CS8091 BIG DATA ANALYTICS

L	T	P	C
3	0	0	3

OBJECTIVES:

- To know the fundamental concepts of big data and analytics.
- To explore tools and practices for working with big data
- To learn about stream computing.
- To know about the research that requires the integration of large amounts of data.

UNIT I INTRODUCTION TO BIG DATA 9

Evolution of Big data - Best Practices for Big data Analytics - Big data characteristics - Validating - The Promotion of the Value of Big Data - Big Data Use Cases- Characteristics of Big Data Applications - Perception and Quantification of Value -Understanding Big Data Storage - A General Overview of High-Performance Architecture - HDFS - MapReduce and YARN - Map Reduce Programming Model

UNIT II CLUSTERING AND CLASSIFICATION 9

Advanced Analytical Theory and Methods: Overview of Clustering - K-means - Use Cases - Overview of the Method - Determining the Number of Clusters - Diagnostics - Reasons to Choose and Cautions .- Classification: Decision Trees - Overview of a Decision Tree - The General Algorithm - Decision Tree Algorithms - Evaluating a Decision Tree - Decision Trees in R - Naïve Bayes - Bayes‘ Theorem - Naïve Bayes Classifier.

UNIT III ASSOCIATION AND RECOMMENDATION SYSTEM 9

Advanced Analytical Theory and Methods: Association Rules - Overview - Apriori Algorithm - Evaluation of Candidate Rules - Applications of Association Rules - Finding Association& finding similarity - Recommendation System: Collaborative Recommendation- Content Based Recommendation - Knowledge Based Recommendation- Hybrid Recommendation Approaches.

UNIT IV STREAM MEMORY 9

Introduction to Streams Concepts – Stream Data Model and Architecture - Stream Computing, Sampling Data in a Stream – Filtering Streams – Counting Distinct Elements in a Stream – Estimating moments – Counting oneness in a Window – Decaying Window – Real time Analytics Platform(RTAP) applications - Case Studies - Real Time Sentiment Analysis, Stock Market Predictions. Using Graph Analytics for Big Data: Graph Analytics

UNIT V NOSQL DATA MANAGEMENT FOR BIG DATA AND VISUALIZATION 9

NoSQL Databases : Schema-less Models‖: Increasing Flexibility for Data Manipulation-Key Value Stores- Document Stores - Tabular Stores - Object Data Stores - Graph Databases Hive - Sharding –

Hbase – Analyzing big data with twitter - Big data for E-Commerce Big data for blogs - Review of Basic Data Analytic Methods using R.

TOTAL: 45 PERIODS

OUTCOMES: Upon completion of the course, the students will be able to:

- Work with big data tools and its analysis techniques
- Analyze data by utilizing clustering and classification algorithms
- Learn and apply different mining algorithms and recommendation systems for large volumes of data
- Perform analytics on data streams
- Learn NoSQL databases and management.

TEXT BOOKS:

1. Anand Rajaraman and Jeffrey David Ullman, "Mining of Massive Datasets", Cambridge University Press, 2012.

2. David Loshin, "Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph", Morgan Kaufmann/El sevier Publishers, 2013.

REFERENCES:

1. EMC Education Services, "Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data", Wiley publishers, 2015.

2. Bart Baesens, "Analytics in a Big Data World: The Essential Guide to Data Science and its Applications", Wiley Publishers, 2015.

3. Dietmar Jannach and Markus Zanker, "Recommender Systems: An Introduction", Cambridge University Press, 2010.

4. Kim H. Pries and Robert Dunnigan, "Big Data Analytics: A Practical Guide for Managers " CRC Press, 2015.

5. Jimmy Lin and Chris Dyer, "Data-Intensive Text Processing with MapReduce", Synthesis Lectures on Human Language Technologies, Vol. 3, No. 1, Pages 1-177, Morgan Claypool publishers, 2010

CS8091- BIG DATA ANALYTICS

UNIT I

INTRODUCTION TO BIG DATA

Evolution of Big data – Best Practices for Big data Analytics – Big data characteristics – Validating –The Promotion of the Value of Big Data – Big Data Use Cases- Characteristics of Big Data Applications – Perception and Quantification of Value -Understanding Big Data Storage – A General Overview of High-Performance Architecture – HDFS – Map Reduce and YARN – Map Reduce Programming Model

COURSE OBJECTIVE: To know the fundamental concepts of big data and analytics.

1. **What is big data approach? (R)**

   Many It tools are available for big data projects. Organizations whose data workloads are constant and predictable are better served by traditional database whereas organizations challenged by increasing data demands will need to take advantage of Hadoop's scalable infrastructure.

2. **List out the applications of big data analytics.(U)**
   a. Marketing
   b. Finance
   c. Government
   d. Healthcare
   e. Insurance
   f. Retail

3. **List the types of cloud environment.(U)**
   a. Public cloud
   b. Private cloud

4. **What is reporting?(R)**

   It is the process of organizing data into informational summaries in order to monitor how different areas of a business are performing.

5. **What is analysis?(R)**

   It is the process of exploring data and reports in order to extract meaningful insights which can be used to better understand and improve business performance.

6. **List out the cross validation technique.(U)**
   a. Simple cross validation
   b. Double cross validation
   c. Multi cross validation

7. **Write short note on MapReduce? (U)**

   MapReduce provides a data parallel programming model for clusters of commodity machines. It is pioneered by google which process 20PB of data per day. MapReduce is popularized by Apache Hadoop project and used by Yahoo, Facebook, Amazon and others.

8. **What is cloud computing?(R)**

   Cloud computing is internet-based computing. It relies on sharing computing resources on-demand rather than having local servers or PCS and other devices. It is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort.

9. **Describe the drawbacks of cloud computing?(U)**

   In cloud computing, cheap nodes fail, especially when you have many of them. Mean time between failures(MTBF) for 1 node = 3 years – MTBF for 1000 nodes = 1 day and commodity network has low bandwidth.

10. **List out the four major types of resampling.(U)**
    a. Randomized exact test
    b. Cross-validation

c. Jackknife
d. Bootstrap

11. How Hadoop MapReduce works?(An)

In MapReduce, during the map phase, it counts the words in each document, while in the reduce phase it aggregates the data as per the document spanning the entire collection. During the map phase, the input data is divided into splits for analysis by map tasks running in parallel across Hadoop framework.

12. Explain what is shuffling in MapReduce?(R)

The process by which the system performs the sort and transfers the map outputs to the reducer as inputs is known as the shuffle

13. Explain what is distributed Cache in MapReduce Framework? (U)

Distributed Cache is an important feature provided by the MapReduce framework. When you want to share some files across all nodes in Hadoop Cluster, Distributed Cache is used. The files could be an executable jar files or simple properties file.

