INSTITUTE OF AERONAUTICAL ENGINEERING

(Autonomous)

Dundigal, Hyderabad - 500 043

COMPUTER SCIENCE AND ENGINEERING

DEFINITIONS AND TERMINOLOGY QUESTION BANK

Course Title	**BIG DATA AND BUSINESS ANALYTICS**				
Course Code	ACS012				
Programme	B.Tech				
Semester	SEVEN				
Course Type	Core				
Regulation	IARE - R16				
Course Structure	**Theory**			**Practical**	
	Lectures	**Tutorials**	**Credits**	**Laboratory**	**Credits**
	3	1	4	3	2
Chief Coordinator	Dr. M Madhu Bala, Professor				

COURSE OBJECTIVES:

The course should enable the students to:	
I	The scope and essentiality of Big Data and Business Analytics.
II	The technologies used to store, manage, and analyze big data in a Hadoop ecosystem.
III	The techniques and principles in big data analytics with scalability and streaming capability.
IV	The hypothesis on the optimized business decisions in solving complex real-world problems.

DEFINITIONS AND TERMINOLOGYQUESTION BANK

S. No	QUESTION	ANSWER	Blooms Level	CO
UNIT -I				
1	What is Data?	The quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media.	Remember	CO 1
2	What is Big Data?	Big Data is a collection of data that is huge in size and yet growing exponentially with time.	Remember	CO 1

S. No	QUESTION	ANSWER	Blooms Level	CO
3	Define Structured data?	Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data.	Remember	CO 2
4	Define Unstructured data?	Unstructured data is not having any data model or any structure.	Remember	CO 2
5	Define Semi-structured data?	Semi-structured data can contain both structured and unstructured data.	Remember	CO 2
6	Define volume?	Volume represents Size of data.	Remember	CO 1
7	Define variety?	Variety represents the different data types like text, audios and videos.	Remember	CO 2
8	Define velocity?	Velocity represents the rate at which data grows.	Remember	CO 1
9	Define variability?	Variability refers to the inconsistency of data.	Remember	CO 1
10	Define five V's of Big Data?	The five V's of Big data are: •volume •variety •velocity •variability •Value	Remember	CO 1
11	List out the applications of big data applications?	•Marketing •Finance •Government •Healthcare •Insurance •Retail	Remember	CO 2
12	Define big data analytics?	**Big data analytics** is the often-complex process of examining **big data** to uncover information -- such as hidden patterns, correlations, market trends and customer preferences -- that can help organizations make informed business decisions.	Remember	CO 2
13	List types of data analytics?	• Descriptive analytics • Diagnostic analytics • Predictive analytics • Prescriptive analytics	Understand	CO 2
15	Define descriptive Analytics?	Descriptive analytics answers the question of what happened.	Remember	CO 2
16	Define diagnostic analytics?	data can be measured against other data to answer the question of why something happened.	Remember	CO 2
17	Define Predictive analytics	Predictive analytics tells what is likely to happen.	Remember	CO 2
18	Define Prescriptive analytics	It prescribes what action to take to eliminate a future problem or take full advantage of a promising trend.	Remember	CO 2
19	What are challenges of big data	• Dealing with data growth • Generating insights in a timer manner • Recruiting and retaining big data talent • Integrating disparate data sources • Validating data	Remember	CO 1
20	What is Batch processing	Efficient way of processing high volumes of data where a group of transactions is collected over a period of time.	Remember	CO 1

S. No	QUESTION	ANSWER	Blooms Level	CO
UNIT-II				
1	What is Hadoop.	Apache **Hadoop** is an open source framework that is **used** to efficiently store and process large datasets ranging in size from gigabytes to petabytes of data. By allowing clustering multiple computers to analyze massive datasets in parallel more quickly.	Remember	CO 4
2	What are the core concepts of the Hadoop framework?	Two core components of hadoop are • Hadoop Distributed File System (HDFS) • MapReduce	Remember	CO 4
3	Define MapReduce	MapReduce Programming is a software framework helps to process massive amounts of data in parallel	Remember	CO 3
4	Name some practical applications of Hadoop.	• Managing street traffic • Fraud detection and prevention • Analyse customer data in real-time to improve customer service • Accessing unstructured medical data from physicians, HCPs, etc., to improve healthcare services.	Remember	CO 4
5	Define RDBMS?	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.	Remember	CO 3
6	What are the vital Hadoop tools that can enhance the performance of Big Data?	The Hadoop tools that boost Big Data performance significantly are Hive, HDFS, HBase, SQL, NoSQL, Oozie, Clouds, Avro, Flume, and ZooKeeper.	Remember	CO 4
7	List out the usecases of hadoop?	• Hadoop in Finance • Hadoop in Healthcare • Hadoop in Telecom • Hadoop in Retail	Remember	CO 3
8	How many Input Formats are there in Hadoop?	There are three input formats in Hadoop.	Remember	CO 4
9	Explain the InputFormats of Hadoop?	• Text Input Format: The text input is the default input format in Hadoop. • Sequence File Input Format: This input format is used to read files in sequence. • Key Value Input Format: This input format is used for plain text files.	Remember	CO 3
10	List out some important features of Hadoop? .	The important features of Hadoop are – • Hadoop framework is designed on Google MapReduce that is based on Google's Big Data File Systems. • Hadoop framework can solve many questions efficiently for Big Data analysis.	Remember	CO 4
11	Compare RDBMS and Hadoop?	• RDBMS is made to store structured data, whereas Hadoop can store any kind of data i.e. unstructured, structured, or semi-structured. • RDBMS follows "Schema on write" policy while Hadoop is based on "Schema on read" policy.	Understand	CO 3

