USN

Internal Assessment Test 1 – March 2019

Sub:	Big Data Analytics					Sub Code:	15CS82	Branch:	CSE
Date:	6/03/2019	Duration:	90 min's	Max Marks:	50	Sem / Sec:	8th A,B,C	OBE	

Answer any FIVE FULL Questions

		MARKS	CO	RBT
1	Define business intelligence and explain the BIDM cycle. List the requirements of a good data warehouse	[10]	CO3	L1
2 (a)	Differentiate between operational and strategic decisions with example.	[05]	CO3	L2
(b)	Explain correlations among data elements and define correlation coefficient.	[05]	CO4	L3
3	Describe the data warehouse architecture? What is ETL? List steps of ETL process.	[10]	CO3	L3
4	What is the concept of data transformation in a data warehouse? What is data visualization? Name some data visualization tools for presenting data reports?	[4+2+4]	CO3	L2
5	What are unsupervised and supervised learning techniques? Explain in detail	[10]	CO3	L2
6	Create a decision tree for the data given below. The Objective is to predict the class category (play tennis or not)	[10]	CO4	L3

Day	Outlook	Temp	Humidity	Wind	Tennis?
D1	Sunny	Hot	High	Weak	*No*
D2	Sunny	Hot	High	Strong	*No*
D3	Overcast	Hot	High	Weak	*Yes*
D4	Rain	Mild	High	Weak	*Yes*
D5	Rain	Cool	Normal	Weak	*Yes*
D6	Rain	Cool	Normal	Strong	*No*
D7	Overcast	Cool	Normal	Strong	*Yes*
D8	Sunny	Mild	High	Weak	*No*
D9	Sunny	Cool	Normal	Weak	*Yes*
D10	Rain	Mild	Normal	Weak	*Yes*
D11	Sunny	Mild	Normal	Strong	*Yes*
D12	Overcast	Mild	High	Strong	*Yes*
D13	Overcast	Hot	Normal	Weak	*Yes*
D14	Rain	Mild	High	Strong	*No*

Then solve the following problem using the model.

Outlook->Sunny Temperature->hot Humidity->Normal Windy->yes Play->?

7. Forecast House prize based on Size in sq feet, use simple linear regression. CO4 L3

House Price	Size (sqft)
229500	1850
273300	2190
247000	2100
195100	1930
261000	2300
179700	1710
168500	1550
234400	1920
168800	1840
180400	1720
156200	1660
288350	2405
186750	1525
202100	2030
256800	2240

Course Outcomes		**Modules covered**	**PO1**	**PO2**	**PO3**	**PO4**	**PO5**	**PO6**	**PO7**	**PO8**	**PO9**	**PO10**	**PO11**	**PO12**	**PSO1**	**PSO2**	**PSO3**	**PSO4**
CO1	Master the concepts of HDFS and MapReduce framework	**1**	1	3	2	3	3	2	0	0	0	0	0	0	3	0	2	3
CO2	Investigate Hadoop related tools for Big Data Analytics and perform basic Hadoop Administration	**1,2**	0	3	2	2	3	2	0	0	0	0	0	0	3	0	2	3
CO3	Recognize the role of Business Intelligence, Data warehousing and Visualization in decision making	**3**	1	2	1	2	1	2	0	0	0	0	0	0	1	0	0	2
CO4	Infer the importance of core data mining techniques for data analytics	**3,4**	2	3	3	3	1	1	0	0	0	0	0	0	1	0	0	2
CO5	Analyze Data Mining Techniques	**3,5**	2	2	3	3	1	3	0	0	0	0	0	0	1	0	2	3

Note: Assignments, study material, Question bank and other course related content would be posted on site mentioned above.

Appendix

Table 01: Cognitive Levels

Cognitive Levels	
Cognitive level	**Revised Blooms Taxonomy Keywords**
L1	List, define, tell, describe, identify, show, label, collect, examine, tabulate, quote, name, who, when, where, etc.
L2	summarize, describe, interpret, contrast, predict, associate, distinguish, estimate, differentiate, discuss, extend
L3	Apply, demonstrate, calculate, complete, illustrate, show, solve, examine, modify, relate, change, classify, experiment, discover.
L4	Analyze, separate, order, explain, connect, classify, arrange, divide, compare, select, explain, infer.
L5	Assess, decide, rank, grade, test, measure, recommend, convince, select, judge, explain, discriminate, support, conclude, compare, summarize.

