DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
QUESTION BANK

UNIT I			
PART-A (2 - MARKS)			
Q. No	**QUESTIONS**	**BT Level**	**Competence**
1.	**Write** the definition of "big data" and under what conditions it is given that name.	Remember	BTL-1
2.	**Demonstrate** the differences between Big data and conventional data.	Apply	BTL-3
3.	**Examine** the various dimensions of growth of big data.	Apply	BTL-3
4.	**Differentiate** between data analysis and data reporting.	Understand	BTL-2
5.	**List** the risks involved in using big data.	Remember	BTL-1
6.	**Explain** the role of big data analytics.	Evaluate	BTL-5
7.	**Identify** the sources of big data.	Remember	BTL-1
8.	**Analyze** the challenges in big data.	Analyze	BTL-4
9.	**Summarize** the reasons for the domain expertise for any type of data analytics.	Evaluate	BTL-5
10.	**Define** the reason behind the phrase "Web data is the most popular big data" .	Remember	BTL-1
11.	**Why** the accuracy in big data is beneficial ?	Remember	BTL-1
12.	**Analyze** the list of data analytical tools.	Analyze	BTL-4

13.	**Predict** about the list of reporting tools.	Understand	BTL-2
14.	**Discuss** about the trends in data analytics tools.	Understand	BTL-2
15.	**Generalize** the role of analytical tools in big data.	Create	BTL-6
16.	**Show** the points of similarities between the data mining and data analysis.	Apply	BTL-3
17.	**Differentiate** between data analytics and big data analytics**.**	Analyze	BTL-4
18.	**Create** a scenario where big Data Analytics be used as a Decision Making tool.	Create	BTL-6
19.	**Summarize** the technologies used to handle big data.	Understand	BTL-2
20.	**State** the definition of "Sand Box".	Remember	BTL-1
	PART-B (13- MARKS)		
1.	**Define** Big Data . Describe the main features of big data in detail. (13)	Remember	BTL-1
2.	(i) **Explain** the main characteristics features and structure of Big data in detail. (4) (ii) **Analyze** big data architecture with a neat schematic diagram. (9)	Analyze	BTL-4
3.	**Describe** the risks involved in handling Big data. (13)	Remember	BTL-1
4.	(i) **Evaluate** ways in which the big data is represented ? (5) (ii) **Assess** the structure of big data representation (8)	Evaluate	BTL-5
5.	(i) **What** are the features of Massive parallel processing system? (5) (ii) **Describe** the use of Massive Parallel Processing system in big data analytics. (7)	Understand	BTL-1
6.	**Discuss** the challenges faced by the traditional system. (13)	Analyze	BTL-4
7.	**Point out** in detail the analysis tools and reporting tools used in Bigdata. (13)	Analyze	BTL-4
8.	**Discuss** in detail about Analytical data set and the types of analytical data set. (13)	Understand	BTL-2
9.	**Discuss** in detail about web data and what does it reveal? (13)	Understand	BTL-2

10.	**Illustrate** in detail how big data are effectively filtered and mixed with the traditional one. (13)	Apply	BTL-3
11.	(i) **Describe** the importance of tools in Big Data. (4) (ii) **Summarize** in detail the trends and technology in big data. (9)	Understand	BTL-2
12.	**Formulate** the difficulties faced by the traditional systems. (13)	Create	BTL-6
13.	**Describe** how the analytical scalability is handled in big data. (13)	Remember	BTL-1
14.	(i) **Explain** in detail about the web data in current action today. (6) (ii) **Examine** the modern tools for big data analytics. (7)	Apply	BTL-3
PART-C (15- MARK)			
1.	**Summarize** in detail about the challenges of the Big Data in Modern Data Analytics. (15)	Evaluate	BTL-5
2.	**Hypothesize** the statement **"**Web Data is the Most Popular Big Data" with reference to data analytic professional." (15)	Create	BTL-6
3.	**Infer** on the statement "Is the "Big" Part or the "Data" Part More Important ". (15)	Analyze	BTL-4
4.	**Formulate** the role of analytic sandbox**,** its benefits and types. (15)	Create	BTL-6
UNIT II			
PART-A (2 - MARKS)			
Q.No	**QUESTIONS**	**BT Level**	**Competence**
1.	**Give** the definition of Hadoop.	Understand	BTL-2
2.	**Define** how Map-Reduce computation is executed .	Remember	BTL-1
3.	**Show** the key advantages in Hadoop.	Apply	BTL-3
4.	**Point out** the meaning of the term " Hadoop YARN".	Analyze	BTL-4

