INSTITUTE OF AERONAUTICAL ENGINEERING

(Autonomous)
Dundigal, Hyderabad - 500 043

COMPUTER SCIENCE AND ENGINEERING

TUTORIAL QUESTION BANK

Course Title	**BIG DATA ND BUSINESS ANALYTICS**				
Course Code	**ACS012**				
Program	**B.Tech**				
Semester	**SEVEN**				
Course Type	**Core**				
Regulations	**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 students will try to learn:	
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.

COURSE OUTCOMES:

At the end of the course the students should be able to:

Course Outcomes		**Knowledge Level (Bloom's Taxonomy)**
CO 1	Explain the evolution of big data with its characteristics and challenges with traditional business intelligence.	Understand
CO 2	Compare big data analysis and analytics in optimizing the business decisions.	Understand
CO 3	Classify the key issues and applications in intelligent business and scientific computing.	Understand

CO 4	Explain the big data technologies used to process and querying the big data in Hadoop, MapReduce, Pig and Hive.	Understand
CO 5	Make use of appropriate components for processing, scheduling and knowledge extraction from large volumes in distributed Hadoop Ecosystem.	Apply
CO 6	Translate the data from traditional file system to HDFS for analyzing big data in Hadoop ecosystem.	Understand
CO 7	Develop a MapReduce application for optimizing the jobs.	Apply
CO 8	Develop applications for handling huge volume of data using Pig Latin.	Apply
CO 9	Explain the importance of big data framework HIVE and its built-in functions, data types and services like DDL.	Understand
CO 10	Demonstrate business models and scientific computing paradigms, and tools for big data analytics.	Understand
CO 11	Categorize Hadoop components for developing real time big data analytics in various applications like recommender systems, social media applications etc.	Analyze

TUTORIAL QUESTION BANK

UNIT – I

INTRODUCTION TO BIG DATA

PART - A (SHORT ANSWER QUESTIONS)

S No	QUESTIONS	Blooms Taxonomy Level	How does this Subsume the Level	Course Outcome
1	**Recall** the term data and show its importance in various data sets.	Remember	---	CO 1
2	**Define** the term information for data analysis.	Remember	---	CO 1
3	**Describe** “BIG DATA” in simple terms and along with it’s significance.	Understand	The learner to **recall** the concept of data and information clearly and **explain** the importance in terms of GB, TB and PB.	CO 2
4	**List** out various data formats that come under Big Data?	Remember	---	CO 1
5	**Compare** structured and unstructured data.	Understand	The learner to **recall** and **Compare** the different data formats i.e., structured and unstructured data in terms of sources, size, speed etc.	CO 2
6	**Relate** the different sources of Big Data, which leads to huge volumes.	Understand	The learner to **recall** the sources of big data and **understand** how it leads to the high and huge volume of data.	CO 2
7	**Illustrate** the characteristics of data and its sensitivity for further enhancements?	Understand	The learner to **list** the characteristics of the data and **explain** the sensitivity characteristics for future enhancement.	CO 2
8	**Explain** about the different approaches to deal with Big Data?	Understand	The learner to **recall** and explain different approaches in bigdata processing.	CO 1

9	**State** few examples of human generated and machine-generated data. Mention in which category your examples belong to?	Remember	---	CO 1
10	**Identify** the benefits and importance of Big Data in this modern world.	Remember	---	CO 1
11	**How** does Big Data assist in Business Decision making?	Remember	---	CO 1
12	**Which** programming language is preferred for specific Big Data Processing among R, Python or other language.	Remember	---	CO 1
13	**Recall** the term Big Data Analytics. What's the need to store Data for Business Analytics?	Remember	---	CO 1
14	**Define** various kinds of projects are better suitable for Big Data? Name top 3 domains where Big Data projects are applicable.	Remember	---	CO 1
15	**Extend** the adoption of Big Data have impact on day to day business operations with different use cases.	Understand	The learner to **recall** bigdata use cases and **relate** the power of Big Data in various domains and at various levels.	CO 2
16	**Define** Big Data insight? How are Big Data and Data Science related?	Remember	---	CO 1
17	**List** the several methodologies to avoid over fitting.	Remember	---	CO 1
18	**Compare** the importance of business analysis and analytics?	Understand	The learner to **recall** the definitions of analysis and analytics in **comparison** with importance of business decision making.	CO 1,CO 2
19	**How** Outliers skew the result in the input data which may affect the behavior of the model.	Remember	---	CO 1
20	**List** the tools for Big Data Visualization?	Remember	---	CO 1
PART - B (LONG ANSWER QUESTIONS)				
1	**Explain** the ETL(Extract-Transform-Load) process concerning Big Data with neat sketch?	Understand	The learner to **recall** all the ETL operations and tools and then **describe** its functionality towards Big Data.	CO 2
2	**Classify** the types of Digital Data and explain the sources of Digital Data with a neat sketch.	Understand	The learner to **define** all the three types of Digital Data and **outline** its sources clearly.	CO 2
3	**Illustrate** the Evolution of Big Data in detail? In perspective of Doug Laney and a Gartner analyst coined the term "Big Data".	Understand	The learner to **recall** the definition of Big Data and **summarize** the evolution of big data from primitive levels.	CO 2
4	**Summarize** the challenges of Big Data in various phases of process with a neat diagram?	Understand	The learner to **recall** the different phases in big data process and **explain** the challenges included in each phase.	CO 1,CO 2
5	**Illustrate** the basic characteristics and sources of Big Data?	Understand	The learner to **recall** the characteristics of big data and **show** the 5V's of big data characteristics along with sources.	CO 2,CO 3
6	Annotate your comments. **Why** we need Big Data?	Remember	---	CO 1

