503208	BIG DATA			L	T	P	C
				3	0	0	3
Programme:	B.E - Computer Science and Engineering	**Sem:**	06	**Category:**			ES
Prerequisite:	Basics of Data Base Management System						
Aim:	To study about the concept of Big data management and learn the different Big data analytics approaches and big data implementation.						

Course Objectives:	
1	To understand the fundamentals of Big Data management
2	To gain knowledge about the technology foundations of Big Data
3	To understand the role of MapReduce and Hadoop in managing Big Data
4	To learn the different Big Data Analytics and text analytics approaches for Big Data
5	Understand the big data implementation and ETL

Course Outcomes: The students will be able to	
CO1:	Know the basics of big data management concepts
CO2:	Know the technology foundation of Big data and its applications
CO3:	Understand the Map reduce fundamentals and Hadoop
CO4:	Understand the big data analytics approaches and building text analytics tools
CO5:	Gain an insight into the various issues in implementation of Big Data

COURSE OUTCOME & PROGRAM OUTCOME MAPPING

CO's NO	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10	PO 11	PO 12	PSO 1	PSO 2
CO.1	3	2			2							2		
CO.2	3		2	2								2		
CO.3	3	2	2	1	1							2		
CO.4	3		2		2							3		
CO.5	3			2	2							3		

UNIT – I	INTRODUCTION	Hrs
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		9
UNIT - II	**TECHNOLOGY FOUNDATION OF BIG DATA**	**Hrs**
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		9
UNIT - III	**BIG DATA MANAGEMENT**	**Hrs**
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 – HadoopMapReduce – Hadoop Foundation and Ecosystem: Building Big Data foundation with Hadoop Ecosystem		9
UNIT – IV	**BIG DATA ANALYTICS**	**Hrs**
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		9
UNIT – V	**BIG DATA IMPLEMENTATION**	**Hrs**
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		9

TEXT BOOK & REFERENCE BOOK

Sl. No	Description	Legend
Text Book(s):		
1	JudithHurwitz, Alan Nugent, Dr. Fern Halper, Marcia Kaufman, “Big Data for Dummies”, John Wiley and Sons Publication, 2013	T1
Reference Book(s):		
1	Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman, “Mining of Massive Datasets”, Cambridge University Press, Second Edition, 2014	R1
2	Michael Frampton, “Big Data Made Easy -A Working Guide to the Complete Hadoop Toolset”, Apress publications, 2015	R2

LESSON PLAN

S.No.	Unit	Topic to be covered	Hours Needed	Mode of Teaching	Text. / Ref
INTRODUCTION					
1	**I**	The evolution of Data Management	1	PPT/BB	T2
2		Understanding the waves of managing data	1	PPT/BB	T2
3		Defining Big Data	1	PPT/BB	T2
4		Building a Successful Big Data Management Architecture	1	PPT/BB	T2
5		Examining Big Data Types: Structured data	1	PPT/BB	T2
6		Time and non-real-time requirements – Putting Big Data together	1	PPT/BB	WS
7		Unstructured data– Looking at real	1	PPT/BB	WS
8		Distributed Computing: A brief history	1	PPT/BB	T1
9		Understanding the basics	1	PPT/BB	T1
10		Learn Hadoop YARN Architechture*	1	PPT/BB	WS
TECHNOLOGY FOUNDATION OF BIG DATA					
11	**II**	Exploring the Big Data stack	1	PPT/BB	T1
12		Redundant Physical Infrastructure	1	PPT/BB	T1
13		Security infrastructure	1	PPT/BB	T1
14		Operational databases – Organizing databases and tools	1	PPT/BB	T1
15		Analytical data warehouses – Big Data analytics	1	PPT/BB	T1
16		Big Data applications – Understanding the basics of virtualization	1	PPT/BB	T1
17		Implementing virtualization to work with Big Data	1	PPT/BB	T1
18		Understanding Cloud deployment and delivery models	1	PPT/BB	T1
19		The cloud as an imperative for Big Data	1	PPT/BB	T1
20		Exploring Hive *	1	PPT/BB	WS
BIG DATA MANAGEMENT					
21	**III**	MapReduce Fundamentals: Origin	1	PPT/BB	T1
22		Understanding map function – adding reduce function	1	PPT/BB	T1
23		putting map and reduce together – Optimizing MapReduce tasks	1	PPT/BB	T1
24		Exploring the world of Hadoop: Explaining Hadoop	1	PPT/BB	Web

S.No.	Unit	Topic to be covered	Hours Needed	Mode of Teaching	Text. / Ref
25		Understanding Hadoop Distributed File System	1	PPT/BB	T1
26		HadoopMapReduce – Hadoop Foundation	2	PPT/BB	T1
27		Ecosystem: Building Big Data foundation with Hadoop	1	PPT/BB	T1
28		Ecosystem	1	PPT/BB	T1
29		Exploring Oozie*	1	PPT/BB	WS
BIG DATA ANALYTICS					
30	**IV**	Big Data Analytics: Modifying business intelligence products to handle Big Data	1	PPT/BB	T1
31		Studying Big Data Analytics Examples	1	PPT/BB	T1
32		Big Data Analytics Solutions	1	PPT/BB	T1
33		Understanding Text Analytics and Big Data: Exploring unstructured data	1	PPT/BB	T1
34		Text analytics – Analysis and extraction techniques	1	PPT/BB	T1
35		Putting results together with structured data – putting Big Data to use	2	PPT/BB	T1
36		Text analytic tools for Big Data	2	PPT/BB	T1
38		Learn NoSQL Data Managemen*	1	PPT/BB	WS
BIG DATA IMPLEMENTATION					
39	**V**	Integrating Data Sources: Identifying needed data	1	PPT/BB	T2
40		Understanding the fundamentals of Big Data Integration	1	PPT/BB	T2
41		Defining traditional ETL Understanding ELT	1	PPT/BB	T2
42		Prioritizing Big Data quality	1	PPT/BB	T2
43		Using Hadoop as ETL	1	PPT/BB	T2
44		Using Complex Event Processing	1	PPT/BB	T2
45		Operationalizing Big Data:	1	PPT/BB	T2
46		Making Big Data a part of Operational Process	1	PPT/BB	T2
47		Understanding Big Data workflows	1	PPT/BB	WS
48		Integrating R and Hadoop and Understanding Hive in Detail *	1	PPT/BB	WS
Total Hours Needed: 45(L) + 5* = 50 Hours					

