MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)

Rasipuram - 637 408, Namakkal Dist., Tamil Nadu.

Estd. 2000

IQAC

MUST KNOW CONCEPTS

MKC

MCA

2021-2022

Course Code & Course Name : 21CAB13 & Big Data Analytics

Year/Sem/Sec : I / II

S.No.	Term	Notation (Symbol)	Concept / Definition / Meaning / Units / Equation / Expression	Units
Unit-I : Introduction to Big Data				
1.	Big data	--	Big data is defined as the voluminous amount of structured, unstructured or semi-structured data that has huge potential for mining but is so large that it cannot be processed using traditional database systems.	I
2.	Big data analytics	--	Big data analytics examines large amounts of data to uncover hidden patterns, correlations and other insights.	I
3.	Types of Big Data	--	Types of Big Data **1.** Structured **2.** Unstructured **3.** Semi-structured	I
4.	Characteristics of Big Data	--	Characteristics of Big Data • Volume • Variety • Velocity • Variability	I
5.	Volume	--	Volume – The name Big Data itself is related to a size which is enormous. Size of data plays a very crucial role in determining value out of data.	I
6.	Variety	--	Variety refers to heterogeneous sources and the nature of data, both structured and unstructured.	I
7.	Velocity	--	The term 'velocity' refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data.	I
8.	Variability	--	Variability – This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.	I
9.	Big data platform	--	Big data platform is a type of IT solution that combines the features and capabilities of several big data application and utilities within a single solution..	I

10.	Intelligent Data Analysis	--	Intelligent Data Analysis (IDA) is one of the hot issues in the field of artificial intelligence and information. Intelligent data analysis reveals implicit, previously unknown and potentially valuable information or knowledge from large amounts of data.	I
11.	Analytical processing	--	Analytical processing involves the interaction between analysts and collections of aggregated data that may have been reformulated into alternate representational forms as a means for improved analytical performance.	I
12.	Business analytics tools	--	Business analytics tools are types of application software that retrieve data from one or more business systems and combine it in a repository, such as a data warehouse, to be reviewed and analyzed.	I
13.	Reporting	--	Reporting is the process of organizing data into informational summaries in order to monitor how different areas of a business are performing.	I
14.	Analysis	--	Analysis is the process of exploring data and reports in order to extract meaningful, actionable insights, which can be used to better understand and improve business performance.	I
15.	R Language	--	R is the leading analytics tool in the industry and widely used for statistics and data modeling. It can easily manipulate your data and present in different ways.	I
16.	Tableau	--	Tableau Public is a free software that connects any data source be it corporate Data Warehouse.	I
17.	Python	--	Python is an object-oriented scripting language which is easy to read, write, maintain and is a free open source tool. It was developed by Guido van Rossum in late 1980's which supports both functional and structured programming methods.	I
18.	Sas	--	Sas is a programming environment and language for data manipulation and a leader in analytics, developed by the SAS Institute in 1966 and further developed in 1980's and 1990's. SAS is easily accessible, managable and can analyze data from any sources.	I
19.	Apache Spark	--	Apache Spark is a fast large-scale data processing engine and executes applications in Hadoop clusters 100 times faster in memory and 10 times faster on disk.	I
20.	Excel	--	Excel is a basic, popular and widely used analytical tool almost in all industries. Whether you are an expert in Sas, R or Tableau, you will still need to use Excel.	I
21.	Sampling distribution	--	A sampling distribution is a	I

			probability distribution of a statistic obtained from a larger number of samples drawn from a specific population.	
22.	Three primary factors of a sampling distribution	--	Three primary factors of a sampling distribution: • The number observed in a population • The number observed in the sample • The method of choosing the sample	I
23.	Resampling	--	Resampling is the method that consists of drawing repeated samples from the original data samples. The method of Resampling is a nonparametric method of statistical inference.	I
24.	Statistical inference	--	*Statistical inference* is the process of using *data* analysis to deduce properties of an underlying distribution of probability. It is assumed that the observed *data* set is sampled from a larger population.	I
25.	Prediction error	--	A prediction error is the failure of some expected event to occur. Applying that type of knowledge can inform decisions and improve the quality of future predictions.	I
Unit-II : Mining Data Streams				
26.	Streaming Applications	--	Streaming Applications Sensor networks – Monitor habitat and environmental parameters – Track many objects, intrusions, trend analysis Utility Companies – Monitor power grid, customer usage patterns etc. – Alerts and rapid response in case of problems	II
27.	Streaming data	--	Streaming data is data that is continuously generated by different sources. Such data should be processed incrementally using Stream Processing techniques without having access to all of the data.	II
28.	Benefits of streaming analytics	--	The top benefits of streaming analytics are: • Improve operational efficiencies. • Reduce infrastructure cost. • Provide faster insights and actions.	II
29.	Streaming	--	Streaming refers to any media content – live or recorded – delivered to computers and mobile devices via the internet and played back in real time.	II
30.	Stream computing	--	The word stream in stream computing is used	II