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Date: 28th July 2018

CO’s to PO’s & PSO’s mapping

Name of the course	:	**Big data analytics**	Sub Code	:	15CS82
Name of the Faculty/s	:	Mrs. Jagdishwari/ Mrs Poonam	Sem & Sec	:	8th A,B,C

SCHEME

Question #	**Description**	**Marks Distribution**		**Max Marks**
1	Business intelligence definition explain the BIDM cycle. List the requirements of a good data warehouse	2M 4M 4M	10M	10M
2 a	Differentiate between operational and strategic decisions with example.	5M	5M	5M
2 b	Explaination on correlations among data elements and define correlation coefficient.	5M	5M	5M
3	Data warehouse architecture ETL Steps of ETL process.	4M 2M 4M	10M	10M
4	concept of data transformation in a data warehouse Data visualization Name some data visualization tools for presenting data reports	2M 4M 2M	10M	10M
5	Unsupervised and supervised learning techniques in detail	5M 5M	10M	10M
6	Solve the following problem using the model.	10M	10M	10M
7	Forecast House prize based on Size in sq feet, use simple linear regression.	10 M	10M	10M

SOLUTION:

1> Business intelligence (BI) is an umbrella term that includes a variety of IT applications that are used to analyze an organization's data and communicate the information to relevant users.

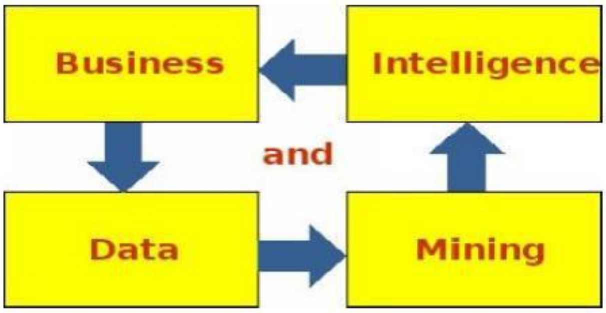

Fig 3.1: BIDM cycle

The nature of life and businesses is to grow. Information is the life-blood of business. Businesses effective than those based on feelings alone. Actions based on accurate data, information, knowledge, experimentation, and testing, using fresh insights, can more likely succeed and lead to sustained growth. One's own data can be the most effective teacher. Therefore, organizations should gather data, sift through it, analyse and mine it, find insights, and then embed those insights into their operating procedures.

There is a new sense of importance and urgency around data as it is being viewed as a new natural resource. It can be mined for value, insights, and competitive advantage. In a hyperconnected world, where everything is potentially connected to everything else, with potentially infinite correlations, data represents the impulses of nature in the form of certain events and attributes. A skilled business person is motivated to use this cache of data to harness nature, and to find new niches of unserved opportunities that could become profitable ventures.

The objective of DW is to provide business knowledge to support decision making. For DW to serve its objective, it should be aligned around those decisions. It should be comprehensive, easy to access, and up-to-date. Here are some requirements for a good DW:

1. *Subject oriented*: To be effective, a DW should be designed around a subject domain, i.e. to help solve a certain category of problems.

2. *Integrated*: The DW should include data from many functions that can shed light on a particular subject area. Thus the organization can benefit from a comprehensive view of the subject area.

3. *Time-variant (time series):* The data in DW should grow at daily or other chosen intervals. That allows latest comparisons over time.

4. *Nonvolatile*: DW should be persistent, that is, it should not be created on the fly from the operations databases. Thus, DW is consistently available for analysis, across the organization and over time.

5. *Summarized*: DW contains rolled-up data at the right level for queries and analysis. The process of rolling up the data helps create consistent granularity for effective comparisons. It also helps reduces the number of variables or dimensions of the data to make them more meaningful for
the decision makers.

6. *Not normalized*: DW often uses a star schema, which is a rectangular central table, surrounded by some look-up tables. The single table view significantly enhances speed of queries.

7. *Metadata*: Many of the variables in the database are computed from other variables in the operational database. For example, total daily sales may be a computed field. The method of its calculation for each variable should be effectively documented. Every element in the DW should be sufficiently well-defined.

8. *Near Real-time and/or right-time (active)*: DWs should be updated in near real-time in many high transaction volume industries, such as airlines. The cost of implementing and updating DW in real time could be discouraging though. Another downside of real-time DW is the possibilities of inconsistencies in reports drawn just a few minutes apart.

2>
a>

There are two main kinds of decisions: strategic decisions and operational decisions. BI can help make both better. Strategic decisions are those that impact the direction of the company. The decision to reach out to a new customer set would be a strategic decision. Operational decisions are more routine and tactical decisions, focused on developing greater efficiency. Updating an old website with new features will be an operational decision.

In strategic decision-making, the goal itself may or may not be clear, and the same is true for the path to reach the goal. The consequences of the decision would be apparent some time later. Thus, one is constantly scanning for new possibilities and new paths to achieve the goals. BI can help with what-if analysis of many possible scenarios. BI can also help create new ideas based on new patterns found from data mining.