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

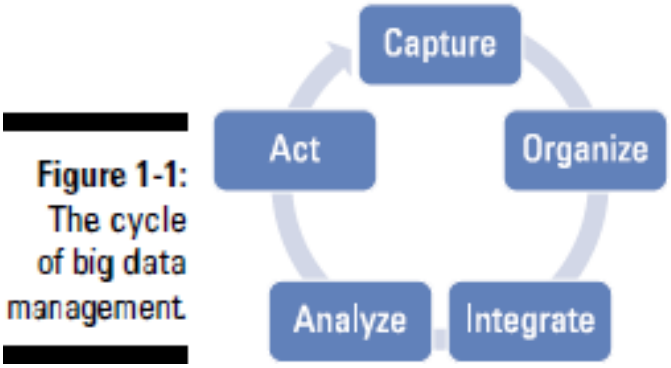
Question Bank

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?** 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 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?** ● Volume: How much data ● Velocity: How fast that data is processed ● Variety: The various types of data	BTL1	Remember

3	**Draw the cycle diagram of Big Data Management and Explain.**  **Figure 1-1:** The cycle of big data management	BTL3	Apply
4	**Write the characteristics operational data sources.** ✓ They represent systems of record that keep track of the critical data required for real-time, day-to-day operation of the business. ✓ They are continually updated based on transactions happening within business units and from the web. ✓ For these sources to provide an accurate representation of the business, they must blend structured and unstructured data.	BTL2	Understand

	✓ These systems also must be able to scale to support thousands of users on a consistent basis. These might include transactional e- commerce systems, custo		
5	**Define MapReduce.** 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 is completed, another function called "reduce" aggregates all the elements back together to provide a result.	BTL1	Remember
6	**What is Hadoop?** 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 data processing across computing nodes to speed computations and hide latency.	BTL1	Remember
7	**Differentiate Structured and Unstructured Data.** 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. Examples: Satellite Image, Scientific Data	BTL4	Analyze
8	**What is Relational Database Management System.** 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 a relational database. RDBMSes store data in the form of tables, with most commercial relational database management systems 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.** ✓ **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. ✓ **Scalability:** Scalability is the capability to sustain a certain level of performance even under increasing loads. ✓ **Versatility:** The system must support both structured and unstructured data streams. ✓ **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?** ● Connectors ● Metadata	BTL1	Remember
11	**Give some examples for Enterprise company which are used in real time (Bigdata Service Provider).** ✔ Cloudera (Hadoop's owner company) ✔ Horton Works (HDP Sand box) ✔ AWS (Amazon Web Services) ✔ MapR ✔ IBM ✔ Microsoft Corporation ✔ Data bricks ✔ Apache Spark	BTL2	Understand
12	**What are the skills required for learning Bigdata?** Linux (Environment – File handling commands) SQL (Database –DDL,DML,DCL commands , MySQL) 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.** ● Performance ● Availability ● Scalability ● Flexibility ● Cost	BTL2	Understand
2	**What are the Security and privacy requirements for big data?** ● Data Access ● Application Access ● Data Encryption ● Thread Detection	BTL1	Remember
3	**Define ACID.** **Atomicity** **Consistency** **Isolation** **Durability**	BTL1	Remember
4	**Write the purpose of Organizing Data Services and Tools.** *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?** **1. A distributed file system** **2. Serialization services** **3. Coordination services** **4. Extract, transform, and load (ETL) tools** **5. Workflow services**	BTL1	Remember
6	**Define Visualization in big data.** 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.** 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?** • ✓ Virtualization of physical resources (such as servers, storage, and networks) enables substantial improvement in the utilization of these resources. • ✓ Virtualization enables improved control over the usage	BTL2	Understand