7	**Recognize** how is traditional Business Intelligent (BI) environment different from the Big Data environment?	Remember	---	CO 1,CO 2
8	**Summarize** a typical data warehouse environment with the Big Data?	Understand	The learner to **define** forces on data warehouse concept with the big data sources examples with a sketch.	CO 2
9	**Describe** the term Big Data Analytics and what is changing in the realms of Big Data?	Understand	The learner to **recall** and **explain** the concept of analytics with its changing scenarios of data.	CO 2
10	**Explain** the various applications of Big Data analytics and why this sudden hype around Big Data Analytics?	Understand	The learner to **list** and **explain** the different analytical applications in the sudden hype.	CO 3
11	**Classify** the Big Data entails with first and second school of thoughts.	Understand	The learner to **recall** the big data terms and **relate** to the thoughts of analytics.	CO 2
12	**Classify** the different analytics types such as Analytics -1.0, Analytics 2.0, Analytics 3.0 with a neat diagram?	Understand	The learner to **recall** the various analytics types in terms of hindsight, insight and foresight and **illustrate** in neat diagram.	CO 3
13	**List** the top challenges facing Big Data in the present scenario along with Hadoop solutions.	Remember	---	CO 1
14	**Describe** what kind of technologies are we looking forward to meeting the challenges posed by Big Data?	Understand	The learner to **recall** and **explain** about various technologies included to meet the Big Data challenges.	CO 3
15	**Outline** the various terminologies used in Big Data environments with a neat diagram?	Understand	The learner to recall the basic key terminology in big data and explain in detail with diagram.	CO 3
16	**Identify** the various types of Analytics along with its impact.	Apply	The learner to **recall** and **summarize** the different types of analytics with respect to predictive and prescriptive analytics. **Recognize** the impact of each in big data processing	CO 3
17	**Outline** the key questions to be answered by all the organizations stepping onto the analytics?	Understand	The learner **recall** and **explain** few key questions to transform from the storage of data to the insights from these analytics.	CO 11
18	**Recall** the CAP theorem and how it is different from ACID properties in a distributed computing environment?	Remember	---	CO 1
19	**Explain** how Big Data Analytics can be useful in the development of smart cities and explain the landscape of Big Data Technology?	Understand	The learner to **recall** the concepts of big data analytics and **explain** the landscape of technology included in real time applications related to smart city development.	CO 2
20	**Compare** SQL, No SQL and New SQL in detail?	Understand	The learner to **recall** SQL databases and **compares** with two important technologies No SQL and New SQL databases.	CO 5
PART - C (PROBLEM SOLVING AND CRITICAL THINKING QUESTIONS)				
1	Enumerate you are a senior faculty at a reputed institute. The HOD has asked you to make a list of the unstructured data that gets generated on the	Apply	The learner to **recall** and **understand** the different data types and types of Digital Data in	CO 5

	institution website that can then be stored and analyzed to improve the website to facilitate and enhance the student's learning.(features: pdf and doc files,forums,blogs,links,.xls sheet, txt files,. wav files, log files) **Identify** appropriate data type for each source of students learning resources.		depth and then **identify** the appropriate data type for each source.	
2	**Interpret** you have just got a book issued from the library. What are the details about the book that can be placed in an RDBMS table?	Understand	The learner to **recall** the concept of RDBMS data and **interpret** what kind of data we do analysis as result.	CO 3
3	**Explain** the process of Data Preparation in Big Data to enhance the better business decision value.	Understand	The learner to **recall** the preprocessing steps and explain the steps to enhance the insight and decision making.	CO 2
4	**Indicate** which V's are satisfied by real time big data case study – Amazon Click Stream with justification.	Apply	The learner to **recall** the definition of big data characteristics - 3V's, **relate** to real time case study.	CO 5
5	**Find** out the same visualization tool that we run over conventional data warehouse, be used in Big Data environment?	Understand	The learner to **recall** and **relate** the visualization tools for traditional BI and Big Data.	CO 2,CO 3
6	**Compare** the traditional analytics architecture and modern database architecture?	Understand	The learner to **recall** the different architectural components of data analytics and **compare** with database architecture.	CO 11
7	**Interpret** stock market predictions a case study, elaborate on the Real Time Analytics Platform (RTAP). Present the assumptions mode.	Apply	The learner to **recall** the predictive analytics concepts and **relate** to the given case study with appropriate assumptions.	CO 5
8	**Explain in detail about the following**: a) Multivariate analysis performed in Big Data. b) Methods of Stochastic search.	Understand	a) The learner to **recall** the different analysis approaches in big data and **explain** multivariate analysis. b) The learner to **recall** the different analysis approaches in big data and **explain** the stochastic search method	CO 11
9	**Identify** the outliers in the given data and write the different issues and challenges associated in data stream query processing.	Understand	The learner to **recall** the outlier identification techniques and **explain** different issues and challenges in query processing.	CO 11
10	**Identify** a cloud-based analytical tools for specific big data processing.	Apply	The learner to **recall** and **summarize** the cloud-based solutions available in market for handling big data. **Identify** cloud for Big Data Development is a good choice with available service providers for specific applications.	CO 11
UNIT – II				
INTRODUCTION TO HADOOP				
PART – A (SHORT ANSWER QUESTIONS)				
1	Memorize why Hadoop is called a Big Data technology? **How** it supports Big Data?	Remember	---	CO 1