14. Explain what is NameNode in Hadoop? (U)

NameNode in Hadoop is the node, where Hadoop stores all the file location information in HDFS (Hadoop Distributed File System). In other words, NameNode is the centerpiece of an HDFS file system. It keeps the record of all the files in the file system and tracks the file data across the cluster or multiple machines

15. Explain what is heartbeat in HDFS? (U)

Heartbeat is referred to a signal used between a data node and Name node, and between task tracker and job tracker, if the Name node or job tracker does not respond to the signal, then it is considered there is some issues with data node or task tracker

16. Explain what combiners are and when you should use a combiner in a MapReduce Job? (U)

To increase the efficiency of MapReduce Program, Combiners are used. The amount of data can be reduced with the help of combiner's that need to be transferred across to the reducers. If the operation performed is commutative and associative you can use your reducer code as a combiner. The execution of combiner is not guaranteed in Hadoop

17. Explain what is Speculative Execution?(R)

In Hadoop during Speculative Execution, a certain number of duplicate tasks are launched. On a different slave node, multiple copies of the same map or reduce task can be executed using Speculative Execution. In simple words, if a particular drive is taking a long time to complete a task, Hadoop will create a duplicate task on another disk. A disk that finishes the task first is retained and disks that do not finish first are killed.

18. Explain what are the basic parameters of a Mapper? (U)

- The basic parameters of a Mapper are
- LongWritable and Text
- Text and Int Writable

19. Explain what is the function of MapReduce partitioner?(U)

The function of MapReduce partitioner is to make sure that all the value of a single key goes to the same reducer, eventually which helps even distribution of the map output over the reducers

20. Explain what is a difference between an Input Split and HDFS Block?(U)

The logical division of data is known as Split while a physical division of data is known as HDFS Block

21. Explain what happens in text format? (U)

In text input format, each line in the text file is a record. Value is the content of the line while Key is the byte offset of the line. For instance, Key: longWritable, Value: text

22. Mention what are the main configuration parameters that user need to specify to run MapReduce Job?(An)

a. The user of the Map Reduce framework needs to specify
b. Job's input locations in the distributed file system
c. Job's output location in the distributed file system
d. Input format
e. Output format
f. Class containing the map function
g. Class containing the reduce function
h. JAR file containing the mapper, reducer and driver classes

23. Explain what is WebDAV in Hadoop? (U)

To support editing and updating files WebDAV is a set of extensions to HTTP. On most operating system WebDAV shares can be mounted as filesystems, so it is possible to access HDFS as a standard filesystem by exposing HDFS over WebDAV.

24. Explain what is Sqoop in Hadoop? (R)

To transfer the data between Relational database management (RDBMS) and Hadoop HDFS a tool is used known as Sqoop. Using Sqoop data can be transferred from RDMS like MySQL or Oracle into HDFS as well as exporting data from HDFS file to RDBMS

25. Explain how Job Tracker schedules a task? (R)

The task tracker sends out heartbeat messages to Job tracker usually every few minutes to make sure that Job Tracker is active and functioning. The message also informs Job Tracker about the number of available slots, so the Job Tracker can stay up to date with wherein the cluster work can be delegated

26. Explain what is Sequence file input format? (R)

Sequence file input format is used for reading files in sequence. It is a specific compressed binary file format which is optimized for passing data between the outputs of one MapReduce job to the input of some other MapReduce job.

27. Explain what does the conf.setMapper Class do?(An)

Conf.setMapperclass sets the mapper class and all the stuff related to map job such as reading data and generating a key-value pair out of the mapper

28. Explain what is Hadoop? (R)

It is an open-source software framework for storing data and running applications on clusters of commodity hardware. It provides enormous processing power and massive storage for any type of data.

29. Mention Hadoop core components? (R)

a. Hadoop core components include,
b. HDFS
c. MapReduce

30. What is NameNode in Hadoop? (R)

NameNode in Hadoop is where Hadoop stores all the file location information in HDFS. It is the master node on which job tracker runs and consists of metadata.

31. Mention what are the data components used by Hadoop? (R)

a. Data components used by Hadoop are
b. Pig
c. Hive

32. Mention what is the data storage component used by Hadoop? (R)

The data storage component used by Hadoop is HBase.

33. Mention what are the most common input formats defined in Hadoop? (R)

a. The most common input formats defined in Hadoop are;
b. TextInputFormat
c. KeyValueInputFormat
d. SequenceFileInputFormat

34. For a Hadoop job, how will you write a custom partitioner? (R)

a. You write a custom partitioner for a Hadoop job, you follow the following path
b. Create a new class that extends Partitioner Class
c. Override method getPartition
d. In the wrapper that runs the MapReduce
e. Add the custom partitioner to the job by using method set Partitioner Class or – add the custom partitioner to the job as a config file

35. For a job in Hadoop, is it possible to change the number of mappers to be created? (R)

No, it is not possible to change the number of mappers to be created. The number of mappers is determined by the number of input splits.

36. Explain what is a sequence file in Hadoop? (R)

To store binary key/value pairs, sequence file is used. Unlike regular compressed file, sequence file support splitting even when the data inside the file is compressed.

37. When Namenode is down what happens to job tracker? (R)

Namenode is the single point of failure in HDFS so when Namenode is down your cluster will set off.

38. Explain how indexing in HDFS is done? (R)

Hadoop has a unique way of indexing. Once the data is stored as per the block size, the HDFS will keep on storing the last part of the data which say where the next part of the data will be.

39. Explain is it possible to search for files using wildcards? (R)

Yes, it is possible to search for files using wildcards.

40. List out Hadoop's three configuration files? (R)

a. The three configuration files are
b. core-site.xml
c. mapred-site.xml
d. hdfs-site.xml

41. Explain how can you check whether Namenode is working beside using the jps command? (R)

- Besides using the jps command, to check whether Namenode are working you can also use
- /etc/init.d/hadoop-0.20-namenode status.

42. Explain what is "map" and what is "reducer" in Hadoop?(U)

a. In Hadoop, a map is a phase in HDFS query solving. A map reads data from an input location, and outputs a key value pair according to the input type.
b. In Hadoop, a reducer collects the output generated by the mapper, processes it, and creates a final output of its own.

43. In Hadoop, which file controls reporting in Hadoop? (R)

In Hadoop, the hadoop-metrics.properties file controls reporting.

44. For using Hadoop list the network requirements? (R)

a. For using Hadoop the list of network requirements are:
b. Password-less SSH connection
c. Secure Shell (SSH) for launching server processes

45. Mention what is rack awareness? (R)

Rack awareness is the way in which the namenode determines on how to place blocks based on the rack definitions.

46. Explain what is a Task Tracker in Hadoop? (R)

A Task Tracker in Hadoop is a slave node daemon in the cluster that accepts tasks from a JobTracker. It also sends out the heartbeat messages to the JobTracker, every few minutes, to confirm that the JobTracker is still alive.

47. Mention what daemons run on a master node and slave nodes? (R)

a. Daemons run on Master node is "NameNode"
b. Daemons run on each Slave nodes are "Task Tracker" and "Data"

48. Explain how can you debug Hadoop code? (R)

a. The popular methods for debugging Hadoop code are:
b. By using web interface provided by Hadoop framework
c. By using Counters

49. Explain what is storage and compute nodes? (R)

The storage node is the machine or computer where your file system resides to store the processing data

The compute node is the computer or machine where your actual business logic will be executed.