	and performance of your IT resources. • ✓ Virtualization can provide a level of automation and standardization to optimize your computing environment. • ✓ 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?** **Partitioning** **Isolation** **Encapsulation**	BTL1	Remember
10	**What is Processor and memory virtualization ?** 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. 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 resources. Processor and memory virtualization can help speed the processing and get your analysis results sooner	BTL2	Understand
11	**What is Streaming data?** Data in motion is called as Streaming data. Streaming data is data that is generated continuously by thousands of data sources, which typically send in the data records simultaneously and in small sizes(order of Kilobytes).	BTL2	Understand
12	**Define ETL** ETL stands for Extract, Transform and Load. 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.** 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?** ✓ Type 1 hypervisors run directly on the hardware platform. They achieve higher efficiency because they're running directly on the platform. ✓ 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.** 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?** **The public cloud** 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. **The private cloud** 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 : i. What are the different classes of tools in Big data Analytics 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 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.** ✓ The processing must be able to expand and contract automatically. ✓ 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 that are easy to leverage by other developers. Therefore, this approach must be independent of where the data and computations have executed.		
2	**What is MapReduce?** 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 capabilities from existing functional computer languages: map and reduce	BTL1	Remember
3	**Why MapReduce Implemented?** 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 would be widely adopted by the computer industry.	BTL4	Analyze
4	**what exactly can you expect from the map function?** 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 series or list of words.	BTL2	Understand
5	**Write the Purpose of Reduce function.** 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 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.** 1. Start with a large number or data or records. 2. Iterate over the data. 3. Use the map function to extract something of interest and create an output list. 4. Organize the output list to optimize for further processing. 5. Use the reduce function to compute a set of results. 6. Produce the final output.	BTL1	Remember
7	**What is Scheduling in MapReduce.** 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. Then the reduce tasks will run with the same behaviors.		
8	**What is Synchronization in MapReduce**. 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.** 1. Keep it warm 2. The bigger the better 3. The long view 4. Keep it secure	BTL4	Analyze
10	Why Hadoop used in Big data. 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.** ✓ **Hadoop Distributed File System:** A reliable, high-bandwidth, low-cost, data storage cluster that facilitates the management of related files across machines. ✓ **MapReduce engine:** A high-performance parallel/distributed dataprocessing mplementation of the MapReduce algorithm.	BTL2	Understand
12	**What is HDFS?** 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.** 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.** 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?** ✓ 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. ✓ Stores the metadata of a block in the local file system based on the metadata template in the NameNode. ✓ Performs periodic validations of file checksums. ✓ Sends regular reports to the NameNode about what blocks are available for file operations. ✓ Provides metadata and data to clients on demand. HDFS supports direct access to the data nodes from client application programs. ✓ Forwards data to other data nodes based on a "pipelining" model.	BTL3	Apply
16	**What are the steps involved in Hadoop MapReduce.** ● Getting the data ready ● Let the mapping begin ● 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?** 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?** 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.** 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?** 1. Predictive modeling 2. Text analytics 3. Other statistical and data-mining algorithms	BTL1	Remember
5	**Where the Predictive modelling used in big data.** 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.** 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.** 1. Advanced Forecasting, 2. Optimization,	BTL1	Remember

	3. Cluster Analysis For Segmentation		
8	**What is data mining?** 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.** 1. Classification trees 2. Logistic regression 3. Neural networks 4. Clustering techniques like K-nearest neighbors	BTL1	Remember
10	**Define Classification trees.** 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.** 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.** 1. It can come from untrusted sources 2. It can be dirty 3. The signal-to-noise ratio can be low 4. It can be real-time	BTL2	Understand
13	**The infrastructure needed to support big data to achieve.** 1. Integrate technologies 2. Store large amounts of disparate data 3. Process data in motion 4. Warehouse data	BTL4	Analyze
14	**The structure of data that might be associated with what?** 1. Documents 2. E-mails 3. Log files 4. Tweets 5. Facebook posts	BTL2	Understand
15	**How Retrieval Insight operation working in Structured and Unstructured data.** ***Retrieval*** ***Insight*** 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?** 1. Terms 2. Facts 3. Events 4. Concepts 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 1. Putting results together with structured data 2. putting Big Data to use 3. Text analytic tools for Big Data	BTL2	Understand
UNIT-5 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?** ✓ Exploratory stage ✓ Codifying stage ✓ Integration and incorporation stage	BLT2	Understand
2	**Define Exploratory stage.** 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.** 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.** 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?** 1. You must create a common understanding of data definitions. 2. You must develop of a set of data services to qualify the data and make it consistent and ultimately trustworthy. 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?** 1. ✓ Extract: Read data from the source database. 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. 3. ✓ Load: Write data to the target database.	BLT2	Understand
7	**What is Data transformation?** 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?** 1. Phase 1: Look for patterns in big data without concern for data quality. 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?** ● ✓ Keep data quality in perspective ● ✓ Consider real-time data requirements ● ✓ Don't create new silos of information	BLT2	Understand
10	**How the flow of data managed in Big data?** 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.** 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 2. ✓ Collecting information about the movement around a secure site 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 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?** 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.** ● IBM InfoSphere Streams ● Twitter's Storm ● Apache S4	BLT1	Remember
14	**How will you differ CEP from Streams?** 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?** 1. ✓ Understand the problem you are trying to solve 2. ✓ Identify the processes involved 3. ✓ Identify the information required to solve the problem 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?** ● ✓ Identify the big data sources you need to use. ● ✓ Map the big data types to your workflow data types. ● ✓ Ensure that you have the processing speed and storage access to support your workflow ● ✓ Select the data store best suited to the data types. ● ✓ 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

MULTIPLE CHOICE QUESTION

1. Data in bytes size is called Big Data.

A. Tera
B. Giga
C. Peta
D. Meta

2. How many V's of Big Data?

A. 2
B. 3
C. 4
D. 5

3. Unprocessed data or processed data are observations or measurements that can be expressed as text, numbers, or other types of media.

A. True
B. False

4. In computers, a is a symbolic representation of facts or concepts from which information may be obtained with a reasonable degree of confidence.

