[Total Marks : 80

(Duration : 3 hrs)

Note: 1) Question No. 1 is **compulsory**

2) Attempt any **Four** questions from the remaining **Six** questions.

Q 1. a) Briefly compare **ANY TWO** following concepts (10)

a)DW and Data mart

Datamart	**Data Warehouse**
Data mart is usually sponsored at the department level and developed with a specific issue or subject in mind, a data mart is a data warehouse with a focused objective.	Data warehouse is a "Subject-Oriented, Integrated, Time-Variant, Nonvolatile collection of data in support of decision making".
A data mart is used on a business division/ department level.	A data warehouse is used on an enterprise level.
A Data Mart is a subset of data from a Data Warehouse. Data Marts are built for specific user groups.	A Data Warehouse is simply an integrated consolidation of data from a variety of sources that is specially designed to support strategic and tactical decision making.
By providing decision makers with only a subset of data from the Data Warehouse, Privacy, Performance and Clarity Objectives can be attained.	The main objective of Data Warehouse is to provide an integrated environment and coherent picture of the business at a point in t
Uses bottom-up approach	Uses top-down approach

(05 marks each)

(1mark for each point)

b) Semi Joins and Bloom Joins

Semi join	**bloom join**
Semi join is a method used for efficient query processing in a distributed database environments.	bloom join is another method used to avoid transferring unnecessary data between sites when executing queries in a distributed database environments
Only some of the attributes (or tuples) that are required for the join need to be transferred between the sites to execute the query efficiently	rather than transferring the join column itself, a compact representation of the join column is transferred between the sites
Semi join is a method that can be used to reduce the amount of data shipped between the sites.	Bloom join is a method that can be used to reduce the amount of data shipped between the sites.
In semi join, only the join column is transferred from one site to the other and then that transferred column is used to reduce the size of the shipped relations between the other sites.	Bloom join uses a bloom filter which employs a bit vector to execute membership queries.

(05 marks each)

(1mark for each point)

No such filters are used in semijoin	a bloom filter is built using the join column and it is transferred between the sites and then the joining operations are performed.		

c) OODBMS and ORDBMS

OODBMS	ORDBMS	
Supports object oriented features extensively	Limited Support to object oriented	**(05 marks each)** **(1mark for each point)**
Relatively less performance	Expected to perform well	
Query language is ODL/OQL	Query language is SQL3	
OODBMSs try to add DBMS functionality to a programming language	ORDBMSs try to add richer data types to a relational DBMS	
OODBMS put more emphasis on the role of the client side This can improve long, process intensive, transactions	ORDBMS SQL is still the language for data definition, manipulation and query.	
OODBMS have been optimized to directly support object-oriented applications and specific OO languages	ORDBMS Most third-party database tools are written for the relational model and will therefore be compatible with SQL3	
We can use OODBMS In applications that generally retrieve relatively few (generally physically large) highly complex objects and work on them for long periods of time.	We can use ORDBMS In applications that process a large number of short-lived (generally ad-hoc query) transactions on data items that can be complex in structure	

b) Write Short Notes on **ANY TWO** (10)

a)Web Mining **(05 marks)**

- Definition of web mining
- Types of web mining
  - Content mining
  - Structure mining
  - Usage mining

b)Neural networks **(05 marks)**

- Definition of neural network
- Structure of neural network (diagram)
- Steps in neural network
- Backpropagation

c)Spatial and Geographic DB **(05 marks)**

- A **spatial database** is a database that is optimized to store and query data that represents objects defined in a geometric space. Most spatial databases allow representing simple geometric objects such as points, lines and polygons. Some spatial databases handle more complex structures such as 3D objects, topological coverages, linear networks.
- A geodatabase (also **geographical database** and geospatial **database**) is a **database** of **geographic** data, such as countries, administrative divisions, cities, and related information. Such **databases** can be useful for websites that wish to identify the locations of their visitors for customization purposes.
- Features of spatial database
- Spatial database elements
- Application of geographic data

