ANNEXURE – I

SRM VALLIAMMAI ENGINEERING COLLEGE

(An Autonomous Institution)

SRM NAGAR, KATTANKULATHUR – 603 203

Register Number	1	4	2	2	1	9	2	0	5	0	8	9
Name of the Candidate	N. Sivakumar											
Degree	Bachelor of Technology											
Branch	Information Technology							**Semester**	5			
Question Paper Code	ST95073											
Subject Code	1908507											
Subject Name	Data warehousing & Data mining											
Date	09 (DD)	02 (MM)	2022 (YYYY)	**Session**	FN ☑	AN ☐						
No. of Pages used	23	**In words**	twenty three									

All particulars given above by me are verified and found to be correct.

Signature of the student with date	N Sivakumar / 9/2/22

FOR OFFICE USE ONLY

Part – A			**Part – B**						
Question No.	**✓**	**Marks**	**Question No.**		**✓**	**Marks**			**Total**
						i	**ii**	**iii**	
1.	✓		11.	a	✓				
2.	✓			b					
3.	✓		12.	a					
4.	✓			b	✓				
5.	✓		13.	a					
6.	✓			b	✓				
7.	✓		14.	a	✓				
8.	✓			b					
9.	✓		15.	a	✓				
10.	✓			b					
Total			**Part – C**						
			16	a	✓	i	ii	iii	Total
				b					
								Grand Total	

Grand Total (in words): .....

Date :

Name & Signature
Of the Examiner

Name & Signature
Of the Chairman / Vice Chairman

1422/19205089
N. Sivakumar

1908507 - DATA WAREHOUSING & DATA MINING

PART-A.

Star schema	Snowflake schema.
* The fact tables & the dimension tables are contained.	* The fact table, dimension table as well as Sub dimension tables are contained.
* Its a top-down model.	* Its a bottom up model.
* It uses more space.	* It uses less space.
* less time to execute queries	* takes more time to execute queries.

2. It is a data about data. It is used to manage, maintain using data warehouse.

eg: author, datae created, date modified & file size are examples of very basic document file metadata. Having the ability to search for a particular element of the metadata makes it much easier for someone to locate a specific document.

3. OLTP	OLAP.
used to control & run fundamental business tasks	used to help with planning, problem solving, & decision support.

 N. Sivakumar

1908507 - DATA WAREHOUSING & DATA MINING

* Short & fast inserts & updates initiated by end users.	* periodic long running batch jobs refresh the data.
* OLTPs are the original source of the data.	* OLAP data comes from the various OLTP Databases.

4. Key features of OLAP:

* Trustworthy data & calculations
* Flexible, self service reporting.
* Business focused calculation.
* Business focused multidimensional data.

5. Associations:

It is the discovery of association rules showing attribute - value condition that occur frequently together in a given set of data.

Correlations:

It is used to study the closeness of the relationship between two or more variables, the degree to which the variables are associated with each other. Suppose in manufacturing firm, they want the relation between - Demand & Supply of commodities.

142219205089
N.Sivakumar

6. major tasks of data pre processing:

* Data cleaning.
* Data integration.
* Data transformation.
* Data reduction.
* Data discretization.

7. correlation:

It is used study the closeness of the relationship between two or more variables. the degree to which the variables are associated with each other.

market basket analysis.

market basket analysis is a data mining technique used by retailers to increase sales by better understanding customer purchasing pattern.

8. Features of decision tree induction:

* Each leaf node holds a class label.
* The topmost node in a tree is the root node.
* Each internal node denotes a test on an attribute.
* Each branch represents an outcome of the test.

9. **Outlier:**

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. Examination of the data for unusual observations that are far removed from the mass of data. These are called outliers.

eg: The data set of N = 90 ordered observation as shown.

To describe data: Two activities are essential for characterizing a set of data.

10. **Classification of hierarchical clustering model:**

i) **Agglomerative clustering -**

It works in bottom up approach & it is also known as AGNES.

ii) **Divisive hierarchical clustering -**

It works in top down manner. It is also known as DIANA.

Part - B.