A. Data
B. Knowledge
C. Program
D. Algorithm

5. In Big Data environments, Velocity refers –

A. Data can arrive at fast speed
B. Enormous datasets can accumulate within very short periods of time
C. Velocity of data translates into the amount of time it takes for the data to be processed
D. All of the mentioned above

6. In Big Data environments, Variety of data includes –

A. Includes multiple formats and types of data
B. Includes structured data in the form of financial transactions,
C. Includes semi-structured data in the form of emails and unstructured data in the form of images
D. All of the mentioned above

7. In Big Data environment, Veracity of data refers -

A. Quality or fidelity of data
B. Large size of the data that cannot be process
C. Small size of the data that can easily process
D. All of the mentioned above

8. Which of the following are Benefits of Big Data Processing?

A. Cost Reduction
B. Time Reductions
C. Smarter Business Decisions
D. All of the mentioned above

9. Structured data conforms to a data model or schema and is often stored in tabular form.

A. True
B. False

10. Data that does not conform to a data model or data schema is known as _.

A. Structured data
B. Unstructured data
C. Semi-structured data
D. All of the mentioned above

11. Amongst which of the following is/are not Big Data Technologies?

A. Apache Hadoop
B. Apache Spark
C. Apache Kafka

D. Apache Pytarch

12. __ involves the simultaneous execution of multiple sub-tasks that collectively comprise a larger task.

A. Parallel data processing
B. Single channel processing
C. Multi data processing
D. None of the mentioned above

13. Amongst which of the following can be considered as the main source of unstructured data.

A. Twitter
B. Facebook
C. Webpages
D. All of the mentioned above

14. Amongst which of the following shows an example of unstructured data,

A. Students roll number, age
B. Videos
C. Audio files
D. Both B and C

15. Scalability, elasticity, resource pooling, self-service, low cost and fault tolerance are the features of,

A. Cloud computing
B. Power BI
C. System development
D. None of the mentioned above

16. Amongst which of the following is/are the cloud deployment models,

A. Public Cloud
B. Private Cloud
C. Hybrid Cloud
D. All of the mentioned above

17. Virtualization separates resources and services from the underlying physical delivery environment.

A. True
B. False

18. What is a Virtual Machine (VM)?

A. Virtual representation of a physical computer
B. Virtual representation of a logical computer
C. Virtual System Integration
D. All of the mentioned above

19. In the given Virtual Architecture, name the missing layer,

A. Virtualization layer
B. Storage layer

C. Abstract layer
D. None of the mentioned above

20. MongoDB is a database.
A. SQL
B. DBMS
C. NoSQL
D. RDBMS

21. MongoDB support cross platform and is written in language.
A. Python
B. C++
C. R
D. Java

22. Amongst which of the following is / are true to run MongoDB?
A. High availability through built-in replication and failover
B. Management tooling for automation, monitoring, and backup
C. Fully elastic database as a service with built-in best practices
D. All of the mentioned above

23. Big data deals with high-volume, high-velocity and high-variety information assets,
A. True
B. False

24. ___hypervisor runs directly on the underlying host system. It is also known as "Native Hypervisor" or "Bare metal hypervisor".
A. TYPE-1 Hypervisor
B. TYPE- 2 Hypervisor
C. Both A and B
D. None of the mentioned above

25 is also known as "Hosted Hypervisor".
A. TYPE-1 Hypervisor
B. TYPE- 2 Hypervisor
C. Both A and B
D. None of the mentioned above

26. In the layered architecture of Big Data Stack, Interfaces and feeds,
A. Internally managed data
B. Data feeds from external sources.
C. It provides access to each and every layer & components of big data stack
D. All of the mentioned above

27 ___is the supporting physical infrastructure is fundamental to the operation and scalability of big data architecture.
A. Redundant physical infrastructure
B. Integrated System

C. Integrated Database
D. All of the mentioned above

28. Data in bytes size is called Big Data. (GATE 2021)

A. Tera
B. Giga
C. Peta
D. Meta

Answer: C) Peta GATE:, 2021

29. In computers, a is a symbolic representation of facts or concepts from which information may be obtained with a reasonable degree of confidence. (GATE 2021)

A. Data
B. Knowledge
C. Program
D. Algorithm

30. Amongst which of the following represents the Use of Hadoop, (GATE 2019)

A. Robust and Scalable
B. Affordable and Cost Effective
C. Adaptive and Flexible
D. All of the mentioned above

31. Nis a platform for developing data flows for the extraction, transformation, and loading (ETL) of huge datasets, as well as for data analysis.

A. Spark
B. HBase
C. Hive
D. Pig

32. In contrast to relational databases, Hive is a query engine that supports the elements of SQL that are specifically designed for querying data.

A. True
B. False

33. Custom extensions built in the programming language are also supported by Hive.

A. Java
B. C#
C. C
D. C++

34. Amongst which of the following is / are correct,

A. Hive is a relational database that supports SQL queries.
B. Pig is a relational database that supports SQL queries.
C. Both A and B
D. None of the mentioned above

35. In order to analyze all of this Big Data, Hive is a tool that has been developed.

A. True
B. False

36. general-purpose model and runtime framework for distributed data analytics.

A. Mapreduce
B. Spark
C. Hive
D. All of the mentioned above

37. Scalability is prioritized over latency in jobs such as _.

A. HBase
B. HDFS
C. Hive
D. Mapreduce

38. ___node serves as the Slave and is responsible for carrying out the Tasks that have been assigned to it by the JobTracker.

A. TaskReduce
B. Mapreduce
C. TaskTracker
D. JobTracker

39. Apache Hive is data storage and that stores and organizes data for study and querying.

A. Querying tool
B. Mapper
C. MapReduce
D. All of the mentioned above

40. The MapReduce framework is responsible for processing one or more pieces of data and producing the output results as _.