Q 2.	a) Ans:	Explain concurrency control and recovery process in Distributed Database Management System. DISTRIBUTED CONCURRENCY CONTROL **(4 marks)** ➢ Lock management can be distributed across sites in many ways: ➢ Centralized: A single site is in charge of handling lock and unlock requests for all objects. ➢ Primary copy: One copy of each object is designated as the primary copy. All requests to lock or unlock a copy of this object are handled by the lock manager at the site where the primary copy is stored, regardless of where the copy itself is stored. ➢ Fully distributed: Requests to lock or unlock a copy of an object stored at a site are handled by the lock manager at the site where the copy is stored. DISTRIBUTED RECOVERY **(4 marks)** ➢ Two-Phase Commit (2PC) protocol ➢ Three-Phase Commit	(08)
	b) Ans:	Explain bitmap index with example. When does it make sense to use bitmap index? What is bitmap index **(3 marks)** ➢ **bitmap index** is a special kind of database index that uses bitmaps. ➢ Bitmap indexes have traditionally been considered to work well for *low-cardinality columns*, which have a modest number of distinct values, either absolutely, or relative to the number of records that contain the data. ➢ The extreme case of low cardinality is Boolean data which has two values, True and False. ➢ Bitmap indexes use bit arrays (commonly called bitmaps) and answer queries by performing bitwise logical operations on these bitmaps. ➢ Bitmap indexes have a significant space and performance advantage over other structures for query of such data. ➢ Their drawback is they are less efficient than the traditional B tree indexes for columns whose data is frequently updated Example of bitmap index **(2 marks)** When does it make sense to use a bitmap index? **(2 marks)** Bitmap indexes are meant to be used on low cardinality columns. A low cardinality column just means that the column has relatively few unique values. For example, a column called Sex which has only “Male” and “Female” as the two possible values is considered low cardinality because there are only two unique values in the column.	(07)
Q 3.	a) Ans:	What is Data Warehouse? Why it is needed? Explain ETL process in data warehousing. Data Warehouse Defination. **(1 marks)** Use of data warehouse **(2 marks)** ETL process in brief **(5 marks)** ➢ Full form of ETL ➢ Extraction: various sources ➢ Transformation : Techniques ➢ Loading : Types of loading	(08)

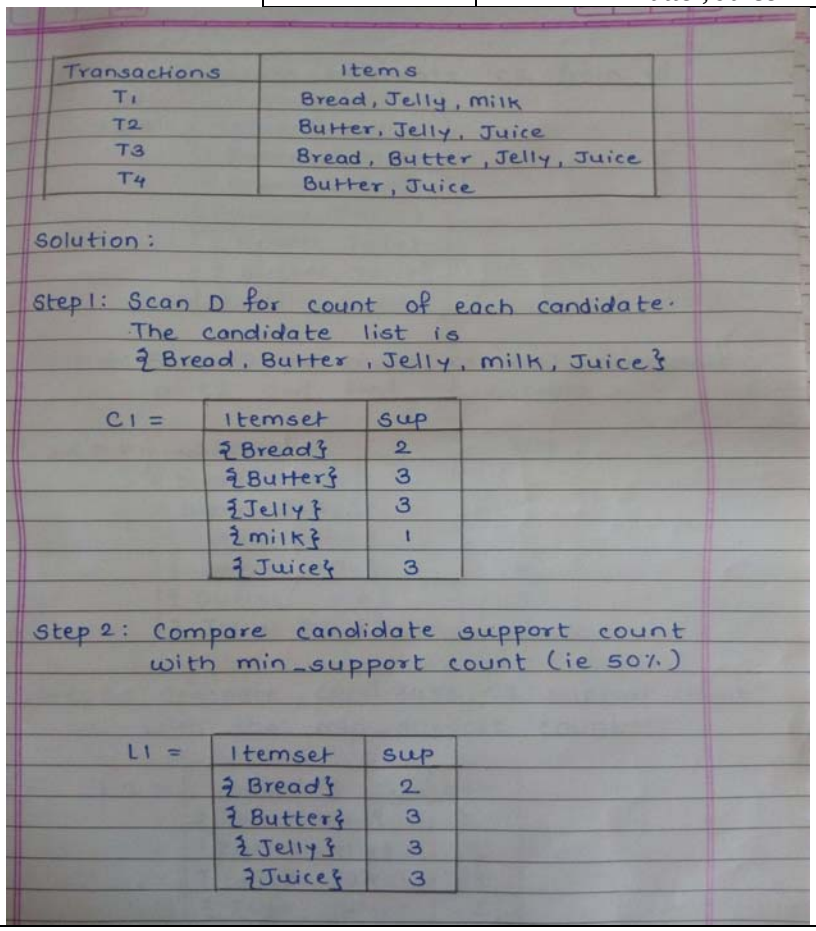
	b)	What are frequent itemsets? Describe an algorithm for finding frequent itemsets.	(07)
	Ans:	Frequent itemsets definition **(2 marks)** Example for frequent itemset **(1 marks)** Apriori algorithm **(4 marks)**	
Q 4.	a)	Define Clustering in data mining. Explain K-Mean clustering with suitable example.	(08)
	Ans:	Defination of clustering **(2 marks)** Algorithm for K-mean clustering **(4 marks)** Example **(2 marks)**	
	b)	Explain ORDBMS Implementation challenges in detail.	(07)
	Ans:	ORDBMS definition **(1 marks)** Implementation challenges are as follows: **(6 marks)** 1) Storage and Access Methods 2) Indexing New Types: 3) QUERY PROCESSING: 4) User-defined aggregation functions: 5) Method Security 6) Method catching	
Q 5.	a)	Define OLAP. Explain MOLAP and ROLAP system with suitable diagram.	(08)
	Ans:	Definition of OLAP (online Analytical Processing) Definition of MOLAP(Multi-dimensional Online Analytical Processing) MOLAP model with suitable diagram **(3 marks)** Definition of ROLAP(Multi-dimensional Online Analytical Processing) ROLAP model with suitable diagram **(3 marks)** Characteristics of MOLAP and ROLAP **(2 marks)**	
	b)	Explain features of XML. Differentiate between XML and HTML.	(07)
	Ans:	XML definition and features **(2 marks)** Features: 1)self descriptive 2)free and extensible 3)platform independent 4)provides domain specific vocabulary 5)it allows data interchange 6)smart searches Difference between HTML and XML **(5 marks)**	