5.	**List** the core concepts of Hadoop.	Remember	BTL-1
6.	**Define** MAP REDUCE concepts.	Remember	BTL-1
7.	How can a key value pair is formed?	Apply	BTL-3
8.	**Develop** the importance of DFS.	Create	BTL-6
9.	**Differentiate** between Hadoop and Map Reduce.	Understand	BTL-2
10.	**Point out** the characteristics of Hadoop.	Analyze	BTL-4
11.	**Distinguish** between Hadoop and Big data.	Understand	BTL-2
12.	**List** the advantages of MaPR.	Remember	BTL-1
13.	**Classify** the classical components of computer.	Analyze	BTL-4
14.	**Express** Shuffle and sort Algorithm.	Understand	BTL-2
15.	**Explain** the goals of HDFS.	Evaluate	BTL-5
16.	**What** are the list of Hadoop applications ?	Remember	BTL-1
17.	**Classify** types of big data.	Apply	BTL-3
18.	**Define** the partitions are shuffled in map reduce.	Remember	BTL-1
19.	**Explain** the steps in map reduce algorithm.	Evaluate	BTL-5
20.	**Generalize** matrix vector multiplication.	Create	BTL-6
	PART-B (13- MARK)		
1.	(i) **Discuss** the features of Hadoop and explain the functionalities of Hadoop cluster? (7) (ii) **Give** the briefly about Hadoop input and output and write a note on data integrity? (7)	Understand	BTL-2
2.	**Express** the various core components of the Hadoop. (13)	Understand	BTL-2
3.	(i) **List** and **explain** significance of Map Reduce. (7) (ii) **List** about Hadoop distributed file system architecture with neat diagram. (6)	Remember	BTL-1
4.	(i) **Demonstrate** the algorithms using Map Reduce. (7) (ii) **Show** the Extensions to Map Reduce. (6)	Apply	BTL-3
5.	(i)**Explain** in detail about Hadoop Distributed File Systems. (7) (ii)**Describe** Hadoop and HadoopR. (6)	Remember	BTL-1
6.	**Analyze** the steps of Map reduce Algorithms. (13)	Analyze	BTL-4
7.	(i) **Point out** the concepts of HDFS. (2)	Remember	BTL-1

	(ii) **State** Map and Reduce Algorithm in detail. (6) (iii) **Elaborate** the map reduce algorithm with an example. (5)		
8.	**Assess** Map Reduce Algorithm in detail. (13)	Evaluate	BTL-5
9.	(i)**What** is the role of combiner and partitioner in a map reduce application? (8) (ii)**List** the algorithms used in Map Reduce. (5)	Create	BTL-6
10.	**List** the details of reducer size and replication rate. (13)	Remember	BTL-1
11.	(i) **Classify** Big Data. (7) (ii) **Demonstrate** on the importance of using HDFS. (6)	Apply	BTL-3
12.	(i) **Discuss** about Matrix Vector Multiplication. (7) (ii) **Describe** about the matrix Multiplication. (6)	Understand	BTL-2
13.	Explain in detail about HADOOP YARN. (13)	Analyze	BTL-4
14.	Consider a collection of literature survey made by a researcher in the form of a text document with respect to cloud and big data analytics. **Analyze** Using Hadoop and Map Reduce, write a program to count the occurrence of pre dominant key words. (13)	Analyze	BTL-4
PART-C (15 -MARKS)			
1.	**Evaluate** a procedure to find the number of occurrence of a word in a document. (15)	Evaluate	BTL-5
2.	(i)**Compile** with a neat sketch about processing of a job in Hadoop. (8) (ii)**Formulate** the various operational modes of Hadoop cluster configuration and explain in detail about configuring/installing the Hadoop in local/standalone mode. (7)	Create	BTL-6
3.	**Analyze** how google file system differs from the Hadoop file system and explains the google file system architecture with a neat sketch. (15)	Analyze	BTL-4
4.	**Create** a Map-Reduce Algorithm to get the Dot Product of two Large Vectors. Assuming Only non-zero elements of those vectors are given in input files and output file should show only non-zero entries(assuming two vectors are same size). (15) ex: v1=[5 4 0 1 2]	Create	BTL-6