50. Mention what is the use of Context Object? (R)

a) The Context Object enables the mapper to interact with the rest of the Hadoop
b) system. It includes configuration data for the job, as well as interfaces which allow it to emit output.

51. Mention what is the next step after Mapper or MapTask? (R)

The next step after Mapper or MapTask is that the output of the Mapper are sorted, and partitions will be created for the output.

52. Mention what is the number of default partitioner in Hadoop? (R)

In Hadoop, the default partitioner is a "Hash" Partitioner.

53. Explain what is the purpose of RecordReader in Hadoop? (R)

In Hadoop, the RecordReader loads the data from its source and converts it into (key, value) pairs suitable for reading by the Mapper.

54. Explain how is data partitioned before it is sent to the reducer if no custom partitioner is defined in Hadoop? (R)

If no custom partitioner is defined in Hadoop, then a default partitioner computes a hash value for the key and assigns the partition based on the result.

55. Explain what happens when Hadoop spawned 50 tasks for a job and one of the task failed?(U)

It will restart the task again on some other TaskTracker if the task fails more than the defined limit.

56. Mention what is the best way to copy files between HDFS clusters? (R)

The best way to copy files between HDFS clusters is by using multiple nodes and the distcp command, so the workload is shared.

57. Mention what is the difference between HDFS and NAS?(An)

HDFS data blocks are distributed across local drives of all machines in a cluster while NAS data is stored on dedicated hardware.

58. Mention how Hadoop is different from other data processing tools? (R)

In Hadoop, you can increase or decrease the number of mappers without worrying about the volume of data to be processed.

59. Mention what job does the conf class do? (R)

Job conf class separate different jobs running on the same cluster. It does the job level settings such as declaring a job in a real environment.

60. Mention what is the Hadoop MapReduce APIs contract for a key and value class?(U)

a. For a key and value class, there are two Hadoop MapReduce APIs contract
b. The value must be defining the org.apache.hadoop.io.Writable interface
c. The key must be defining the org.apache.hadoop.io.WritableComparable interface

61. Mention what are the three modes in which Hadoop can be run? (R)

a. The three modes in which Hadoop can be run are
b. Pseudo distributed mode
c. Standalone (local) mode
d. Fully distributed mode

62. Mention what does the text input format do? (R)

The text input format will create a line object that is an hexadecimal number. The value is considered as a whole line text while the key is considered as a line object. The mapper will receive the value as 'text' parameter while key as 'longwriteable' parameter.

63. Mention what is distributed cache in Hadoop? (R)

Distributed cache in Hadoop is a facility provided by MapReduce framework. At the time of execution of the job, it is used to cache file. The Framework copies the necessary files to the slave node before the execution of any task at that node.

64. Explain how does Hadoop Classpath plays a vital role in stopping or starting in Hadoop daemons?(U)

Classpath will consist of a list of directories containing jar files to stop or start daemons.

65. List the main characteristics of Big Data (U) (NOVEMBER /DECEMBER 2020/ APRIL / MAY 2021)

66. Veracity, Variety, Velocity, Volume, Validity, Variability, Volatility, Visualization and Value.

67. Why HDFS preferred to RDBMS (U) (NOVEMBER /DECEMBER 2020/ APRIL / MAY 2021)

It is more flexible in storing, processing, and managing data than traditional RDBMS. Unlike traditional systems, Hadoop enables multiple analytical processes on the same data at the same time. It supports scalability very flexibly.

PART-B

1. Analyse in detail about the challenges of the Big Data in Modern Data Analytics. **(An)**
2. Justify the Statement "Web Data is the Most Popular Big Data" with reference to data analytic professional. **(E)**
3. Comment on the statement "Is the "Big" Part or the "Data" Part More Important ".**(E)**
4. Develop the role of Analytic Sandbox and its benefits in the Analytic Process.**(C)**
5. List the features of Hadoop and explain the functionalities of Hadoop cluster? **(U)**
6. Describe briefly about Hadoop input and output and write a note on data integrity? **(U)**
7. Discuss the various core components of the Hadoop.**(U)**
8. Assess the significances of MapReduce .**(U)**
9. Explain about Hadoop distributed file system architecture with neat diagram.**(U)**
10. Summarize briefly on
11. Algorithms using MapReduce. **(U)**
12. Extensions to MapReduce. **(U)**
13. Compare and Contrast the Hadoop and MapR**(U)**
14. Analyse the steps of Map Reduce Algorithms. **(An)**
15. Describe the concepts of HDFS. **(U)**
16. (i) Explain the management of computing resources and the management of the data across the network of storage nodes in High performance architecture (**4) (U) NOVEMBER /DECEMBER 2020/ APRIL / MAY 2021**
17. Write short notes on the following programming model:(**9) (R) NOVEMBER /DECEMBER 2020/ APRIL / MAY 2021**

1. HDFS
2. MapReduce
3. YARN

18. (i) Brief about the characteristics of Big data Applications (5)**(U)NOVEMBER /DECEMBER 2020/ APRIL / MAY 2021**

19. Explain the role of Big Data Analytics in the following: (8)**(U)NOVEMBER /DECEMBER 2020/ APRIL / MAY 2021**

1. Credit Fraud Detection
2. Clustering and data segmentation
3. Recommendation engines
4. Price modeling

COURSE OUTCOMES:

Students can able to work with big data tools and its analysis techniques

UNIT II
CLUSTERING AND CLASSIFICATION

Advanced Analytical Theory and Methods: Overview of Clustering – K-means – Use Cases – Overview of the Method – Determining the Number of Clusters – Diagnostics – Reasons to Choose and Cautions .- Classification: Decision Trees – Overview of a Decision Tree – The General Algorithm – Decision Tree Algorithms – Evaluating a Decision Tree – Decision Trees in R – Naïve Bayes – Bayes‘ Theorem – Naïve Bayes Classifier.

COURSE OBJECTIVE:

To explore tools and practices for working with big data

1. What are the three stages of IDA process? (R)

- Data preparation
- Data mining and rule finding
- Result validation and interpretation

2. What is linear regression? (R)

Linear regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted X. The case of one explanatory variable is called simple **linear regression**.

3. Explain Bayesian Inference ?(U)

Bayesian inference is a method of statistical **inference** in which **Bayes**' theorem is used to update the probability for a hypothesis as more evidence or information becomes available. **Bayesian inference** is an important technique in statistics, and especially in mathematical statistics.

4. What is meant by rule induction? (R)

Rule induction is an area of machine learning in which formal rules are extracted from a set of observations. The rules extracted may represent a full scientific model of the data, or merely represent local patterns in the data.

5. What are the two strategies in Learn-One-Rule Function. (R)

- General to specific
- Specific to general

6. Write down the topologies of Neural Network. (U)

- Single layer
- Multi layer
- Recurrent
- Self-organized

7. What is meant by fuzzy logic. (R)

More than data mining tasks such as prediction, classification, etc., fuzzy models can give insight to the underlying system and an be automatically derived from system's dataset. For achieving this, the technique used is grid based rule set.