HTML	XML
HTML is an abbreviation for HyperText Markup Language.	XML stands for eXtensible Markup Language.
HTML was designed to display data with focus on how data looks.	XML was designed to be a software and hardware independent tool used to transport and store data, with focus on what data is.
HTML is a markup language itself.	XML provides a framework for defining markup languages.
HTML is a presentation language.	XML is neither a programming language nor a presentation language.

			HTML is not case sensitive.	XML is case sensitive.	
			HTML is used for designing a web-page to be rendered on the client side.	XML is used basically to transport data between the application and the database.	
			HTML has it own predefined tags.	While what makes XML flexible is that custom tags can be defined and the tags are invented by the author of the XML document.	
			HTML is not strict if the user does not use the closing tags.	XML makes it mandatory for the user the close each tag that has been used.	
			HTML does not preserve white space.	XML preserves white space.	
			HTML is about displaying data, hence static.	XML is about carrying information,hence dynamic.	

Q 6. a) Find out the association rules with support 50% and confidence 70% from the following sample data. (08)

Transactions	**Items**
T1	Bread, Jelly, Milk
T2	Butter, Jelly, Juice
T3	Bread, Butter, Jelly, Juice
T4	Butter, Juice



Transactions	Items
T1	Bread, Jelly, milk
T2	Butter, Jelly, Juice
T3	Bread, Butter, Jelly, Juice
T4	Butter, Juice

Solution :

Step 1: Scan D for count of each candidate. The candidate list is {Bread, Butter, Jelly, milk, Juice}

C1 =

Itemset	sup
{Bread}	2
{Butter}	3
{Jelly}	3
{milk}	1
{Juice}	3

Step 2: Compare candidate support count with min_support count (ie 50%)

L1 =

Itemset	sup
{Bread}	2
{Butter}	3
{Jelly}	3
{Juice}	3

Step 3: Generate candidate C2 from L1

C2 =

Itemset
{Bread, Butter}
{Bread, Jelly}
{Bread, Juice}
{Butter, Jelly}
{Butter, Juice}
{Jelly, Juice}

Step 4: Scan D for count of each candidate in C2 and find the support.

C2 =

Itemset	Sup
{Bread, Butter}	1
{Bread, Jelly}	2
{Bread, Juice}	1
{Butter, Jelly}	2
{Butter, Juice}	3
{Jelly, Juice}	2

Step 5: Compare candidate C2 support count with the min-support count.

L2 =

Itemset	Sup
{Bread, Jelly}	2
{Butter, Jelly}	2
{Butter, Juice}	3
{Jelly, Juice}	2

Step 6: Generate candidate C3 from L2

C3 =

Itemset
{Bread, Jelly, Juice}
{Butter, Jelly, Juice}
{Bread, Butter, Jelly}

Step 7: Scan D for count of each candidate in C3

C3 =

Itemset	Sup
{Bread, Jelly, Juice}	1
{Butter, Jelly, Juice}	2
{Bread, Butter, Jelly}	1

Step 8: Compare candidate C3 support count with the min_support count.

