**Subject Name**: **BIG DATA** **Class** : III CSE
**Subject Code** : **503208** **Semester** : VI

**Question Bank**

## Unit – I

The evolution of Data Management – Understanding the waves of managing data – Defining Big Data – Building a Successful Big Data Management Architecture – Examining Big Data Types: Structured data – Unstructured data– Looking at real-time and non-real-time requirements – Putting Big Data together –
Distributed Computing: A brief history – Understanding the basics

**Part – A (Two Marks)**

| Q.No | Question | BT Level* | Competence* |
|---|---|---|---|
| 1 | **What is Big Data?**<br>Big data refers to the large, diverse sets of information that grow at ever-increasing rates. It encompasses the volume of information, the velocity or speed at which it is created and<br>collected, and the variety or scope of the data points being covered (known as the "three v's" of big data). | BTL1 | Remember |
| 2 | **What are the characteristics of Big Data?**<br>● Volume: How much data<br>● Velocity: How fast that data is processed<br>● Variety: The various types of data | BTL1 | Remember |
| 3 | **Draw the cycle diagram of Big Data Management and Explain.**<br><br>**Figure 1-1:** The cycle of big data management | BTL3 | Apply |
| 4 | **Write the characteristics operational data sources.**<br>✓ They represent systems of record that keep track of the critical data<br>required for real-time, day-to-day operation of the business.<br>✓ They are continually updated based on transactions happening within<br>business units and from the web.<br>✓ For these sources to provide an accurate representation of the business,<br>they must blend structured and unstructured data. | BTL2 | Understand |

| | | | |
|---|---|---|---|
| | ✓ These systems also must be able to scale to support thousands of users<br>on a consistent basis. These might include transactional e- commerce systems, custo | | |
| 5 | **Define MapReduce.**<br>MapReduce was designed by Google as a way of efficiently executing a set of functions against a large amount of data in batch mode. The "map" component distributes the programming problem or tasks across a large number of systems and handles the placement of the tasks in a way that balances the load and manages recovery from failures. After the distributed computation<br>is completed, another function called "reduce" aggregates all the elements back together to provide a result. | BTL1 | Remember |
| 6 | **What is Hadoop?**<br>Hadoop is an Apache-managed software framework derived from MapReduce and Big Table. Hadoop allows applications based on MapReduce to run on large clusters of commodity hardware. The project is the foundation for the computing architecture supporting Yahoo!'s business. Hadoop is designed to parallelize<br>data processing across computing nodes to speed computations and hide latency. | BTL1 | Remember |
| 7 | **Differentiate Structured and Unstructured Data.**<br>The term structured data generally refers to data that has a defined length and format. Examples of structured data include numbers, dates, and groups of words and numbers called strings Unstructured data is data that does not follow a specified format.<br>Examples: Satellite Image, Scientific Data | BTL4 | Analyze |
| 8 | **What is Relational Database Management System.**<br>A relational database management system (RDBMS) is a collection of programs and capabilities that enable IT teams and others to create, update, administer and otherwise interact with<br>a relational database. RDBMSes store data in the form of tables, with most commercial relational database management systems<br>using Structured Query Language (SQL) to access the database. | BTL2 | Understand |

| | | | |
|---|---|---|---|
| 9 | **What are the things you need to consider regarding a system's capability to ingest data, process it, and analyze it in real time.**<br>✓ **Low latency:** Latency is the amount of time lag that enables a service to execute in an environment. Some applications require less latency, which means that they need to respond in real time. A real-time stream is going to require low latency. So you need to be thinking about compute power as well as network constraints.<br>✓ **Scalability:** Scalability is the capability to sustain a certain level of performance even under increasing loads.<br>✓ **Versatility:** The system must support both structured and unstructured data streams.<br>✓ **Native format:** Use the data in its native form. Transformation | BTL1 | Remember |

