Q1. What is Big Data?

Answer: Big Data is a collection of data that is huge in volume, yet growing exponentially with time. It is data with such large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also data but with huge size.

Q2. Write Big Data Characteristics.

Answer: Characteristics:

1. Volume –
   - The name Big Data itself is related to a size which is enormous. Size of data plays a very crucial role in determining value out of data.
2. Variety –
   - The next aspect of Big Data is its variety.Variety refers to heterogeneous sources and the nature of data, both structured and unstructured.
3. Velocity –
   - The term 'velocity' refers to the speed of generation of data.
4. Variability –
   - This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

Q3. Types of Big Data.

Answer: Types:

1. Structured
    - Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data.
2. Unstructured
    - Any data with unknown form or the structure is classified as unstructured data.
    - A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc.
3. Semi-structured
    - Semi-structured data can contain both the forms of data. We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational DBMS
    - . Example of semi-structured data is data represented in an XML file.

Q4. Traditional vs. Big Data business approach

Answer:

Traditional	Big data business
Traditional data is generated at the enterprise level.	Big data is generated outside and at the enterprise level.
Traditional database tools are required to perform any database operation.	Special kinds of database tools are required to perform any database operation.
Its volume ranges from Gigabytes to Terabytes.	Its volume ranges from Petabytes to Zettabytes or Exabytes.
Data integration is very easy.	Data integration is very difficult.

Q5. write the concept of Hadoop.

Answer: Hadoop is designed to scale up from a single server to thousands of machines, each offering local computation and storage.

- Hadoop runs code across a cluster of computers.Hadoop framework allows the user to quickly write and test distributed systems.

Q6. What is the Hadoop Ecosystem?

Answer: Apache Hadoop ecosystem refers to the various components of the Apache Hadoop software library; it includes open source projects as well as a complete range of complementary tools.

- Some of the most well-known tools of the Hadoop ecosystem include HDFS, Hive, Pig, YARN, Spark, HBase, Oozie, Sqoop, Zookeeper, etc.
- 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.

Q7. What is HDFS?

Answer:HDFS stands for Hadoop Distributed File System

- Hadoop Distributed File System , is one of the largest Apache projects and primary storage system of Hadoop.
- It employs a NameNode and DataNode architecture.
- It is a distributed file system able to store large files running over the cluster of commodity hardware.

Q8. What are the map tasks ?

Answer: Map Task is a single instance of a MapReduce app. These tasks determine which records to process from a data block.

- The input data is split and analyzed, in parallel, on the assigned compute resources in a Hadoop cluster.
- It takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key-value pairs).

Q9. Discuss MapReduce.

Answer:It is a Programming based Data Processing

- Hadoop MapReduce is the processing unit of Hadoop. In the MapReduce approach, the processing is done at the slave nodes, and the final result is sent to the master node.
- A data containing code is used to process the entire data. This coded data is usually very small in comparison to the data itself.
- You only need to send a few kilobytes worth of code to perform a heavy-duty process on computers.

Q11. Discuss Computing Selections by MapReduce.

Answer: Apply a condition c to each table in the relation and produce as output only those tuples that satisfy c.

- The result of this selection is denoted by 6c(R)
- Selection really does not need the full power of MapReduce. They can be done most conveniently in the map portion alone, although they could also be done in the reduced portion also.
- The pseudo code is as follows :

  Map (key, valve)
  for tuple in valve :
  if tuple satisfies C :
  emit (tuple, tuple)
  Reduce (key, valves)

MapReduce is a framework using which we can write applications to process huge amounts of data, in parallel, on large clusters of commodity hardware in a reliable manner.

emit (key, key)

Q12. What are data visualizations?

Answer: Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

Q13. What are the hadoop limitations ?

Answer: Hadoop limitations

1. Support for Batch Processing only
    - Hadoop supports batch processing only, it does not process streamed data, and hence overall performance is slower.
    - The MapReduce framework of Hadoop does not leverage the memory of the Hadoop cluster to the maximum.
2. No Delta Iteration
    - Hadoop is not so efficient for iterative processing, as Hadoop does not support cyclic data flow
3. Latency
    - In Hadoop, MapReduce framework is comparatively slower, since it is for supporting different format, structure and huge volume of data.
4. Not Easy to Use
    - In Hadoop, MapReduce developers need to hand code for each and every operation which makes it very difficult to work.
5. Security
    - Hadoop is challenging in managing complex applications. If the user doesn't know how to enable a platform who is managing the platform, your data can be a huge risk.

6. No Abstraction
    - Hadoop does not have any type of abstraction so MapReduce developers need to hand code for each and every operation which makes it very difficult to work.
7. Vulnerable by NatureNo Caching
    - Hadoop is not efficient for caching. In Hadoop, MapReduce cannot cache the intermediate data in memory for a further requirement which diminishes the performance of Hadoop.
8. Uncertainty
    - Hadoop only ensures that the data job is complete, but it's unable to guarantee when the job will be complete.

