INSTITUTE OF AERONAUTICAL ENGINEERING

(Autonomous)
Dundigal, Hyderabad - 500 043

INFORMATION TECHNOLOGY

DEFINITIONS AND TERMINOLOGY QUESTION BANK

Course Title	**BIG DATA AND BUSINESS ANALYTICS**				
Course Code	ACS012				
Programme	B.Tech				
Semester	VII	CSE			
Course Type	Core				
Regulation	IARE - R16				
Course Structure	**Theory**			**Practical**	
	Lectures	**Tutorials**	**Credits**	**Laboratory**	**Credits**
	3	1	4	3	2
Chief Coordinator	Ms. G Sulakshana, Assistant Professor				
Course Faculty	Mr. D Rahul, Assistant Professor				

COURSE OBJECTIVES:

The course should enable the students to:	
I	Optimize business decisions and create competitive advantage with Big data analytics.
II	Understand several key big data technologies used for storage, analysis and manipulation of data.
III	Recognize the key concepts of Hadoop framework, map reduce.
IV	Demonstrate the concepts in Hadoop for application development.

DEFINITIONS AND TERMINOLOGY QUESTION BANK

S.No	QUESTION	ANSWER	Blooms Level	CO	CLO	CLO Code
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.	Understand	CO 1	CLO 1	ACS012 .01
2	What is Big Data?	Big Data is a collection of data that is huge in size and yet growing exponentially with time.	Understand	CO 1	CLO 1	ACS012 .01
3	Define Structured data?	Any data that can be stored, accessed and processed in the form of fixed format is termed	Remember	CO 1	CLO 1	ACS012 .01

S.No	QUESTION	ANSWER	Blooms Level	CO	CLO	CLO Code
		as a 'structured' data.				
4	Define Unstructured data?	Any data with unknown form or the structure is classified as unstructured data	Remember	CO 1	CLO 1	ACS012 .01
5	Define Semi-structured data?	Semi-structured data can contain both structured and unstructured data.	Remember	CO 1	CLO 1	ACS012 .01
6	Define volume?	Volume represents Size of data.	Understand	CO 1	CLO 2	ACS012 .02
7	Define variety?	Variety represents the different data types like text, audios and videos.	Understand	CO 1	CLO 2	ACS012 .02
8	Define velocity?	Velocity represents the rate at which data grows.	Understand	CO 1	CLO 2	ACS012 .02
9	Define variability?	Variability refers to the inconsistency of data.	Understand	CO 1	CLO 2	ACS012 .02
10	Define four V's of Big Data?	The four V's of Big data are: a) volume b) variety c) velocity d) variability	Understand	CO 1	CLO 2	ACS012 .02
11	List out the applications of big data applications?	1. Marketing 2. Finance 3. Government 4. Healthcare 5. Insurance 6. Retail	Remember	CO 1	CLO 4	ACS012 .04
12	List the types of cloud environment.	1. Public cloud 2. Private cloud	Remember	CO 1	CLO 4	ACS012 .04
13	List types of data analytics	Descriptive analytics Diagnostic analytics Predictive analytics Prescriptive analytics	Understand	CO 1	CLO 3	ACS012 .03
15	Define descriptive analytics	Descriptive analytics answers the question of what happened.	Remember	CO 1	CLO 3	ACS012 .03
16	Define diagnostic analytics	data can be measured against other data to answer the question of why something happened.	Remember	CO 1	CLO 3	ACS012 .03
17	Define Predictive analytics	Predictive analytics tells what is likely to happen.	Remember	CO 1	CLO 3	ACS012 .03
18	Define Prescriptive analytics	It prescribe what action to take to eliminate a future problem or take full advantage of a promising trend..	Remember	CO 1	CLO 3	ACS012 .03
19	What are challenges of big data	1. Dealing with data growth 2. Generating insights in a timer manner 3. Recruiting and retaining big data talent 4. Integrating disparate data sources 5. Validating data	Remember	CO 1	CLO 3	ACS012 .03
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	CLO 3	ACS012 .03