11. a) **Data warehouse:**

A data warehouse is a repository of multiple heterogeneous data sources organized under a unified schema at a single site to facilitate management decision making (or) A data warehouse is a subject - oriented, time - variant & non - volatile collection of data in support for management's decision

Construction of data warehouse:

Business users want to make decision quickly & correctly using all available data.

Technological factors:

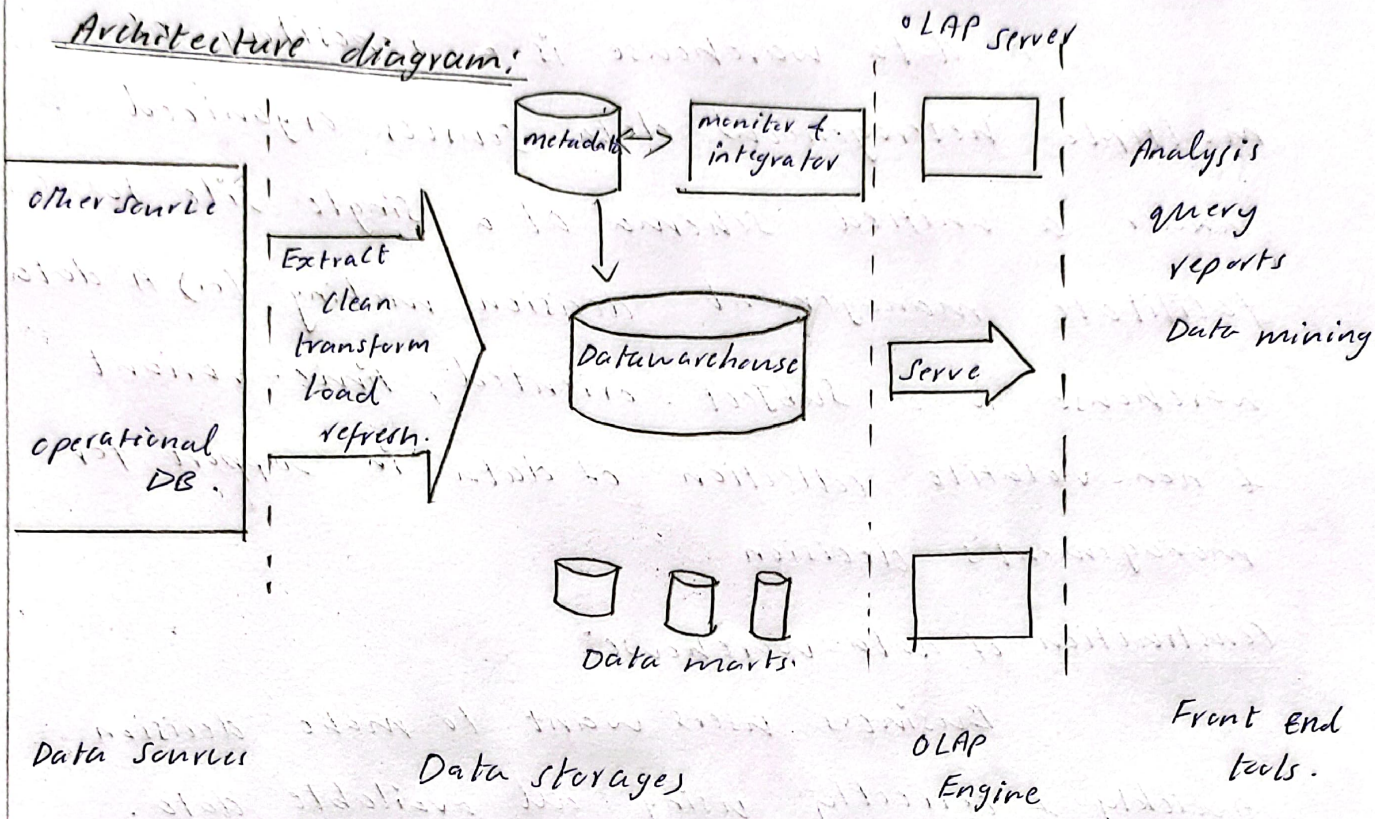
* To address the incompatibility of operational data stores.

