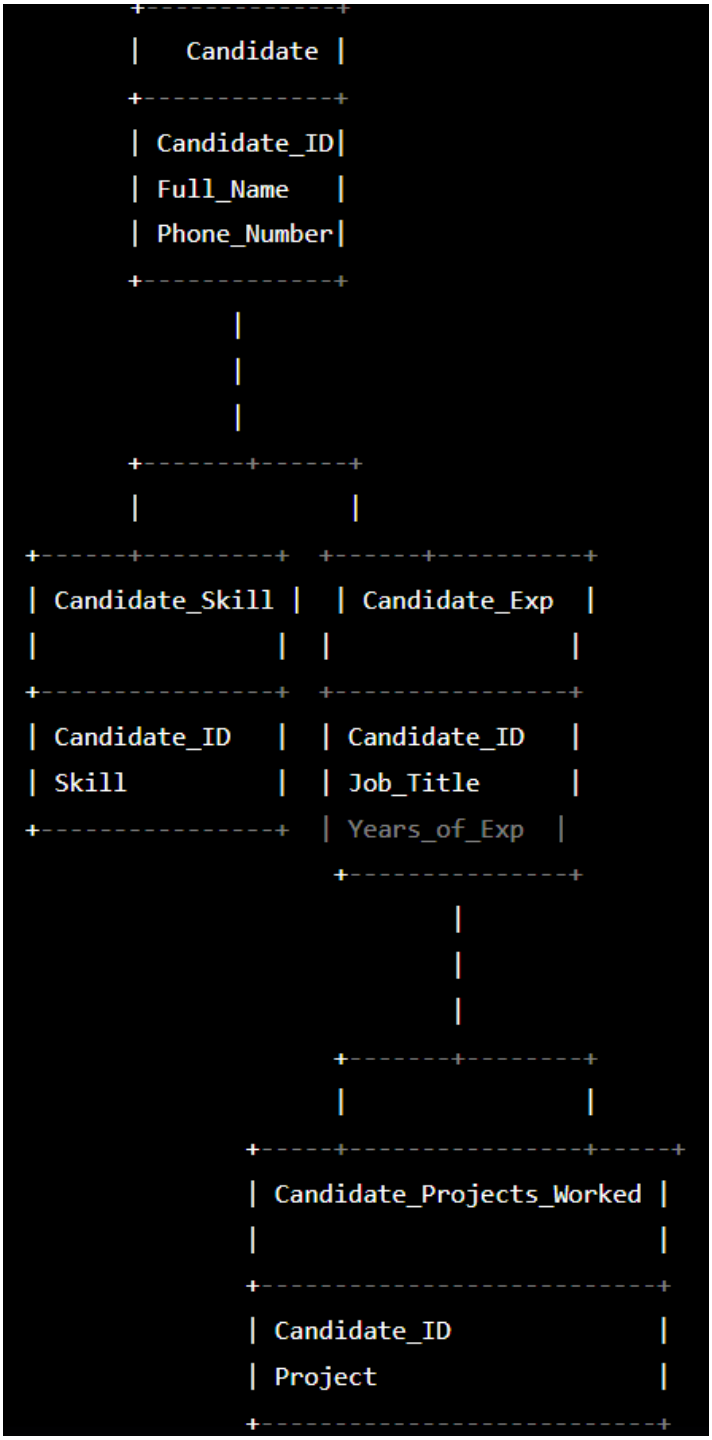
Once the important data has been extracted from the resumes, we can perform the following transformations on the data:

- Standardize the formatting of the data to a common format across all resumes.
- Normalize the data to remove any redundancies and inconsistencies.
- Use data cleaning techniques to remove any irrelevant or incomplete data.
- Use data aggregation techniques to group similar data together.

Full Name	Phone number	Skill 1	Skill 2	Skill 3	Job Title	Years of Experience	Project 1	Project 2
John Smith	555-123-4567	Python	Java	SQL	Software Engineer	3	Inventory Management System	CRM System
Sarah Lee	555-987-6543	HTML	CSS	JavaScript	Front-end Developer	2	E-commerce Website	Online Learning Portal
Michael Johnson	555-555-1212	C++	Python	R	Data Scientist	4	Customer Churn Prediction Model	Fraud Detection Model
Emily Wang	555-111-2222	Java	JavaScript	React	Full-stack Developer	1	Restaurant Ordering System	Online Marketplace
Kevin Kim	555-222-3333	Python	SQL	AWS	Cloud Architect	5	Migration to AWS	Implementation of Kubernetes

3.Entity Relationship Model:

This diagram shows four entities - Candidate, Candidate_Skill, Candidate_Exp, and Candidate_Projects_Worked - with one-to-many relationships between Candidate and each of the other entities. The Candidate entity has attributes such as Candidate_ID, Full_Name, and Phone_Number. The Candidate_Skill entity has attributes such as Candidate_ID and Skill. The Candidate_Exp entity has attributes such as Candidate_ID, Job_Title, and Years_of_Exp. The Candidate_Projects_Worked entity has attributes such as Candidate_ID and Project.



4.what technologies you would use to process them?

1. Web scraping tools: Web scraping tools such as Scrapy, Beautiful Soup, and Selenium can be used to automatically extract data from job sites. These tools can be used to navigate web pages, identify relevant data fields, and extract data in a structured format.

2. API integration: Many job sites offer APIs that allow developers to access data in a structured format. These APIs can be integrated into data pipelines to extract data in a programmatic way.

3. ETL tools: ETL (Extract, Transform, Load) tools such as Apache Nifi, Talend, and Apache Airflow can be used to automate the process of extracting, cleaning, and transforming data from job sites. These tools provide a graphical interface for building data pipelines and can handle large volumes of data.

4. 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.

5. 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.