A. Maptask
B. Task execution
C. Mapper
D. All of the mentioned above

41. Apache Hive is a data infrastructure that is built on top of the Hadoop platform.

A. Warehouse
B. Map
C. Reduce
D. None of the mentioned above

42. The Hadoop framework is built in Java, which means that MapReduce applications do not need to be written in _.

A. C#
B. C
C. Java
D. None of the mentioned above

43. ___maps input key/value pairs to a set of intermediate key/value pairs.

A. Reducer
B. Mapper
C. File system
D. All of these

44. HQL is a query language that is used to construct the custom map-reduce framework in Hive, which is written in _.

A. Java
B. PHP
C. C#
D. None of the mentioned above

45. The is the default partitioned in Hadoop, and it offers a method called Get Partition that allows us to partition data.

A. Hash Partitioner
B. Map function
C. Reduce function
D. All of the mentioned above

46. Hadoop is a framework that can be used in conjunction with a number of related products. Among the most common cohorts are _.

A. MapReduce, Hive and HBase
B. Hive, Spark and HBase
C. Spark, Hive and ZooKeeper
D. Spark, HBase and Hive

47. __ is best described as a programming model that is used to construct Hadoop-based applications that can be scaled up and down.

A. Oozie
B. Zookepper
C. MapReduce
D. All of the mentioned above

48. Amongst which of the following is/are the Hive function Meta commands.

A. Show functions
B. Describe function
C. Both A and B
D. None of the mentioned above

Answer: C) Both A and B

49. ___is a shell utility that can be used to run Hive queries in either interactive or batch mode, depending on the situation.

A. $HIVE_HOME/bin/hive
B. $HIVE/bin/
C. $HIVE_HOME/hive
D. All of the mentioned above

50. The tool has the capability of listing all of the possible database schemas.

A. sqoop-list-databases
B. Hbase-list
C. hive schema
D. sqoop-list-columns

51. Amongst which of the following is/are true with reference to User-defined Functions of Hive.

A. function that fetches one or more columns from a row as arguments
B. It returns a single value
C. Both A and B
D. None of the mentioned above

52. Amongst which of the following is/are correct.

A. Default location of Hadoop configuration is in $HADOOP /conf/ HOME
B. If $HADOOP HOME is specified, Sqoop will utilise the default installation location
C. default location of Hadoop configuration is in $HADOOP HOME/conf/
D. Sqoop command-line tool serves as a wrapper for the bin/hadoop script that is included with Hadoop as a base.

53. A serves as the master, and each cluster has just one NameNode. (GATE 2020)

A. Data Node
B. Block Size
C. Data block
D. NameNode

54. HDFS always needs to work with large data sets.

A. True
B. False

55. HDFS operates in a manner.

A. Master-slave architecture
B. Master-worker architecture
C. Worker-slave architecture
D. All of the mentioned above

56. HDFS follows the write-once, read-many.

A. True
B. False

57. Amongst which of the following is not aligns as a characteristic of HDFS? (GATE 2020)

A. HDFS file system is well suited for storing data associated with applications that require low latency data access.
B. HDFS is well-suited for storing data connected to applications that require low- latency data access to be performed.
C. HDFS is not suited for instances in which multiple/simultaneous writes to the same file are required.
D. None of the mentioned above

58. In order to interact with HDFS, a command line interface named is provided. (GATE 2019)

A. HDFS Shell
B. DFS Shell
C. K Shell
D. FS Shell

59. HDFS stores data in a distributed manner, the data can be processed in parallel on a ___ ___of nodes.

A. Cluster
B. Data Node
C. Master Node
D. None of the mentioned above

60. With reference to HDFS, Name Node is the prime node which contains metadata.

A. True
B. False

61. The database which is used to manage and store data in real time is called ___.

A. Traditional database
B. Operational database
C. Database Management System
D. None of the mentioned above

62. Database requirements for operational data includes ___.

A. Indexing and Cataloging, Replication
B. File Storage and Structure, Query Processing
C. Transactions Support
D. All of the mentioned above

63. Indexing and Cataloguing refers to efficiently store data that can be retrieved?

A. True
B. False

64. File Storage and structure is an important function of the operational database to robust enough to sort and store files at relevant locations?

A. True
B. False

Answer: A) True

65. Query processing system refers to the entire process from translating a ___ to the database system.

A. Query
B. Statement
C. Function
D. None of the mentioned above

66. Operational Database with distributed systems and ___ based system can harness the true potential with big data.

A. SQL
B. NoSQL
C. PL / SQL

D. None of the mentioned above

67. ___ a record is created for every search key valued in the database.

A. Primary Index
B. Secondary Index
C. Complex Index
D. None of the mentioned above

68. A non-clustered index tells us where the data lies?

A. True
B. False

69. Data warehouse modeling is the initial stage of building a data warehouse wherein the ___ is designed.

A. Schema
B. Table
C. Both A and B
D. None of the mentioned above

70. An operational database is designed to run the day-to-day operations or transactions of your business?

A. True
B. False

71. As companies move past the experimental phase with Hadoop, many cite the need for additional capabilities, including _____-_-_-_

a) Improved data storage and information retrieval
b) Improved extract, transform and load features for data integration
c) Improved data warehousing functionality
d) Improved security, workload management, and SQL support

72. Point out the correct statement.

a) Hadoop do need specialized hardware to process the data
b) Hadoop 2.0 allows live stream processing of real-time data
c) In the Hadoop programming framework output files are divided into lines or records
d) None of the mentioned

73. According to analysts, for what can traditional IT systems provide a foundation when they're integrated with big data technologies like Hadoop? (GATE 2022)

a) Big data management and data mining
b) Data warehousing and business intelligence
c) Management of Hadoop clusters
d) Collecting and storing unstructured data