* IT infrastructure is changing rapidly, Its capacity is increasing & cost is decreasing so that building a data warehouse is easy.

organization intrested in development of a datawarehouse can choose one of the following two approaches.

* top - down Approach
* Bottom - up approach.

Architecture diagram:



Bottom Tier (Data Sources & data storage):

i) The bottom tier usually consists of data sources & data storage.

ii) It is a ware house database server. For eg RDBMS.

iii) In Bottom Tier, using the application program interface (called gateways), data is extracted from operational & external sources.

N Sivakumar

4. iv) Application program interface likes ODBC, OLE-DB.

middle Tier:

The middle tier is an OLAP server that is typically implemented using either. A relational OLAP model. extended relational DBMS that maps operations from standard data or multidimensional OLAP model.

Top tier:

The top tier is a front-end client layer, which includes query & reporting tools, analysis tools & or data mining tools.

12. b) Features of OLTP & OLAP.

i) Users & System orientation:

An OLTP is an customer oriented & is used for transaction & query processing by clerks, clients, & information technology professionals.

An OLAP system is system is market oriented & is used for data analysis by

knowledge workers, including managers, executives & analysis.

ii) Data contents:

An OLTP system manages current data that typically are too detailed to be easily used for decision making. An OLAP system manages large amount of historical data, provides facilities for summarization & aggregation, & stores & manages information at different level of granularity. These features make the data easier for use in informed decision making.

iii) Database design:

An OLTP system usually adopts an entity-relationship (ER) data model & an application-oriented database design. An OLAP system typically adopts either a star or snowflake model & a subject oriented database schema OLAP.

iv) view:

An OLTP system focuses mainly on the current data within an enterprise without referring to historical data or data in different organization. In contrast, an OLAP system often spans multiple version of a data base schema.

OLAP systems also deal with information that originates from different organizations, integration information from many data stores. Because of their huge volumes, OLAP data are stored on multiple storage media.

v) Access pattern:

The access patterns of an OLTP system consist mainly of short, atomic transactions such a systems requires concurrency control & recovery mechanisms.

However accesses to OLAP systems are mostly read-only operations although many could

be complex queries:

response time:

OLPP only took less response time like in milliseconds.

OLAP takes more response minutes or hours depending on the amount of data to process.

Basic operations:

OLPP is based on Insert, update, delete commands.

OLAP is based on select commands to aggregate data for reporting.

---

13. b) Five primitives for specifying a data mining task:

* Task - relevant data.
* Background knowledge.
* visualization of discovered patterns.
* knowledge type to be mined.
* pattern interestingness measure.

Task relevant data:

This primitive specifies the data upon which mining is to be performed. It involves specifying the database & tables or data warehouse containing the relevant data, conditions for selecting the revelant data, the relevant attributes or dimension for exploiting & instruction regarding the ordering or grouping of the data retrieved.

Background knowledge.

This primitive allows users to specify knowledge they have about the domain to be mined. Such knowledge can be used to guide the knowledge discovery process & evaluate the patterns that are found. of the several kinds of background knowledge, this chapter focuses on concept hierarchies.

visualization of discoverded patterns:

This primitive refers to the form in which discovered patterns are to be displayed. In order for data mining to be effective in conveying knowledge to users, data mining systems should be able to display the discovered pattern in multiple forms such as rules, tables, cross tabs, pie or bar charts or other visual representation.

knowledge type to be mined:

The primitive specific data mining function to be performed, such as characterization, discrimination, association, classification, clustering or evolution analysis. As well, the user can be more specific & provide pattern templates that all discovered pattern must match. These templates or meta pattern can be used to guide the discovery process.

Pattern Interestingness measure:

This primitive allows users to specify functions that are used to seperate uninteresting pattern from knowledge & may be used to guide the mining process, as well as to evaluate the discovered patterns..

---

14. c) i) **Bayes Theorem:**

$$P(C|x) = P(x|C) - P(C)/P(x)$$

* $P(x)$ is constant for all classes.
* $p(c)$ - relative freq of class C samples.
* C such that $P(C|x)$ is maximum = C such that $P(x|C) - P(C)$ is maximum.
* Problem: computing $P(x|C)$ is unfeasible.