Operational decisions can be made more efficient using an analysis of past data. A classification system can be created and modeled using the data of past instances to develop a good model of the domain. This model can help improve operational decisions in the future. BI can help automate operations level decision-making and improve efficiency by making millions of microlevel operational decisions in a model-driven way. For example, a bank might want to make decisions about making financial loans in a more scientific way using data-based models. A decision-tree-based model could provide a consistently accurate loan decisions. Developing such decision tree models is one of the main applications of data mining techniques.

Effective BI has an evolutionary component, as business models evolve. When people and organizations act, new facts (data) are generated. Current business models can be tested against the new data, and it is possible that those models will not hold up well. In that case, decision models should be revised and new insights should be incorporated. An unending process of generating fresh new insights in real time can help make better decisions, and thus can be a significant competitive advantage.

2>
b>

Statistical relationships are about which elements of data hang together, and which ones hang separately. It is about categorizing variables that have a relationship with one another, and categorizing variables that are distinct and unrelated to other variables. It is about describing significant positive relationships and significant negative differences.

The first and foremost measure of the strength of a relationship is co-relation (or correlation). The strength of a correlation is a quantitative measure that is measured in a normalized range between 0 (zero) and 1. A correlation of 1 indicates a perfect relationship, where the two variables are in perfect sync. A correlation of 0 indicates that there is no relationship between the variables.

The relationship can be positive, or it can be an inverse relationship, that is, the variables may move together in the same direction or in the opposite direction. Therefore, a good measure of correlation is the correlation coefficient, which is the square root of correlation. This coefficient, called r, can thus range from −1 to +1. An r value of 0 signifies no relationship. An r value of 1 shows perfect relationship in the same direction, and an r value of −1 shows a perfect relationship but moving in opposite directions.

3>

DW Architecture

DW has four key elements .The first element is the data sources that provide the raw data. The second element is the process of transforming that data to meet the decision needs. The third element is the methods of regularly and accurately loading of that data into EDW or data marts. The fourth element is the data access and analysis part, where devices and applications use the data from DW to deliver insights and other benefits to users.

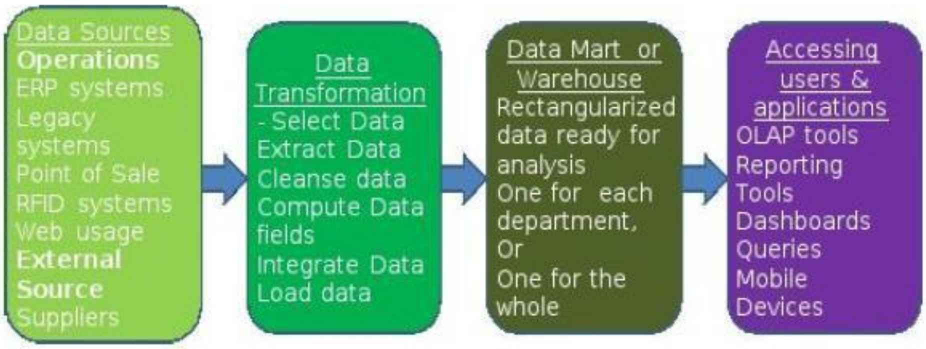

Fig 3.2: Data Warehousing Architecture

Data Sources

Data Warehouses are created from structured data sources. Unstructured data such as text data would need to be structured before inserted into the DW.

1. *Operations data:* This includes data from all business applications, including from ERPs systems that form the backbone of an organization‘s IT systems. The data to be extracted will depend upon the subject matter of the data warehouse. For example, for a sales/marketing data mart, only the data about customers, orders, customer service, and so on would be extracted.

2. *Specialized applications*: This includes applications such as Point of Sale (POS) terminals, and e-commerce applications, that also provide customer-facing data. Supplier data could come from Supply Chain Management systems. Planning and budget data should also be added as needed for making comparisons against targets.

3. *External syndicated data*: This includes publicly available data such as weather or economic activity data. It could also be added to the DW, as needed, to provide good contextual information to decision makers.

Data Loading Processes

The heart of a useful DW is the processes to populate the DW with good quality data. This is called the Extract-Transform-Load (ETL) cycle.

1. Data should be extracted from the operational (transactional) database sources, as well as from other applications, on a regular basis.

2. The extracted data should be aligned together by key fields and integrated into a single data set. It should be cleansed of any irregularities or missing values. It should be rolled-up together to the same level of granularity. Desired fields, such as daily sales totals, should be computed. The entire data should then be brought to the same format as the central table of DW.

