Big Data 2 marks

1) What is the role of big data analytics?

-> - Big data significantly reduce costs when it comes to storing large amounts of data, finding more efficient ways of doing businesses

- Making faster, better decisions with the ability to analyze new sources of data.

- Developing & marketing new products and services

2) What are the challenges in Big Data?

-> ① Storage : With a vast amount of data generator daily, the greatest challenge is storage with legacy-systems and unstructured data cannot be stored in traditional databases.

② Processing : It refers to reading, transforming, extraction and formatting of useful information from raw data.

③ Security : It is a big concern for organizations where non-encrypted information is at risk of theft or damaged by cyber criminals therefore data security professionals must balance to data against maintaining strict security protocols.

3) Generalize the role of analytical tools in bigdata.

→ The various analytical tools of bigdata help to uncover the patterns from raw data and derive valuable insights from it. It helps businesses to get real time insights about sales, marketing, finance to achieve better results.

4) What is Zookeeper? What are the services provided by it?

→ Zookeeper is an open-source Apache project that provides a centralized service for providing configuration information. The services provided by it are as follows:

① Configuration information, naming and group services.
② Cluster Management
③ Leader Election
④ Locking & synchronization service
⑤ ~~Hia~~ Highly reliable data registry.

5) What is Znode? Mention its types?

→ Every node in a Zookeeper tree is referred to as a Znode.

Types of Znode:
① Persistence Znode
② Ephemeral Znode
③ Sequential Znode

6) Write about the flume application and its components.

→ It is a distributed, reliable and available service for efficiently collecting, aggregating and moving large amounts of log data.

Components of flume application:

1) Source

2) Channel

3) Sink

7) Define HBase. Write about its storage mechanism.

→ HBase is a distributed column-oriented database, built on top of the Hadoop file system. It is an open-source project and is horizontally scalable.

Storage Mechanism:

i) It is a column oriented database and the tables in it are sorted by row.

ii) Subsequent column values are sorted contiguously on the disk.

iii) Each cell value of the table has a time stamp.

8) What is Apache Pig and what is the need of it?

→ It is a platform for analyzing large datasets that consists of high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs.

9) Write about the applications of kafka.

→ Kafka is used to build real time streaming data pipelines & real time streaming applications.

Example: Twitter, LinkedIn, Netflix.

10) Define Sampling Distribution.

→ It is all possible values of statistics and their probabilities of occuring for a sample of a particular size. It can be used to calculate the probability that the sample statistic have occurred by chance and thus to decide by something that is true of a sample statistic, is also likely true of a population parameter.