S. No	QUESTION	ANSWER	Blooms Level	CO
		• The schema of data is already known in RDBMS that makes Reads fast, whereas in HDFS, writes no schema validation happens during HDFS write, so the Writes are fast.		
12	What is difference between regular file system and HDFS?	**Regular File Systems** a) Small block size of data (like 512 bytes) b) Multiple disk seeks for large files **HDFS** a) Large block size (orders of 64mb) b) Reads data sequentially after single seek	Remember	CO 4
13	What is HDFS block size and what did you choose in your project?	By default, the HDFS block size is 64MB. It can be set to higher values as 128MB or 256MB. 128MB is acceptable industry standard.	Remember	CO 3
14	What are different hdfs dfs shell commands to perform copy operation?	$hadoop fs -copyToLocal $ hadoop fs -put	Remember	CO 3
15	What is the default replication factor?	Default replication factor is 3	Remember	CO 4
16	How to keep HDFS cluster balanced?	Balancer is a tool that tries to provide a balance to a certain threshold among data nodes by copying block data distribution across the cluster.	Remember	CO 3
17	What are the daemons of HDFS?	1. NameNode 2. DataNode 3. Secondary NameNode.	Remember	CO 3
18	Command to format the NameNode?	$ hdfs namenode –format	Remember	CO 3
19	What is a DataNode?	• A DataNode stores data in the Hadoop File System HDFS is a slave node. • On startup, a DataNode connects to the NameNode. • DataNode instances can talk to each other mostly during replication.	Remember	CO 4
20	What is the command for printing the topology?	It displays a tree of racks and DataNodes attached to the tracks as viewed by the hdfs dfsadmin –printTopology	Remember	CO 4
UNIT -III				
1	What is Secondary NameNode?	A Secondary NameNode is a helper daemon that performs check pointing in HDFS.	Remember	CO 6
2	What is a block?	Blocks are the smallest continuous location on your hard drive where data is stored.	Remember	CO 6
3	What is a block scanner in HDFS?	Block scanner runs periodically on every DataNode to verify whether the data blocks stored are correct or not.	Remember	CO 6
4	Define DFS?	Stands for "Distributed File System." A DFS manages files and folders across multiple computers. It serves the same purpose as a traditional file system, but is designed to provide file storage and controlled access to files over local and wide area networks.	Remember	CO 6
5	What are the two files associated with Name node?	**FsImage:** It contains the complete state of the file system namespace since the start of the NameNode. **EditLogs:** It contains all the recent modifications	Remember	CO 6