	v2=[4 2 1 0 6] file1: file2: output: (0,5) (0,4) (0,20) (1,4) (1,2) (1,8) (3,1) (2,1) (4,12) (4,2) (4,6)		
UNIT III			
PART-A (2 - MARKS)			
1.	**Express** data analysis.	Understand	BTL-2
2.	**Discuss** how would do classification.	Understand	BTL-2
3.	**Define** support-vector machines.	Remember	BTL-1
4.	**What** do you mean by Regression ?	Remember	BTL-1
5.	**Give** the different types of regression.	Understand	BTL-2
6.	**Differentiate** regression and correlation.	Understand	BTL-2
7.	**What** are the different types of data analysis ?	Remember	BTL-1
8.	**Summarize** the importance of classification in data analysis.	Evaluate	BTL-5
9.	**Examine** about Kernel.	Remember	BTL-1
10.	**Write** short note on Multiple Linear Regression.	Create	BTL-6
11.	**Point** out rule mining.	Analyze	BTL-4
12.	**Illustrate** multivariate analysis.	Apply	BTL-3
13.	**Evaluate** the concept of data analysis.	Evaluate	BTL-5
14.	**What** are the different types of kernel?	Remember	BTL-1
15.	**Discover** the importance of clustering along with its different types.	Apply	BTL-3
16.	**Explain** the difference between data analytics and data science.	Analyze	BTL-4
17.	**Name** the importance of rule induction.	Remember	BTL-1
18.	**Illustrate** the Factors relating classification and clustering.	Apply	BTL-3
19.	**Analyze** importance of clustering in data analytics.	Analyze	BTL-4
20.	**Formulate** what is R.	Create	BTL-6
PART-B (13- MARKS)			

1.	(i) **Discuss** about regression. List the types of regression. (7) (ii) **List** the purpose of using Regression Modelling in Data Analysis. (6)	Remember	BTL-1
2.	**Explain** in detail about various Multivariate Analysis and in what conditions do we use it. (13)	Evaluate	BTL-5
3.	**Analyze** the role of SVM in detail and try to provide some of its applications in detail. (13)	Analyze	BTL-4
4.	**Illustrate** in detail about Kernel Methods. (13)	Apply	BTL-3
5.	(i) **Give** a short note of Data Analysis . (7) (ii) **Discuss** the importance of data analysis. (6)	Understand	BTL-2
6.	(i) **List** the important points with regards to rule mining. (7) (ii)**Describe** in detail about association rule mining. (6)	Remember	BTL-1
7.	(i) **Discuss** about Clustering in data analysis. (7) (ii) **Discuss** about density based and Grid based clustering in detail. (6)	Remember	BTL-2
8.	**Describe** how clustering is used in real time applications. (13)	Understand	BTL-2
9.	**Examine** the function cost of exact counts and what approaches are used to estimate the moments. (13)	Apply	BTL-3
10.	**Explain** model based clustering and high dimensional clustering in detail. (13)	Analyze	BTL-4
11.	**Describe** partitioning method of clustering in detail. (13)	Remember	BTL-1
12.	(i) **Generalize** k-means clustering algorithm with an example. (7) (ii) **Develop** different hierarchical clustering techniques and explain any one. (6)	Create	BTL-6
13.	(i) **Analyze** the meaning of the term “prediction”. (7) (ii) **Explain** in detail about Predictive Analysis. (6)	Analyze	BTL-4
14.	(i)**What** is meant by data analysis using R? (7) (ii)**Discuss** the concept of R Programming. (6)	Remember	BTL-1
PART-C(15 -MARKS)			
1.	**Evaluate** the statement in detail : “Data Analysis is not a decision-making system, but a decision supporting system”. (15)	Evaluate	BTL-5