			to mean pulling in streams of data; processing the data and streaming it back out as a single flow.	
31.	Stream sampling	--	Stream sampling is the process of collecting a representative sample of the elements of a data stream.	II
32.	Stream sampling	--	Stream sampling is the process of collecting a representative sample of the elements of a data stream.	II
33.	Four main types of probability sample	--	Four main types of probability sample • Simple random sampling • Systematic sampling • Stratified sampling • Cluster sampling	II
34.	Filtering stream	--	Filtering condition of a stream item is independent of other items of the same stream or any other data stream.	II
35.	Count-distinct problem	--	In computer science, the count-distinct problem is the problem of finding the number of distinct elements in a data stream with repeated elements.	II
36.	Different streaming data types	--	Different streaming data types – Permutations, Graph Data, Geometric Data (Location Streams)	II
37.	Different streaming processing models	--	Different streaming processing models – Sliding Windows, Exponential and other decay, Duplicate sensitivity, Random order streams, Skewed streams	II
38.	Different streaming scenarios	--	Different streaming scenarios – Distributed computations, sensor network computations	II
39.	Pattern finding	--	Pattern finding: finding common patterns or features – Association rule mining, Clustering, Histograms, Wavelet & Fourier Representations	II
40.	Data Quality Issues	--	Data Quality Issues – Change Detection, Data Cleaning, Anomaly detection, Continuous Distributed Monitoring	II
41.	Learning and Predicting	--	Learning and Predicting – Building Decision Trees, Regression, Supervised Learning	II
42.	Six rules to represent a stream by buckets	--	Six rules to represent a stream by buckets • The right end of a bucket is always a position with a 1. • Every position with a 1 is in some bucket. • No position is in more than one bucket. • There are one or two buckets of any given size, up to some maximum size. • All sizes must be a power of 2. • Buckets cannot decrease in size as we move to the left (back in time).	II

43.	Decaying window	--	In a decaying window, you assign a score or weight to every element of the incoming data stream. Further, you need to calculate the aggregate sum for each distinct element by adding all the weights assigned to that element. The element with the highest total score is listed as trending or the most popular.	II
44.	Real-time analytics	--	Real-time analytics • Refers to finding meaningful patterns in data at the actual time of receiving • Real-Time Analytics Platform (RTAP) analyses the data, correlates, and predicts the outcomes in the real time.	II
45.	Benefits of RTAP	--	Benefits of RTAP • Manages and processes data and helps timely decision-making • Helps to develop dynamic analysis applications • Leads to evolution of business intelligence	II
46.	Widely used RTAPs	--	Widely used RTAPs • Apache Spark Streaming—a Big Data platform for data stream analytics in real time. • Cisco Connected Streaming Analytics (CSA)—a platform that delivers insights from high-velocity streams of live data from multiple sources and enables immediate action.	II
47.	IBM Stream Computing	--	IBM Stream Computing —a data streaming tool that analyzes a broad range of streaming data — unstructured text, video, audio, geospatial, sensor — helping organizations spot the opportunities and risks and make decisions in real time	II
48.	Sentiment Analysis other names	--	Sentiment Analysis other names • Opinion extraction • Opinion mining • Sentiment mining • Subjectivity analysis	II
49.	Why Sentiment analysis?	--	Why Sentiment analysis? Movie: Is this review positive or negative? Products: What do people think about the new iPhone? Public sentiment: How is consumer confidence? Is despair increasing? Politics: What do people think about this candidate or issue? Prediction: Predict election outcomes or market trends from sentiment	II