S. No	QUESTION	ANSWER	Blooms Level	CO
		made to the file system with respect to the most recent FsImage.		
6	What are the different goals of HDFS	Fault detection and recovery Huge datasets Hardware at data	Remember	CO 6
7	What is daemon?	Daemon is the process that runs in background in the UNIX environment. In Windows it is 'services' and in DOS it is 'TSR'.	Remember	CO 6
8	What is the function of 'job tracker'?	Job tracker is one of the daemons that runs on name node and submits and tracks the MapReduce tasks in Hadoop.	Remember	CO 6
9	What is the process of indexing in HDFS?	Once data is stored HDFS will depend on the last part to find out where the next part of data would be stored.	Remember	CO 6
10	What type of data is processed by Hadoop?	Hadoop processes the digital data only.	Remember	CO 6
11	Define Namenode	NameNode is the master node in the Apache Hadoop HDFS Architecture that maintains and manages the blocks present on the DataNodes (slave nodes). NameNode is a very highly available server that manages the File System Namespace and controls access to files by clients.	Remember	CO 6
12	What is Block Management	Block Management maintains the membership of datanodes in the cluster. It supports block-related operations such as creation, deletion, modification and getting location of the blocks. It also takes care of replica placement and replication.	Understand	CO 6
13	Define the Block Pool	A Block Pool is a set of blocks that belong to a single namespace. Datanodes store blocks for all the block pools in the cluster.	Remember	CO 6
14	Define the programming interface	A programming interface, consisting of the set of statements, functions, options, and other ways of expressing program instructions and data provided by a program or language for a programmer to use.	Remember	CO 6
15	What is FSI image?	It contains the complete state of the file system namespace since the start of the NameNode.	Remember	CO 5
16	Define Replication Factor?	Replication Factor: It is basically the number of times Hadoop framework replicate each and every Data Block. Block is replicated to provide Fault Tolerance.	Remember	CO 5
17	Describe the Coherency Model	A coherency model for a filesystem describes the data visibility of reads and writes for a file. HDFS trades off some POSIX requirements for performance, so some operations may behave differently than you expect them to.	Remember	CO 6
18	Define Limitations of HAR files	There are a few limitations to be aware of with HAR files. Creating an archive creates a copy of the original files, so you need as much disk space as the files you are archiving to create the archive (although you can delete the originals once you have created the archive).	Remember	CO 6
19	What is Edit log image?	It contains all the recent modifications made to the file system with respect to the most recent FsImage.	Understand	CO 6
20	Define HDFS?	The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware.	Remember	CO 6

S. No	QUESTION	ANSWER	Blooms Level	CO
UNIT-IV				
1	Define the MapReduce	Hadoop MapReduce (Hadoop Map/Reduce) is a software framework for distributed processing of large data sets on compute clusters of commodity hardware. It is a sub-project of the Apache Hadoop project. The framework takes care of scheduling tasks, monitoring them and re-executing any failed tasks.	Remember	CO 7
2	Explain the Serialization	Serialization is the process of turning structured objects into a byte stream for transmission over a network or for writing to persistent storage. Deserialization is the reverse process of turning a byte stream back into a series of structured objects.	Understand	CO 7
3	What is Avro MapReduce	Avro provides a number of classes for making it easy to run MapReduce programs on Avro data. For example, AvroMapper and AvroReducer in the org.apache.avro.mapred package are specializations of Hadoop's (old style) Mapper and Reducer classes	Remember	CO 8
4	Define the Sequence File	Imagine a log file, where each log record is a new line of text. If you want to log binary types, plain text isn't a suitable format. Hadoop's Sequence File class fits the bill in this situation, providing a persistent data structure for binary key-value pairs.	Remember	CO 8
5	Define the Sorting and merging SequenceFiles	The most powerful way of sorting (and merging) one or more sequence files is to use MapReduce. MapReduce is inherently parallel and will let you specify the number of reducers to use, which determines the number of output partitions	Remember	CO 7
6	Define the Map File	A Map File is a sorted Sequence File with an index to permit lookups by key. Map File can be thought of as a persistent form of java.util.Map (although it doesn't implement this interface), which is able to grow beyond the size of a Map that is kept in memory.	Remember	CO 7
7	Define the Reducer	The reducer has to find the maximum value for a given key	Remember	CO 7
8	What is Packaging	package provides a mechanism for using different serialization frameworks in Hadoop.	Remember	CO 7
9	Describe the Debugging a Job	The time-honored way of debugging programs is via print statements, and this is certainly possible in Hadoop. However, there are complications to consider: with programs running on tens, hundreds, or thousands of nodes, how do we find and examine the output of the debug statements, which may be scattered across these nodes? For this particular case, where we are looking for (what we think is) an unusual case, we can use a debug statement to log to standard error, in conjunction with a message to update the task's status message to prompt us to look in the error log.	Remember	CO 7
10	Explain the JobControl	When there is more than one job in a MapReduce workflow, the question arises: how do you manage the jobs so they are executed in order? There are several approaches, and the main consideration is whether you	Understand	CO 7