| | | | |
|---|---|---|---|
| | takes time and money. The capability to use the idea of processing complex interactions in the data that trigger events may be transformational. | | |
| 10 | **What are the components need to integrating data types into a big data environment?**<br>● Connectors<br>● Metadata | BTL1 | Remember |
| 11 | **Give some examples for Enterprise company which are used in real time ( Bigdata Service Provider).**<br>✔ Cloudera (Hadoop's owner company)<br>✔ Horton Works ( HDP Sand box)<br>✔ AWS (Amazon Web Services)<br>✔ MapR<br>✔ IBM<br>✔ Microsoft Corporation<br>✔ Data bricks<br>✔ Apache Spark | BTL2 | Understand |
| 12 | **What are the skills required for learning Bigdata?**<br>Linux (Environment – File handling commands)<br>SQL (Database –DDL,DML,DCL commands , MySQL)<br>Programming Language ( Core Java, Core Python, Scala) | BTL2 | Understand |

**Part – B (16 Marks)**

| Q.No | Question | BT Level* | Competence* |
|---|---|---|---|
| 1 | Explain in detail about Big Data Management Architecture. | BTL1 | Remember |
| 2 | Explain in detail evolution of Data Management. | BTL1 | Remember |
| 3 | What is Distributed Computing? Explain | BTL2 | Understand |
| 4 | How will you Examining Big Data Types. | BTL4 | Analyzing |
| 5 | Explain about the different categories of sources of Structured Data. | BTL2 | Understand |
| 7 | How to integrate data types into big data environment? | BTL3 | Apply |
| 8 | How to explore the sources of Unstructured Data. | BTL4 | Analyzing |
| 9 | Why we need Distributed computing for Bigdata? | BTL3 | Apply |

## Unit – II

Exploring the Big Data stack – Redundant Physical Infrastructure – Security infrastructure – Operational databases – Organizing databases and tools – Analytical data warehouses – Big Data analytics – Big Data applications – Understanding the basics of virtualization – Implementing virtualization to work with Big Data – Understanding Cloud deployment and delivery models – The cloud as an imperative for Big Data

**Part - A (Two Marks)**

| Q.N | Questio | BT | Competence |
|---|---|---|---|

| o | n | Level* | * |
|---|---|---|---|
| 1 | **Write the principals needed for big data implementation.**<br>● Performance<br>● Availability<br>● Scalability<br>● Flexibility<br>● Cost | BTL2 | Understand |
| 2 | **What are the Security and privacy requirements for big data?**<br>● Data Access<br>● Application Access<br>● Data Encryption<br>● Thread Detection | BTL1 | Remember |
| 3 | **Define ACID.**<br>**Atomicity**<br>**Consistency**<br>**Isolation**<br>**Durability** | BTL1 | Remember |
| 4 | **Write the purpose of Organizing Data Services and Tools.**<br>*Organizing data services and tools* capture, validate, and assemble various big data elements into contextually relevant collections. Because big data is massive, techniques have evolved to process the data efficiently and seamlessly | BTL4 | Analyze |
| 5 | **What are the technologies included in Organizing Data Services and Tools?**<br>**1. A distributed file system**<br>**2. Serialization services**<br>**3. Coordination services**<br>**4. Extract, transform, and load (ETL) tools**<br>**5. Workflow services** | BTL1 | Remember |
| 6 | **Define Visualization in big data.**<br>These tools are the next step in the evolution of reporting. The output tends to be highly interactive and dynamic in nature. Another important distinction between reports and visualized output is animation. Business users can watch the changes in the data utilizing a variety of different visualization techniques, including mind maps, heat maps, infographics, and connection diagrams | BTL4 | Analyze |
| 7 | **How virtualization concepts working in big data Analytic.**<br>Virtualization — the process of using computer resources to imitate other resources — is valued for its capability to increase IT resource utilization, efficiency, and scalability. One primary application of virtualization is server consolidation which helps organizations increase the utilization of physical servers and potentially save on infrastructure costs | BTL1 | Remember |
| 8 | **What are the benefits of Virtualization?**<br>• ✓ Virtualization of physical resources (such as servers, storage, and networks) enables substantial improvement in the utilization of these resources.<br>• ✓ Virtualization enables improved control over the usage | BTL2 | Understand |