74. Hadoop is a framework that works with a variety of related tools. Common cohorts include _____-_-_

a) MapReduce, Hive and HBase
b) MapReduce, MySQL and Google Apps
c) MapReduce, Hummer and Iguana
d) MapReduce, Heron and Trumpet

75. Point out the wrong statement.

a) Hardtop processing capabilities are huge and its real advantage lies in the ability to process terabytes & petabytes of data

b) Hadoop uses a programming model called "MapReduce", all the programs should conform to this model in order to work on the Hadoop platform

c) The programming model, MapReduce, used by Hadoop is difficult to write and test

d) All of the mentioned

76. What was Hadoop named after? (GATE 2020)

a) Creator Doug Cutting's favorite circus act

b) Cutting's high school rock band

c) The toy elephant of Cutting's son

d) A sound Cutting's laptop made during Hadoop development

77. All of the following accurately describe Hadoop, EXCEPT _____

a) Open-source

b) Real-time

c) Java-based

d) Distributed computing approach

78. __ can best be described as a programming model used to develop Hadoop-based applications that can process massive amounts of data.

a) MapReduce

b) Mahout

c) Oozie

d) All of the mentioned

79. __ has the world's largest Hadoop cluster.

a) Apple

b) Datamatics

c) Facebook

d) None of the mentioned

80. Facebook Tackles Big Data With __based on Hadoop. (GATE 2021)

a) 'Project Prism'

b) 'Prism'

c) 'Project Big'

d) 'Project Data'

81. __ __is a platform for constructing data flows for extract, transform, and load (ETL) processing and analysis of large datasets.

a) Pig Latin

b) Oozie

c) Pig

d) Hive

82. Point out the correct statement.

a) Hive is not a relational database, but a query engine that supports the parts of SQL specific to querying data

b) Hive is a relational database with SQL support

c) Pig is a relational database with SQL support

d) All of the mentioned

83. __ hides the limitations of Java behind a powerful and concise Clojure API for Cascading.

a) Scalding

b) HCatalog

c) Cascalog

d) All of the mentioned

84. Hive also support custom extensions written in _____

a) C#

b) Java

c) C

d) C++

85. Point out the wrong statement.

a) Elastic MapReduce (EMR) is Facebook's packaged Hadoop offering

b) Amazon Web Service Elastic MapReduce (EMR) is Amazon's packaged Hadoop offering

c) Scalding is a Scala API on top of Cascading that removes most Java boilerplate

d) All of the mentioned

86. __ __is the most popular high-level Java API in Hadoop Ecosystem

a) Scalding

b) HCatalog

c) Cascalog

d) Cascading

87. __ __is general-purpose computing model and runtime system for distributed data analytics.

a) Mapreduce

b) Drill

c) Oozie

d) None of the mentioned

88. The Pig Latin scripting language is not only a higher-level data flow language but also has operators similar to _____._._

a) SQL
b) JSON
c) XML
d) All of the mentioned

89. __ jobs are optimized for scalability but not latency.

a) Mapreduce
b) Drill
c) Oozie
d) Hive

90. __ is a framework for performing remote procedure calls and data serialization.

a) Drill
b) BigTop
c) Avro
d) Chukwa

91. Which one is not in Basic analytics for insight

a) Slicing and Dicing of data
b) Reporting & Basic monitoring.
c) Simple Visualizations
d) The business process.

92. Find out which one is under Advanced analytics for insight (GATE 2020)

a) predictive modeling
b) drive revenue.
c) transparency
d) business intelligence

93. Which Analytics type become part of the business process.

a) Operationalized analytics
b) Basic analytics
c) Advanced analytics
d) Monetized analytics

94. Which Analytics type is utilized to directly drive revenue.

a) Basic analytics
b) Operationalized analytics
c) Monetized analytics
d) Advanced analytics

95. Slicing and dicing refers

a) to breaking down your data into smaller sets of data
b) to monitor large volumes of datain real time
c) to identify anomalies
d) provides algorithms for complex analysis

96. Basic monitoring refers

a) to breaking down your data into smaller sets of data
b) to monitor large volumes of datain real time
c) to identify anomalies
d) provides algorithms for complex analysis

97. Anomaly identification refers

a) to breaking down your data into smaller sets of data
b) to monitor large volumes of datain real time
c) to identify anomalies an event where the actual observation differs from what you expected
d) provides algorithms for complex analysis

98. Predictive modeling used to (GATE 2021)

a) to determine future outcomes
b) to find patterns in thatdata
c) calculates the distances between the record and points
d) broke the data into training data anda test data set

99. Advanced analytics can be deployed to find patterns in

a) data & prediction
b) forecasting
c) complex event processing
d) All of the above

100. The process of analyzing unstructured text, extracting relevant information, and transforming it into structured information is called

a) Text analytics
b) data-mining
c) segmentation
d) cluster analysis

101. The potential characteristics of your data

a) It can come from untrusted sources
b) It can be real-time
c) It can be dirty
d) All of the above

102. Big data consists of (GATE 2019)

a) Structured data
b) Semi-structureddata
c) Unstructured data
d) All of the above

103. Dirty data refers

a) inaccurate data
b) incompletedata
c) erroneous data
d) All of the above

104. What are the Infrastructure needed to support big data

a) Integrate technologies
b) Process data in motion
c) Warehouse data
d) All of the above

105. Users of Orbitz perform

a) the company collects hundreds of gigabytes of raw data each day from these searches
b) useful information inthe web log files that it was collecting from its web analytics software
c) Both of (A) & (B)
d) None of these