2	**Identify** Big Data is encountered as a problem in the real time scenario?	Understand	The learner to **recall** adverse the term Big Data and **relate** the reasons for this phenomenon.	CO 2
3	**List** out various technologies came into an existence in processing Big Data?	Remember	---	CO 1
4	**Recall** the introduction of Hadoop over Big Data?	Remember	---	CO 1
5	**Summarize** the challenges of Big Data with Hadoop environment.	Understand	The learner to **recall** and **relate** the challenges of big data in Hadoop environment	CO 2,CO 3
6	Cite. **Why** the name "Hadoop" into the big ocean of Data streams?	Remember	---	CO 2,CO 3
7	Correlate the situation necessity for Hadoop arises and **why** do we need Hadoop.	Remember	---	CO 1
8	**Name** some of the characteristics of Hadoop framework?	Remember	---	CO 2
9	**Compare** the traditional RDBMS and Hadoop data bases?	Understand	The learner to **recall** and **compare** the differences between traditional DBMS system and Hadoop.	CO 2
10	**Recall** the basic requirements that are to be fulfilled with the structured and unstructured data?	Remember	---	CO 1
11	**Recall** the basic core components of Hadoop for analyzing the data.	Remember	---	CO 2
12	**Explain** in detail about the Hadoop Cluster?	Understand	The learner to **recall** and **explain** the Hadoop Cluster components.	CO 3
13	**List** the fundamentals concepts of distributed computing components in hadoop.	Remember	---	CO 4
14	**Name** the Distributed Computing challenges over Big Data in Hadoop?	Remember	---	CO 3
15	**What** are the various Hadoop Distributors for processing Big Data?	Remember	---	CO 4
16	**List** the various use cases of Hadoop?	Remember	---	CO 1
17	**Demonstrate** in detail about the journey of Doug Cutting with Hadoop?	Understand	The learner to **recall** the evolution of hadoop and **explain** the Doug Cutting journey in Hadoop from Lucere to yahoo.	CO 4
18	**Recall** the key distinctions of Hadoop?	Remember	---	CO 1
19	**Recall** the term Commodity hardware in Hadoop and how those are easily replaced?	Remember	---	CO 1
20	**Define** the philosophy of Big Data problem that resolves through Hadoop.	Remember	---	CO 1
PART - B (LONG ANSWER QUESTIONS)				
1	**Explain** Hadoop architecture and its components with diagram.	Understand	The learner to recall the components and explain Hadoop architecture clearly.	CO 4
2	**Summarize** the Hadoop Ecosystem role in different use cases.	Understand	The learner to **recall** the Different components in hadoop and **explain** the role in the solutions of specific use cases efficiently.	CO 4
3	**With** the help of Hadoop explain the processing of Big Data and challenges in distributed and parallel computing environment?	Understand	The learner to **recall** the hadoop ecosystem and **relate** each component to process the big data.	CO 4

4	**Explain** interacting process with Hadoop Ecosystem in terms of various big data processing technologies.	Understand	The learner to **recall** the big data technologies and **relate** with the identified interactions to process the data.	CO 4
5	**Find** out the 5 basic problems facing in Big Data and how to overcome the challenges in Hadoop through HDFS.	Remember	----	CO 4
6	**Illustrate** with neat diagram about Hadoop and its features?	Understand	The learner to **recall** the hadoop features and **show** with diagram.	CO 4
7	**Recall** the concept of Divide and conquer philosophy to enrich the jobs efficiently.	Remember	---	CO 1
8	**Demonstrate** Distributed processing is non-trivial.	Understand	The learner to **recall** and **explain** about distributed environment.	CO 4
9	**Find** out the Big Data storage as a challenge and find a solution to overcome through Hadoop system.	Remember	---	CO 1
10	**Demonstrate** in detail about the history of Hadoop with a neat sketch.	Understand	The learner to **recall** the hadoop evolutions and **show** in diagram.	CO 4
11	**How** Apache Hadoop Ecosystem technologies draw a distributed efficient responsibility.	Remember	----	CO 1
12	**Explain** in detail about the Animal planet for Hadoop and list out the reasons for the specific animal.	Understand	The learner to **recall** the hadoop versions and **explain** the history behind logos.	CO 4
13	**Recall** all the Apache Hadoop Ecosystem technologies to map each other with a neat sketch.	Remember	---	CO 1
14	**Recall** all the Big Data storage and processing elements and justify whether Hadoop tackles these challenges.	Remember	---	CO 1
15	**Extend** the core components of Hadoop with workflows in detail?	Understand	The learner to **recall** the core components and **explain** the components briefly.	CO 4
16	**Explain** the data locality optimization and heterogeneous cluster?	Understand	The learner to **recall** the hadoop features and **describe** the Hadoop cluster properly.	CO 4
17	**Explain** the key distinctions of Hadoop which are very flexible to handle the huge volume of data?	Understand	The learner to **recall** and **relate** the hadoop key distinctions to handle huge volume of the data.	CO 4
18	**Discuss** the overview of Hadoop distributors and its use cases in depth?	Understand	The learner to **recall** the hadoop distributors and **relate** to solve use cases.	CO 4
19	**Extend** the scalabilty and efficiency of Hadoop's global market in present scenario?	Understand	The learner to **recall** the hadoop features and **relate** to the present scenario in global markets.	CO 4
20	**Describe** the following terms: Data Science Hadoop Developer Hadoop Administrator Big Data Architect and Engineer	Understand	The learner to **define** the various terms of Big Data and **explain** each role in big data development.	CO 4
PART – C (PROBLEM SOLVING AND CRITICAL THINKING)				
1	**Describe** the concept of Distributed and parallel computing challenges with a neat diagram?	Understand	The learner to **recall** and **explain** the big data challenges in distributed environment.	CO 4
2	**Compare** between the Hadoop1.0 and Hadoop 2.0	Understand	The learner to recall the basic	CO 4

