**Question-1: Difference between MapReduce and Spark?**

|  |  |
| --- | --- |
| **Map Reduce** | **Spark** |
| It uses Batch processing for processing data i.e. It takes the data as an input from the disk and perform the required computations and uploading the updated data back into the disk. Batch Processing is Slower in comparison to Spark’s Memory Processing Technique. | It uses Memory processing for processing data i.e. The data is directly stored into the memory of the computer and the computations are done directly in the memory without having access of the disk. This process makes the computation faster as compared to the batch processing in Map Reduce. |
| It has Low-level API which makes it difficult for the developers to work with. But is Cheap therefore can be used my mass. | It has High-level API which makes it easier for developer to understand and work. But this makes it expensive. |
| It uses Replication technique for handling Fault tolerance. It stores the data across multiple nodes for availability which increases the overhead. | Spark uses Resilient distributed datasets (RDDs) for handling fault tolerance. The data is distributed across multiple nodes in the cluster. If any node fails during processing data can be automatically reconstructed from remaining nodes. |
| It is closely tied to the Hadoop ecosystem. | It can be used independently of the Hadoop. |
| It doesn’t support Machine Learning. | It Supports Machine Learning Models as a built-in support. Therefore, Spark is more popular among Data Science and Machine Learning applications. |

**Question-2: Difference between Flume and Sqoop?**

|  |  |
| --- | --- |
| **Flume** | **Sqoop** |
| It uses Wide range of data sources i.e., log files, Social Media Files, and all streaming sources. | It is designed to take the data just from the relational database in the Hadoop. |
| It supports data in the different format including Json , text and Avro. | It does not support all formats. It supports binary and text format. |
| In Apache Flume data flows to HDFS through multiple channels | Apache Sqoop HDFS is the destination for importing data. |
| It has agent-based architecture i.e., the code written in flume is known as agent which is responsible for fetching | The architecture is based on connectors. The connectors in Sqoop know how to connect with the various data sources and fetch data accordingly. |
| It is mainly used for parallel data transfers, for data imports as it copies data quickly. | It is used for collecting and aggregating data because of its distributed, reliable nature and highly available backup routes. |