S.No	QUESTION	ANSWER	Blooms Level	CO	CLO	CLO Code
		UNIT-II				
1	What is Hadoop.	Hadoop is an infrastructure equipped with relevant tools and services required to process and store Big Data. To be precise, Hadoop is the 'solution' to all the Big Data challenges. Furthermore, the Hadoop framework also helps organizations to analyze Big Data and make better business decisions.	Remember	CO 2	CLO 6	ACS012 .06
2	What are the core concepts of the Hadoop framework?	Hadoop. is fundamentally based on two core concepts. They are: **HDFS:** HDFS or Hadoop Distributed File System is a Java-based reliable file system used for storing vast datasets in the block format. The Master-Slave Architecture powers it. **MapReduce:** MapReduce is a programming structure that helps process large datasets. This function is further broken down into two parts – while 'map' segregates the datasets into tuples, 'reduce' uses the map tuples and creates a combination of smaller chunks of tuples.	Understand	CO 2	CLO 5	ACS012 .05
3	What are the primary components.	The primary components of Hadoop are: i. HDFS ii. Hadoop MapReduce iii. Hadoop Common iv. YARN v. PIG and HIVE – The Data Access Components. vi. HBase – For Data Storage vii. Ambari, Oozie and ZooKeeper – Data Management and Monitoring Component viii. Thrift and Avro – Data Serialization components ix. Apache Flume, Sqoop, Chukwa – The Data Integration Components x. Apache Mahout and Drill – Data Intelligence Components	Understand	CO 2	CLO 5	ACS012 .05
4	Name some practical applications of	a) Managing street traffic b) Fraud detection and prevention	Understand	CO 2	CLO 5	ACS012 .05

S.No	QUESTION	ANSWER	Blooms Level	CO	CLO	CLO Code
	Hadoop.	c) Analyse customer data in real-time to improve customer service d) Accessing unstructured medical data from physicians, HCPs, etc., to improve healthcare services.				
5	What are the modes in which Hadoop can run?	The modes in which Hadoop can run are: •Standalone mode – This is a default mode of Hadoop that is used for debugging purpose. It does not support HDFS. •Pseudo-distributed mode – This mode required the configuration of mapred-site.xml, core-site.xml, and hdfs-site.xml files. Both the Master and Slave Node are the same here. •Fully-distributed mode – Fully-distributed mode is Hadoop's production stage in which data is distributed across various nodes on a Hadoop cluster. Here, the Master and the Slave Nodes are allotted separately.	Understand	CO 2	CLO 6	ACS012 .06
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 2	CLO 7	ACS012 .07
7	What is the purpose of RecordReader in Hadoop?	Hadoop breaks data into block formats. RecordReader helps integrate these data blocks into a single readable record. For example, if the input data is split into two blocks – Row 1 – Welcome to Row 2 – UpGrad RecordReader will read this as "Welcome to UpGrad."	Remember	CO 2	CLO 5	ACS012 .07
8	How many Input Formats are there in Hadoop?	There are three input formats in Hadoop.	Remember	CO 2	CLO 5	ACS012 .05
9	Explain the InputFormats of Hadoop?	1. Text Input Format: The text input is the default input format in Hadoop. 2. Sequence File Input Format: This input format is used to read files in sequence. 3. Key Value Input Format: This input format is used for plain text files.	Remember	CO 2	CLO 8	ACS012 .08
10	State some of the important features of	The important features of Hadoop are – • Hadoop framework is	Remember	CO 2	CLO 7	ACS012 .07