| | | | |
|---|---|---|---|
| | and performance of your IT resources.<br>• ✓ Virtualization can provide a level of automation and standardization to optimize your computing environment.<br>• ✓ Virtualization provides a foundation for cloud computing. | | |
| 9 | **What are the three characteristics that support the scalability and operating efficiency required for big data environments?**<br>**Partitioning**<br>**Isolation**<br>**Encapsulation** | BTL1 | Remember |
| 10 | **What is Processor and memory virtualization ?**<br>Processor virtualization helps to optimize the processor and maximize performance. Memory virtualization decouples memory from the servers. In big data analysis, you may have repeated queries of large data sets and the creation of advanced analytic algorithms, all designed to look for patterns and trends that are not yet understood.<br>These advanced analytics can require lots of processing power (CPU) and memory (RAM). For some of these computations, it can take a long time without sufficient CPU and memory<br>resources. Processor and memory virtualization can help speed the processing and get your analysis results sooner | BTL2 | Understand |
| 11 | **What is Streaming data?**<br>Data in motion is called as Streaming data.<br>Streaming data is data that is generated continuously by thousands of data sources, which typically send in the data records simultaneously<br>and in small sizes(order of Kilobytes). | BTL2 | Understand |
| 12 | **Define ETL**<br>ETL stands for Extract, Transform and Load.<br>It is a integration process that combines data from multiple data sources into a single a data store that is loaded into a data warehouse or other target system. | BTL2 | Understand |
| 13 | **Define Abstraction.**<br>IT resources and services to be virtualized, they are separated from the underlying physical delivery environment. The technical term for this act of separation is called abstraction. | BTL2 | Understand |
| 14 | **What are the types of hypervisors?**<br>✓ Type 1 hypervisors run directly on the hardware platform. They achieve higher efficiency because they're running directly on the platform.<br>✓ Type 2 hypervisors run on the host operating system. They are often used when a need exists to support a broad range of I/O devices. | BTL1 | Remember |

| | | | |
|---|---|---|---|
| 15 | **Define Cloud Computing.**<br>Cloud computing is a method of providing a set of shared computing resources that include applications, computing, storage, networking, development, and deployment platforms, as well as business processes. Cloud computing turns traditional | BTL2 | Understand |

| | | | |
|---|---|---|---|
| | siloed computing assets into shared pools of resources based on an underlying Internet foundation. | | |
| 16 | **What are the types of Cloud deployment models?**<br>**The public cloud**<br>The public cloud is a set of hardware, networking, storage, services, applications, and interfaces owned and operated by a third party for use by other companies and individuals. These commercial providers create a highly scalable data center that hides the details of the underlying infrastructure from the consumer.<br>**The private cloud**<br>A private cloud is a set of hardware, networking, storage, services, application, and interfaces owned and operated by an organization for the use o its employees, partners, and customers. A private cloud can be create and managed by a third party for the exclusive use of one enterprise. The private cloud is a highly controlled environment not open for public consumption. Thus, the private cloud sits behind a firewall. | BTL2 | Understand |
| **Part - B (16 Marks)** | | | |
| **1** | What is Virtualization? Explain various types of Virtualization. | BTL1 | Remember |
| 2 | How the layers of big data technology stack working. | BTL4 | Analyze |
| 3 | Explain in detail about Cloud computing in Big data. | BTL2 | Understand |
| 4 | Analyse why Cloud as an Imperative for Big Data. | BTL4 | Analyze |
| 5 | Elaborate various Cloud delivery models. | BTL1 | Remember |
| 6 | Explain about the following :<br>i. What are the different classes of tools in Big data Analytics<br>ii. Explain Bigdata applications. | BTL4 | Analyze |
| 7 | How the Big data analytics is used in real time either medical or banking usage? | BTL 4 | Analyze |
| **UNIT-3**<br>MapReduce Fundamentals: Origin – Understanding map function – adding reduce function – putting map and reduce together – Optimizing MapReduce tasks – Exploring the world of Hadoop: Explaining Hadoop – Understanding Hadoop Distributed File System – Hadoop MapReduce – Hadoop Foundation and Ecosystem: Building Big Data foundation with Hadoop Ecosystem | | | |
| **Part – A (Two Marks)** | | | |
| **Q.No** | **Question** | **BT Level*** | **Competence*** |
| 1 | **Why distribution of work must be performed in parallel.**<br>✓ The processing must be able to expand and contract automatically.<br>✓ The processing must be able to proceed regardless of failures in the network or the individual systems. | BTL4 | Analyze |

