Que.1 Justify the need of MAP and REDUCE in Hadoop Ecosystem?

Ans. MapReduce is a software framework and programming model used for processing huge amounts of data. MapReduce program work in two phases, namely, Map and Reduce. Map tasks deal with splitting and mapping of data while Reduce tasks shuffle and Reduce the data.

Hadoop is capable of running MapReduce programs written in various languages: Java, Ruby, Python, and C++. The programs of Map Reduce in cloud computing are parallel in nature, thus are very useful for performing large-scale data analysis using multiple machines in the cluster.

The input to each phase is key-value pairs. In addition, every programmer needs to specify two functions: map function and reduce function.

Que. 2 Decode the concept of Hadoop Ecosystem.

Ans. Apache Hadoop is an open source framework intended to make interaction with big data. Hadoop ecosystem is a plateform or a suite which provides various services to solve the big data problems. It includes Apache projects and various commercial tools and solutions. There are four major elements of Hadoop i.e. HDFS, MapReduce, YARN, and Hadoop Common. Most of the tools or solutions are used to supplement or support these major elements. All these tools work collectively to provide services such as absorption, analysis, storage and maintenance of data etc.

Following are the components that collectively form a Hadoop ecosystem:

- HDFS :- Hadoop Distributed File System
- YARN :- Yet Another Resource Negotiator
- MapReduce :- Programming based Data processing
- Spark :- In-memory data processing
- PIG, HIVE :- Query based processing of data services
- HBase :- NOSQL Database
- Mahout, Spark MLLib :- Machine Learning algorithm libraries
- Solar, Lucene :- Searching and Indexing
- Zookeeper :- Managing cluster
- Oozie :- Job Scheduling