S.No	QUESTION	ANSWER	Blooms Level	CO	CLO	CLO Code
	Hadoop. .	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.				
11	How can you differentiate RDBMS and Hadoop?	1. RDBMS is made to store structured data, whereas Hadoop can store any kind of data i.e. unstructured, structured, or semi-structured. 2. RDBMS follows "Schema on write" policy while Hadoop is based on "Schema on read" policy. 3. 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. 4. RDBMS is licensed software, so one needs to pay for it, whereas Hadoop is open source software, so it is free of cost. 5. RDBMS is used for Online Transactional Processing (OLTP) system whereas Hadoop is used for data analytics, data discovery, and OLAP system as well.	Remember	CO 2	CLO 8	ACS012 .08
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 2	CLO 5	ACS012 .05
13	What is HDFS block size and what did you chose 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 2	CLO 5	ACS012 .05
14	What are	$ hadoop fs -copyToLocal	Remember	CO 2	CLO 5	ACS012 .05

S.No	QUESTION	ANSWER	Blooms Level	CO	CLO	CLO Code
	different hdfs dfs shell commands to perform copy operation?	$ hadoop fs -copyFromLocal $ hadoop fs -put				
15	What is the default replication factor?	Default replication factor is 3	Remember	CO 2	CLO 5	ACS012 .05
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.	Understand	CO 2	CLO 7	ACS012.07
17	What are the daemons of HDFS?	1. NameNode 2. DataNode 3. Secondary NameNode.	Understand	CO 2	CLO 6	ACS012.06
18	Command to format the NameNode?	$ hdfs namenode -format	Remember	CO 2	CLO 8	ACS012.08
19	What is a DataNode?	1. A DataNode stores data in the Hadoop File System HDFS is a slave node. 2. On startup, a DataNode connects to the NameNode. 3. DataNode instances can talk to each other mostly during replication.	Remember	CO 2	CLO 5	ACS012.05
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 2	CLO 8	ACS012.08
UNIT -III						
1	What is Secondary NameNode?	A Secondary NameNode is a helper daemon that performs checkpointing in HDFS.	Understand	CO 3	CLO 9	ACS012 .09
2	What is a block?	Blocks are the smallest continuous location on your hard drive where data is stored.	Remember	CO 3	CLO 10	ACS012 .10
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 3	CLO 9	ACS012 .09
4	Define Data Integrity?	Data Integrity is about the correctness of the data.	Remember	CO 3	CLO 11	ACS012 .11
5	What is a heartbeat in HDFS?	Heartbeats in HDFS are the signals that are sent by DataNodes to the NameNode to indicate that it is functioning properly (alive).	Remember	CO 3	CLO 12	ACS012 .12
6	What is meant by streaming access?	Steaming access refers to reading the complete data instead of retrieving single record from the database.	Understand	CO 3	CLO 9	ACS012 .09
7	What is daemon?	Daemon is the process that runs	Remember	CO 3	CLO 10	ACS012 .10

S.No	QUESTION	ANSWER	Blooms Level	CO	CLO	CLO Code
		in background in the UNIX environment. In Windows it is 'services' and in DOS it is 'TSR'.				
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.	Understand	CO 3	CLO 9	ACS012 .09
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 3	CLO 12	ACS012 .12
10	What type of data is processed by Hadoop?	Hadoop processes the digital data only.	Remember	CO 3	CLO 10	ACS012 .10
11	Explain the HDFS Federation	HDFS Federation improves the existing HDFS architecture through a clear separation of namespace and storage, enabling generic block storage layer. It enables support for multiple namespaces in the cluster to improve scalability and isolation. Federation also opens up the architecture, expanding the applicability of HDFS cluster to new implementations and use cases.	Remember	CO 3	CLO 11	ACS012 .11
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 3	CLO 9	ACS012 .09
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 3	CLO 11	ACS012 .11
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.	Understand	CO 3	CLO 9	ACS012 .09
15	What is unstructured data?	The phrase unstructured datausually refers to information that doesn't reside in a traditional row-column database. As you might expect, it's the opposite of structured data — the data stored in fields in a database.	Remember	CO 3	CLO 10	ACS012 .10
16	Define Replica	A replica is an exact	Remember	CO 3	CLO 9	ACS012 .09