106. Which one is false statement in the following

a) Hadoop providedthe distributed file system
b) Hive provided an SQL-type interface
c) A series of steps to put the data into Hive. After the data was in Hive, the company used machine learning.
d) None of these

107. Nokia provides

a) multi petabyte platform
b) wireless communication devices and services
c) improve customer retention
d) All of the above

108. A number of vendors on the market today support big data solutions

a) IBM , Oracle
b) SAS , Pentaho

c) Tableau
d) All of the above

109. InfoSphere Streams product is tightly integrated with

a) its Statistical Package for the Social Sciences (SPSS) statistical software to support real-time predictive analytics
b) capability to dynamically update models based on real-time data
c) Both of (A) & (B)
d) None of the above

110. The different kinds of unstructured data are

a) Documents, E-mails
b) Log files , Tweets
c) Face book posts
d) All of the above

111. NLP abbreviated as

a) Numerous Language Processing
b) Natural Language Processing
c) Numerous Language Project
d) Natural Language Project

112. What are the methods exist for analyzing unstructured data

a) Natural Language Processing
b) knowledge discovery & data mining
c) Information retrieval & statis tics
d) All of the above

113. Search is about retrieving a document based on

a) what end users already know they are looking for.
b) discovering information
c) classification of documents
d) None of these above

114. A goal of NLP is

a) To derive meaning from text.
b) Generally makes use of linguistic concepts such as grammatical structures and parts of speech
c) To determine who did what to whom, when, where, how, and why.
d) All of the above

115. Lexical/morphological analysis

a) examines the characteristics of an individual word

b) uses grammatical structure to dissect the text and put individual words into context
c) determines the possible meanings of a sentence.
d) to determine the meaning of text beyond the sentence level.

116. Which one is uses grammatical structure to dissect the text and put individual words into context.

a) Lexical analysis
b) Semantic analysis
c) Syntactic analysis
d) Discourse-level analysis

117. To extract information from various document sources, organiza tions sometimes need to develop rules. These rules can be

a) The name of a person must start with a capital letter.
b) Every course on the college website must follow a three-digit course number and a semicolon.
c) A logo must appear in a certain location on every page.
d) All of the above

118. Sentiment analysis is used to

a) identify viewpoints or emotions in the underlying text
b) organizing information into hierarchical relationships
c) None of these
d) All of the above

119. Text Analytics Tools for Big Data

a) Attensity
b) Clarabridge , OpenText
c) IBM , SAS
d) All of the above

120. NASA is using predictive models to

a) analyze safety data on aircrafts
b) understand whether the introduction of a new technology into an aircraft
c) dealing with a massive amount of data
d) All of the above

121. Which are the major categories of big data integration?

a) Theintegration of multiple big data sources in big data environments
b) The integration of unstructured big data sources with structured enterprise data.
c) Both of these (A) & (B)
d) None of these

122. Which one is not in the stages of Big data analysis?

a) Exploratory stage
b) Codifying stage
c) Integration and incorporation stage
d) None of the above

123. To complete your analysis, you need to move large amounts of data from

a) log files
b) Twitter feeds , RFID tags
c) weather data feeds
d) All of these above.

124. Which is widely used as an underlying building block for capturing and processing big data

a) Hadoop
b) Twitter feeds , RFID tags
c) weather data feeds
d) All of these above.

125. Which are two primary components of Hadoop

a) Hadoop Distributed File System (HDFS)
b) MapReduce
c) Both (A) & (B)
d) None of the above

126. Traditional integration tools

a) E
T
L
b) S
S
L
c) P
S
L
d) T
T
L

127. Which one is not true for the following?

a) Traditional integration tools such as ETL would not be fast enough to move the large streams ofdata in time to deliver results for analysis.
b) Flume is used to collect large amounts of log data from distributed servers.

c) Flume is designed for scalability and can continually add more resources to a system to handle extremely large amounts of data in an efficient way.
d) None of the above.

128. To codify the relationship between your big data analytics and your operational data, you need to

a) Integrate the data.
b) Split the data
c) Divide the data
d) Explore the data

129. Traditionally, data integration has focused on themovement of data through

a) middleware
b) specifications on message passing
c) application programming interfaces (APIs).
d) All of the above

130. Traditional tools for data integration are evolving to handle the increasing variety of

a) Unstructured data
b) The growing volume
c) Velocity of big data
d) All of the above

131. What are the basic principles apply from specific to individual systems/applications

a) You must create a common understanding of data definitions
b) You must develop of a set of data services to qualify the data and make it consistent and ultimately trustworthy
c) You need a streamlined way to integrate your big data sources and systems of record
d) All of the above

132. What are the new tools used to support integrationof big data environments (GATE 2021)

a) Sqoop
b) Scribe
c) Both (A) & (B)
d) None of the above

133. What are the important functions of ETL which required to get data from one data environment and put it into another data environment.

a) **Extract**: Read data from the source database.
b) **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.

c) **Load**: Write data to the target database
d) All of the above.