Q14. Describe NoSQL Business Drivers?

Answer: There are 4 major business drivers for SQL as:

1. Volume
    - The key factor pushing organizations to look at alternatives to their current RDBMSs is a need to query big data using clusters of commodity processors.
2. Velocity
    - Though big data problems are a consideration for many organizations moving away from RDBMSs, the ability of a single processor system to rapidly read and write data is also key.
3. Variability
    - Companies that want to capture and report on exception data struggle when attempting to use rigid database schema structures imposed by RDBMSs.
4. Agility
    - The most complex part of building applications using RDBMSs is the process of putting data into and getting data out of the database.

Q15. What is noSQL?

Answer: NoSQL database technology stores information in JSON documents instead of columns and rows used by relational databases. NoSQL stands for “not only SQL” rather than “no SQL” at all. the four most-popular types of NoSQL database.

Q16. Write NoSQL systems to handle big data Problems.

Answer: NoSQL systems to handle big data problems:

1. The queries should be moved to the data rather than moving data to queries:
    - At this point, when an overall query is needed to be sent by a customer to all hubs/nodes holding information, the more proficient way is to send a query to every hub than moving a huge set of data to a central processor.
2. Hash rings should be used for even distribution of data:
    - To figure out a reliable approach to allocating a report to a processing hub/node is perhaps the most difficult issue with databases that are distributed.
3. For scaling read requests, replication should be used:
    - In real-time, replication is used by databases for making data’s backup copies of data.
4. Distribution of queries to nodes should be done by the database:
    - Separation of concerns of evaluation of query from the execution of the query is important for getting more increased performance from queries traversing numerous hubs/nodes.

Q17. What is A Data-Stream-Management System?

Answer: Data Stream Management System deals with stream data.It is based on Data Driven processing model i.e called push based model.DSMS is based on adaptive query plans.DSMS uses bounded main memory means limited main memory.

Q18. Variations of NoSQL architectural patterns.

Answer:

1. Key-Value Store Database:
2. Column Store Database:
3. Document Database:
4. Graph Databases:

Q19. NoSQL master-slave versus peer-to-peer.

Answer: Peer-to-peer models may be more resilient to failure than master-slave models. Some master-slave distribution models have single points of failure that might impact your system availability, so you might need to take special care when configuring these systems.

- In the master-slave model, one node is in charge (master). When there's no single node with a special role in taking charge, you have a peer-to-peer distribution model.

Q20. Write Issues in Stream Processing.

Answer:

1. Scalability
2. Integration
3. Fault-tolerance
4. Timeliness
5. Consistency
6. Heterogeneity and incompleteness
7. Load balancing
8. High throughput
9. Privacy
10. Accuracy

Q21. What are the sampling data techniques?

Answer: Sampling technique uses randomization to make sure that every element of the population gets an equal chance to be part of the selected sample. It's alternatively known as random sampling.

Simple Random Sampling:

- Every element has an equal chance of getting selected to be the part sample.

Cluster Sampling

- Our entire population is divided into clusters or sections and then the clusters are randomly selected.

Systematic Clustering

- Here the selection of elements is systematic and not random except the first element.

Multi-Stage Sampling

- It is the combination of one or more methods described above.

Non-Probability Sampling

- It does not rely on randomization. This technique is more reliant on the researcher's ability to select elements for a sample.

Q22. Discuss Bloom Filter with Analysis.

Answer:A Bloom filter is defined as a data structure designed to identify an element's presence in a set in a rapid and memory efficient manner.

- A specific data structure named as probabilistic data structure is implemented as a bloom filter. This data structure helps us to identify that an element is either present or absent in a set.

Appending objects to the Bloom filter

- Compute hash values for the object to append;
- implement these hash-values to set certain bits in the Bloom filter state (hash value is the position of the bit to set).

Verifying whether the Bloom filter contains an object -

- compute hash values for the object to append;

- Next we verify whether the bits indexed by these hash values are set in the Bloom filter state.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Each empty cell in that table specifies a bit and the number below it its index or position. To append an element to the Bloom filter, we simply hash it a few times and set the bits in the bit vector at the position or index of those hashes to 1.

Q23. What is the Count Distinct Problem ?

Answer: The count-distinct problem; (also known in applied mathematics as the cardinality estimation problem) is the problem of finding the number of distinct elena data stream with repeated elements,

- This is a well-known problem with numerous applications.

➢ **Problem:**

- Data stream consists of a universe of elements chosen from a set of size *N*
- Maintain a count of the number of distinct elements seen so far

Obvious approach :

- Maintain the set of elements seen so far
- That is, keep a hash table of all the distinct elements seen so far

Example :

- As a useful example of this problem, consider a Web site gathering statistics on how many unique users it has seen in each given month. The universal set is the set of logins for that site, and a stream element is generated each time someone logs in. This measure is appropriate for a site like Amazon,where the typical user logs in with their unique login name.