8. Write short note on fuzzy qualitative modeling. (R)

The fuzzy modeling can be interpreted as a qualitative modeling scheme by which the system behavior is qualitatively described using a natural language. A fuzzy qualitative model is a generalized fuzzy model consisting of linguistic explanations about system behavior in the framework of fuzzy logic instead of mathematical equations with numerical values or conventional logical formula with logical symbols.

9. What are the steps for Bayesian data analysis. (R)

- Setting up the prior distribution
- Setting up the posterior distribution
- Evaluating the fit of the model

10. Write short notes on time series model. (R)

A time series is a sequential set of data points, measured typically at successive times. It is mathematically defined as a set of vectors x(t), t=0,1,2,… where t represents the time elapsed. The Variable x9t0 is treated as a random variable.

11. Define Bayes' Theorem (R)

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred.

12. Multinomial Naive Bayes(R)

Feature vectors represent the frequencies with which certain events have been generated by a **multinomial distribution**. This is the event model typically used for document classification.

13. Bernoulli Naive Bayes(R)

In the multivariate Bernoulli event model, features are independent booleans (binary variables) describing inputs. Like the multinomial model, this model is popular for document classification tasks, where binary term occurrence (i.e. a word occurs in a document or not) features are used rather than term frequencies (i.e. frequency of a word in the document).

14. What are the branches of subspace clustering based on search strategy

There are two branches of subspace clustering based on their search strategy.

- Top-down algorithms find an initial clustering in the full set of dimension and evaluate the subspace of each cluster.
- Bottom-up approach finds dense region in low dimensional space then combine to form clusters.

15. Define Classification(R)

It is a Data analysis task, i.e. the process of finding a model that describes and distinguishes data classes and concepts. Classification is the problem of identifying to which of a set of categories (subpopulations), a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known.

16. What is Discriminative(R)

It is a very basic classifier and determines just one class for each row of data. It tries to model just by depending on the observed data, depends heavily on the quality of data rather than on distributions.

Example: Logistic Regression

17. Define Generative (R)

It models the distribution of individual classes and tries to learn the model that generates the data behind the scenes by estimating assumptions and distributions of the model. Used to predict the unseen data. **Example**: Naive Bayes Classifier

18. List out Classifiers Of Machine Learning(U)

- Decision Trees
- Bayesian Classifiers
- Neural Networks
- K-Nearest Neighbour
- Support Vector Machines
- Linear Regression
- Logistic Regression

19. List out an Associated Tools and Languages Used to mine/ extract useful information from raw data.(U)

a) Main Languages used: R, SAS, Python, SQL
b) Major Tools used: RapidMiner, Orange, KNIME, Spark, Weka
c) Libraries used: Jupyter, NumPy, Matplotlib, Pandas, ScikitLearn, NLTK, TensorFlow, Seaborn, Basemap, etc.

20. List out Real Life Examplesof classification (U)

a) **Market Basket Analysis:**

It is a modeling technique that has been associated with frequent transactions of buying some combination of items.

Example: Amazon and many other Retailers use this technique. While viewing some product, certain suggestions for the commodities are shown that some people have bought in the past.

b) **Weather Forecasting:**

Changing Patterns in weather conditions needs to be observed based on parameters such as temperature, humidity, wind direction. This keen observation also requires the use of previous records in order to predict it accurately.

21. List out an Advantages of classification (U)

- Mining Based Methods are cost effective and efficient
- Helps in identifying criminal suspects
- Helps in predicting risk of diseases

- Helps Banks and Financial Institutions to identify defaulters so that they may approve Cards, Loan, etc.

22. **List out disadvantages of classification (U)**

Privacy: When the data is either are chances that a company may give some information about their customers to other vendors or use this information for their profit.

Accuracy Problem: Selection of Accurate model must be there in order to get the best accuracy and result.

23. List out an Applications of classification (U)

- Marketing and Retailing
- Manufacturing
- Telecommunication Industry
- Intrusion Detection
- Education System
- Fraud Detection

24. State Bayes theorem (R) NOVEMBER /DECEMBER 2020/ APRIL / MAY 2021

Bayes Theorem is the extension of Conditional probability. Conditional probability helps us to determine the probability of A given B, denoted by P(A|B). So Bayes' theorem says if we know P(A|B) then we can determine P(B|A), given that P(A) and P(B) are known to us.

25. What is the application of clustering in medical domain? (R) NOVEMBER /DECEMBER 2020/ APRIL / MAY 2021

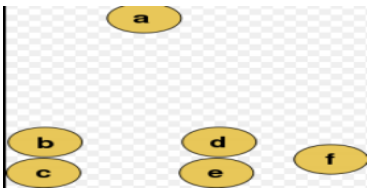
Clustering is a powerful machine learning tool for detecting structures in datasets. In the medical field, clustering has been proven to be a powerful tool for discovering patterns and structure in labeled and unlabeled datasets

PART-B

1.Analyze the statement in detail : "Data Analysis is not a decision-making system, but a decision-supporting system" **(An)**

2.Create a Regression Model for " happy people get many hours of sleep" using your own data and what kind of inferences it provides.**(C)**

3.Summarize hierarchical clustering in detail. Analyse the given diagram and draw the dendrogram using hierarchical clustering algorithm . **(U)**



4.Compose the K-means partitioning algorithm using the given data. **(C)**

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)

5.Two Cluster the following eight points into three clusters using K means clustering algorithm and use Euclidean distance. Ala(2,10), A2-(2,5), ??-64), A4-(5,8), A5-(7,5), A6 (6,4), A7-(1,2), A8 (4,9). a) Create distance matrix by calculating Euclidean distance between each pair of points

(C)

6.Brief about K-means clustering with example**(8) (U) NOVEMBER /DECEMBER 2020/ APRIL / MAY 2021**

7.Explain the several decision that the practitioner must make for the following parameters in K-means clustering: (**5)(U) NOVEMBER /DECEMBER 2020/ APRIL / MAY 2021**

1. Object attributes
2. Units of measures
3. Rescaling

8.Write short notes on the following Decision Tree **(C)** (**13) NOVEMBER /DECEMBER 2020/ APRIL / MAY 2021**

1. ID3 algorithm
2. C4.5
3. CART
4. Evaluation of Decision Tree

COURSE OUTCOME:

Students can able to Analyze data by utilizing clustering and classification algorithms

UNIT III
ASSOCIATION AND RECOMMENDATION SYSTEM

Advanced Analytical Theory and Methods: Association Rules – Overview – Apriori Algorithm – Evaluation of Candidate Rules – Applications of Association Rules – Finding Association& finding similarity – Recommendation System: Collaborative Recommendation- Content Based Recommendation – Knowledge Based Recommendation- Hybrid Recommendation Approaches.

COURSE OBJECTIVE:

To learn about association and recommendation system.