50.	Sentiment Analysis	--	Sentiment Analysis is the process of determining whether a piece of writing is positive, negative or neutral. Sentiment analysis helps data analysts within large enterprises gauge public opinion, conduct nuanced market research, monitor brand and product reputation, and understand customer experiences.	II
Unit-III : Hadoop Environment				
51.	Hadoop features	--	Hadoop features**:** • Open Source • Highly Scalable • Runs on Commodity Hardware • Has a good ecosystem	III
52.	YARN components	--	YARN components : Resource Manager: Runs on a master daemon and manages the resource allocation in the cluster. Node Manager: They run on the slave daemons and are responsible for the execution of a task on every single Data Node.	III
53.	YARN application components	--	YARN application components: Client ApplicationMaster(AM) Container	III
54.	Hosts View	--	Hosts View The host name, IP address, number of cores, memory, disk usage, current load average, and Hadoop components are listed in this window in tabular form.	III
55.	HDFS in Safe Mode - command	--	HDFS in Safe Mode - command**:** To Enter hdfs dfsadmin -safemode enter To Leave hdfs dfsadmin -safemode leave	III
56.	fsck	--	fsck stands for File System Check. It is a command used by HDFS. This command is used to check inconsistencies and if there is any problem in the file. For example, if there are any missing blocks for a file, HDFS gets notified through this command.	III
57.	Components of HDFS	--	NameNode – This is the master node for processing metadata information for data blocks within the HDFS DataNode/Slave node – This is the node which acts as slave node to store the data, for processing and use by the NameNode	III
58.	NameNode	--	NameNode – This is the master node for processing metadata information for data blocks within the HDFS	III

59.	DataNode/Slave node	--	DataNode/Slave node – This is the node which acts as slave node to store the data, for processing and use by the NameNode	III
60.	BackupNode	--	BackupNode- It is a read-only NameNode which contains file system metadata information excluding the block locations	III
61.	What happens when two users try to access the same file in the HDFS	--	HDFS NameNode supports exclusive write only. Hence, only the first user will receive the grant for file access and the second user will be rejected.	III
62.	Rack Awareness	--	It is an algorithm applied to the NameNode to decide how blocks and its replicas are placed. Depending on rack definitions network traffic is minimized between DataNodes within the same rack.	III
63.	HDFS Block Vs Input Split	--	The HDFS divides the input data physically into blocks for processing which is known as HDFS Block. Input Split is a logical division of data by mapper for mapping operation	III
64.	Common input formats in Hadoop	--	Text Input Format Sequence File Input Format Key Value Input	III
65.	Pseudo-Distributed Mode	--	Pseudo-Distributed Mode – In the pseudo-distributed mode, Hadoop runs on a single node just like the Standalone mode. In this mode, each daemon runs in a separate Java process. As all the daemons run on a single node, there is the same node for both the Master and Slave nodes.	III
66.	Standalone (Local) Mode	--	Standalone (Local) Mode – By default, Hadoop runs in a local mode i.e. on a non-distibuted, single node. This mode uses the local file system to perform input and output operation.	III
67.	Fully – Distributed Mode	--	Fully – Distributed Mode – In the fully-distributed mode, all the daemons run on separate individual nodes and thus forms a multi-node cluster. There are different nodes for Master and Slave nodes.	III
68.	Hadoop default block size	--	Hadoop default block size The default block size in Hadoop 1 is: 64 MB The default block size in Hadoop 2 is: 128 MB	III
69.	Distributed Cache	--	Distributed Cache is a feature of Hadoop MapReduce framework to cache files for applications. Hadoop framework makes cached files available for every map/reduce tasks running on the data nodes.	III

70.	core-site.xml	--	core-site.xml – This configuration file contains Hadoop core configuration settings, for example, I/O settings, very common for MapReduce and HDFS. It uses hostname a port.	III
71.	mapred-site.xml	--	mapred-site.xml – This configuration file specifies a framework name for MapReduce by setting mapreduce.framework.name	III
72.	hdfs-site.xml	--	hdfs-site.xml – This configuration file contains HDFS daemons configuration settings. It also specifies default block permission and replication checking on HDFS.	III
73.	yarn-site.xml	--	yarn-site.xml – This configuration file pecifies configuration settings for ResourceManager and NodeManager	III
74.	MapReduce	--	MapReduce is a programming model in Hadoop for processing large data sets over a cluster of computers, commonly known as HDFS. It is a parallel programming model.	III
75.	Two phases of MapReduce operation	--	Map phase – In this phase, the input data is split by map tasks. The map tasks run in parallel. These split data is used for analysis purpose. Reduce phase- In this phase, the similar split data is aggregated from the entire collection and shows the result.	III
Unit-IV : Data Analysis Systems and Visualization				
76.	Link Analysis	--	Link Analysis deals with mining useful information from linked structures like graphs. Graphs have vertices representing objects and links among those vertices representing relationships among those objects.	IV
77.	Link mining	--	Link mining works with graph structures that have nodes with defined set of properties. These nodes may be of the same type (homogeneous) or different (heterogeneous).	IV
78.	Hyperlink	--	The most common interpretation of the word link today is hyperlink—a means of connecting two web documents wherein activating a special element embedded in one document takes you to the other.	IV
79.	Link	--	A link represents a relationship and connects two objects that are related to each other in that specific way	IV
80.	Network, or graph	--	A collection of links representing the same kind of relationship form a network, or graph, where the objects being related correspond to the graph vertices and the links themselves are the edges.	IV
81.	Homogeneous network	--	When two objects being related by a link are of the same kind, then the network formed by such links is termed a homogeneous network	IV