| | | | |
|---|---|---|---|
| | ✓ Developers leveraging this approach must be able to create services<br>that are easy to leverage by other developers. Therefore, this approach<br>must be independent of where the data and computations have executed. | | |
| 2 | **What is MapReduce?**<br>MapReduce was designed as a generic programming model. Some of the initial implementations provided all the key requirements of parallel execution, fault tolerance, load balancing, and data manipulation. The engineers in charge of the project named the initiative MapReduce because it combines two<br>capabilities from existing functional computer languages: map and reduce | BTL1 | Remember |
| 3 | **Why MapReduce Implemented?**<br>Google engineers designed MapReduce to solve a specific practical problem. Therefore, it was designed as a programming model combined with the implementation of that model — in essence, a reference implementation. The reference implementation was used to demonstrate the practicality and effectiveness of the concept and to help ensure that this model<br>would be widely adopted by the computer industry. | BTL4 | Analyze |
| 4 | **what exactly can you expect from the map function?**<br>It applies a function to each element (defined as a key-value pair) of a list and produces a new list. Suppose that you wanted to create a program that counts the number of characters in a<br>series or list of words. | BTL2 | Understand |
| 5 | **Write the Purpose of Reduce function.**<br>The reduce function takes the output of a map function and "reduces" the list in whatever fashion the programmer desires. The first step that the reduce function requires is to place a value in something called an accumulator, which holds an initial value. After storing a starting value in the accumulator, the reduce function then processes each element of the list and performs the operation you need across the list. At the end of the list, the<br>reduce function returns a value based on what operation you wanted to perform on the output list | BTL2 | Understand |
| 6 | **What are the step by step procedure for MapReduce.**<br>1. Start with a large number or data or records.<br>2. Iterate over the data.<br>3. Use the map function to extract something of interest and create an output list.<br>4. Organize the output list to optimize for further processing.<br>5. Use the reduce function to compute a set of results.<br>6. Produce the final output. | BTL1 | Remember |
| 7 | **What is Scheduling in MapReduce.**<br>MapReduce jobs get broken down into individual tasks for the map and the reduce portions of the application. Because the | BTL2 | Understand |