$P(H)$, $P(x|H)$ & $P(x)$ may be estimated from the given data,

Baye's theorem is useful in that it provides a way of calculating the posterior probability $P(H|x)$, from $P(H)$, $P(x|H)$ & $P(x)$

& the theorem is classified as.

$$P(H/x) = \frac{P(x|H)\,P(H)}{P(x)}$$

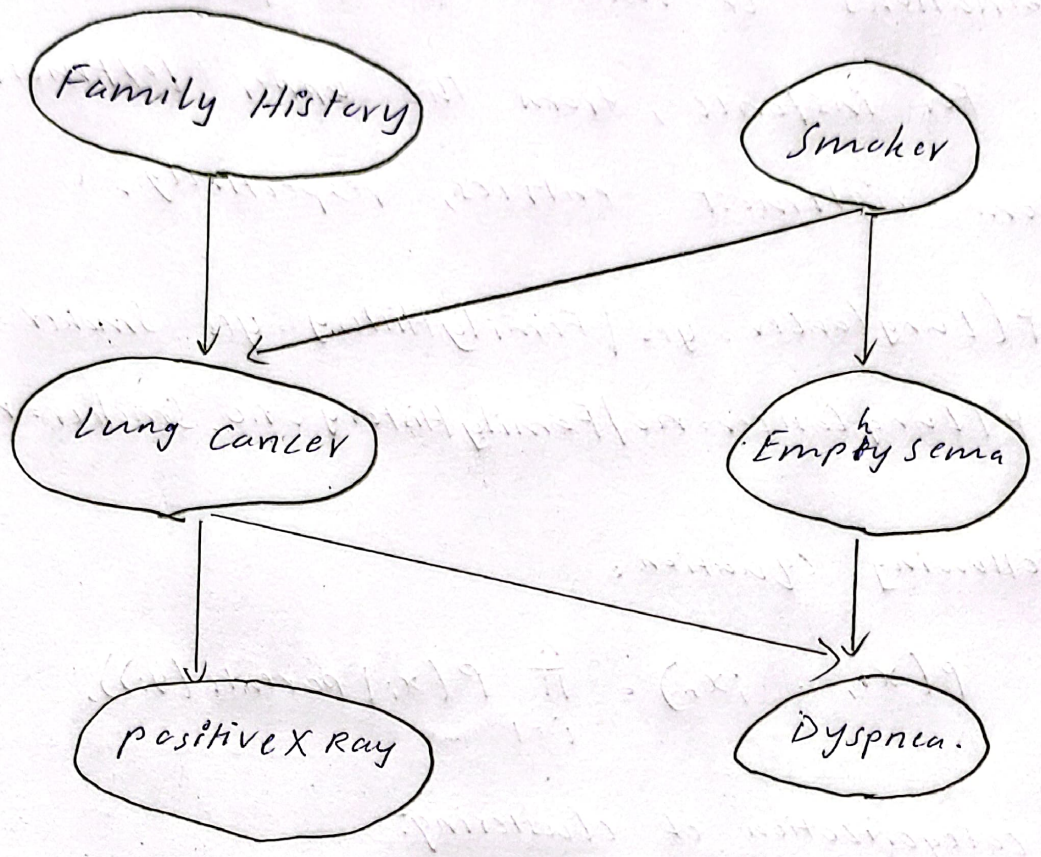
ii)

A belief network is defined by 2 components a directed acyclic graph & a set of conditional probability tables. Each nodes in the directed acyclic graph represents a random variable.

The variable may be discrete or continuous valued. They may correspond to actual attributes given in the data or to hidden variables. beleived to form a relationship

Each arc represents a probabilistic dependence If an arc is drawn from a node Y to node Z, then Y is a parent or immediate predecessor of Z, Z is a descendant of Y. Each variable is conditionally independent of its non descendants in the graph.

a)



b)

	FH,S	FH, ~S	~FH, S	~FH, ~S
LC	0.8	0.5	0.7	0.1
~LC	0.2	0.5	0.3	0.9

The CPT for a variable Y specifies the conditional distributions $P(Y_j \mid Parents(Y))$; where parents (Y) are the parents of Y.