S. No	QUESTION	ANSWER	Blooms Level	CO
		have a linear chain of jobs, or a more complex directed acyclic graph (DAG) of jobs		
11	Define the Piggybacking	Piggybacking, in a wireless communications context, is the unauthorized access of a wireless LAN. Piggybacking is sometimes referred to as "Wi-Fi squatting."	Remember	CO 8
12	Explain the Task Failure	The most common way that this happens is when user code in the map or reduce task throws a runtime exception. If this happens, the child JVM reports the error back to its parent tasktracker, before it exits. The error ultimately makes it into the user logs. The tasktracker marks the task attempt as failed, freeing up a slot to run another task.	Remember	CO 7
13	Describe the Jobtracker Failure	Failure of the jobtracker is the most serious failure mode. Hadoop has no mechanism for dealing with failure of the jobtracker—it is a single point of failure—so in this case the job fails. However, this failure mode has a low chance of occurring, since the chance of a particular machine failing is low.	Understand	CO 7
14	What is Task JVM Reuse	Hadoop runs tasks in their own Java Virtual Machine to isolate them from other running tasks. The overhead of starting a new JVM for each task can take around a second, which for jobs that run for a minute or so is insignificant.	Remember	CO 7
15	Describe the Multiple Outputs	FileOutputFormat and its subclasses generate a set of files in the output directory. There is one file per reducer, and files are named by the partition number: part-r-00000, partr-00001, etc.	Understand	CO 08
UNIT-V				
1	Define Apache Pig	Apache Pig is a platform that is used to analyze large data sets. It consists of a high-level language to express data analysis programs, along with the infrastructure to evaluate these programs. One of the most significant features of Pig is that its structure is responsive to significant parallelization.	Remember	CO 8
2	List Pig components	Pig consists of two components: Pig Latin, which is a language A runtime environment, for running PigLatin programs.	Remember	CO 8
3	Define the Map Reduce mode	In this mode, queries written in Pig Latin are translated into MapReduce jobs and are run on a Hadoop cluster (cluster may be pseudo or fully distributed). MapReduce mode with the fully distributed cluster is useful of running Pig on large datasets.	Remember	CO 8
4	List different Execution Types	Pig has two execution types or modes: local mode and MapReduce mode.	Remember	CO 8
5	Define Local mode	Local mode in local mode, Pig runs in a single JVM and accesses the local filesystem. This mode is suitable only for small datasets and when trying out Pig.	Remember	CO 8

S. No	QUESTION	ANSWER	Blooms Level	CO
6	Describe the Map Reduce mode in pig	In MapReduce mode, Pig translates queries into MapReduce jobs and runs them on a Hadoop cluster. The cluster may be a pseudo- or fully distributed cluster. MapReduce mode (with a fully distributed cluster) is what you use when you want to run Pig on large datasets.	Understand	CO 8
7	Define the Grunt	Grunt has line-editing facilities like those found in GNU Readline (used in the bash shell and many other command-line applications). For instance, the Ctrl-E key combination will move the cursor to the end of the line. Grunt remembers command history, too,1 and you can recall lines in the history buffer using Ctrl-P or Ctrl-N (for previous and next) or, equivalently, the up or down cursor keys.	Remember	CO 8
8	What is meant by Load function?	A function that specifies how to load data into a relation from external storage.	Remember	CO 9
9	What is meant by Store function?	A function that specifies how to save the contents of a relation to external storage. Often, load and store functions are implemented by the same type. For example, PigStorage, which loads data from delimited text files, can store data in the same format	Remember	CO 9
10	Define Macros	Macros provide a way to package reusable pieces of Pig Latin code from within Pig Latin itself.	Remember	CO 9
11	Define Parallelism	When running in MapReduce mode it's important that the degree of parallelism matches the size of the dataset. By default, Pig will set the number of reducers by looking at the size of the input, and using one reducer per 1GB of input, up to a maximum of 999 reducers. You can override these parameters by setting pig.exec.reduc ers.bytes.per.reducer (the default is 1000000000 bytes) and pig.exec.reducers.max (default 999).	Remember	CO 9
12	Explain the Hive Shell	The shell is the primary way that we will interact with Hive, by issuing commands in HiveQL. HiveQL is Hive's query language, a dialect of SQL. It is heavily influenced by MySQL, so if you are familiar with MySQL you should feel at home using Hive.	Remember	CO 9
13	What is the HiveQL	Hive's SQL dialect, called HiveQL, does not support the full SQL-92 specification. There are a number of reasons for this. Being a fairly young project, it has not had time to provide the full repertoire of SQL-92 language constructs.	Remember	CO 9
14	Define Tables	A Hive table is logically made up of the data being stored and the associated metadata describing the layout of the data in the table. The data typically resides in HDFS, although it may reside in any Hadoop filesystem, including the local filesystem or S3.	Remember	CO 9

S. No	QUESTION	ANSWER	Blooms Level	CO
15	Explain the Storage Formats	Two dimensions that govern table storage in Hive: the row format and the file format. The row format dictates how rows, and the fields in a particular row, are stored. In Hive parlance, the row format is defined by a SerDe, a portmanteau word for a Serializer-Deserializer.	Remember	CO 9

Signature of the Faculty **HOD,CSE**