VISVESVARAYA TECHNOLOGICAL UNIVERSITY
BELAGAVI

BIG DATA ANALYTICS

(17CS82)

(As per Visvesvaraya Technological University Syllabus)

Complied By:

Prof. Akshatha Ballal.

Assistant Professor, Dept. of ISE

DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING
ACHARYA INSTITUTE OF TECHNOLOGY

(Affiliated to VTU, Belagavi, Approved by AICTE, New Delhi and Accredited by NAAC, New Delhi)
Acharya Dr. Sarvepalli Radhakrishnan Road, Bangalore-560107**.**
URL: www.acharya.ac.in

2020-21

Disclaimer

Trademark

Edition: 2020- 21

Document Owner

The primary contact for questions regarding this document is:

Author(s):	Prof. Akshatha Ballal.
Department:	Information Science & Engineering
Contact Email(s):	akshathab@acharya.ac.in

MODULE-1

1. Write a note on HDFS Design Features.
2. Mention the areas where HDFS do not work well.
3. What is HDFS? Explain various components of HDFS. **(10 Marks)[July 2019/Jan 2020]**
4. Explain various components of HDFS.
5. Explain HDFS block replication with example.**(6 Marks)[July 2019](8 Marks) [Sep 2020]**
6. Write a note on a. HDFS safe mode b. Rack awareness.
7. Explain namenode high availability.
8. Explain HDFS Namenode federation, NFS gateway, snapshots and e. checkpoints and backups. **(8 Marks)[July 2019]**
9. Define any 8 HDFS user commands. **(7 Marks) [Jan 2020]**
10. What are the various system roles in an HDFS development? Explain with a neat diagram. **(8 Marks) [Sep 2020]**
11. Write the command to

    a. Make a Directory in HDFS
    b. Copy Files to HDFS
    c. Copy Files from HDFS
    d. Copy Files within HDFS
    e. Delete a File within HDFS
    f. Delete a Directory in HDFS
    g. Get an HDFS Status Report

11. What is MapReduce model? What are the features of MapReduce model?
12. Explain Parallel MapReduce dataflow model with neat diagram. **(7 Marks) [Jan 2020](8Marks)[Sep 2020]**
13. Explain a simple three-node MapReduce process with a neat diagram.
14. Write a short notes on a. Fault Tolerance and Speculative Execution b. Hadoop mapreduce hardware.
15. Write the code for simple mapper and simple reducer script. **(8 Marks)[Sep 2020**]
16. Write the Java code for MAP and REDUCE of word count problem. Describe the steps for compiling and removing the mapreduce program. **(9 Marks) [Jan 2020]**
17. How does Hadoop Mapreduce Data flow work for a word count program? Give example. **(8 Marks)[July 2019]**

MODULE-2

1. Write a short note on Apache Pig.
2. What is Apache Sqoop? Explain Apache Sqoop Import and Export Method with a neat diagram **(10 Marks)[July 2019] (8 Marks)[Jan 2020/ Sep 2020]**
3. Explain HBase Architecture.
4. How do you run Map Reduce and Message Passing Interface(MPI) on YARN architecture. Discuss. **(10 Marks)[July 2019]**
5. Explain the different frameworks that run under YARN with a neat diagram. **(8 Marks)[Jan 2020/ Sep 2020]**
6. What do you understand by YARN Distributed shell. **(6 Marks)[July 2019]**

7. Discuss the different views supported by Apache Ambari. **(6 Marks)[Jan 2020]**
8. Explain the Apache Ambari dashboard view of a hadoop cluster. **(8 Marks)[Sep 2020]**
9. Describe the various features of Hadoop YARN Administration. **(4 Marks)[Jan 2020]**
10. Explain the different HDFS administration features. **(6 Marks)[Jan 2020]**
11. How the basic YARN administration is carried out? Explain. **(8 Marks)[Sep 2020]**
12. Explain with diagram, the Apache Oozie workflow for Hadoop Architecture. **(6 Marks)[July 2019]**

MODULE-3

1. What is BI? List the applications and explain any 5 in detail. **(10 Marks)[Jan 2020]**
2. What are the applications of BI for various sectors? **(8 Marks)[July 2019/Sep 2020]**
3. What are the different types of decision? How does BI help take better decision?
4. Write a note on BI tools and BI skills.
5. What is data mining? What are supervised and unsupervised learning techniques?
6. What are various Tools and Platforms for Data Mining?
7. What are the range of data mining platforms available in the market today?
8. Compare Excel, IBM SPSS, Modeler ,Weka
9. How do you Evaluate Data Mining Results? Explain with Confusion Matrix. **(8 Marks)[Sep 2020]**
10. Explain the data cleaning techniques
11. What are the best practices of Data Mining?
12. What are the Myths of DM?
13. Describe the key steps in the data mining process. Why is it important to follow these processes?
14. What is a confusion matrix? **(2 Marks)[July 2019]**
15. Why is data preparation so important and time consuming? Describe the common data mining mistakes. **(4 Marks)[Jan 2020]**
16. What are the key requirements for a skilled data analyst?
17. What is Data Visualization? Explain it's importance in Big Data Analytics. **(5 Marks)[July 2019]**
18. What are the various Tips for Data Visualization?
19. Differentiate between Data Mining and Data warehousing. **(3 Marks)[July 2019]**
20. Describe the different types of charts used for Data Visualization. **(4 Marks)[Jan 2020]**
21. Explain with diagram the different types of graphs. **(8 Marks)[Sep 2020]**
22. What are some key requirements for good visualization.
23. What is Data Warehousing ? what are the requirements of a good Data Warehousing.
24. Compare Data Mart and Data Warehouse.
25. Explain the data warehouse architecture with a neat diagram **(6 Marks)[Jan 2020]**
26. Explain the star schema design of data warehousing with an example. **(6 Marks)[July 2019] (8 Marks)[Sep 2020]**
27. What are the different types of Data Sources?
28. Explain Crisp-DM cycle with a neat diagram. **(8 Marks)[July 2019/Jan 2020]**
29. Explain Extract-Transform-Load (ETL) cycle.
30. Write a note on a. DW Design, b. DW Access, c. DW Best Practices.
31. How will data warehousing evolve in the age of social media?

