1. Difference between Map Reduce and Spark

|  |  |
| --- | --- |
| **Map Reduce** | **Spark** |
| It is written in Java. | It can be written in Python, Scala, Java, R. |
| It can be used for batch processing. | It can be used for real time and batch processing. |
| It is slower as it performs operations on disk but is useful for large data batch processing. | It is faster as it stores data and intermediate output on RAM and does necessary processing or computation. |
| It is more failure tolerant as it depends on hard disks, if process crashes then it will start from where it left. | It is less failure tolerant as it relies on RAM, if there is a crash then it must start processing from beginning. |
| It is more secure as it can use different Hadoop elements to maintain security. | It is more vulnerable to attack as security is set to ‘off’ by default. |

1. Difference between Flume and Sqoop

|  |  |
| --- | --- |
| **Flume** | **Sqoop** |
| It collects, aggregates, and moves large quantities of streaming data from different sources to central repository (HDFS). | It is used for bulk of data between Hadoop and relational databases which has JDBC. |
| It is used with semi-structured and unstructured data. | It is used with structured data and for parallel data transfer. |
| Flume is completely event driven. | Sqoop is not driven by events. |
| It follows agent-based architecture. Agent is responsible for fetching data. | It follows connector-based architecture. Connectors know how to connect to a different data source. |
| Only ingestion/import of log data to HDFS is possible. | We can import/export data to/from databases. |

1. 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 transformations you perform, last one Entity Relationship model.

Source | Full Name | Address | Phone Number | Email Id | Skills | Experience | Projects Worked

[Extraction, Transfom, ERD Excel File](https://wgcp-my.sharepoint.com/personal/saijyosthna_yenumula_blend360_com/Documents/Desktop/COHORT10/big%20data/Big%20Data_Sai%20Jyosthna/BigDataQ3.xlsx)

1. What technologies you would use to process them?

For extraction of semi structured and unstructured data, we can use Flume, for structured data we can use Sqoop. OpenNLP can also be used to identify and extract specific information from resumes. Tika can also be used to extract text and metadata from pdf, doc etc and analyse text to get relevant information.

HDFS stores the information. Hive can be used for querying and analysing the extracted data to generate insights on factors like skills, experience etc. Map Reduce and Spark can also be used for processing and analysing the extracted data. Spark can be used to process and analyse resume data stored in Hadoop and extract relevant information like skills, experience etc.

|  |  |
| --- | --- |
| Extraction | Flume, Sqoop, OpenNLP, Tika |
| Processing | Spark, Map Reduce |
| ER Diagram design | Draw.io, SmartDraw, LucidChart |