| | | | |
|---|---|---|---|
| | mapping must be concluded before reducing can take place, those tasks are prioritized according to the number of nodes in the cluster. If you have more tasks than nodes, the execution framework will manage the map tasks until all are complete.<br>Then the reduce tasks will run with the same behaviors. | | |
| 8 | **What is Synchronization in MapReduce**.<br>When multiple processes execute concurrently in a cluster, you need a way to keep things running smoothly. Synchronization mechanisms do this automatically. Because the execution framework knows that the program is mapping and reducing, it keeps track of what has run and when. When all the mapping is complete, the reducing begins. Intermediate data is copied over the network as it is produced using a mechanism called "shuffle and sort." This gathers and prepares all the mapped data for reduction. | BTL2 | Understand |
| 9 | **For implementing MapReduce function what are the four things need to consider.**<br>1. Keep it warm<br>2. The bigger the better<br>3. The long view<br>4. Keep it secure | BTL4 | Analyze |
| 10 | Why Hadoop used in Big data.<br>Hadoop was developed because it represented the most pragmatic way to allow companies to manage huge volumes of data easily. Hadoop allowed big problems to be broken down into smaller elements so that analysis could be done quickly and cost-effectively. | BTL4 | Analyze |
| 11 | **What are the two components in Hadoop.**<br>✓ **Hadoop Distributed File System:** A reliable, high-bandwidth, low-cost, data storage cluster that facilitates the management of related files across machines.<br>✓ **MapReduce engine:** A high-performance parallel/distributed dataprocessing<br>mplementation of the MapReduce algorithm. | BTL2 | Understand |
| 12 | **What is HDFS?**<br>The Hadoop Distributed File System is a versatile, resilient, clustered approach to managing files in a big data environment. HDFS is not the final destination for files. Rather, it is a data service that offers a unique set of capabilities needed when data volumes and velocity are high. Because the data is written once and then read many times thereafter, rather than the constant read-writes of other file systems, | BTL1 | Remember |
| 13 | **Define Name Nodes.**<br>HDFS works by breaking large files into smaller pieces called blocks. The blocks are stored on data nodes, and it is the responsibility of the NameNode to know w hat blocks on which data nodes make up the complete file. TheNameNode also acts | BTL2 | Understand |

| | | | |
|---|---|---|---|
| | as a "traffic cop," managing all access to the files, including reads, writes, creates, deletes, and replication of data blocks on the data nodes | | |
| 14 | **Define Metadata.**<br>Metadata is defined as "data about data." Software designers have been using metadata for decades under several names like data dictionary, metadata directory, and more recently, tags. | BTL2 | Understand |
| 15 | **What exactly does a block server do?**<br>✓ Stores (and retrieves) the data blocks in the local file system of the server. HDFS is available on many different operating systems and behaves the same whether on Windows, Mac OS, or Linux.<br>✓ Stores the metadata of a block in the local file system based on the metadata template in the NameNode.<br>✓ Performs periodic validations of file checksums.<br>✓ Sends regular reports to the NameNode about what blocks are available for file operations.<br>✓ Provides metadata and data to clients on demand. HDFS supports direct access to the data nodes from client application programs.<br>✓ Forwards data to other data nodes based on a "pipelining" model. | BTL3 | Apply |
| 16 | **What are the steps involved in Hadoop MapReduce.**<br>● Getting the data ready<br>● Let the mapping begin<br>● Reduce and combine | BTL2 | Understand |
| | **Part - B (16 Marks)** | | |
| 1 | How MapRefuce function working in Big data. Explain with suitable example | BTL4 | Analyze |
| 2 | Explain in detail about foundational behaviors of MapReduce. | BTL2 | Understand |
| 3 | How will you optimize MapReduce task? Explain | BTL4 | Analyze |
| 4 | Explain in detail about Hadoop. | BTL1 | Remember |
| 5 | Elaborate HDFS with suitable diagram. | BTL2 | Understand |
| 6 | Explain about Hadoop MapReduce | BTL1 | Remember |
| 7 | Give brief explanation on Big Data foundation with Hadoop Ecosystem. | BTL2 | Understand |
| 8 | Explain the Origins of MapReduce. | BTL1 | Remember |
| **UNIT-4** | | | |

Big Data Analytics: Modifying business intelligence products to handle Big Data – Studying Big Data Analytics Examples – Big Data Analytics Solutions
Understanding Text Analytics and Big Data: Exploring unstructured data – Text analytics – Analysis and extraction techniques – Putting results together with structured data – putting Big Data to use – Text analytic tools for Big Data