3. This transformed data should then be uploaded into the DW.This ETL process should be run at a regular frequency. Daily transaction data can be extracted from ERPs, transformed, and uploaded to the database the same night. Thus, the DW is up to date every morning. If a DW is needed for

near-real-time information access, then the ETL processes would need to be executed more frequently. ETL work is usually done using automated using programming scripts that are written, tested, and then deployed for periodically updating the DW.

4>
The extracted data should be aligned together by key fields and integrated into a single data set. It should be cleansed of any irregularities or missing values. It should be rolled-up together to the same level of granularity. Desired fields, such as daily sales totals, should be computed. The entire data should then be brought to the same format as the central table of DW.

This transformed data should then be uploaded into the DW

Visualization
Data can be presented in the form of rectangular *tables*, or it can be presented in colorful graphs of various types. "Small, non-comparative, highly-labeled data sets usually belong in tables" – (Ed Tufte, 2001, p 33). However, as the amount of data grows, graphs are preferable. Graphics help give shape to data. Tufte, a pioneering expert on data visualization, presents the following objectives for graphical excellence:
1. *Show, and even reveal, the data*: The data should tell a story, especially a story hidden in large masses of data. However, reveal the data in context, so the story is correctly told.
2. *Induce the viewer to think of the substance of the data*: The format of the graph should be so natural to the data, that it hides itself and lets data shine.
3. *Avoid distorting what the data have to say*: Statistics can be used to lie. In the name of simplifying, some crucial context could be removed leading to distorted communication.
4. *Make large data sets coherent*: By giving shape to data, visualizations can help bring the data together to tell a comprehensive story.
5. *Encourage the eyes to compare different pieces of data*: Organize the chart in ways the eyes would naturally move to derive insights from the graph.
6. *Reveal the data at several levels of detail*: Graphs leads to insights, which raise further curiosity, and thus presentations should help get to the root cause.
7. *Serve a reasonably clear purpose* – informing or decision-making.
8. *Closely integrate with the statistical and verbal descriptions of the dataset*: There should be no separation of charts and text in presentation. Each mode should tell a complete story. Intersperse text with the map /graphic to highlight the main insights. Context is important in interpreting graphics.
There are many kinds of data as seen in the caselet above. Time series data is the most popular form of data. It helps reveal patterns over time. However, data could be organized around alphabetical list of things, such as countries or products or salespeople.

5>
There are two primary kinds of data mining processes: supervised learning and unsupervised learning. In supervised learning, a decision model can be created using past data, and the model can then be used to predict the correct answer for future data instances. Classification is the main category of supervised learning activity. There are many techniques for classification, decision trees being the most popular one. Each of these techniques can be implemented with many algorithms. A common metric for all of classification techniques is predictive accuracy.
Predictive Accuracy = (Correct Predictions) / Total Predictions

6>
Please refer text book 2 :page no 77
Not able to copy images

7>

y	x	A	B	A*B		
House	Price Size (sqft)	x-xmean	y-y mean	(x-xmean)(y-ymean)	(x-xmean)2	(y-ymean)2
229500	1850	-81.33	14306.7	-1163563.911	6614.5689	204681664.9
273300	2190	258.67	58106.7	15030460.09	66910.1689	3376388585
247000	2100	168.67	31806.7	5364836.089	28449.5689	1011666165
195100	1930	-1.33	-20093.3	26724.089	1.7689	403740704.9
261000	2300	368.67	45806.7	16887556.09	135917.569	2098253765
179700	1710	-221.33	-35493.3	7855732.089	48986.9689	1259774345
168500	1550	-381.33	-46693.3	17805556.09	145412.569	2180264265
234400	1920	-11.33	19206.7	-217611.911	128.3689	368897324.9
168800	1840	-91.33	-46393.3	4237100.089	8341.1689	2152338285
180400	1720	-211.33	-34793.3	7352868.089	44660.3689	1210573725
156200	1660	-271.33	-58993.3	16006652.09	73619.9689	3480209445
288350	2405	473.67	73156.7	34652134.09	224363.269	5351902755
186750	1525	-406.33	-28443.3	11557366.09	165104.069	809021314.9
202100	2030	98.67	-13093.3	-1291915.911	9735.7689	171434504.9
256800	2240	308.67	41606.7	12842740.09	95277.1689	1731117485
215193.3	1931.333	0.05	0.033333	146946633.3	1053523.33	25810264333

r	0.891131
r^2	0.794114

y=a+bx

b=r Sy/Sx

SDy=sqrt((y-ymean)^2/(n-1))	42937.05
SDy=sqrt((x-xmean)^2/(n-1))	274.3204
b=rSDy/SDx	139.4811
y intercept a=ymean-b xmean	-54191.2