Q24. Write cost of exact counts.

Answer: To begin, suppose we want to be able to count exactly the number of 1's in the last k bits for any k ≤ N.

- Then we claim it is necessary to store all N bits of the window, as any representation that used fewer than N bits could not work.
- In proof, suppose we have a representation that uses fewer than N bits to represent the N bits in the window.
- Since there are 2^N sequences of N bits, but fewer than 2^N representations, there must be two different bit strings w and x that have the same representation.
- Since w 6= x, they must differ in at least one bit. Let the last k - 1 bits of w and x agree, but let them differ on the kth bit from the right end.

Example :

- If w = 0101 and x = 1010, then k = 1, since scanning from the right, they first disagree at position 1. If w = 1001 and x = 0101, then k = 3, because they first disagree at the third position from the right. Suppose the data representing the contents of the window is whatever sequence of bits represents both w and x. Ask the query "how many 1's are in the last k bits?" The query-answering algorithm will produce the same answer, whether the window contains w or x, because the algorithm can only see their representation. But the correct answers are surely different for these two bit-strings. Thus, we have proved that we must use at least N bits to answer queries about the last k bits for any k.

Q25. What are decaying windows ?

Answer: The decaying window algorithm not only tracks the most recurring elements in an incoming data stream, but also discounts any random spikes or spam requests that might have boosted an element's frequency.Sudden spikes or spam data is taken care.
New elements are given more weight by this mechanism, to achieve right trending output.

Q26. Describe a model for a recommendation system.

Answer: The recommendation system model uses the utility matrix and the concept of "long-tail" which explains the advantage of online vendors over conventional, brick-and-mortar vendors.

Applications of Recommendation Systems:

1. Movie Recommendations: Netflix offers its customers recommendations of movies they might like. These recommendations are based on ratings provided by users.
2. Product Recommendations: Perhaps the most important use of recommendation systems is at on-line retailers. We have noted how Amazon or similar on-line vendors strive to present each returning user with some suggestions of products that they might like to buy. These suggestions are not random, but are based on the purchasing decisions made by similar customers or on other techniques.

Q27. Discuss Content-Based Recommendations.

Answer:A Content-Based Recommender works by the data that we take from the user, either explicitly (rating) or implicitly (clicking on a link).

- By the data we create a user profile, which is then used to suggest to the user, as the user provides more input or takes more actions on the recommendation, the engine becomes more accurate.Factors considered are user profile,item profile,utility matrix.

Q28. What are decaying windows ?

Answer: The decaying window algorithm not only tracks the most recurring elements in an incoming data stream, but also discounts any random spikes or spam requests that might have boosted an element's frequency.

Q29. Write about Clustering of Social-Network Graphs.

Answer: Clustering of the graph is considered as a way to identify communities. Clustering of graphs involves following steps:

1. Applying Standard
2. Clustering Methods
3. Betweenness
4. The Girvan-Newman Algorithm
5. Using betweenness to find communities

Q30. Write down Variables in R.

Answer: A variable in R can store an atomic vector, a group of atomic vectors or a combination of many Robjects.

- The variable name starts with a letter or the dot not followed by a number.The variables can be assigned values using leftward, rightward and equal to operator.
- The values of the variables can be printed using print() or cat() function. The cat() function combines multiple items into a continuous print output.

Q31. Describe built-in functions in R.

Answer: Math Functions

- R provides the various mathematical functions to perform the mathematical calculation. These mathematical functions are very helpful to find absolute value, square value and much more calculations.

String Function

- R provides various string functions to perform tasks. These string functions allow us to extract substring from string, search pattern etc.Statistical Probability Functions

- R provides various statistical probability functions to perform statistical tasks. These statistical functions are very helpful to find normal density, normal quantile and many more calculations.

Q32. Data visualizations types and Applications.

Answer: Applications:

1. Healthcare Industries
    - A dashboard that visualizes a patient's history might aid a current or new doctor in comprehending a patient's health.
2. Business intelligence
    - Multiple datasets can be correlated using analytics/BI tools, which allow for searches using a common set of filters and/or parameters.
3. Military
    - It's a matter of life and death for the military; having clarity of actionable data is critical, and taking the appropriate action requires having clarity of data to pull out actionable insights.
4. Data science
    - Data scientists generally create visualizations for their personal use or to communicate information to a small group of people.
5. Marketing
    - In marketing analytics, data visualization is a boon. We may use visuals and reports to analyze various patterns and trends analysis, such as sales analysis, market research analysis, customer analysis, defect analysis, cost analysis, and forecasting.