82.	Link analysis	--	Link analysis is a data-analysis technique used to evaluate relationships (connections) between nodes. Relationships may be identified among various types of nodes (objects), including organizations, people and transactions.	IV
83.	LOC	--	LOC (Link-based Object Classification) is a technique used to assign class labels to nodes according to their link characteristics.	IV
84.	PageRank	--	PageRank is an algorithm that addresses the Link-based Object Ranking (LOR) problem. The objective is to assign a numerical rank or priority to each web page by exploiting the "link" structure of the web.	IV
85.	importance of a web page rating	--	The importance of a web page can be rated based on the number of backlinks to that page and the importance of the web pages that provide these backlinks, i.e., a web page referred to by important and reliable web pages, is important and reliable.	IV
86.	Backlink	--	A backlink of a page Pu is a citation to Pu from another page	IV
87.	In-degree , out-degree	--	deg (P) − The number of links coming into a page P (in-degree of P) deg (P) + The number of links going out of a page P (outdegree of P)	IV
88.	HITS	--	The Hyperlink-Induced Topic Search **(**HITS**)** algorithm was originally proposed by Kleinberg (1999) as a method of filtering results from web page search engines in order to identify results most relevant to a user query.	IV
89.	Recommender system	--	Recommender system − The objective is to develop a system that recommends choices based on user behavior. Netflix is the characteristic example of this data product, where based on the ratings of users, other movies are recommended	IV
90.	Dashboard	--	Dashboard − Business normally needs tools to visualize aggregated data. A dashboard is a graphical mechanism to make this data accessible.	IV
91.	content based recommender	--	A content based recommender works with data that the user provides, either explicitly (rating) or implicitly (clicking on a link). Based on that data, a user profile is generated, which is then used to make suggestions to the user.	IV
92.	Core components of recommender system	--	Data collection and processing Recommender model Recommendation post-processing Online modules User interface	IV
93.	Collaborative filtering	--	Collaborative filtering is a technique that can filter out items that a user might like on	IV

			the basis of reactions by similar users. It works by searching a large group of people and finding a smaller set of users with tastes similar to a particular user	
94.	Dimensionality reduction in recommender systems	--	There are two ways of using dimensionality reduction in recommender systems: The first is creating latent factor models which reduce the dimensions of both users and items simultaneously, and produce a dense matrix, which can generate rating predictions.	IV
95.	Data visualization	--	Data visualization is the graphical representation of information and data. In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.	IV
96.	VR	--	Virtual reality is going to have a huge impact on the potential for data visualizations, allowing people to interact with data in the third dimension for the first time.	IV
97.	Common general types of data visualization	--	Common general types of data visualization**:** • Charts • Tables • Graphs • Maps • Info graphics • Dashboards	IV
98.	Big Data visualization	--	Big Data visualization involves the presentation of data of almost any type in a graphical format that makes it easy to understand and interpret.	IV
99.	Interaction techniques	--	Interaction techniques essentially involve data entry and manipulation, and thus place greater emphasis on input than output. Output is merely used to convey affordances and provide user feedback.	IV
100.	Four stages of Visualization	--	Four stages of Visualization • Exploration • Analysis • Synthesis • Presentation	IV
Unit-V : Frameworks and Applications				
101.	Hbase	--	HBase is a distributed column-oriented database built on top of the Hadoop file system.	V
102.	Hive	--	Hive: It is a platform used to develop SQL type scripts to do MapReduce operations	V
103.	Features of Hive	--	It stores schema in a database and processed data into HDFS. It provides SQL type language for querying	V