MODULE- 4

1. What is a decision tree? What are the key factors of a good decision tree?
2. Explain with a dataset how to construct a decision tree. **(8 Marks)[Sep 2020]**
3. Explain the different steps in constructing a decision tree for the example given in Page 116. **(8 Marks)[Jan 2020]**
4. Compare decision tree and look up table.
5. Write a pseudo-code for DT algorithm. What are the 3 key elements of DT algorithm based on which they differ?
6. What is a splitting variable? Describe the criteria for choosing a splitting variable. **(4 Marks)[July 2019]**
7. What is regression and its objective ? what are the key steps ?
8. Write a note on a.Logistic regression b. correlations and relationships c.scatter plot for regression
9. Create a Regression model for the dataset given to predict Test 2 score from Test 1 score. Then predict the score for one who got 46 in Test 1.

Test 1	Test 2
59	56
52	63
44	55
51	50
42	66
42	48
41	58
45	36
27	13
63	50
54	81
44	56
50	64
47	50

10. Create a decision tree for the data set given (Refer 4th Review question page 130-131) **(8 Marks)[July 2019]**
11. What are the advantages and disadvantages of a.Regression b.ANN c.Cluster analysis ?
12. Describe the advantages and disadvantages of Regression model. **(4 Marks)[July 2019](8 Marks)[Jan 2020]**
13. What is ANN? What are the applications of ANN?
14. Explain the Design Principles of an Artificial Neural Network . **(8 Marks)[July 2019/Sep 2020]**
15. How do you represent a Neural network ?
16. Describe the advantages of ANN **(3 Marks)[Jan 2020]**
17. Explain the steps required to build an ANN. **(5 Marks)[Jan 2020]**
18. Define Cluster and cluster analysis. What are the aplications of CA?
19. Write the pseudo-code for generic cluster analysis.
20. Explain K-Means algorithm with example.
21. Write the advantages and disadvantages of K-Means algorithm. **(4 Marks)[Sep 2020]**
22. What are the different techniques of CA?
23. What is association rule mining? What are the applications of it?
24. How association rules are represented? **(4 Marks)[Sep 2020]**
25. Write a short note on Apriori algorithm

26. Describe the different steps for forming Association Rules using Apriori Algorithm for the example in Page 197.
27. How does the Apriori algorithm work. Apply the same for the example.

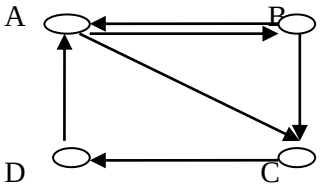
TID	List of item-ID
T_{100}	I1,I2,I5
T_{200}	I2,I4
T_{300}	I2,I3
T_{400}	I1,I2,I4
T_{500}	I1,I3
T_{600}	I2,I3
T_{700}	I1,I3
T_{800}	I1,I2,I3,I5
T_{900}	I1,I2,I3,

Assume the support count=2 **(8 Marks)[July 2019]**

28. Given are a dozen sales transactions. The objective is to use this transaction data to find affinities between products, i.e. which products sell together often by creating the association rule.

MODULE-5

1. Discuss the differences between Text Mining and Data Mining. **(6 Marks)[Jan 2020]**
2. What are the types of web mining? Explain. **(6 Marks)[July 2019]**
3. Explain the 3 step process of Text Mining? **(3 Marks)[July 2019]**
4. Explain with a neat diagram the text mining process. **(8 Marks)[Sep 2020]**
5. Discuss the applications and practical considerations of Social Network analysis ? **(8 Marks)[Jan 2020]**
6. Describe the SVM model with diagram **(4 Marks)[Jan 2020/Sep 2020]**
7. What is support vector machine? Explain its model. **(8 Marks)[July 2019]**
8. Explain different techniques and algorithms used in SNA?
9. What are the practical considerations to guard against the pitfall of SNA?
10. Compare SNA and traditional data analytics.
11. Explain the applications of SNA. **(8 Marks)[Sep 2020]**
12. Explain kernel method.
13. Compute the rank values for the nodes of the following network shown in figure. Which is the highest rank node? Solve with eight iterations. **(10 Marks)[July 2019]**



14. What is web mining? What are the different types of web mining? **(8 Marks)[Jan 2020]**
15. Explain with diagram the web usage mining architecture. **(8 Marks)[Sep 2020]**
16. What is click stream analysis?
17. What are the ways by which website can become popular?
18. What is Naïve Bayes technique? Explain its model? **(5 Marks)[July 2019]**

19. Explain Naive Bayes model to classify text data into right class using the dataset given below

Training set	Document ID	Keyword in the document	Class= H(healthy)
	1	Love Happy Joy Joy Love	Yes
	2	Happy Love KICK JOY Happy	Yes
	3	Love Move Joy Good	Yes
	4	Love Happy Joy Pain Love	Yes
	5	Joy Love Pain Kick Pain	No
	6	Pain Pain Love Kick	No
Test Data	7	Love Pain Joy Love Kick	?

(6 Marks)[Jan 2020]

20. Write the advantages and disadvantages of Naïve Bayes technique. **(4 Marks)[Sep 2020]**
21. What are the different network topologies considered in SNA?