1.What is data stream model?(R)

A data stream is a real-time, continuous and ordered sequence of items. It is not possible to control the order in which the items arrive, nor it is feasible to locally store a stream in its entirety in any memory device.

2.Define Data Stream Mining. (R)

Data Stream Mining is the process of extracting useful knowledge from continuous, rapid data streams. Many traditional data mining algorithms can be recast to work with larger datasets, but they cannot address the problem of a continuous supply of data.

3.Write short note about sensor networks. (R)

Sensor networks are a huge source of data occurring in streams. They are used in numerous situations that require constant monitoring of several variables, based on which important decisions are made. in many cases, alerts and alarms may be generated as a response to the information received from a series of sensors.

4.what is meant by one-time queries? (R)

One-Time queries are queries that are evaluated once over a point-in-time snapshot of the

data set, with the answer returned to the user.

Eg: A stock price checker may alert the user when a stock price crosses a particular price point.

5.Define biased reservoir sampling. (R)

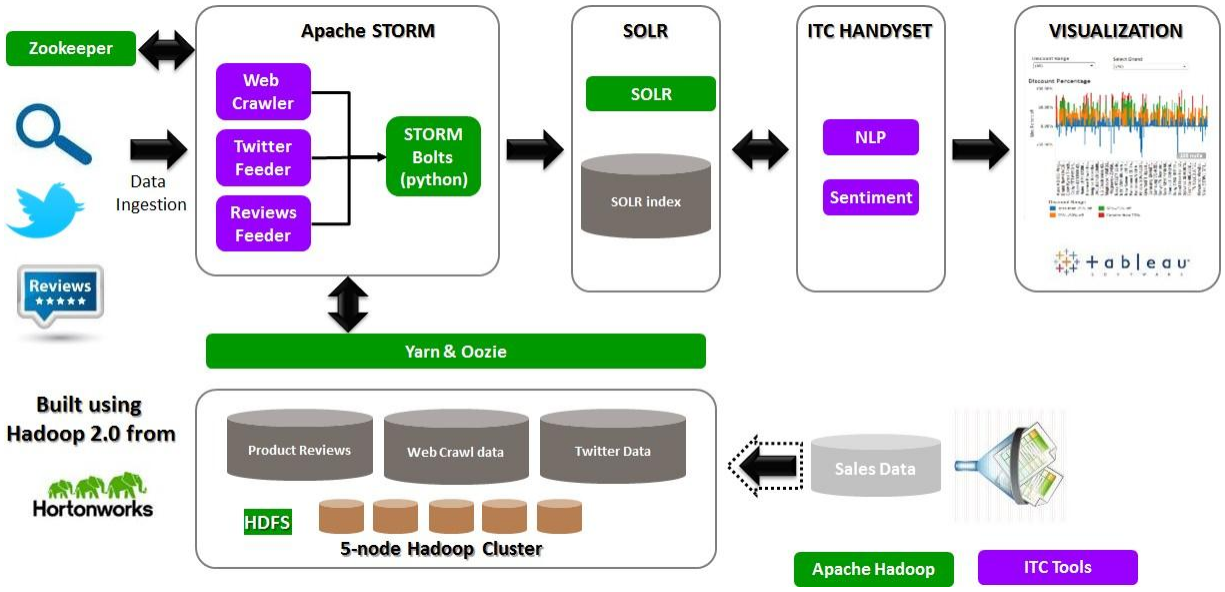
Biased reservoir sampling is defined as bias function to regulate the sampling from the stream. The bias gives a higher probability of selecting data points from recent parts of the stream as compared to distant past.

6.What is Bloom Filter? (R)

A Bloom Filter is a space-efficient probabilistic data structure, conceived by Burton Howard Bloom in 1970, that is used to test whether an element is a member of set. False Positive matches are possible but false negative are not, thus a Bloom filter has a 100% recall rate.

7.List out the applications of RTAP. (R)

- Financial services
- Government
- E-Commerce sites



8.What are the three layers of Lambda architecture. (R)

- Batch Layer- for batch processing of all data.
- Speed Layer- for real-time processing of streaming data.
- Serving Layer- for responding to queries.

9. What is RTSA? (R)

Real-Time Sentiment analysis (also known as opinion mining) refers to the use of natural language processing text analysis and computational linguistics to identify and extract subjective information in source materials.

10. What is the purpose of Data mining Technique? (R)

It provides a way to use various data mining tasks.

11. Define Predictive model. (R)

It is used to predict the values of data by making use of known results from a different set of sample data.

12. Data mining tasks that are belongs to predictive model(R)

- Classification
- Regression
- Time series analysis

13. Define descriptive model(R)

It is used to determine the patterns and relationships in a sample data. Data mining tasks that Belongs to descriptive model: Clustering Summarization Association rules Sequence discovery

14. List out the advanced database systems. (R)

- Extended-relational databases
- Object-oriented databases
- Deductive databases
- Spatial databases
- Temporal databases
- Multimedia databases
- Active databases
- Scientific databases
- Knowledge databases

15. Classifications of Data mining systems. Based on the kinds of databases mined: According to model(U)

- Relational mining system
- Transactional mining system
- Object-oriented mining system
- Object-Relational mining system
- Data warehouse mining system o Types of Data
- Spatial data mining system
- Time series data mining system
- Text data mining system
- Multimedia data mining system

16. Explain Association rule in mathematical notations. (R)

Let I-{i1,i2,…..,im} be a set of items

Let D, the task relevant data be a set of database transaction T is a set of items

An association rule is an implication of the form A=>B where A C I, B C I,

17. Define support and confidence in Association rule mining. (R)

Support S is the percentage of transactions in D that contain AUB.

Confidence c is the percentage of transactions in D containing A that also contain B.

Support (A=>B)= P(AUB)

Confidence (A=>B)=P(B/A)

18. How are association rules mined from large databases? (R)

I step: Find all frequent item sets:

II step: Generate strong association rules from frequent item sets

19. What is the purpose of Apriori Algorithm? (R)

Apriori algorithm is an influential algorithm for mining frequent item sets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent item set properties.

20. Define anti-monotone property. (R)

If a set cannot pass a test, all of its supersets will fail the same test as well.

21. How to generate association rules from frequent item sets?(An)

Association rules can be generated as follows,For each frequent item set1, generate all non empty subsets of 1. For every non empty subsets s of 1, output the rule "S=>(1-s)"if Support count(=min_conf, Support_count(s),Where min_conf is the minimum confidence threshold.

22. Give few techniques to improve the efficiency of Apriori algorithm. (R)

Hash based technique, Transaction Reduction, Portioning, Sampling & Dynamic item counting.

23. What is frequent item set? (R) (NOVEMBER /DECEMBER 2020/ APRIL / MAY 2021)

Apriori algorithm uses frequent itemsets to generate association rules. It is based on the concept that a subset of a frequent itemset must also be a frequent itemset. Frequent Itemset is an itemset whose support value is greater than a threshold value (support).

