**Big Data Assignment First 2 Questions**

**1)Difference between Map reduce and Spark?**

|  |  |
| --- | --- |
| Map Reduce | Spark |
| Map reduce relies on rom | Spark relies on ram |
| It is used for batch processing | It can handle all types of data processing |
| It supports only java | It can support python, java, scala, sql,r |
| Map reduce is cheaper in cost | It is expensive due to it’s in memory processing power and ram requirement |
| Dependent on external job scheduler like oozie | Can schedule all tasks itself |
| It requires HDFS, YARN to process | We can use it on any platform like cloud,  Some external storages. |
| Not faster compared to Spark | It is 100 times faster |

**2)Difference between Sqoop and Flume?**

|  |  |
| --- | --- |
| Sqoop | Flume |
| Sqoop is used to ingest structured data into HDFS | Flume is used to ingest unstructured and semi structured data in HDFS |
| Sqoop is not driven by events | Flume is completely even driven |
| Sqoop follows connector-based architecture | Flume follows agent-based architecture |
| Sqoop is used for parallel data transfer for this reason the output could be in multiple files | Flume is used for collecting aggregating data because is distributed in nature |
| Apache Sqoop reduces the processing loads and excessive storage transferring them to the other systems. Thus, have fast performance. | Apache Flume is highly robust, fault-tolerant, and has a tuneable reliability mechanism for failover and recovery. |