2.	Create a Regression Model for " happy people get many hours of sleep" using your own data and what kind of inferences it provides. (15)	Create	BTL-6
3.	Summarize hierarchical clustering in detail. (15)	Analyze	BTL-4
4.	Compose the K-means partitioning algorithm using the given data. Consider five points { X1, X2,X3, X4, X5} with the following coordinates as a two dimensional sample for clustering: X1 = (0,2.5); X2 = (0,0); X3= (1.5,0); X4 = (5,0); X5 = (5,2) (15)	Create	BTL-6
UNIT IV			
PART-A (2 -MARKS)			
1.	**List** the main characteristics of stream sources.	Remember	BTL-1
2.	**Analyze** the term "Data Stream".	Analyze	BTL-4
3.	**Differentiate** between RTAP and RTSA.	Understand	BTL-2
4.	**Identify** the examples related to "Data Stream".	Remember	BTL-1
5.	**Show** the sampling data in a stream.	Remember	BTL-1
6.	**Measure** out the few challenges of data stream mining algorithms.	Evaluate	BTL-5
7.	**What** are the differences between DBMS and DSMS?	Remember	BTL-1
8.	**Point out** the statement "Filtering a Data Stream".	Analyze	BTL-4
9.	**What** are the applications of DSMS ?	Remember	BTL-1
10.	**Summarize** Real-Time Analysis.	Understand	BTL-2
11.	**Examine** how to deal with infinite streams.	Remember	BTL-1
12.	**Classify** Data Stream Management System.	Apply	BTL-3
13.	**Identify** what examples can you find for stream sources.	Apply	BTL-4
14.	**Give** the definition of the term "Data Mining".	Understand	BTL-2
15.	**Generalize** the statement "Filtering a data Stream".	Create	BTL-6
16.	**Examine** the need for RTAP.	Apply	BTL-3
17.	**Show** the features of what is Real Time Analytics Platform (RTAP).	Apply	BTL-3
18.	**Generalize** the surprise number (second moment) for the stream 3, 1, 4, 1, 3, 4, 2, 1, 2. What is the third moment of this stream?	Create	BTL-6
19.	**Give** real time data.	Understand	BTL-2

20.	**Assess** what information is used to substitute the view of streams over databases.	Evaluate	BTL-5
	PART-B (13- MARKS)		
1.	**Describe** Data Stream Model with a neat architecture diagram. (13)	Remember	BTL-1
2.	**Analyze** briefly about the sources of data stream	Analyze	BTL-4
3.	(i) **Write** issues in data stream queries. (7) (ii) **Explain** the issues in data streaming . (6)	Remember	BTL-1
4.	**Give** a description the stream data model and its architecture. (13)	Understand	BTL-2
5.	**Describe** Aurora system model. (13)	Remember	BTL-1
6.	(i) **Summarize** sampling in Data Streams. (7) (ii) **Describe** the sampling types in detail. (6)	Understand	BTL-2
7.	**Explain** in detail about examples in Aurora query model. (13)	Apply	BTL-3
8.	**Analyze** how mining is done with data streams. (13)	Analyze	BTL-4
9.	(i) **Evaluate** briefly how to count the distinct elements in a stream. (7) (ii) **Summarize** the count–distinct problem. (6)	Evaluate	BTL-5
10.	Write a **generalized** role of the following: (i). Sliding window concept. (7) (ii). Land mark window concept. (6)	Create	BTL-6
11.	**Illustrate** how would you describe the various windowing approach to data stream mining . (13)	Remember	BTL-1
12.	(i) **Analyze** the methods for analyzing time series data. (7) (ii) **Assess** the several types of motivation and data analysis available for time series? (6)	Analyze	BTL-4
13.	**Summarize** about time series in detail and discuss its significance. (13)	Understand	BTL-2
14.	(i) **Examine** real time sentiment analysis. (7) (ii) **Explain** how the mining concept is used in real time sentiment analysis. (6)	Apply	BTL-3
	PART-C (15-MARKS)		

1.	**Evaluate** the process of Data Stream Mining with suitable examples. (15)	Evaluate	BTL-5
2.	**Analyze** key stream mining problems and discuss the challenges associated with each problem and also summarize data streaming algorithms. (15)	Evaluate	BTL-5
3.	**Generalize** how is data analysis used in (i). stock market predictions. (8) (ii). weather forecasting predictions. (7)	Create	BTL-6
4.	**Prepare** a generic design for Realtime Analytics Platform(RTAP). Discuss your answer related to real time sentiment analysis. (15)	Create	BTL-6

UNIT V

PART-A (2 -MARKS)