24. Differentiate between Collaborative and Content based Recommendation (An) (NOVEMBER /DECEMBER 2020/ APRIL / MAY 2021)

Collaborative recommendation system:

Recommender systems that recommend items through consumer collaborations and are the most widely used and proven method of providing recommendations.

There are two types:

- User-to-user collaborative filtering based on user-to-user similarity
- Item-to-item collaborative filtering based on item-to-item similarity.

Content based Recommendation:

Content-based recommenders treat recommendation as a user-specific classification problem and learn a classifier for the user's likes and dislikes based on an item's features. In this, keywords are used to describe the items, and a user profile is built to indicate the type of item this user likes.

PART-B

1. Summarize data streaming algorithms in detail. **(U)**

2. Analyse key stream mining problems and discuss the challenges associated with each problem.**(An)**

3. Explain Advanced Analytical Theory and Methods**(U)**

4. Explain Ariori Algorithm **(U)**

5. List out an Applications of Association Rules**(U)**

6. Explain in detail about Recommendation System**(U)**

7. Explain Knowledge based and Hybrid Recommendation system in detail **(U) NOVEMBER /DECEMBER 2020/ APRIL / MAY 2021**

8. Explain the Apriori algorithm for mining frequent item sets with an example **(U) NOVEMBER /DECEMBER 2020/ APRIL / MAY 2021**

COURSE OUTCOME:

Students able to learn and apply different mining algorithms and recommendation systems for large volumes of data

UNIT IV
STREAM MEMORY

Introduction to Streams Concepts – Stream Data Model and Architecture – Stream Computing, Sampling Data in a Stream – Filtering Streams – Counting Distinct Elements in a Stream – Estimating moments – Counting oneness in a Window – Decaying Window – Real time Analytics Platform(RTAP) applications – Case Studies – Real Time Sentiment Analysis, Stock Market Predictions. Using Graph Analytics for Big Data: Graph Analytics

COURSE OBJECTIVE:

To learn about stream computing

1.What is Association Rule Mining?(R)

The Association Rule Mining is main purpose to discovering frequent itemsets from a large dataset is to discover a set of if-then rules called Association rules. The form of an association rules is $I \rightarrow j$, where I is a set of items(products) and j is a particular item.

2.List any two algorithms for Finding Frequent Itemset.(R)

- Apriori Algorithm
- FP-Growth Algorithm
- SON algorithm
- PCY algorithm

3.What is meant by curse of dimensionality? (R)

Points in high-dimensional Euclidean spaces, as well as points in non-Euclidean spaces often behave unintuitively. Two unexpected properties of these spaces are that the random points are almost always at about the same distance, and random vectors are almost always orthogonal.

4.Define Toivonen's Algorithm(R)

Toivonen's algorithm makes only one full pass over the database. The algorithm thus produces exact association rules in one full pass over the database. The algorithm will give neither false negatives nor positives, but there is a small yet non-zero probability that it will fail to produce any answer at all. Toivonen's algorithm begins by selecting a small sample of the input dataset and finding from it the candidate frequent item sets.

5.List out some applications of clustering. (R)

- Collaborative filtering
- Customer segmentation
- Data summarization
- Dynamic trend detection
- Multimedia data analysis
- Biological data analysis
- Social network analysis

6.What are the types of Hierarchical Clustering Methods? (R)

- Single-link clustering

- Complete-link clustering
- Average-link clustering
- Centroid link clustering

7.Define CLIQUE(R)

CLIQUE is a subspace clustering algorithm that automatically finds subspaces with high-density clustering in high dimensional attribute spaces. CLIQUE is a simple grid-based method for finding density-based clusters in subspaces. The procedure for this grid-baased clustering is relatively simple.

8.What is meant by k-means algorithm? (R)

The family of algorithms is of the point-assignment type and assumes a Euclidean space. It is assumed that there are exactly *k* clusters for some known *k*. After picking *k* initial cluster centroids, the points are considered one at a time and assigned to the closest centroid.

9.List out the types of Data streaming operators (U)

- Stateless operators
- Stateful operators

10. List out the steps to be followed to deploy a Big Data solution (U)

- Data Ingestion
- Data Storage
- Data Processing

11. What is FSCK?

FSCK (File System Check) is a command used to detect inconsistencies and issues in the file.

12. What are the real-time applications of Hadoop?

Some of the real-time applications of Hadoop are in the fields of:

- Content management.
- Financial agencies.
- Defense and cyber security.
- Managing posts on social media.

13. What is the function of HDFS? (R)

The HDFS (Hadoop Distributed File System) is Hadoop's default storage unit. It is used for storing different types of data in a distributed environment.

14. What is commodity hardware?(R)

Commodity hardware can be defined as the basic hardware resources needed to run the Apache Hadoop framework.

15. Name a few daemons used for testing JPS command.(U)

- NameNode
- NodeManager
- DataNode
- ResourceManager

16. What are the most common input formats in Hadoop?(R)

- Text Input Format
- Key Value Input Format

- Sequence File Input Format

17. Name a few companies that use Hadoop.(U)

Yahoo, Facebook, Netflix, Amazon, and Twitter.

18. What is the default mode for Hadoop?

Standalone mode is Hadoop's default mode. It is primarily used for debugging purpose.

19. What is the role of Hadoop in big data analytics?(R)

By providing storage and helping in the collection and processing of data, Hadoop helps in the analytics of big data.

20. What are the components of YARN?(R)

The two main components of YARN (Yet Another Resource Negotiator) are:

- Resource Manager
- Node Manager

21. Name a few data management tools used with Edge Nodes?(U)

Oozie, Flume, Ambari, and Hue are some of the data management tools that work with edge nodes in Hadoop.

22. What are the steps to deploy a Big Data solution?(U)

The three steps to deploying a Big Data solution are:

1. Data Ingestion
2. Data Storage and
3. Data Processing

23. How many modes can Hadoop be run in?(An)

Hadoop can be run in three modes— Standalone mode, Pseudo-distributed mode and fully-distributed mode.

26. Name the core methods of a reducer (U)

The three core methods of a reducer are,

1. setup()
2. reduce()
3. cleanup()

27. Define decaying window (R) NOVEMBER /DECEMBER 2020/ APRIL / MAY 2021

Decaying window is to identify the most popular elements (trending, in other words) in an incoming data stream.

The aggregate sum of the decaying exponential weights can be calculated using the following formula:

$\sum t-1 i=0 at-i(1-c)i$

28. What are the technical complexities of analyzing graphs? (U) NOVEMBER /DECEMBER 2020/ APRIL / MAY 2021

The technical complexities of analyzing graphs are:

(1) Geometric (deviation of the data's approximator from some "idealized" configuration, such as deviation from harmonicity)

(2) Structural (how many elements of a principal graph are needed to approximate the data

(3) Construction complexity (how many applications of elementary graph transformations are needed to construct the principal object starting from the simplest one).