	architectures in detail?		features of different versions and compare at architecture level.	
3	**Identify** which technology is used to import or export the data from RDBMS to file systems?	Apply	The learner to **recall** and **relate** the different file systems, data import and export tools and **identify** the appropriate tool for translating from one to another file systems.	CO 5
4	**Explain** the four modules that make up the Apache Hadoop framework?	Understand	The learner to **recall** and **explain** the different modules to the specified framework.	CO 4
5	**Describe** the architecture of Hadoop technology and Justify how it satisfies the business insights now-a-days?	Understand	The learner to **recall** the components of Hadoop framework and its functionality of each and **discuss** how those are helpful for describing business insights.	CO 11
6	**Illustrate** “Big Data is a buzz word!” and list a few statistics to explore big data which is generated every day?	Understand	The learner to **recall** the various statistics related to bigdata creation and **show** in diagrammatic way.	CO 5
7	**Explain** the flow of data generated from the IoT devices towards Big Data through cloud computing services.	Understand	The learner to **recall** different data types and **explain** the streamed data process in importance of current technologies.	CO 4
8	Accommodate the 500GB of data file within a cluster of commodity hardware and how the hadoop can overcome the challenge of storing and processing the data?	Apply	The learner to **recall** and **explain** the Hadoop storage structure and block size constraints and **construct** a cluster for given data hardware requirement.	CO 5
9	**Explain** in detail about the Hadoop YARN with an example?	Understand	The learner to **recall** the Hadoop controlling components and **explain** the functionalities of YARN components with example.	CO 5
10	**Interpret** the integrated Hadoop systems offered by leading market vendors with Cloud-based Hadoop solutions.	Understand	The learner to **recall** and **explain** glimpse of the leading market vendors offering integrated Hadoop system.	CO 4
UNIT-III				
THE HADOOP DISTRIBUTED FILESYSTEM				
PART - A (SHORT ANSWER QUESTIONS)				
1	**List** the between the Linux file system and Hadoop distributed file system?	Remember	---	CO 4
2	**List** the Hadoop’s three configuration files?	Remember	---	CO 4
3	**List** the file formats that support Hadoop?	Remember	---	CO 1
4	**What** is the main purpose of HDFS fsck command and its usage/command?	Remember	---	CO 1
5	**Define** the term HDFS as a primary storage system of Hadoop?	Remember	---	CO 4
6	**Show** the default HDFS block size and default replication factor?	Remember	---	CO 1
7	**Explain** the HDFS error – “File could only be replicated to 0 nodes, instead of 1”?	Understand	The learner to **recall** the default replication factor and **relate** to explain the HDFS error.	CO 4

8	**What** are the key features of HDFS?	Remember	---	CO 1
9	**Recall** the terms Fault tolerance and streaming access?	Understand	The learner to relate the files stored in a system and what are the problems encountered.	CO 4
10	**Define** the term block in HDFS.	Remember	---	CO 1
11	**Choose** the block size and replication factor to configure HDFS?	Remember	---	CO 1
12	**What** are the benefits of block transfer?	Remember	---	CO 6
13	**Recall** the term daemon and mention the 5 daemons in the Hadoop cluster?	Understand	The learner to **recall** the daemon and **list** out the various daemons in the cluster.	CO 6
14	**Define** various modes of Hadoop?	Remember	---	CO 6
15	**Illustrate** the client communication with HDFS?	Understand	The learner to **recall** the workflow concept and **show** the communications with the Hadoop cluster.	CO 6
16	**Explain** about the file permissions and data integrity in HDFS?	Understand	The learner to **recall** the file permissions of HDFS.	CO 6
17	**What** mechanism does Hadoop framework provide to synchronize changes made in Distribution Cache during runtime of the application?	Remember	---	CO 1
18	**Suppose** Hadoop spawned 100 tasks for a job and one of the tasks failed. What will Hadoop do?	Apply	The learner to **recall** the Hadoop features and **relate** to the replication feature and **apply** on task monitoring.	CO 7
19	**What is** an Input Split and HDFS block?	Remember	---	CO 6
20	**What** is the difference between MapReduce engine and HDFS cluster? What is “Key-Value pair” in HDFS?	Remember	---	CO 1
PART – B (LONG ANSWER QUESTIONS)				
1	**Explain** in brief bout the Hadoop’s rack topology with the following terms: • Rack Awareness • Fault Tolerance	Understand	The learner to **recall** the concept of rack and **show** the topology with the collection of multiple servers based on the requirements.	CO 6
2	**Outline** the different ways to overwrite the replication factors in HDFS?	Understand	The learner should **recall** the Hadoop file system commands and **relate** to file writing.	CO 6
3	**Explain** the importance of Input Format and Record Reader in Hadoop? What are the various Input Formats in Hadoop?	Understand	The learner to **recall** the input formats and relate the record readers and different formats of Hadoop.	CO 6
4	**Discuss** the HDFS Architecture and HDFS Commands in brief. Write down the goals of HDFS.	Understand	The learner to **define** and **discuss** the architecture of HDFS.	CO 6
5	**How** does HDFS ensure data Integrity in a Hadoop Cluster?	Understand	The learner to **recall** and **explain** the data integrity in Hadoop.	CO 6
6	**Discuss** racks in Hadoop Cluster? Explain how Hadoop Clusters are arranged in several racks with a real time example?	Understand	The learner to **define** the concept of racks in a cluster and **explain** the cluster arrangement.	CO 6
7	Create a file in HDFS. **Explain** the Anatomy of a	Understand	The learner to **recall** the anatomy	CO 6