S.No	QUESTION	ANSWER	Blooms Level	CO	CLO	CLO Code
		reproduction executed by the original artist or a copy or reproduction, especially one on a scale smaller than the original. A replica is a copying closely resembling the original concerning its shape and appearance.				
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.	Understand	CO 3	CLO 12	ACS012 .12
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). There is currently no support for archive compression, although the files th	Remember	CO 3	CLO 10	ACS012 .10
19	Define Compression	File compression brings two major benefits: it reduces the space needed to store files, and it speeds up data transfer across the network, or to or from disk. When dealing with large volumes of data, both of these saving	Understand	CO 3	CLO 9	ACS012 .09
20	Describe the Codecs	A codec is the implementation of a compression-decompression algorithm. In Hadoop, a codec is represented by an implementation of the CompressionCodec interface. So, for example, GzipCodec encapsulates the compression and decompression algorithm for gzip.	Remember	CO 3	CLO 12	ACS012 .12
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	Remember	CO 4	CLO 15	ACS012 .015

S.No	QUESTION	ANSWER	Blooms Level	CO	CLO	CLO Code
		tasks.				
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 4	CLO 16	ACS012 .016
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 4	CLO 16	ACS012 .016
4	Describe 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 4	CLO 15	ACS012 .015
5	Describe 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	Understand	CO 4	CLO 14	ACS012 .014
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 4	CLO 16	ACS012 .016
7	Define the Reducer	The reducer has to find the maximum value for a given key	Remember	CO 4	CLO 13	ACS012 .013
8	What is Packaging	Packaging We don't need to make any modifications to the program to run on a cluster rather than on a single machine, but we do need to package the program as a JAR file to send to the cluster.	Remember	CO 4	CLO 16	ACS012 .016
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	Remember	CO 4	CLO 14	ACS012 .014

S.No	QUESTION	ANSWER	Blooms Level	CO	CLO	CLO Code
		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.				
10	Explain the Job Control	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 have a linear chain of jobs, or a more complex directed acyclic graph (DAG) of jobs	Understand	CO 4	CLO 16	ACS012 .016
11	Define the Piggy backing	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 4	CLO 15	ACS012 .015
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 4	CLO 14	ACS012 .014
13	Describe the Job tracker 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 4	CLO 16	ACS012 .016
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	Remember	CO 4	CLO 15	ACS012 .015

S.No	QUESTION	ANSWER	Blooms Level	CO	CLO	CLO Code
		insignificant.				
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 4	CLO 13	ACS012 .013
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.	Understand	CO 5	CLO 20	ACS012 .020
2	What are Pig components	Pig consists of two components: 1. **Pig Latin,** which is a language 2. **A runtime environment,** for running PigLatin programs.	Remember	CO 5	CLO 19	ACS012 .019
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.	Understand	CO 5	CLO 18	ACS012 .018
4	What are the Execution Types	Pig has two execution types or modes: local mode and MapReduce mode.	Remember	CO 5	CLO 17	ACS012 .017
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.	Understand	CO 5	CLO 18	ACS012 .018
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 5	CLO 20	ACS012 .020
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	Understand	CO 5	CLO 20	ACS012 .020

S.No	QUESTION	ANSWER	Blooms Level	CO	CLO	CLO Code
		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.				
8	Explain the Load function	A function that specifies how to load data into a relation from external storage.	Remember	CO 5	CLO 18	ACS012 .018
9	Describe the 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	Understand	CO 5	CLO 19	ACS012 .019
10	Define Macros	Macros provide a way to package reusable pieces of Pig Latin code from within Pig Latin itself.	Understand	CO 5	CLO 17	ACS012 .017
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 sets 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 5	CLO 20	ACS012 .020
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 5	CLO 20	ACS012 .020
13	Describe 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 5	CLO 17	ACS012 .017
14	Define Tables	A Hive table is logically made	Understand	CO 5	CLO 18	ACS012 .018

S.No	QUESTION	ANSWER	Blooms Level	CO	CLO	CLO Code
		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.				
15	Explain the Storage Formats	There are 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 5	CLO 19	ACS012 .019

Signature of the Faculty **HOD, IT**