**Module 1**

1. Write the features of HDFS design.  
   2. Explain all the components of HDFS with diagram.  
   3. Explain HDFS block replication.  
   4. Explain HDFS safe mode and rack awareness.  
   5. Explain name node high availability design.  
   6. Explain HDFS snapshots and HDFS NFS gateway.  
   7. Write any five HDFS user commands.

8. Write all the steps to execute terasort basic hadoop benchmark.  
9. Explain mapreduce parallel data flow with near diagram.  
10. Write a program using streaming interface to count the words of file.  
11. Write a java mapreduce program to count the words of file.

**Module 2**

1. Explain Apache pig with latin script commands.  
2. Explain Apache hive with minimum 5 of hive query language commands.   
3. Explain Apache sqoop import and export methods with neat diagram.  
4. Explain how to import data from local file system to HDFS using sqoop commands.   
5. Explain Apache flume to acquire real data streams with diagrams.  
6. Explain managing hadoop workflows with Apache oozie.  
7. Explain Apache Hbase with example commands.  
8. Explain distributed shell with commands.  
9. Explain structure of YARN applications.  
10. Explain YARN application framework.  
11. Explain Apache spark and Apache REEF.  
12. Write a short note on Apache Ambari.  
13. Explain the setting of container memory and container cores.

**Module 3**

1. Explain BIDM cycle with diagram.  
2.  Explain BI types and tools.ss  
3. Explain the applications of BI in CRM and healthcare and wellness.  
4. Define data warehouse and write the design considerations for DW.  
5. Explain DW architecture with DW development approaches.  
6. Define data mining and explain gathering and selecting data.  
7. Explain data cleansing and preparation.  
8. Explain evaluating data mining results.  
9. Explain 5 important data mining techniques.  
10. Define data visualization and explain the types of  charts.

**Module 4**

1. Define decision tree and write the algorithm to construct it.  
2. Create the decision trees with data set given below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Outlook | Temp | Humidity | Windy | Play |
| Sunny | Hot | High | False | No |
| Sunny | Hot | High | True | No |
| Overcast | Hot | High | False | Yes |
| Rainy | Mild | High | False | Yes |
| Rainy | Cool | Normal | False | Yes |
| Rainy | Cool | Normal | True | No |
| Overcast | Cool | Normal | True | Yes |
| Sunny | Mild | High | False | No |
| Sunny | Cool | Normal | False | Yes |
| Rainy | Mild | Normal | False | Yes |
| Sunny | Mild | Normal | True | Yes |
| Overcast | Mild | High | True | Yes |
| Overcast | Hot | Normal | False | Yes |
| Rainy | Mild | High | True | No |

3. Explain simple and non linear regression with graphs.  
4.  Explain logistic regression.  
5. Define ANN and explain the design principles of an artificial neural networks.  
6. Define cluster analysis and explain K-means for clustering with algorithm.  
7. Explain association rule mining with business applications.  
8. Explain representing association rules.  
9. Using Apriori algorithm create the association rules with following data set.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Transaction List | | | | |
| 1 | Milk | Egg | Bread | Butter |
| 2 | Milk | Butter | Egg | Ketchup |
| 3 | Bread | Butter | Ketchup |  |
| 4 | Milk | Bread | Butter |  |
| 5 | Bread | Butter | Cookies |  |
| 6 | Milk | Bread | Butter | Cookies |
| 7 | Milk | Cookies |  |  |
| 8 | Milk | Bread | Butter |  |
| 9 | Bread | Butter | Egg | Cookies |
| 10 | Milk | Butter | Bread |  |
| 11 | Milk | Bread | Butter |  |
| 12 | Milk | Bread | Cookies | Ketchup |