	File Read and Write?		of file read and write and **explain** the workflow in creating file.	
8	**Explain** the following terms in detail: Name Node Secondary Name Node Data Node Job Tracker Task Tracker	Understand	The learner to **recall** and **explain** the important nodes in the Hadoop cluster.	CO 6
9	**Demonstrate** the Streaming access pattern of HDFS Hadoop Cluster?	Understand	The learner to **name** and **explain** the different access patterns and parameters.	CO 6
10	**Differentiate** between the basic File System and HDFS?	Understand	The learner to **recall** the basic file system and **explain** with other file systems.	CO 6
11	**Explain** in detail about Hadoop Cluster and the Master – Slave architecture?	Understand	The learner to **define** the important nodes in the cluster. **Explain** each	CO 6
12	**Describe** in detail about the two types of “writes” n HDFS?	Understand	The learner to **recall** the concept of HDFS and **explain** the anatomy of file read /write?	CO 6
13	**Which** modes can Hadoop be run in? List out the few features for each mode.	Understand	The learner to **recall** the different modes of Hadoop and **explain** feature of each	CO 6
14	The default block size is 64MB and the replication factor is 3.**Calculate** no. of blocks allocated for a file having the size of 300MB?	Apply	The learner to **recall** the block and replication concepts in Hadoop and **explain** the block allocation process. **Apply** on the given file.	CO 7
15	**What** is Name Node and Data Node? Explain how many Name Nodes and Data Nodes can run on a single Hadoop cluster?	Remember	---	CO 1
16	**Define** metadata and commodity hardware? Does commodity hardware include RAM? Is Name Node also commodity?	Understand	The learner to **recall** the basic terminologies of Hadoop cluster and **explain** metadata and commodity hardware.	CO 6
17	**Explain** how the NameNode gets to know all the available data node in the Hadoop cluster?	Understand	The learner to **recall** the various nodes in hadoop and **explain** all the available nodes in the cluster.	CO 6
18	**Briefly** explain HDFS Name Node Federation, NFS Gateway, Snapshots, Checkpoint and Backups.	Understand	The learner to **define** and **explain** the various terms of HDFS.	CO 6
19	**Bring** out the concepts of HDFS block replication, with an example?	Understand	The learner to **recall** the Hadoop Cluster block replications and **explain** with example.	CO 6
20	**Illustrate** for each YARN job, the Hadoop framework generates a task log file, where are Hadoop task log files stored?	Understand	The learner to **recall** and **determine** the container logs to be stored in the nodes.	CO 6
	PART – C (PROBLEM SOLVING AND CRITICAL THINKING)			
1	**Demonstrate** as per the configuration, HDFS is in high availability mode with automatic failover. Explain in brief about the daemon which will take care of the failover.	Understand	The learner to **recall** the automatic failover maintenance and explain configuration details.	CO 6
2	**Compare** the setup of YARN cluster where the application memory available is 30GB with two companies Wipro and TCS. Wipro queue has 15GB allocated and TCS queue has 5GB allocated. Each map task requires 25 GB allocation. How	Understand	The learner to **know** the resource allocation within the queues is controlled separately. **Compare** with different schedules.	CO 6

	does the fair scheduler assign the available memory resources under the Dominant Resource Fairness (DRF) scheduler?			
3	**Illustrate** the usual block size on an HDFS? Can we make it much larger say 1GB and what are the advantages that a block provides over a file system?	Apply	The learner to **recall** the concept of block size in HDFS and **explain** the limitations with the size variations. **Illustrate** the advantages over distributed file system.	CO 7
4	**Demonstrate** what do you mean by High Availability in HDFS? What are failover and fencing, and what role do they play in making the system highly available.	Understand	The learner to **recall** the single point of failure without any manual intervention and **demonstrate** its features in the available system.	CO 6
5	**Examine** the number of spilled records from map tasks far exceeds the number of map output records. The child heap size is 1GB and your io.sort.mb value is set to 1000MB. How would you tune your io.sort? MB value to achieve maximum memory to I/0 ratio.	Apply	The learner to **recall** the total amount of memory size in a buffer and **explain** the concept of heaps. **Apply** to tune io.sort. to maximize the heap size in memory while sorting files.	CO 7
6	**What** are the components and characteristics of HDFS?	Remember	---	CO 1
7	**Illustrate** the anatomy of File Read in HDFS with a neat sketch and elaborate the workflow from client to Hadoop framework.	Understand	The learner to **recall** the architecture of Hadoop and **illustrate** the workflow of Hadoop and client communication.	CO 11
8	**Illustrate** the anatomy of File Write in HDFS with a neat sketch and elaborate the File Writes processing methodology in the distributed file system.	Understand	The learner to **recall** the architecture of Hadoop and **Illustrate** the workflow of Hadoop and client communication for file writing.	CO 11
9	**Identify** the mechanism of heartbeat in HDFS and justify the Name Node handles data nodes failures?	Apply	The learner to **recall** the concept of heartbeat in the cluster and **explain** the data storage nodes in the framework. **Identify** the heartbeat signals to the data nodes in a regular time stamp.	CO 7
10	Examine if we want to copy 10 blocks from one machine to another, but another machine can copy only 8.5 blocks, can the blocks be broken at the time of replication?	Apply	The learner to **recall** the concept of block replication and **explain** the fail over management and then **apply** the given blacks to the framework.	CO 7
UNIT-IV				
UNDERSTANDING MAPREDUCE FUNDAMENTALS				
PART – A (SHORT ANSWER QUESTIONS)				
1	**Recall** the term MapReduce? Explain about life cycle of MapReduce?	Remember	---	CO 1
2	**Visualize** the terms Map Phase &Reducer Phase and Differentiate the measures in Sort and shuffle?	Remember	---	CO 1
3	**Show** the differences between Block & Input split?	Remember	---	CO 9
4	**Tabulate** what are the main classes of MR Job?	Remember	---	CO 1