1.	**What** is NoSQL?	Remember	BTL-1
2.	**Analyze** the reason behind why do we need NoSQL ?	Analyze	BTL-1
3.	**Assess** the components of Hadoop framework	Evaluate	BTL-5
4.	**Assess** the categories of NoSQL.	Analyze	BTL-4
5.	**List** is the advantages of NoSQL.	Remember	BTL-1
6.	**Discuss** the disadvantages of NoSQL.	Understand	BTL-2
7.	**Give** is HBase.	Understand	BTL-2
8.	**Compare** Cassandra vs Hadoop.	Evaluate	BTL-5
9.	**Predict** who is generating the big data and also name the ecosystems projects used for processing.	Understand	BTL-2
10.	**Examine** the differences between HBase and Hive.	Remember	BTL-1
11.	**Describe** the aggregate data models.	Remember	BTL-1
12.	**Give** what is Pig in Hadoop.	Understand	BTL-2
13.	**Examine** the need for Apache pig.	Apply	BTL-3

14.	**Generalize** the difference between Pig and Hive.	Create	BTL-6
15.	**List** out the usages of Pig, Hive and HBase.	Remember	BTL-1
16.	**Classify** the features of Hive.	Analyze	BTL-4
17.	**Infer** about Pig, Hive and HBase.	Analyze	BTL-4
18.	**Explain** hive in Big Data.	Apply	BTL-3
19.	**Show** the important features of Cassandra Client.	Apply	BTL-3
20.	**Generalize** types of built in operator in HIVE.	Create	BTL-6
PART-B(13 MARKS)			
1.	(i)**Describe** what is NoSQL. (7) (ii)**Identify** the advantages and disadvantages of NoSQL. (6)	Remember	BTL-1
2.	(i) **Analyze** in detail about Hive data manipulation, queries, and data types. (7) (ii) **Explain** data definition in Hive. (6)	Analyze	BTL-4
3.	**Write** about the system architecture and components of Hive and Hadoop. (13)	Create	BTL-5
4.	**Develop** in detail about briefly on aggregate data models. (13)	Evaluate	BTL-6
5.	**Discuss** about : (i) Pig (7) (ii) Pig Latin (6)	Understand	BTL-2
6.	**Demonstrate** about HBase and Hbase clients in detail. (13)	Apply	BTL-3
7.	**Describe** how Cassandra is integrated with Hadoop and also the tools related to Hadoop. (13)	Remember	BTL-1
8.	**Explain** briefly on Hbase architecture with neat diagram (13)	Analyze	BTL-4
9.	**Quote** the short notes on (i)Conceptual data modeling. (4) (ii)Logical data modeling. (4) (iii)Physical data modeling (5)	Understand	BTL-2
10.	**Discuss** the about Cassandra clients. (13)	Remember	BTL-1
11.	(i) **Explain** Compare and Contrast the Hadoop and MapR. (13)	Analyze	BTL-4
12.	(i) **Show** the method of invoking the Grunt shell. (7)	Apply	BTL-3

	(ii) **Explain** about Grunt in detail. (6)		
13.	**Predict** about Pig data model in detail with neat diagram. (13)	Understand	BTL-2
14.	(i) **Describe** the difference between hive and map reduce. (7) (ii) How is Hive used ? **Describe** in detail. (6)	Remember	BTL-1
	PART-C(15 MARKS)		
1.	**Perform** analysis on web server report (15) Sample Data: teleman.pr.mcs.net,-,-,[01/Jul/2005:00:03:57,0400], "GET,/images/KSC-logosmall.gif,HTTP/1.0", 304,0teleman.pr.mcs.net,-,-,[01/Jul/2005:00:03:57,0400],"GET,/images/KSC-logosmall.gif,HTTP/1.0",304,0. The data is comma separated. It consists of the user IP address, time at which the request is received, time zone, request type, requested link, request details, response code and bytes transferred. Usually the scale of these datasets is quite huge and running queries in a conventional method is not possible. Hence use Pig programming on this dataset to retrieve the necessary statistics which helps us to understand the load and usage of the server, user visit frequency, webpage popularity and the total bytes transferred	Create	BTL-6
2.	**Prepare** Formulate a Hbase table from the following data (15) Data_file.txt contains the below data 1. 1,India,Bihar,Champaran,2009,April,P1,1,5 2. 2,India, Bihar,Patna,2009,May,P1,2,10 3. 3,India, Bihar,Bhagalpur,2010,June,P2,3,15 4. 4,United States,California,Fresno,2009,April,P2,2,5 5. 5,United States	Create	BTL-6
3.	**Recommend** a procedure to find the number of occurrence of a word in a document using Hive. (15)	Evaluate	BTL-5
4.	How will you **Orde**r the use of Hive. How Does Hive Interact With Hadoop explain in detail? (15)	Evaluate	BTL-5