Data Analytics (IT – 3006)

Assignment

Unit 1

Q1. Why traditional databases cannot store big data?

Q2. A finance company wants to evaluate their users, on the basis of loans they have taken. They have hired you to find the number of cases per location and categorize the count with respect to the reason for taking a loan. Next, they have also tasked you to display their average risk score. Discuss and then model your views concerning descriptive and predictive analytics.

Q3. A retail company wants to enhance their customer experience by analysing the customer reviews for different products, so that they can inform the corresponding vendors and manufacturers about the product defects and shortcomings. You have been tasked to analyse the complaints filed under each product & the total number of complaints filed based on the geography, type of product, etc. You also have to figure out the complaints which have no timely response. Discuss and then model your views concerning descriptive, diagnostic and predictive analytics.

Q4. A mobile health organisation captures patient's physical activities by attaching various sensors on different body parts. These sensors measure the motion of diverse body parts like the acc., the rate of turn, the magnetic field orientation etc. A model will be built for effectively deriving information about the motion of different body parts like chest, ankle etc. Discuss and then model your views concerning diagnostic and predictive analytics.

Q5. Discuss some effective ways to measure the fault tolerance technique to increase the reliability of a private, public and hybrid cloud.

Unit 2

Q1. The management of a chain of medical store wants to investigate the relationship between the daily sales volume of its store and the number of competitor medical stores within 1mile radius. The data shown in Table 1 has been collected. Draw a scatter diagram to examine whether a relationship exists between the number of competitors and the volume of sales. Once the scatter diagram has been produced, the solution should interpret it.

Table 1: Competitor with sale volume

Competitor	Sale
1	3600
1	3300
2	3100
3	2900
3	2700

5	2300
5	2000
6	1800

Q2. In reference to Table 1, use correlation analysis to examine whether a relationship exists between the number of competitors and the volume of sales. The solution should interpret the findings.

Q3. In reference to Table 1, develop a linear regression model that would relate the volume of sales to the number of competitors. What is the accuracy of the model?

Q4. Consider the hypothetical data (shown in Table 2) concerning student characteristics whether or not each student should be hired. Use Naive Bayes Classifier (NBC) to determine whether or not someone with excellent attendance, poor GPA and lots of effort should be hired.

Table 2: Hypothetical data

Name	GPA	Effort	Hirable
Sarah	Poor	Lots	Yes
Dana	Average	Some	No
Alex	Average	Some	No
Annie	Average	Some	Yes
Emily	Excellent	Lots	Yes
Pete	Excellent	Lots	No
John	Excellent	Lots	No
Kathy	Poor	Some	No

Q5. In reference to the Table 3, exponential smoothing is used to forecast automobile battery sales. Tow values of α are examined, α = 0.8 and α = 0.5. Evaluate the accuracy for each smoothing constant and which is preferable and why?

Table 3: Battery Sales

Month	Actual	Forecasted
January	20	22
February	2	
March	15	
April	14	

May	13	
June	16	
July	17	
August	18	
September	21	
October	20	
November	22	

A consumer electronics company has adopted an aggressive policy to increase sales of a newly launched product. The company has invested in advertisements as well as employed salesmen for increasing sales rapidly. Table 7 presents the sales, the number of employed salesmen, and advertisement expenditure for 12 randomly selected months. Develop a regression model to predict the impact of advertisement and the number of salesmen on sales.

Table 7: Cable wire company's sales and advertisement expenses

Month No	1	2	3	4	5	6	7	8	9	10	11	12
Sales	5000	5200	5700	6300	6000	6400	6100	6400	6900	7300	6950	7350
Salesmen	25	35	15	27	20	11	8	11	29	31	6	10
Advertisement	180	250	150	240	185	160	177	315	170	240	184	218

Q.6. A company wants to test the effect of age and gender on the productivity (in terms of units produced by the employees per month) of its employees. The HR manager has taken a random sample of 15 employees and collected information about their age and gender. Table 8 provides data about the productivity, age, and gender of 15 randomly selected employees. Fit a regression model considering productivity as the dependent variable and age and gender as the independent variables. The, predict the productivity of male and female at 45 years of age.

Table 8: Productivity random sample

Employee	Productivity	Age	Gender
1	850	40	Male
2	760	34	Female
3	750	28	Female
4	860	34	Male

5	800	38	Female
6	710	26	Male
7	760	31	Male
8	860	38	Male
9	770	31	Male
10	800	30	Male
11	870	38	Male
12	800	28	Male
13	750	31	Female
14	840	37	Male
15	760	31	Female

Unit 3

Q1. With respect to data stream querying, give at least 3 examples of - One-time queries - Continuous queries - Predefined queries - Ad hoc queries

Q2. For each of the following applications, explain whether a Bloom filter would be an appropriate choice. Explain in required details. - Checking whether an object is stored in cache in a distributed system. - Storing a list of individuals on a “Cannot-Visit” list at a museum. - Testing whether a required web page has been listed in blocked list of web site. - In an e-voting system checking whether an attempt is made to case a duplicate vote.

Q3. A bloom filter with a size of 1000 slots is used to store the information of 100 items using 4 hash functions. Calculate the false positive probability of this instance. Will the performance improve by increasing the number of hash functions from 4 to 5? Explain in required details.

