Sr.No.	Problem Statement
1	Design a distributed application using MapReduce which processes a log file of a system. List out the users who have logged for maximum period on the system. Use simple log file from the Internet and process it using a pseudo distribution mode on Hadoop platform.
2	Design and develop a distributed application to find frequency of words from sample text data. Use sample text data and process it using MapReduce.
3	Design a distributed application using MapReduce which processes Music dataset. List out the number of unique listeners and no of times the track was shared with others. Use music dataset and process it using a pseudo distribution mode on Hadoop platform.
4	Design a distributed application using MapReduce which processes Music dataset. List out the number of times the track was listened on Radio and no of times the track was skipped. Use music dataset and process it using a pseudo distribution mode on Hadoop platform.
5	Design a distributed application using MapReduce which processes Movie dataset. Recommend the Movie based on the user ratings. Use Movie dataset and process it using a pseudo distribution mode on Hadoop platform.
6	Write an application using HBase and HiveQL for flight information system which will include a. Create Flight Info Hbase Table(with Flight information, schedule, and delay) b. Demonstrate Creating, Dropping, and altering Database tables in Hbase c. Creating an external Hive table to connect to the HBase for Flight Information Table d. Find the total departure delay in Hive e. Find the average departure delay in Hive f. Create index on Flight information Table
7	Write an application using HBase and HiveQL for Customer information system which will include a. Creation of –Cutomer_info(Cust-ID,Cust-Name,orderID), order_info(OrderID,ItemID,Quantity), item_info(Item-ID,Item-Name,ItemPrice) tables in Hive b. Load table with data from local storage in Hive. c. Perform Join tables with Hive d. Create Index on Customer information system in Hive. e. Find the total, average sales in Hive f. Find Order details with maximum cost. g. Creating an external Hive table to connect to the HBase for Customer Information System. h. Display records of Customer Information Table in Hbase.
8	Write an application using HBase and HiveQL for OnlineRetail Dataset which will include i. Create and Load table with Online Retail data in Hive. j. Create Index on Online Retail Table in Hive. k. Find the total, average sales in Hive l. Find Order details with maximum cost. m. Find Customer details with maximum order total.

	n. Find the Country with maximum and minimum sale. o. Creating an external Hive table to connect to the HBase for OnlineRetail. p. Display records of OnlineRetail Table in Hbase.
9	Perform the following operations using Python on the Facebook metrics data sets a. Create data subsets for type of post b. Merge two subsets c. Sort Data on Page total likes d. Transposing Data e. Melting Data to long format f. Casting data to wide format
10	Perform the following operations using Python on the Iris data sets g. Create data subsets for different species h. Merge two subsets i. Sort Data Petal Length j. Transposing Data k. Melting Data to long format l. Casting data to wide format
11	Perform the following operations using Python on Movie data sets m. Create data subsets for different languages(Original Language). n. Merge two subsets o. Sort Data using customer ratings. p. Transposing Data q. Melting Data to long format r. Casting data to wide format
12	Perform the following operations using Python on census bureau databset(Adult data sets). s. Create data subsets for different Country, Sex, race. t. Merge two subsets u. Sort Data using customer ratings. v. Transposing Data w. Melting Data to long format x. Casting data to wide format
13	Perform the following operations using Python on Heart Diseases data sets a. Data cleaning(Remove NA, ?, Negative values etc.) b. Error correcting(Outlier detection and removal) c. Data transformation d. Build Data model using regression and kNN methods and compare accuracy of heart disease prediction.
14	Perform the following operations using Python on Iris data sets e. Data cleaning(Remove NA, ?, Negative values etc.) f. Error correcting(Outlier detection and removal) g. Data transformation

	h. Build Data model using regression and Naïve Bayes methods and compare accuracy of Iris Species Prediction.
15	Perform the following operations using Python on Breast Cancer data sets i. Data cleaning(Remove NA, ?, Negative values etc.) j. Error correcting(Outlier detection and removal) k. Data transformation l. Build Data model using regression and Naïve Bayes methods and compare accuracy of benign and malignant tumors in Breast Cancer Dataset.
16	Perform the following operations using Python on census bureau databset(Adult data sets) m. Data cleaning(Remove NA, ?, Negative values etc.) n. Error correcting(Outlier detection and removal) o. Data transformation p. Build Data model using regression and Naïve Bayes methods for prediction of income category (>=50k or <=50k) and compare accuracy Prediction.
17	Visualize the Heart disease dataset by plotting the following graphs using Python. (Define objective for every graph) a. Histograms b. Dot Plots c. Bar Plots d. Line Charts e. Add Histogram and Scatter plot to box plot.
18	Visualize the Heart disease dataset by plotting the following graphs using Python. (Define objective for every graph) a. Histograms b. Pie Charts c. Box Plots d. Scatter Plots e. Add boxplots to a scatterplot
19	Perform the data visualization operations using Tableau to get answers to various business questions on Retail dataset. a. Find and Plot top 10 products based on total sale b. Find and Plot product contribution to total sale c. Find and Plot the month wise sales in year 2010 in descending order d. Find and Plot most loyal customers based on purchase order e. Find and Plot yearly sales comparison f. Find and Plot country wise total sales price and show on Geospatial graph
20	Perform the data visualization operations using Tableau to get answers to various business questions on Retail dataset. a. Find and Plot country wise popular product b. Find and Plot bottom 10 products based on total sale c. Find and Plot top 5 purchase order d. Find and Plot most popular products based on sales e. Find and Plot half yearly sales for the year 2011

	f. Find and Plot country wise total sales quantity and show on Geospatial graph
21	Visualize the census bureau databset(Adult data sets)by plotting the following graphs using Python. (Define objective for every graph) f. Histograms g. Dot Plots h. Bar Plots i. Line Charts j. Add Histogram and Scatter plot to box plot.
22	Visualize the census bureau databset(Adult data sets)by plotting the following graphs using Python. (Define objective for every graph) a. Histograms b. Pie Charts c. Box Plots d. Scatter Plots e. Add boxplots to a scatterplot
23	Perform the data visualization operations using Tableau to get answers to various questions on the census bureau databset(Adult data sets). a. Find and Plot Income class of People whose education is master's and doctorate. b. Find and Plot Income class of people who have private jobs. c. Find and Plot yearly sales comparison d. Find and Plot country wise statistics on Geospatial graph e. Plot agewise- education vs salary statistics. f. Plot Countrywise male female ratio. g. Plot Income class based on workclass(Government and other)
24	Perform the following operations using Python on ForestFires Dataset. a. Create data subsets by making classes for amount of region affected.(e.g. NotAffected, Partially affected, Mostlyaffected). b. Merge two subsets c. Sort Data using Temperature, wind and area. d. Transposing Data e. Melting Data to long format f. Casting data to wide format
25	Perform the following operations using Python on Hepatitis Dataset. a. Create data subsets for different sex. b. Merge two subsets c. Sort Data using age, SGOT, PROTIME. d. Transposing Data e. Melting Data to long format f. Casting data to wide format

26	Perform the following operations using Python on Hepatitis dataset. q. Data cleaning(Remove NA, ?, Negative values etc.) r. Error correcting(Outlier detection and removal) s. Data transformation t. Build Data model using regression and Naïve Bayes methods for prediction class DIE, LIVE and compare accuracy Prediction.