PART-B

1. **Describe** the Data Stream model with a neat architecture diagram **(U)**
2. **Illustrate** briefly about the sources of data stream. **(U)**
3. **Explain** issues in data stream queries .**(U)**
4. **(i) List** the issues in data streaming . **(R)**
5. **(ii) Summarize** the stream data model and its architecture. **(U)**
6. **Analyse** and write a short note on Aurora system model.**(An)**
7. Explain Sampling in Data Streams . **(U)**
8. Explain the sampling types in detail **(U)**
9. Describe about Aurora query model. **(U)**
10. Generalize how mining is done with data streams. **(U)**
11. Describe briefly how to count the distinct elements in a stream. **(U)**

What do you meant by count–distinct problem . **(R)**

12. Quote short notes on
    a) Sliding window concept **(R)**
    b) Land mark window concept **(U)**
13. Illustrate how would you describe the various windowing approach to data stream mining**(U)**
14. List the methods for analyzing time series data. **(R)**
15. What are the several types of motivation and data analysis available for time series? **(R)**
16. **Explain** about time series in detail and discuss its significance **(R)**
17. **Evaluate the** process of Data Stream Mining with suitable examples. **(E)**
18. Summarize data streaming algorithms in detail. **(U)**
19. Discuss the challenges associated with each problem. (U)
20. **Generalize** how is data analysis used in **(U)**
    a) stock market predictions.
    b) weather forecasting predictions.
21. Explain the Count distinct problem and Flajolet-Martin algorithm in stream **(7) (U) NOVEMBER /DECEMBER 2020/ APRIL / MAY 2021**
22. Explain in detail about the Alon–Matias-Szegedy Algorithm for estimating second moments in stream **(6) (U) NOVEMBER /DECEMBER 2020/ APRIL / MAY 2021)**
23. Explain in detail about the Sampling in Data stream **(9) (U) NOVEMBER /DECEMBER 2020/ APRIL / MAY 2021**
24. Brief about the features of a graph analytics platform to be considered for various Big data applications **(4) (U) NOVEMBER /DECEMBER 2020/ APRIL / MAY 2021**

COURSE OUTCOME:

Students can able to perform analytics on data streams

UNIT V

NOSQL DATA MANAGEMENT FOR BIG DATA AND VISUALIZATION

NoSQL Databases : Schema-less Models‖: Increasing Flexibility for Data Manipulation-Key Value Stores- Document Stores – Tabular Stores – Object Data Stores – Graph Databases Hive – Sharding –- Hbase – Analyzing big data with twitter – Big data for E-Commerce Big data for blogs – Review of Basic Data Analytic Methods using R.

COURSE OBJECTIVE:To know about the research that requires the integration of large amounts of data.

PART-A

1. **Compare NoSQL & RDBMS (An)**

Criteria	NoSQL	RDBMS
Data format	Does not follow any order	Organized and structured
Scalability	Very Good	Average
Querying	Limited as no Join Clause	Using SQL
Storage mechanism	Key-Value Pair, document, column storage, etc.	Data & relationship stored in different tables

2. **What is NoSQL?(R)**

   NoSQL encompasses a wide variety of different database technologies that were developed in response to a rise in the volume of data stored about users, objects and products. The frequency in which this data is accessed, and performance and processing needs. Relational databases, on the other hand, were not designed to cope with the scale and agility challenges that face modern applications, nor were they built to take advantage of the cheap storage and processing power available today.

3. **What are the features of NoSQL? (R)**

   When compared to relational databases, NoSQL databases are more scalable and provide superior performance, and their data model addresses several issues that the relational model is not designed to address:

   a) Large volumes of structured, semi-structured, and unstructured data
   b) Agile sprints, quick iteration, and frequent code pushes
   c) Object-oriented programming that is easy to use and flexible
   d) Efficient, scale-out architecture instead of expensive, monolithic architecture

4. **Why NoSQL? (An)**

   The concept of NoSQL databases became popular with Internet giants like Google, Facebook, Amazon, etc. who deal with huge volumes of data. The system response time becomes slow when you use RDBMS for massive volumes of data.

5. **Brief History of NoSQL Databases (R)**

- 1998- Carlo Strozzi use the term NoSQL for his lightweight, open-source relational database
- 2000- Graph database Neo4j is launched
- 2004- Google BigTable is launched
- 2005- CouchDB is launched
- 2007- The research paper on Amazon Dynamo is released
- 2008- Facebooks open sources the Cassandra project

6. List out Features of NoSQL(R)

- NoSQL databases never follow the relational model
- Never provide tables with flat fixed-column records
- Work with self-contained aggregates or BLOBs
- Doesn't require object-relational mapping and data normalization
- No complex features like query languages, query planners,
- referential integrity joins, ACID

7. Explain about Schema-free (U)

- NoSQL databases are either schema-free or have relaxed schemas
- Do not require any sort of definition of the schema of the data
- Offers heterogeneous structures of data in the same domain

8. Write about Simple API (U)

- Offers easy to use interfaces for storage and querying data provided
- APIs allow low-level data manipulation & selection methods
- Text-based protocols mostly used with HTTP REST with JSON
- Mostly used no standard based query language
- Web-enabled databases running as internet-facing services

9. List out types of NoSQL Databases(R)

- Key-value Pair Based
- Column-oriented Graph
- Graphs based
- Document-oriented

10. Query Mechanism tools for NoSQL (R)

- The most common data retrieval mechanism is the REST-based retrieval of a value based on its key/ID with GET resource
- Document store Database offers more difficult queries as they understand the value in a key-value pair. For example, CouchDB allows defining views with MapReduce

11. What is the CAP Theorem?(R)

- CAP theorem is also called brewer's theorem. It states that is impossible for a distributed data store to offer more than two out of three guarantees
- Consistency
- Availability
- Partition Tolerance

12. List out an Advantages of NoSQL(R)

- Can be used as Primary or Analytic Data Source
- Big Data Capability
- No Single Point of Failure
- Easy Replication
- No Need for Separate Caching Layer
- It provides fast performance and horizontal scalability.
- Can handle structured, semi-structured, and unstructured data with equal effect
- Object-oriented programming which is easy to use and flexible
- NoSQL databases don't need a dedicated high-performance server
- Support Key Developer Languages and Platforms
- Simple to implement than using RDBMS

13. List out disadvantages of NoSQL **(R)**

- No standardization rules
- Limited query capabilities
- RDBMS databases and tools are comparatively mature
- It does not offer any traditional database capabilities, like consistency when multiple transactions are performed simultaneously.
- When the volume of data increases it is difficult to maintain unique values as keys become difficult
- Doesn't work as well with relational data
- The learning curve is stiff for new developers
- Open source options so not so popular for enterprises.

14. Explain the data import in R language.(U)

- R provides to import data in R language. To begin with the R commander GUI, user should type the commands in the command Rcmdr into the console. Data can be imported in R language in 3 ways such as:
- Select the data set in the dialog box or enter the name of the data set as required.
- Data is entered directly using the editor of R Commander via Data->New Data Set. This works good only when the data set is not too large.
- Data can also be imported from a URL or from plain text file (ASCII), or from any statistical package or from the clipboard.