The conditional probability for each known value of Lung Cancer is given for each possible

combinations of values of its parents.

For instance, from the upper leftmost bottom rightmost entries, respectively.

$P(\text{LungCancer} = \text{yes} \mid \text{FamilyHistory} = \text{yes}, \text{Smoker} = \text{yes}) = 0.8$

$P(\text{LungCancer} = \text{no} \mid \text{FamilyHistory} = \text{no}, \text{Smoker} = \text{no}) = 0.9$

Following equation:

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid \text{parents}(y_i)),$$

15. a) i) Categorization of clustering:

ii) K-means:

* partition objects into k non empty subsets.

* Compute seed points as the centroids of the clusters of the current partition.

* Assign each object to the cluster with the nearest seed point.

* Go back to step 2, stop when no more new assignment.

b) Partition method:

~~Suppose~~ Each partition will represent a cluster & k≤n. It means that it will classify the data into k groups, which satisfy the following requirements.

* Each group contain atleast one object.
* Each object must belong to exactly one group.

c) ~~Grid~~ Constraint based method:

* Clustering by considering user-specified or application-specific constraints.

* Typical methods: COD (obstacles), constrained clustering.

* need user feedback: users know their application the best.

* Less parameters but more user-desired constraints, eg: an ATM allocation problem: obstacle & desired clusters.

* Clustering in applications: desirable to have user guided cluster analysis.

d) model-based methods:

* Attempt to optimize the fit between the given data & some mathematical model.

* Based on the assumption: Data are generated by a mixtures of underlying probability distribution.

* In this method a model is hypothesize for each cluster & find the best fit of data to the given model.

ii) Applications of clustering:

* market research, pattern recognition, data analysis & image processing.

* characterize their customer group based on purchasing pattern.

* clustering also helps in classifying documents on the web for information discovery.

* Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in a city according house type, value & geographic location.

* Clustering is also used in outlier detection application such as detection of credit card fraud.

* As a data mining function cluster analysis serve as a tool to gain insight into the distribution of data to observe characteristics of each cluster.

* In field of biology: It can be used to derive plant & animal taxonomies, categorize genes with similar functionality & gain insight into structures inherent in populations.

PART - C

16. a) **major issues in data warehousing & data mining:**

i) ~~mining different kinds of knowledge in~~

Data mining issues:

The types of issues in Data mining are.

* mining methodology & user interaction.
* performance issue.
* Diverse data type.

i) **mining methodology & user interaction:**

* mining different kinds of knowledge in database.
* Interactive mining of knowledge at multiple levels of abstraction.
* Incorporation of background knowledge.

* Data mining query language & ad hoc data mining.

* presentation & visualization of data mining results.

* Handling noisy or incomplete data.

* pattern evaluation.

ii) performance issue:

* Efficiency & scalability of data mining algorithms.

* parallel, distributed & incremental mining algorithms.

iii) Diverse Data types issues.

* Handling of relational & complex type of data.

* mining information from heteregeneous databases & global information systems.

Issues of data warehousing:

i) rigid, inflexible architecture:

Data warehouse is not flexible & can become a bottleneck ~~& can become~~ in meeting business requirements today.

ii) High complexity & redundancy:

Due to the inflexible structure, most organizations purchase hardware addons & tools to facilitate their data needs more quickly. This leads to a complex yet redundant architecture with several data silos.

iii) Slow & degrading performance:

The volume of data the business need to store, process & analyze has grown exponentially over the last decade.

Such great volumes can affect a traditional data warehouse's performance leading to slow performance leading to slow performance & significant delays in reporting.

This can can be caused by a number of reasons but the most common are inefficient & redundant methods.

iv) <u>High costs & failure rates:</u>

Data warehouse has high failure rate, This is not just because of the complex architecture or technical challenges, but because these project often fail to meet user requirements.

The same challenges exist when a data warehouse needs to be updated or changed to meet new reporting requirements or data needs.