5	**What** are the basic parameters of a mapper and reducer?	Remember	---	CO 7
6	**Summarize** the naming conventions for output files from Map phase and Reduce Phase?	Understand	The learner to **recall** the MapReduce phases and **explain** the naming conventions in different phases	CO 6
7	**Recall** the terms identity Mapper and Reducer and state its computation?	Remember	---	CO 1
8	**Illustrate** in detail is it mandatory to set input and output type/format in MapReduce?	Understand	The learner to **recall** the Map and Reduce jobs and **infer** the input and output formats.	CO 6,CO 7
9	What do you understand by TextInputFormat, KeyValueTextInputFormat and NLineOutputFormat?	Remember	---	CO 6
10	**What is** RecordReader in MapReduce	Remember	---	CO 6
11	**Describe** the term Combiner?	Remember	---	CO 6,CO 7
12	**Define** the Null Writable and how is it special from other Writable data types?	Remember	---	CO 1
13	**Describe** about the Mapper Output (intermediate key-value data) stored?	Remember	---	CO 1
14	What does a MapReduce partitioner do?	Remember	---	CO 6
15	**Generalize** the use of Context object?	Understand	The learner to recall the concept of containers and name the use cases of Context object.	CO 6
16	**What is role of** distributed Cache in MapReduce Framework?	Remember	---	CO 6
17	**Define** Custom Writable? What is a Writable in Hadoop?	Remember	---	CO 1
18	**Recall** about Data Locality in MapReduce?	Remember	---	CO 1
19	**Explain** in what scenario can the container be killed by the node manager?	Understand	The learner to **recall** the concept of container and **explain** the node manager responsibility.	CO 6
20	**Express** how does a map task partition the output in the case of multiple reducers?	Understand	The learner to **recall** the partitions and **explain** map tasks in the concept of reducers.	CO 6
PART – B (LONG ANSWER QUESTIONS)				
1	**Explain** Map-reduce framework in brief and Draw the architectural diagram for Physical Organization of Compute Nodes.	Understand	The learner to **recall** the architecture of cluster and **explain** the framework of MapReduce.	CO 6
2	**Infer** out the main features of MapReduce and its significance?	Understand	The learner to **recall** the concepts of MapReduce and **explain** the features and its significance.	CO 6
3	**Describe** the working of the MapReduce algorithm?	Understand	The learner to **recall** and **relate** the working principles of MapReduce.	CO 6
4	**Explain** working of following phases of MapReduce with one common example. (i) Map Phase (ii) Combiner Phase (iii) Shuffle and Sort Phase (iv)Reducer Phase.	Understand	The learner to **recall** all the definitions of MapReduce and **explain** in detail.	CO 6
5	**Estimate** the entire process of data analysis conducted in the MapReduce programming model?	Understand	The learner to know the process of analytics and understand the programming model.	CO 6

6	**Explain** the description of MapReduce process for a specific case?	Understand	The learner to **recall** and **explain** the process for analyzing and understand in specific case.	CO 6
7	**Describe** the uses of MapReduce? Define what conditions must be met to implement MapReduce application?	Understand	The learner to **recall** and **explain** the uses and conditions in MR jobs.	CO 6
8	**Extend** the MapReduce be used to solve any kind of computational problems? if not, explain the cases where MapReduce is not applicable?	Understand	The learner to **recall** the MapReduce framework and **solve** the computational problems.	CO 6
9	**Discuss** some techniques to optimize MapReduce jobs and the points you need to consider while designing a file system in MapReduce?	Understand	The learner to **define** the optimized techniques and **explain** the designing of a file.	CO 6
10	**Illustrate** a short note on Input Split and Explain the MapReduce application?	Understand	The learner to **recall** Input Split concepts and **relate to** the applications of MapReduce.	CO 6
11	**Classify** a short note on Input Format and the File Input Format class?	Understand	The learner to **recall** the input File formats and **demonstrate** different types for specific needs.	CO 11
12	**Explain** the anatomy of a map-reduce job run?	Understand	The learner to **recall** and **explain** a clear assumptions of data transformations.	CO 6
13	**Illustrate** with diagram about how Hadoop uses HDFS staging directory as well as local directory during a job run?	Understand	The learner to **recall** the concept of HDFS and **state** the directories in an MR jobs.	CO 6
14	**Demonstrate** the map side join by comparing with a reduced side join in MapReduce programming?	Understand	The learner to **define** the map side join and **explain** in detail while comparing with reduce side join in fulfilling a job.	CO 11
15	**Explain** in detail about the few interesting facts about MapReduce.	Understand	The learner to **recall** and **relate** the basic facts MapReduce and understand the applications.	CO 6
16	**Illustrate** how MapReduce Engine Works in a step by step procedure?	Understand	The learner to **recall** the MR framework and **relates** the step by step procedure of MapReduce.	CO 6
17	**Explain** how MapReduce Works on Parallel Programming Concept?	Understand	The learner to **recall** the Parallel Programming and **explain** MR phases.	CO 6
18	**Describe** in detail about the Driver class, map and reducer phases with a real time example?	Understand	The learner to **recall** the 3 classes and **relate** them to the phases in the MapReduce.	CO 6
19	**Interpret** the Data Locality Optimization in MR jobs?	Apply	The learner to **recall** and explain Data Locality Optimization features and **apply** them in the cluster.	CO 7
20	**Discuss** the workflow in a basic word count MapReduce program to understand MapReduce Paradigm.	Understand	The learner to **recall** the MR framework and **explain** steps to implement the MapReduce job.	CO 6
PART – C (PROBLEM SOLVING AND CRITICAL THINKING)				
1	**Discuss** briefly about the job or application ID. How job history server is handling the job details and brief about logging and log files.	Understand	The learner to **define** derivative and **explain** the formula on log files.	CO 6
2	**Explain** the role of a combiner and partitioner in a Map-Reduce job? Is the combiner triggered first or the partitioner?	Understand	The learner to **recall** the different derivatives and **describe** its jobs.	CO 6