15. Difference between library () and require () functions in R language.(An)

library()	**require()**
Library () function gives an error message display, if the desired package cannot be loaded.	Require () function is used inside function and throws a warning messages whenever a particular package is not Found

It loads the packages whether it is already loaded or not	It just checks that it is loaded, or loads it if it isn't (use in functions that rely on a certain package). The documentation explicitly states that neither function will reload an already loaded package.

16. What is R? (R)

R is a programming language which is used for developing statistical software and data analysis. It is being increasingly deployed for machine learning applications as well.

17. How R commands are written? (An)

By using # at the starting of the line of code like #division commands are written.

18. What is t-tests() in R? (R)

It is used to determine that the means of two groups are equal or not by using t.test() function.

19. What are the disadvantages of R Programming? (R)

The disadvantages are:-

- Lack of standard GUI
- Not good for big data.
- Does not provide spreadsheet view of data.

20. What is the use of With () and By () function in R? (R)

with() function applies an expression to a dataset.

```
#with(data,expression)
```

By() function applies a function t each level of a factors.

```
#by(data,factorlist,function)
```

21. What is the use of subset() and sample() function in R?(R)

Subset() is used to select the variables and observations and sample() function is used to generate a random sample of the size n from a dataset.

22. Explain what is transpose. (U)

Transpose is used for reshaping of the data which is used for analysis. Transpose is performed by t() function.

23. What are the advantages of R? (R)

- The advantages are:-
- It is used for managing and manipulating of data.
- No license restrictions
- Free and open source software.
- Graphical capabilities of R are good.
- Runs on many Operating system and different hardware and also run on 32 & 64 bit processors etc.

26. What is the function used for adding datasets in R? (R)

For adding two datasets rbind() function is used but the column of two datasets must be same.

Syntax: rbind(x1,x2……) where x1,x2: vector, matrix, data frames.

27. How you can produce co-relations and covariances? (An)

Cor-relations is produced by cor() and covariances is produced by cov() function.

28. What is difference between matrix and dataframes? (An)

Dataframe can contain different type of data but matrix can contain only similar type of data.

29. What is difference between lapply and sapply? (U)

lapply is used to show the output in the form of list whereas sapply is used to show the output in the form of vector or data frame

30. What is the difference between seq(4) and seq_along(4)? (U)

Seq(4) means vector from 1 to 4 (c(1,2,3,4)) whereas seq_along(4) means a vector of the length(4) or 1(c(1)).

31. Explain how you can start the R commander GUI.(U)

rcmdr command is used to start the R commander GUI.

32. What is the memory limit of R?(R)

In 32 bit system memory limit is 3Gb but most versions limited to 2Gb and in 64 bit system memory limit is 8Tb.

33. **What is Key value data store? (R)** NOVEMBER /DECEMBER 2020/ APRIL / MAY 2021

A key-value database is a type of non-relational database that uses a simple key-value method to store data. A key-value database stores data as a collection of key-value pairs in which a key serves as a unique identifier. Both keys and values can be anything, ranging from simple objects to complex compound objects.

34. **What is Graph database? (R) NOVEMBER /DECEMBER 2020/ APRIL / MAY 2021**

A graph database is defined as a specialized, single-purpose platform for creating and manipulating graphs. Graph analytics is another commonly used term, and it refers specifically to the process of analyzing data in a graph format using data points as nodes and relationships as edges.

PART-B

1. **Describe** what is NoSQL. & **List** the advantages and disadvantages of NoSQL. **(U)**
2. **Illustrate** in detail about Hive data manipulation, queries, and data types **(R)**
3. **Describe** the system architecture and components of Hive and Hadoop. **(R)**
4. **Explai**n briefly on aggregate data models. **(U)**
5. **Generalize** Pig and Pig Latin in detail**(R)**
6. **Describe** about HBase in detail**.(U)**
7. Explain Hbase clients in detail**.(U)**
8. **Analyse** how Cassandra is integrated with Hadoop**.(An)**
9. **Explain** the tools related to Hadoop. **(U)**
10. **Summarize** briefly on Hbase architecture with neat diagram **(U)**
11. **Quote** short notes on **(U)**
    a) Conceptual data modeling
    b) Logical data modeling.
    c) Physical data modeling

12. **Discuss** about Cassandra clients. **(U)**
13. **Compare** and Contrast the Hadoop and MapR**.(An)**
14. **Explain** about Grunt in detail. **(U)**
15. **Describe** about Pig data model in detail with neat diagram**.(U)**
16. **Assess** the difference between hive and map reduce**(An)**
17. **Explain about** Hive in detail**.(U)**
18. **Analyse** the use of Hive. How Does Hive Interact With Hadoop explain in detail? **(An)**
19. Write short notes on features of Hive and Sharding **(8) (U) NOVEMBER /DECEMBER 2020/ APRIL MAY 2021**
20. Explain the impact of Big data on the Blogs **(5) (U) NOVEMBER /DECEMBER 2020/ APRIL / MAY 2021**
21. Explain the following Statistical Methods for Evaluation in R **(13) (U) NOVEMBER /DECEMBER 2020/ APRIL / MAY 2021**
    1. Hypothesis Testing
    2. Difference of means
    3. Type I and Type II errors
    4. Power and Sample size
    5. ANOVA
22. Consider the E-Commerce Recommendation System. Analyze and indicate suitability of the type of Recommendation system and explain the same. **(15) (An) NOVEMBER /DECEMBER 2020/ APRIL / MAY 2021**
23. Analyze and explain the Real time analytics platform for Sentiment analysis in Tweets (**15) (An) NOVEMBER /DECEMBER 2020/ APRIL / MAY 2021**

COURSE OUTCOME: Students can able to learn NoSQL databases and management.

COURSE NAME : CS8091 – BIG DATA ANALYTICS

YEAR/SEMESTER : III/ VI

YEAR OF STUDY : 2021 –2022 EVEN (R – 2017)

On Completion of this course student will gain

CS8091.1	An ability to work with big data tools and its analysis techniques
CS8091.2	An ability to analyse data by utilizing clustering and classification algorithms
CS8091.3	An ability to learn and apply different mining algorithms and recommendation systems for large volumes of data
CS8091.4	An ability to perform analytics on data streams
CS8091.5	An ability to learn NoSQL databases and management.

CO-PO MATRIX:

CO	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12
CS8091.1	3	3	3	3		2	-	-	-	-	-	-
CS8091.2	3	3	3	3		2	-	-	-	-	-	-
CS8091.3	3	3	3	3		2	-	-	-	-	-	-
CS8091.4	3	3	3	3		2	-	-	-	-	-	-
CS8091.5	3	3	3	3		2	-	-	-	-	-	-
CS8091	3	3	3	3		2	-	-	-	-	-	-

CO-PSO MATRIX:

CO	PSO1	PSO2	PSO3
CS8091.1	2	-	2
CS8091.2	2	-	2
CS8091.3	2	-	2
CS8091.4	2	-	2
CS8091.5	2	-	2
CS8091	2	-	2