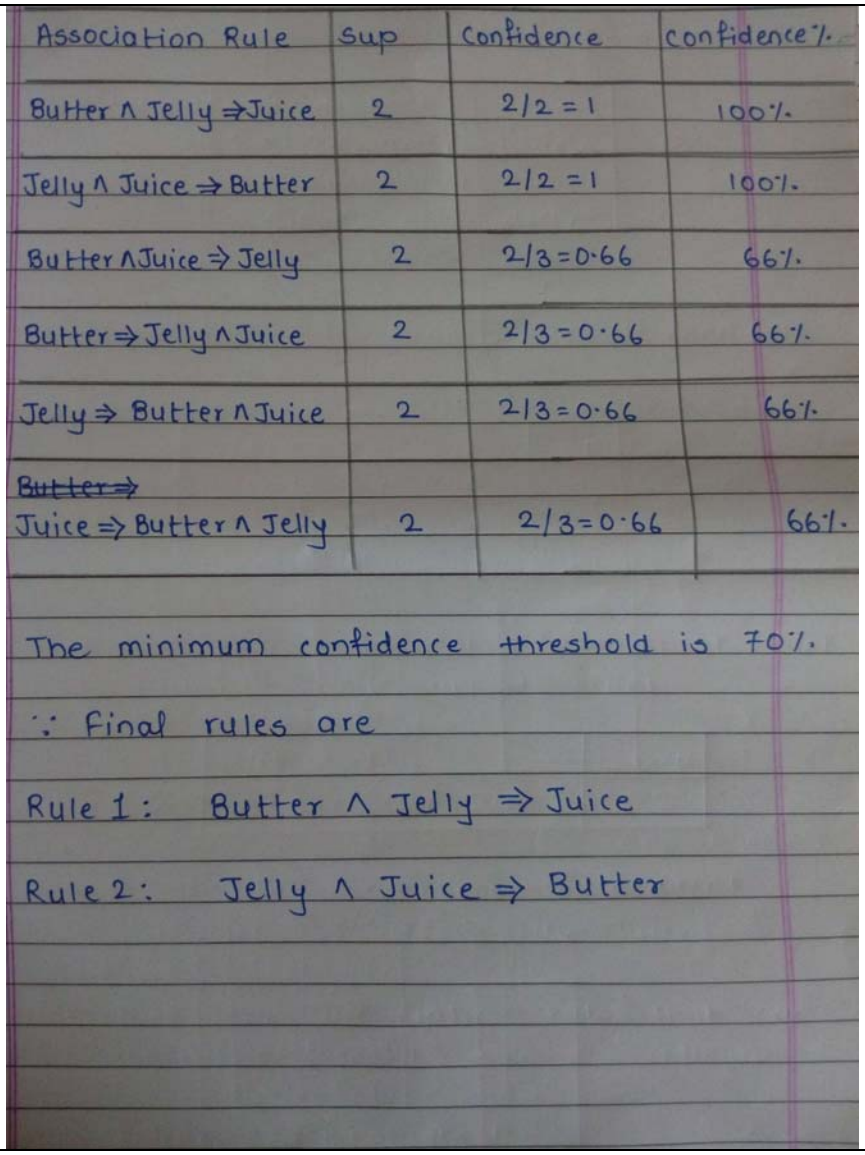
L3 =

Itemset	Sup
{Butter, Jelly, Juice}	2

Step 9: So data contain the frequent itemset L {Butter, Jelly, Juice}

Therefore the association rule that can be generated from L3 are as follows.

$$\text{confidence } (A \Rightarrow B) = \frac{A \cup B}{A}$$

			
	b)	Define the term fragmentation and replication in terms of where data is stored and also how the objects are uniquely identified in distributed database.	(07)
	Ans:	Definition of fragmentation and replication **(3 marks)** Types of fragmentation ➢ Horizontal and vertical Distributed catalog management **(2 marks)** ➢ Naming object ➢ Local name field ➢ Birth site field ➢ Global name field Distributed data independence **(2 marks)**	
Q 7.	a)	What are different complex data types available in ORDBMS? Explain with example.	(08)
	Ans:	There are many complex data types available in ORDBMS: Some of them are listed below-- (any four can be explained) **each carry (2 marks**) with suitable example and syntax. ➢ Collection data types: LIST, SET, MULTISET ➢ Abstract data type (ADT) ➢ User defined types (UDT) ➢ References (REF and DEREF), OID ➢ ROW type	

		➢ VARRAY and NESTED TABLE	
	b) Ans:	Explain Star schema, snowflake schema and fact constellation with suitable example. Star schema **(2 marks)** ➢ Structure of star schema ➢ Example snowflake schema **(2 marks)** ➢ Structure of snowflake schema ➢ Example Advantages and disadvantages of star and snow flake schema **(1 marks)** fact constellation **(2 marks)** ➢ Structure of snowflake schema ➢ Example	(07)