Q4. Identify at least 5 platforms for real-time analytics and provide comparison chart for the same.

Q5. Suppose the stream is a, b, c, b, d, a, c, d, a, b, d, c, a, a, a, b. Calculate the distinct value using - Traditional method. Outline the necessary precondition. - Flajolet-Martin algorithm

Unit-4

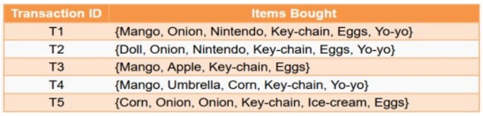
Q1. Suppose there are 100 items, numbered 1 to 100, and also 100 baskets, also numbered 1 to 100. Item i is in basket b if and only if i divide b with no remainder. Thus, item 1 is in all the baskets, item 2 is in all fifty of the even-numbered baskets, and so on. Basket 12 consists of items {1, 2, 3, 4, 6, 12}, since these are all the integers that divide 12. Answer the following questions: - If the support threshold is 5, which items are frequent? - If the support threshold is 5, which pairs of items are frequent? - What is the sum of the sizes of all the baskets? - Which basket is the largest?

Q2. Suppose there are 100 items, numbered 1 to 100, and also 100 baskets, also numbered 1 to 100. Item i is in basket b if and only if b divides i with no remainder. For example, basket 12 consists of items {12, 24, 36, 48, 60, 72, 84, 96}. Repeat the same exercise of

Q3 For the data of Q2, what is the confidence of the following association rules? - {5, 7} → 2. - {2, 3, 4} → 5. Q16. For the data of Q2, what is the confidence of the following association rules? - {24, 60} → 8 - {2, 3, 4} → 5.

Q4. Describe association rules that have 100% confidence for the market-basket data of Q3 and Q2.

Q4. Consider the following transactional data in which minimum support is 2 and minimum confidence is 50%. Find frequent item sets and generate association rules for them. Illustrate it with step-by-step process.



Transaction ID	Items Bought
T1	{Mango, Onion, Nintendo, Key-chain, Eggs, Yo-yo}
T2	{Doll, Onion, Nintendo, Key-chain, Eggs, Yo-yo}
T3	{Mango, Apple, Key-chain, Eggs}
T4	{Mango, Umbrella, Corn, Key-chain, Yo-yo}
T5	{Corn, Onion, Onion, Key-chain, Ice-cream, Eggs}

Unit 5.

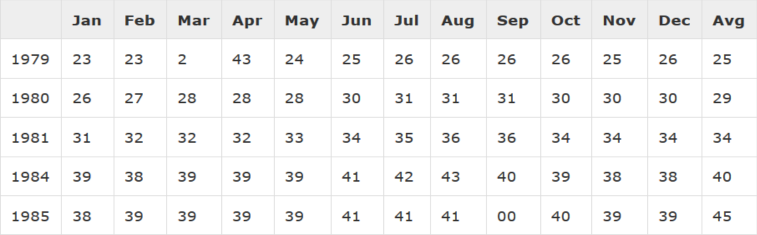
Q1. Describe the working of Map Reduce model. Perform the Map Reduce task for the following input files containing the following data.

Input File 1	Input File 2	Input File 3
Apple Orange Mango Orange Grapes Plum	Apple Plum Mango Apple Apple Plum	Apple Orange Mango Plum Apple Grapes

Q2. Draw the Euler diagram and Venn diagram for the sets, X = {1, 2, 5, 8}, Y = {1, 6, 9} and Z = {4, 7, 8, 9}.

Q3. Consider the following data. Draw the Map Reduce process to find the maximum electrical consumption for:

- Each year.
- Each month.



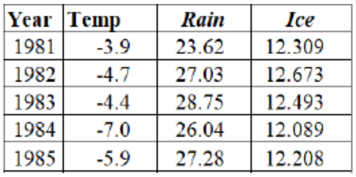
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Avg
1979	23	23	2	43	24	25	26	26	26	26	25	26	25
1980	26	27	28	28	28	30	31	31	31	30	30	30	29
1981	31	32	32	32	33	34	35	36	36	34	34	34	34
1984	39	38	39	39	39	41	42	43	40	39	38	38	40
1985	38	39	39	39	39	41	41	41	00	40	39	39	45

Q.4. Explain CAP Theorem (also called as Brewer's Theorem) and prove it.

Q.5. What is NoSQL? How it is different from traditional SQL? Explain the four types of schemas for NOSQL along with its examples.

Q.6. Explain data locality with suitable examples. Explain the difference between "moving computation" and "moving data" in a cluster.

Q.7. Consider the following multivariate climate data observed from 1981 to 1990.



Year	Temp	*Rain*	*Ice*
1981	-3.9	23.62	12.309
1982	-4.7	27.03	12.673
1983	-4.4	28.75	12.493
1984	-7.0	26.04	12.089
1985	-5.9	27.28	12.208

- Draw the parallel coordinate plot, wherein min value of temp is -5.0 and max value is -2.0.
- Draw the chronological display of rain.
- Draw an ordinogram illustrating the relationship between rain and ice.