**Part – A (Two Marks)**

| Q.No | Question | BT Level* | Competence* |
|---|---|---|---|
| 1 | **What is Slicing and dicing?**<br>Slicing and dicing refers to breaking down your data into smaller sets of data that are easier to explore. For example, you might have a scientific data set of water column data from many different locations that contains numerous variables captured from multiple sensors. | BTL1 | Remember |
| 2 | **What is Basic monitoring in big data analytic?**<br>You might also want to monitor large volumes of data in real time. For example, you might want to monitor the water column attributes in the preceding example every second for an extended period of time from hundreds of locations and at varying heights in the water column. This would produce a huge data set. | BTL2 | Understand |
| 3 | **Define Anomaly identification.**<br>You might want to identify anomalies, such as an event where the actual observation differs from what you expected, in your data because that may clue you in that something is going wrong with your organization, manufacturing process, and so on. For example, you might want to analyze the records for your manufacturing operation to determine whether one kind of machine | BTL2 | Understand |
| 4 | **What are all advanced analytics for big data?**<br>1. Predictive modeling<br>2. Text analytics<br>3. Other statistical and data-mining algorithms | BTL1 | Remember |
| 5 | **Where the Predictive modelling used in big data.**<br>Predictive modeling is one of the most popular big data advanced analytics use cases. A predictive model is a statistical or data-mining solution consisting of algorithms and techniques that can be used on both structured and unstructured data (together or individually) to determine future outcomes. | BTL4 | Analyze |
| 6 | **What is Text analytics.**<br>Unstructured data is such a big part of big data, so text analytics — the process of analyzing unstructured text, extracting relevant information, and transforming it into structured information that can then be leveraged in various ways — has become an important component of the big data ecosystem**.** | BTL2 | Understand |
| 7 | **Write some statistical and data-mining algorithms.**<br>1. Advanced Forecasting,<br>2. Optimization, | BTL1 | Remember |

| | | | |
|---|---|---|---|
| | **3.** Cluster Analysis For Segmentation | | |
| 8 | **What is data mining?**<br>Data mining involves exploring and analyzing large amounts of data to find patterns in that data. The techniques came out of the fields of statistics and artificial intelligence (AI), with a bit of database management thrown into the mix. Generally, the goal of the data mining is either classification or prediction. In classification, the idea is to sort data into groups. | BTL2 | Understand |
| 9 | **What are all the algorithm used in Typical algorithms used in data mining.**<br>1. Classification trees<br>2. Logistic regression<br>3. Neural networks<br>4. Clustering techniques like K-nearest neighbors | BTL1 | Remember |
| 10 | **Define Classification trees.**<br>A popular datamining technique that is used to classify a dependent categorical variable based on measurements of one or more predictor variables. The result is a tree with nodes and links between the nodes that can be read to form if-then rules. | BTL2 | Understand |
| 11 | **How Operationalized analytics.**<br>When you operationalize analytics, you make them part of a business process. For example, statisticians at an insurance company might build a model that predicts the likelihood of a claim being fraudulent. The model, along with some decision rules, could be included in the company's claims-processing system to flag claims with a high probability of fraud. These claims would be sent to an investigation unit for further review. | BTL3 | Apply |
| 12 | **What are all the potential characteristics of your data.**<br>1. It can come from untrusted sources<br>2. It can be dirty<br>3. The signal-to-noise ratio can be low<br>4. It can be real-time | BTL2 | Understand |
| 13 | **The infrastructure needed to support big data to achieve.**<br>1. Integrate technologies<br>2. Store large amounts of disparate data<br>3. Process data in motion<br>4. Warehouse data | BTL4 | Analyze |
| 14 | **The structure of data that might be associated with what?**<br>1. Documents<br>2. E-mails<br>3. Log files<br>4. Tweets<br>5. Facebook posts | BTL2 | Understand |
| 15 | **How Retrieval Insight operation working in Structured and Unstructured data.**<br>***Retrieval*** ***Insight***<br>Structured Query: Returns data from structured data Data mining: Insight | BTL4 | Analyze |