			called HiveQL or HQL. It is familiar, fast, scalable, and extensible	
104.	Hive - Data Types	--	• Column Types • Literals • Null Values • Complex Types	V
105.	Hive - Complex Types	--	Arrays: Arrays in Hive are used the same way they are used in Java. Maps: Maps in Hive are similar to Java Maps. Structs: Structs in Hive is similar to using complex data with comment	V
106.	Where to Use HBase	--	● Apache HBase is used to have random, real-time read/write access to Big Data. ● It hosts very large tables on top of clusters of commodity hardware. ● Apache HBase is a non-relational database modeled after Google's Bigtable. Bigtable acts up on Google File System, likewise Apache	V
107.	YARN components	--	YARN components : Resource Manager: Runs on a master daemon and manages the resource allocation in the cluster. Node Manager: They run on the slave daemons and are responsible for the execution of a task on every single Data Node.	V
108.	YARN application components	--	YARN application components: • Client • ApplicationMaster(AM) • Container	V
109.	Key components of HBase	--	Region- This component contains memory data store and Hfile. Region Server-This monitors the Region. HBase Master-It is responsible for monitoring the region server. Zookeeper- It takes care of the coordination between the HBase Master component and the client. Catalog Tables-The two important catalog tables are ROOT and META.ROOT table tracks where the META table is and META table stores all the regions in the system.	V
110.	Region	--	Region- This component contains memory data store and Hfile.	V
111.	Zookeeper	--	Zookeeper- It takes care of the coordination between the HBase Master component and the client.	V
112.	Operational commands in HBase	--	Record Level Operational Commands in HBase are –put, get, increment, scan and delete. Table Level Operational Commands in HBase are-describe, list, drop, disable and scan.	V
113.	RDBMS data model Vs HBase data model	--	RDBMS is a schema based database whereas HBase is schema less data model.	V

			RDBMS does not have support for in-built partitioning whereas in HBase there is automated partitioning. RDBMS stores normalized data whereas HBase stores de-normalized data.	
114.	Catalog tables in HBase	--	The two important catalog tables in HBase, are ROOT and META. ROOT table tracks where the META table is and META table stores all the regions in the system.	V
115.	HBase Vs Hive	--	HBase and Hive both are completely different hadoop based technologies-Hive is a data warehouse infrastructure on top of Hadoop whereas HBase is a NoSQL key value store that runs on top of Hadoop.	V
116.	MongoDB features	--	• Licence based (also Open Source) • NoSQL Database • Document Oriented • Aggregation Pipeline etc.	V
117.	Cassandra features	--	• Open Source • NoSQL Database • Log-Structured Storage • Includes Cassandra Structure Language (CQL)	V
118.	NoSQL Database	--	NoSQL Database is a non-relational Data Management System, that does not require a fixed schema. It avoids joins, and is easy to scale. The major purpose of using a NoSQL database is for distributed data stores with humongous data storage needs	V
119.	Scaleup or Vertical Scaling	--	Scaleup or Vertical Scaling: Increase of RAM, CPU, and HDD	V
120.	Scaleout or Horizontal Scaling	--	Scaleout or Horizontal Scaling: Increase of Commodity hardware	V
121.	Types of NoSQL Databases	--	Types of NoSQL Databases: • Key-value Pair Based • Column-oriented Graph • Graphs based • Document-oriented	V
122.	Key Value Pair Based	--	Key Value Pair Based Data is stored in key/value pairs. It is designed in such a way to handle lots of data and heavy load. Key-value pair storage databases store data as a hash table where each key is unique, and the value can be a JSON, BLOB (Binary Large Objects), string, etc. eg. DynamoDB, Redis, etc.	V
123.	Column-based	--	Column-oriented databases work on columns and are based on BigTable paper by Google. Every column is treated separately. Values of single column databases are stored contiguously. eg.Cassandra, HBase, etc.	V
124.	Documents-Oriented	--	Document-Oriented NoSQL DB stores and retrieves data as a key value pair but the value part is stored as a document. The document is	V