3	**Discuss** MapReduce runs on top of yarn and utilizes YARN containers to schedule and execute its map and reduce tasks. When configuring MapReduce resource utilization on YARN, what are the aspects to be considered?	Understand	The learner to **define** and **explain** how much maximum memory each map and reduce task will take.	CO 6
4	**Examine** every hour Hadoop runs 100 jobs in parallel. Now currently, single job is running. How much of the resource capacity of the cluster will be used by this running single job?	Apply	The learner to **recall** the Hadoop cluster and **relate** the scheduler when the single application is running may request entire cluster. **Identify** the resource capacity for running single job.	CO 7
5	**Construct** the MapReduce job, under what scenario does a combiner get triggered? What are the various options to reduce the shuffling of data in a map – reduce job?	Apply	The learner to **recall** and **relate** the concept of MapReduce jobs and options for the MapReduce jobs in minimizing. **Identify** the optimal scenario.	CO 11
6	**Examine** MapReduce job you consistently see that map tasks on your cluster are running slowly because of excessive garbage collection of JVM. How do you increase JVM heap size property to 3GB to optimize performance?	Apply	The learner to **recall** and **relate** the MapReduce jobs and **make** consistently see that MapReduce map tasks on your cluster.	CO 7
7	**Explain** the concept of joins in MR jobs? Compare the various join processing methods?	Understand	The learner to **recall** the concept of joins in MapReduce jobs and **explain** the mapper and reducer methods.	CO 11
8	**Summarize** the reason why we can't perform "aggregation" (addition) in mapper? Why do we need the "reducer" for this?	Understand	The learner to **recall** the basic idea of mapper and reducer and **explain** the analytical process.	CO 6
9	**Write** a MapReduce program that mines weather data. Hint: Weather sensors collecting data every hour at many locations across the globe gather a large volume of log data, which is a good candidate for analysis with MapReduce, since it is semi structured and record oriented.	Apply	The learner to **recall** the basic constructs of MapReduce program and **Summarize** the data and **identify** suitable map and reduce functions to perform analysis.	CO 7
10	**Make use of** Hadoop MapReduce functions for implementing matrix multiplication.	Apply	The learner to **recall** the MapReduce programming functions and **explain** the applicability and develop map and reduce functions to perform matrix multiplication.	CO 11

UNIT-V

INTRODUCTION TO PIG and HIVE

PART – A (SHORT ANSWER QUESTIONS)

1	List the advantages and uses of PIG.	Remember	---	CO 1
2	**List** out the features of PIG and different modes of execution in PIG.	Remember	---	CO 1
3	**What is** the need of MapReduce during PIG programming?	Remember	---	CO 9
4	**Why** should we use 'distinct' keyword in PIG scripts?	Remember	---	CO 9
5	**What** is the importance of PIG use cases?	Remember	---	CO 1

6	**List** the custom Data types in PIG and define briefly.	Remember	---	CO 1
7	**Describe** the term inner bag and PIG in embedded mode.	Understand	The learner to **recall** the scripts use in PIG and **explain** them in embedded mode.	CO 9
8	**Illustrate** the co-group representations in PIG?	Understand	The learner to **know** the groups concept and **show** the elements in the field.	CO 9
9	**Discuss** the keyword 'DEFINE' like a function name?	Understand	The learner to **recall** the functions and **explain** the parameters of the function.	CO 9
10	**Discuss** the keyword 'FUNCTIONAL' a User Defined Function (UDF)?	Understand	The learner to **recall** the user's perspectives and **explain** the keywords in UDF.	CO 9
11	**Illustrate** PIG Latin language is case-sensitive or not? What does FOREACH do?	Understand	The learner to **recall** the semantics of PIG Latin and **infer** for all applications.	CO 9
12	**List** out the relational operations in PIG Latin?	Remember	---	CO 1
13	**Recall** the importance of partioning and bucketing in Hive? .	Remember	---	CO 1
14	**Illustrate** the OrderBy and SortBy with an example in Hive?	Understand	The learner to **recall** the functions in hive and **relate** the orders in examples.	CO 9
15	**Explain** the different kinds of tables in Hive?	Understand	The learner to **recall** the concept of tables in hive and **explain** the tables.	CO 9
16	**How** to create external table in hive?	Remember	---	CO 1
17	In Hive, explain the term 'aggregation' and its uses?	Understand	The learner to **recall** the concept of DDL and **explain** the aggregation and its uses.	CO 9
18	**Interpret** joins with an example?	Understand	The learner to recall the joins in hive and explain with example.	CO 9
19	**List** out the Data types in Hive?	Remember	---	CO 9
20	**List** out the Hive services with a neat sketch?	Remember	---	CO 1
PART - B (LONG ANSWER QUESTIONS)				
1	**Explain** briefly the difference between MapReduce and PIG?	Understand	The learner to **recall** the concept of PIG and **relate** the parameters and functions of both frame works.	CO 9
2	**Discuss** and explain PIG structure and architecture in brief?	Understand	The learner to **define** and **discuss** the automatic optimizations.	CO 9
3	**Compare** logical and physical plans in Pig Latin?	Understand	The learner to **recall** the plans and **compare** logical and physical plans in Pig Latin.	CO 9
4	**Compare** PIG and SQL for query optimization and significance?	Understand	The learner to **recall** the query optimization concepts and **compare** PIG and SQL in query optimization and significance.	CO 10
5	**Outline** the conditions and Data Types in PIG?	Understand	The learner to **recall** the list of data types and **explain** the conditions.	CO 9

