***Q1. Define Big Data. Explain the Evolution of Big Data and their characteristics.***

***Ans.***

*1. Big data refers to very large data sets that are too big to be processed on a regular computer.*

*2. Large network-based systems generate big data in both standard and non-standard formats.*

*3. Cloud computing services have arisen mainly due to big data.*

*4. Two big data tools are Apache Hadoop and Apache Storm.*

*5. Amazon and Redshift are big data sources.*

*6. Big data offers benefits such as cost savings, time reduction, and faster decision-making.*

*7. The characteristics of big data include volume, velocity, and variety.*

*8. Big data can be structured, unstructured, semi-structured and hybrid.*

*9. Big data is useful in many applications, such as recommendation engines, fraud detection, and market basket analysis.*

***Q2. Describe the MapReduce execution steps with neat diagram***

***Ans.***

*1. MapReduce is a software framework that processes huge data sets through two phases: Map and Reduce, where Map splits and maps data and Reduce shuffles and reduces it.*

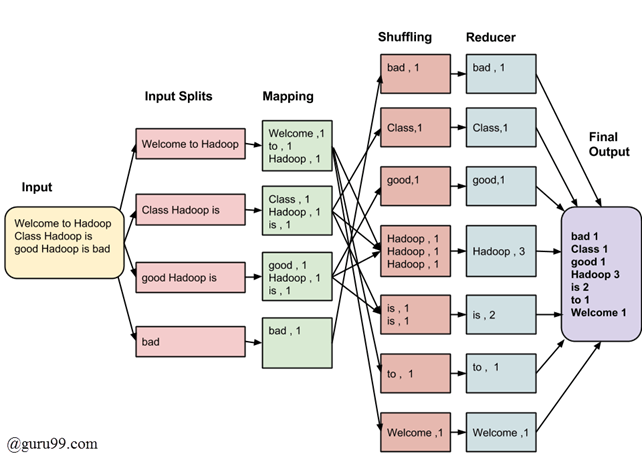
*2. Hadoop can run MapReduce programs in multiple languages and they are parallel, making them useful for large-scale data analysis on multiple machines in a cluster.*

*3. Input to a MapReduce job is split into fixed-size pieces called input splits in Big Data.*

*4. The mapping phase is the first step in executing a map-reduce program where each split's data is processed using a mapping function.*

*5. The shuffling phase in MapReduce combines relevant records from the output of the mapping phase.*

*6. The reducing phase in MapReduce aggregates output values from the shuffling phase and returns a single output value.*



***Q3. Illustrate the Hadoop core components with neat diagram.***

***Ans:***

*1. Hadoop is a framework used by data analysts to manage big data, which has three components: HDFS, MapReduce, and YARN.*

*2. Hadoop is user-friendly and useful for data analysts.*

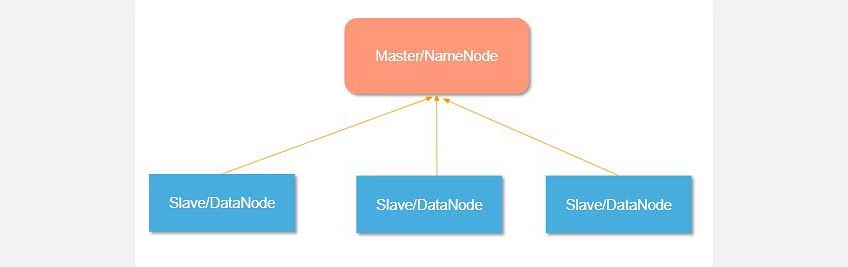
*3. Hadoop can store and process huge amounts of data quickly.*

*4. Hadoop improves operational decision-making and batch workloads.*

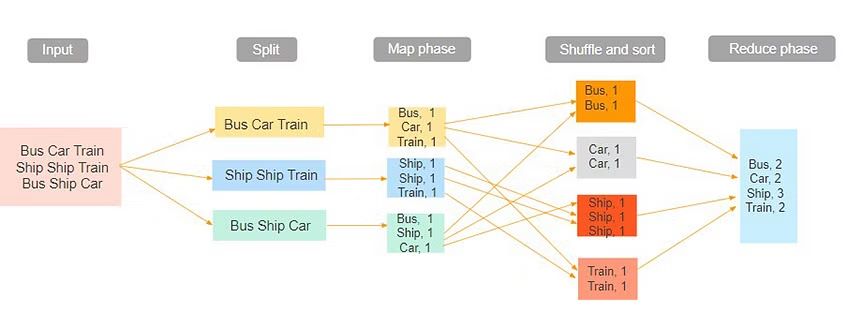
*5. Hadoop is scalable, allowing for the addition of more nodes and picking up more data.*

*6. Companies such as British Airways, Uber, Netflix, and Twitter use Hadoop.*

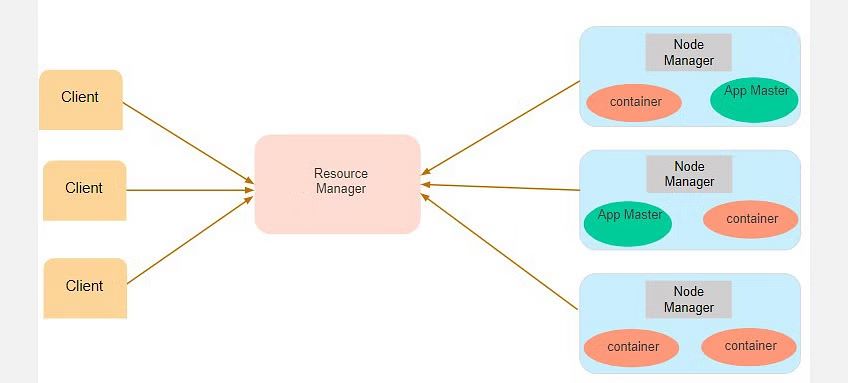
*7. Hadoop HDFS distributes data storage, with the name node as the master and data nodes as slaves forming the HDFS cluster.* *HDFS distributes data storage across multiple nodes, with one name node and many data nodes. HDFS clusters are composed of master (name) and slave (data) nodes. The name node manages the data nodes and stores metadata. Data nodes read, write, process, and replicate data, sending heartbeats to the name node to report their status.*



*8. Hadoop MapReduce processes data at slave nodes and sends the final result to the master node. Refer Q2…..*



*9. Hadoop YARN is the resource management unit of Hadoop available in version 2. It acts like an OS to Hadoop. It is a file system that is built on top of HDFS. It performs job scheduling to make sure that the jobs are scheduled in the right place.*



*10. Hadoop offers advantages such as speed, cost-effectiveness, and scalability.*

*11. Hadoop is used in various sectors such as finance, healthcare, and advertising.*

***Q4. Describe the structure of HDFS in a Hadoop ecosystem using a diagram.***

***Ans.****1. HDFS distributes data storage across multiple nodes, with one name node and many data nodes.*

*2. HDFS provides distributed storage and data security, with automatic data replication.*

*3. HDFS clusters are composed of master (name) and slave (data) nodes.*

*4. The name node manages the data nodes and stores metadata.*

*5. Data nodes read, write, process, and replicate data, sending heartbeats to the name node to report their status.*