			stored in JSON or XML formats. The value is understood by the DB and can be queried. eg. CouchDB, MongoDB,etc.	
125.	Graph-Based	--	A graph type database stores entities as well the relations amongst those entities. The entity is stored as a node with the relationship as edges. An edge gives a relationship between nodes. Every node and edge has a unique identifier. eg. Neo4j, OrientDB,etc.	V
Placement Questions				
126.	Text mining	--	Text mining is the art and science of discovering knowledge, insights, and patterns from an organized collection of textual databases.	
127.	Naïve Bayes technique	--	Naïve Bayes technique is a supervised machine learning technique that that uses probability theory based analysis.	
128.	Support Vector Machine	--	Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges.	
129.	Web mining	--	Web mining is the art and science of discovering patterns and insights from the World Wide Web so as to improve it.	
130.	Business Intelligence	--	Business Intelligence (BI) is an umbrella term that includes a variety of IT applications that are used to analyze an organization's data and communicate the information to relevant users.	
131.	Applications of BI and data mining	--	Retail, Telecom, Customer Relationship Management, Healthcare and Wellness, Education, Banking, Financial Services, Insurance, Manufacturing, and Public Sector	
132.	Data warehouse	--	A data warehouse (DW) is an organized collection of integrated, subject oriented databases designed to support decision support functions	
133.	Data mining	--	Data mining is the art and science of discovering knowledge, insights, and patterns in data.	
134.	Classification techniques	--	Classification techniques are called supervised learning as there is a way to supervise whether the model’s prediction is right or wrong.	
135.	Decision tree	--	A decision tree is a hierarchically organized branched, structured to help make decision in an easy and logical manner.	
136.	Regression	--	Regression is a relatively simple and the most popular statistical data mining technique. The goal is to fit a smooth well-defined curve to	

			the data. Regression analysis techniques, for example, can be used to model and predict the energy consumption as a function of daily temperature.	
137.	Artificial neural network	--	Artificial neural network (ANN) is a sophisticated data mining technique from the Artificial Intelligence stream in Computer Science. It mimics the behavior of human neural structure: Neurons receive stimuli, process them, and communicate their results to other neurons successively, and eventually a neuron outputs a decision.	
138.	Cluster analysis	--	Cluster analysis is an exploratory learning technique that helps in identifying a set of similar groups in the data. It is a technique used for automatic identification of natural groupings of things.	
139.	Association rules	--	Association rules are a popular data mining method in business, especially where selling is involved. Also known as market basket analysis, it helps in answering questions about cross-selling opportunities	
140.	NFS Vs HDFS	--	NFS (Network File System) is one of the oldest and popular distributed file storage systems whereas HDFS (Hadoop Distributed File System) is the recently used and popular one to handle big data.	
141.	Structured Data	--	Data which can be stored in traditional database systems in the form of rows and columns, for example the online purchase transactions can be referred to as Structured Data.	
142.	Semi structured data.	--	Data which can be stored only partially in traditional database systems, for example, data in XML records can be referred to as semi structured data.	
143.	Unstructured data	--	Unorganized and raw data that cannot be categorized as semi structured or structured data is referred to as unstructured data. Facebook updates, Tweets on Twitter, Reviews, web logs, etc. are all examples of unstructured data.	
144.	Two ways of Big Data processing	--	Two ways of Big Data processing 1. Batch processing 2. Stream processing	
145.	Data Science Vs Big Data	--	Data Science Vs Big Data • Data science is a broad spectrum of activities involving analysis of Big Data, finding patterns, trends in data, interpreting statistical terms and predicting future trends. • Big Data is just one part of Data Science. Though Data Science is a	

			broad term and very important in the overall Business operations, it is nothing without Big Data. • All the activities we perform in Data Science are based on Big Data. Thus Big Data and Data Science are interrelated and cannot be seen in isolation.	
146.	Cloud computing	--	Cloud computing is internet-based computing. It relies on sharing computing resources on-demand rather than having local servers or PCS and other devices.	
147.	Rule induction	--	Rule induction is an area of machine learning in which formal rules are extracted from a set of observations.	
148.	Sensor networks	--	Sensor networks are a huge source of data occurring in streams. They are used in numerous situations that require constant monitoring of several variables, based on which important decisions are made.	
149.	Bloom Filter	--	A Bloom Filter is a space-efficient probabilistic data structure, conceived by Burton Howard Bloom in 1970, that is used to test whether an element is a member of set.	
150.	Reservoir sampling	--	Biased reservoir sampling is defined as bias function to regulate the sampling from the stream.	

Dr.M.Moorthy

Faculty Prepared **Signature**

HoD