6	**Explain** Pig features for allowing grouping on expressions?	Understand	The learner to **recall** and **relate** the concepts of grouping and the pig expressions.	CO 8
7	**Describe** in detail about the scalar and complex data types in PIG?	Understand	The learner to **recall** the data types in specific ways and **explain** the data types in PIG.	CO 9
8	**Explain** multi query execution in PIG and its operations?	Understand	The learner to **recall** the all the operations and functions and **explain** the operations.	CO 8
9	**Describe** the Functions that can be used in PIG and PIG latin Schemas?	Understand	The learner to **recall** and **relate** the functions and schemas used in PIG and PIG Latin.	CO 9
10	**Explain** the UDF functions used in PIG with its description?	Understand	The learner to **recall** the functions in PIG and **explain** the UDF functions and its descriptions.	CO 8
11	**Explain** in brief the architecture of Hive?	Understand	The learner to **recall** the architecture of Hive and explain its components.	CO 9
12	**Discuss** the various Hive services with an example?	Understand	The learner to **recall** the Hive services and **explain** with examples.	CO 9
13	**Describe** the various Hive Data types?	Understand	The learner to **recall** the data types in HIVE and **explain** each in detail.	CO 9
14	**Explain** the Built-in Functions in Hive?	Understand	The learner to **recall** the basic built in functions and **explain** with example.	CO 9
15	**Discuss** the user defined functions in hive?	Understand	The learner to **recall** and **relate** the user parameters and understand its functions.	CO 9
16	**Explain** about Collection data types in hive?	Understand	The learner to **recall** the hive data types and **explain** collection data types in detail.	CO 9
17	**Compare** HIVE and PIG in detail?	Understand	The learner to **recall** the concepts of PIG and HIVE and compare in detail.	CO 9
18	**Explain** the procedure to load data in manage tables?	Understand	The learner to recall the tables and **explain** the data transformations clearly.	CO 9
19	**Explain** architecture of Apache Hive and various data insertion techniques in Hive with example.	Understand	The learner to recall the Apache hive architecture and outline the various data insertion techniques.	CO 9
20	**Describe** Hive SQL Data Definition Language.	Understand	The learner to recall the DDL concepts and explain queries in HIVE SQL DDL.	CO 9
PART – C (PROBLEM SOLVING AND CRITICAL THINKING)				
1	On what scenarios MapReduce jobs will be more useful than PIG. Categorize the problems which can only be solved by MapReduce and cannot be solved by PIG?	Analyze	The learner to **recall** the concept of PIG Latin and MapReduce, **relate** different scenarios and **categorize** each problem based on Hadoop appropriate component to solve.	CO 11
2	I already register my LoadFunc / StoreFunc jars in "register" statement, but why I still get "Class Not Found" exception? Explain the situation briefly.	Understand	The learner to **recall** the functions of PIG and **relate** to the given situation.	CO 10
3	A file employee.txt in the HDFS directory with	Apply	The learner to **recall** and **relate** the	CO 11

	100 records. To see only the first 10 records from the employee.txt file. **Illustrate** the results with appropriate command.		PIG commands and **apply** on the given file to get the result data.	
4	**Solve** a statistical problem by calculating percentage (partial aggregate / total aggregate) in PIG?	Apply	The learner to **recall** the concepts of PIG and **explain** the aggregations and solve a given proven problem.	CO 8
5	**Illustrate** different types of joins in Pig Latin with examples on different data types.	Apply	The Learner to **recall** the types of joins and **relate** the PIG Latin joins and **apply** on different data types.	CO 11
6	**Discuss** the Hive commands to create a table with four columns: First name, last name, age, and income?	Understand	The Learner to **recall** the commands of Hive and extend with creation of table.	CO 9
7	A start-up company wants to use Hive for storing its data. **Discuss** a shell command in Hive to list all the files in the current directory?	Understand	The Learner to **recall** the Hive concepts in detail and **explain** the storage capabilties.	CO 9
8	**Explain** a shell command in Hive to list all the files in the current directory?	Understand	The Learner to **recall** the commands in hive and **relate** to find the list of files.	CO 9
9	**Develop a** PIG Latin program for an application of word count in a given file.	Analyze	The Learner to **recall** and **relate** the programming in PIG and develop an application for word count.	CO 11
10	**Describe** the importance of partitions in Hive with an example?	Understand	The Learner to **recall** the concept of partitions and **explain** the importance of partitions.	CO 9

Prepared by:
Dr. M Madhu Bala, Professor

HOD, CSE