134. Customer relationship management [CRM]) used to

a) analyze and report on data relevant to their specific business focus
b) batch processing in data warehouse environments
c) to consolidate information across disparate sources
d) None of the above

135. What are the statements true about Data Transformation?

a) Data transformation is the process of changing the format of data so thatit can be used by different applications.
b) The process of data transformation is made far more complex because of the staggering growth in the amount of unstructured data.
c) Data transformation tools are not designed to work well with unstructured data
d) All of the above

136. Which tools can transform the data in the source or target database without requiring an ETL server? (GATE 2020)

a) ELT (extract, load, and transform)
b) ELT uses structured query language (SQL) to transform the data
c) ETL tools extracted the data to an intermediary location to perform the transformation before loading the data to the data warehouse.
d) Massively parallel processing systems and columnar databases

137. What are the different phase approach followed for Data Quality?

a) Look for patterns in big data without concern for data quality.
b) 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.
c) Both (A) & (B)
d) None of the above

138. The quality of data refers to characteristics about the data, including

a) consistency, accuracy
b) reliability, completeness, timeliness
c) reasonableness, and validity

139. All of the aboveData quality software can be used to

a) identify all the variations of the com- pany name in your different data stores
b) ensure that you know everything that this customer purchases from your business.
c) Cleans up or removes redundant data
d) All of the above

140. Data profiling tools are used in the data quality process to help you to understand

a) The content , Structure & .Condition of your data
b) Analyze the data to identify errors and inconsistencies
c) you can ensure that your big data is complete and consistent.
d) All of the above

141. Which Hadoop tools can be used for the transformation process?

a) HiveQL
b) Pig Latin
c) Both (A) & (B)
d) None of the above

142. What are the two techniques for managing the flow of data.

a) **Streaming technology** is closely tied to the volume of the data
b) **Complex event processing** of the volume of data is secondary to the capabilityto match data to rules.
c) Both (A) & (B)
d) None of the above

143. Which statements are true about Complex event processing?

a) CEP is dependent on data streams.
b) CEP is not required for streaming data ,. Like streaming data, CEP relies on analyzing streams of data in motion.
c) Streaming computing is designed to handle a continuous stream of a large amount of unstructured data.
d) All of the above

144. In the Hadoop cluster, data is collected in which mode and then processed. (GATE 2019)

a) Batch
b) Streaming
c) Real-time calculation
d) Cluster

145. Which statement is false in streaming of data?

a) Implicit metadata from unstructured data, it is possible to parse the information using eXtensible Markup Language (XML)

b) XML is a technique for presenting unstructured text files with meaningful tags.
c) Examples of products for streaming data include IBM's InfoSphere Streams, Twitter's Storm, and Yahoo's S4
d) None of the above.

146. IBM InfoSphere Streams used to

a) perform complex analytics of heterogeneous data types
b) perform text, images, audio, voice, VoIP, video, web traffic, e-mail, GPS data, financial transaction data, satellite data, and sensors.
c) provides continuous analysis of massive data volumes.
d) All of the above

147. Twitter's Storm is an open source real-time analytics engine developed bya company called

a) BackType
b) InfoSphere
c) Apache S4
d) None of the above

148. Companies using Storm in their big data implementations include

a) Groupon
b) RocketFuel
c) Navisite and Oolgala
d) All of the above

149. Which statement is not true in CEP?

a) Streams are intended to analyze large volumes of data in real time
b) Complex Event Processing is a technique for tracking, analyzing, and processing data as an event happens
c) CEP is an advanced approach based on simple event processing that collects and combines data from different relevant sources to discover events and patterns that can result in action
d) None of the above

150. **The set of "V" characteristics that are key to operationalizing big data includes**

a) Validity: Is the data correct and accurate for the intended usage?
b) Veracity: Are the results meaningful for the given problem space?
c) Volatility: How long do you need to store this data?
d) All of the above.

ANSWERS:

1	**C**	16	**D**	31	**D**	46	**A**	61	**B**	76	**C**	91	**D**	106	**D**	121	**C**	136	**A**
2	**D**	17	**A**	32	**A**	47	**C**	62	**D**	77	**B**	92	**A**	107	**D**	122	**D**	137	**C**
3	**A**	18	**A**	33	**A**	48	**C**	63	**A**	78	**A**	93	**A**	108	**D**	123	**D**	138	**D**
4	**A**	19	**A**	34	**C**	49	**A**	64	**A**	79	**C**	94	**C**	109	**C**	124	**D**	139	**D**
5	**D**	20	**C**	35	**A**	50	**A**	65	**A**	80	**A**	95	**A**	110	**D**	125	**C**	140	**D**
6	**D**	21	**B**	36	**A**	51	**C**	66	**B**	81	**C**	96	**B**	111	**B**	126	**A**	141	**C**
7	**D**	22	**D**	37	**D**	52	**D**	67	**B**	82	**A**	97	**C**	112	**D**	127	**D**	142	**C**
8	**D**	23	**A**	38	**C**	53	**D**	68	**A**	83	**C**	98	**A**	113	**A**	128	**A**	143	**D**
9	**A**	24	**A**	39	**A**	54	**A**	69	**B**	84	**B**	99	**D**	114	**D**	129	**D**	144	**A**
10	**B**	25	**B**	40	**A**	55	**B**	70	**A**	85	**A**	100	**A**	115	**A**	130	**D**	145	**D**
11	**D**	26	**D**	41	**A**	56	**A**	71	**D**	86	**D**	101	**D**	116	**C**	131	**D**	146	**D**
12	**A**	27	**A**	42	**C**	57	**C**	72	**B**	87	**A**	102	**D**	117	**D**	132	**C**	147	**D**
13	**D**	28	**C**	43	**B**	58	**D**	73	**A**	88	**A**	103	**D**	118	**A**	133	**D**	148	**D**
14	**D**	29	**A**	44	**A**	59	**A**	74	**A**	89	**D**	104	**D**	119	**D**	134	**A**	149	**D**
15	**A**	30	**D**	45	**A**	60	**A**	75	**C**	90	**C**	105	**C**	120	**D**	135	**D**	150	**D**