| | | | |
|---|---|---|---|
| | Unstructured Search: Returns documents Text analytics: Insight from text | | |
| 16 | **Write the purpose of Lexical/morphological analysis.** Examines the characteristics of an individual word — including prefixes, suffixes, roots, and parts of speech (noun, verb, adjective, and so on) — information that will contribute to understanding what the word means in the context of the text provided. Lexical analysis depends on a dictionary, thesaurus, or any list of words that provides information about those words. | BTL1 | Remember |
| 17 | **What is NLP?** Natural Language Processing (NLP) techniques to extract information from unstructured data. NLP is a broad and complex field that has developed over the last 20 years. A primary goal of NLP is to derive meaning from text. Natural Language Processing generally makes use of linguistic concepts such as grammatical structures and parts of speech. Often, the idea behind this type of analytics is to determine who did what to whom, when, where, how, and why | BTL1 | Remember |
| 18 | **What are the information needed to text documents extract?**<br>1. Terms<br>2. Facts<br>3. Events<br>4. Concepts<br>5. Sentiments | BTL2 | Understand |
| | **Part - B (16 Marks)** | | |
| 1 | Explain in detail about types of analytics of Big data. | BTL1 | Remember |
| 2 | Elaborate the concepts of Data mining. | BTL2 | Understand |
| 3 | How will you Modifying Business Intelligence Products to Handle Big Data. | BTL4 | Analyze |
| 4 | Explain Big Data Analytics with example. | BTL3 | Apply |
| 5 | Elaborate vendors support big data solutions. | BTL1 | Remember |
| 6 | Explain various Analysis and Extraction Techniques. | BTL3 | Apply |
| 7 | Analyze the information needed to extract text documents | BTL4 | Analyze |
| 8 | Explain the concepts of<br>1. Putting results together with structured data<br>2. putting Big Data to use<br>3. Text analytic tools for Big Data | BTL2 | Understand |
| **UNIT-5**<br>Integrating Data Sources: Identifying needed data – Understanding the fundamentals of Big Data Integration – Defining traditional ETL – Understanding ELT – Prioritizing Big Data quality – Using Hadoop as ETL – Data Streams: Using streaming data – Using Complex Event Processing – Operationalizing Big Data: Making Big Data a part of Operational Process – Understanding Big Data workflows | | | |

| Part – A (Two Marks) | | | |
|---|---|---|---|
| **Q.No** | **Question** | **BT Level*** | **Competence*** |
| 1 | **What are the three stages of big data Analysis?**<br>✓ Exploratory stage ✓ Codifying stage ✓ Integration and incorporation stage | BLT2 | Understand |
| 2 | **Define Exploratory stage.**<br>It is the early stages of your analysis, you will want to search for patterns in the data. It is only by examining very large volumes (terabytes and petabytes) of data that new and unexpected relationships and correlations among elements may become apparent. | BLT2 | Understand |
| 3 | **Write the purpose of Codifying stage.**<br>It helps to codify it and make it a part of your business process. You need to make the connection between your big data analytics and your inventory and product systems. | BLT1 | Remember |
| 4 | **How the Integration and incorporation stage is working in big data Analysis.**<br>Big data is having a major impact on many aspects of data management, including data integration. Traditionally, data integration has focused on the movement of data through middleware, including specifications on message passing and requirements for application programming interfaces (APIs). These concepts of data integration are more appropriate for managing data at rest rather than data in motion | BLT4 | Analysis |
| 5 | **What are the three basic principles needed for delivering Information to the business in a trusted, controlled, consistent, and flexible way?**<br>1. You must create a common understanding of data definitions.<br>2. You must develop of a set of data services to qualify the data and make it consistent and ultimately trustworthy.<br>3. You need a streamlined way to integrate your big data sources and systems of record. | BLT1 | Remember |
| 6 | **What are the Three important functions in ETL?**<br>1. ✓ Extract: Read data from the source database.<br>2. ✓ Transform: Convert the format of the extracted data so that it conforms to the requirements of the target database. Transformation is done by using rules or merging data with other data.<br>3. ✓ Load: Write data to the target database. | BLT2 | Understand |
| 7 | **What is Data transformation?**<br>Data transformation is the process of changing the format of data so that it can be used by different applications. This may mean a change from the format the data is stored in into the format needed by the application that will use the data. This process also includes mapping instructions so that applications are told how | BLT1 | Remember |

| | | | |
|---|---|---|---|
| | to get the data they need to process The process of data transformation is made far more complex because of the staggering growth in the amount of unstructured data. | | |
| 8 | **Write the two-phase approach to data quality?**<br>1. Phase 1: Look for patterns in big data without concern for data quality.<br>2. Phase 2: After you locate your patterns and establish results that are important to the business, apply the same data quality standards that you apply to your traditional data sources. You want to avoid collecting and managing big data that is not important to the business and will potentially corrupt other data elements in Hadoop or other big data platforms | BLT2 | Understand |
| 9 | **What are the key issues of big data management?**<br>● ✓ Keep data quality in perspective<br>● ✓ Consider real-time data requirements<br>● ✓ Don't create new silos of information | BLT2 | Understand |
| 10 | **How the flow of data managed in Big data?**<br>Streaming data and complex event processing helps to manage flow of data in Big data. | BLT4 | Analysis |
| 11 | **Write key principles needed for streams is most appropriate.**<br>1. ✓ When it is necessary to determine a retail buying opportunity at the point of engagement, either via social media or via permission-based messaging<br>2. ✓ Collecting information about the movement around a secure site<br>3. ✓ To be able to react to an event that needs an immediate response, such as a service outage or a change in a patient's medical condition<br>4. ✓ Real-time calculation of costs that are dependent on variables such as usage and available resources | BLT1 | Remember |
| 12 | **Why metadata need in streams?**<br>Most data management professionals are familiar with the need to manage metadata in structured database management environments. These data sources are strongly typed and designed to operate with metadata. So metadata playing important role in streams. | BLT4 | Analysis |
| 13 | **Write the examples for streaming data.**<br>● IBM InfoSphere Streams<br>● Twitter's Storm<br>● Apache S4 | BLT1 | Remember |
| 14 | **How will you differ CEP from Streams?**<br>While stream computing is typically applied to analyzing vast amounts of data in real time, CEP is much more focused on solving a specific use case based on events and actions. However, a streaming data technique is often used as an integral part of a CEP application. | BLT4 | Analysis |

| | | | |
|---|---|---|---|
| 15 | **How will you identify the "right" sources of data?**<br>1. ✓ Understand the problem you are trying to solve<br>2. ✓ Identify the processes involved<br>3. ✓ Identify the information required to solve the problem<br>4. ✓ Gather the data, process it, and analyse the results | BLT4 | Analysis |
| 16 | **What are the best practice for understanding workflows and the effect of big data?**<br>● ✓ Identify the big data sources you need to use.<br>● ✓ Map the big data types to your workflow data types.<br>● ✓ Ensure that you have the processing speed and storage access to support your workflow<br>● ✓ Select the data store best suited to the data types.<br>● ✓ Modify the existing workflow to accommodate big data or create new big data workflow | BLT1 | Remember |
| | **Part - B (16 Marks)** | | |
| 1 | Explain in detail about three stages of big data Analysis | BTL2 | Understand |
| 2 | Explain How Big Data Integration working. | BTL4 | Analysis |
| 3 | Elaborate Tradition ETL. | BTL2 | Understand |
| 4 | How will you Prioritizing Big Data Quality. | BTL4 | Analysis |
| 5 | Explain in detail how streaming data and complex event processing impact big data. | BTL4 | Analysis |
| 6 | Explain in detail about Streaming Data and Complex Event Processing. | BTL2 | Understand |
| 7 | Elaborate the concepts of Operationalizing Big Data. | BTL2 | Understand |
| 8 | Explain Big Data Workflows. | BTL1 | Remember |