2017-310-001

DELTA Pg No. Date / /

Sec-A

Ans-1 Sec A :- Discuss the key characteristics of a Data warehouse.

⇒ The key characterstics of Data warehouse are :-

i) Subject Oriented
ii) Integrated
iii) Time-Variant
iv) Non-Volalite

→ Subject-oriented :- A data warehouse is always a subject oriented as it delivers information about a theme insted of organizations current operation.

These subjects can be product, customers, suppliers, sales, revenue, etc. A data warehouse dose not focus on the ongoing operations, rather it focusses on modelling and analysis of data for decision making.

→ Integrated :- A data warehouse is constructed by integrating data from heterogeneous sources such as relational databases, flat files, etc. This integration enhances the effective analysis of data.

A database is built by integrating data from various sources of data such that a mainframe and relational database. In addition, it must have reliable naming conventions, format and codes. Integration of data warehouse benifits in effective analysis of data.

→ Time-Variant :- Data is maintained via different intervals of time such as weekly, monthly or annually etc.

The data resided in data warhouse is predictable with a specific interval of time and deliver information from the historical prespective.

Non - Volatile :- As the name defines the data resided in data Warehouse is permanent. It also means that data is not erased or deleted when new data is inserted.

Data is read-only and refreshed at particular intervals. This is benificial in analysing historical data and in comprehension and functionality.

Ans. No 20 -

Sec-A. Constraint Based association mining :-

A data mining process may uncover thousand of rules from the given set of data, most of which end up being unrelated or uninteresting to the other. Often, user good sense of which observations of mining may lead to intresting pattern and the 'form' of the pattern or rules they would like to find. Thus a good habit is to have

Sign.

user specify such intentions or expectations as constraints to confine the search space the strategy is known as constraint based mining.
The constraint can include the following

→ Knowledg type constraints :- These specifies the type of knowledge to be mined, such as association or correlation.

→ Data constraints :- These specifies the set of task-relevant data.

→ Dimention constraints :- These specifies the desired dimentions of the data, or levels of the concept heriarchies to be used in mining.

→ Intrestingness constraints :- These specifies threashold on statistical researhes of rule intrestingness.

→ Rule Constraints :- These specifies the form of rules to be mined such ~~constrains~~ may be expressed as parameters, as the max & min no. of predictors can be occur in the rule antecedent.

DELTA Pg No. Date / /

Ans 3 :- Differentiate b/w OLAP and OLTP
Sec A

⇒ Data WareHouse (OLAP)	Operational Database (OLTP)
1. It involves historical processing of information.	It involves day-to-day processing.
2. OLAP systems are used by knowledge workers such as executives, managers, and analyst.	OLTP system are used by clerks, DBAs, or database professionals.
3. It is used to analyze the business.	It is used to run the business.
4. It focuses on information out.	It focuses on Data in.
5. It is based on star schema, snowflake schema, and fact constellation schema.	It based on Entity Relationship Model.
6. It provides summarized and consolidate data.	It provide primitive and highly detailed data.

My NoteBook

Sign.

7.	It provide summarized and multidimentional view of data	It provides premitive and highly detailed data
8.	The number of user is in hundreds.	The no. of user are in thousands.
9.	The no. of records accessed in millions	the no. of records accessed in tens.
10.	These are highly flexible	It provides high performance.
11.	The database size from 100 gb too 100 Tb	The database size is from 100 Mb to 100 GB.

Ans. No-4:- Discuss the limitations of K-means clustering:-

=> limitations of K-means clustering

- Predefined no. of clusters:-

The no. of groups must be defined before creating clustering.

DELTA Pg No.
Date / /

- Distorted by outliners :-

In case of many outliners, K-means clustering may not create an optimal grouping because the outliners will be assigned to many of the allocated groups.

- The K-means meathod is not suitable for discovering cluster with non convex shapes and clusters of very different size.

Ans. No 5:- What are Data marts

Sec A:-

→ A Data marts is focussed on a single functional area of an organization and contains a subset of data stored in a data warehouse. A data mart is condensed version of data warehouse and is designed for use by specific department, unit or set of users in an organization. Ex. Marketing, HR, Sales or finance. It is often controlled by a single department in an organization.

→ Datamart usually draw data from only a few sources compared to a data warehouse. Data mart are small in size and more flexible compared to a Data warehouse.

→ It help to enhance user's response time due to reduction in volume of data.

→ It provide easy access to frequently requested data.

→ Data mart are simpler to implemet when compare to computation sale data warehouse. At the same time, the cost of implementing data mart is certainly lower compared, with implimenting full data warehouse.

→ Data is partitioned & allows very gradudy allow controll.

→ There are three types of data mart - ~~dependent~~.

i) Dependent

ii) Independent

iii) Hybrid.

DELTA Pg No.
Date / /

Qus. What are the limitations of Apriori Algorithm.

Ans: Apriori algorithms suffers from some weakness in spite of being clear and simple. The main limitation is costly wasting of time to host a vast no. of candidate sets with much frequent itemsets, low minimum support or large itemset. for example, if there 10^4 ~~more~~ from 1-itemset, it need to generate more then 10^7 candidate into 2-length which in turn they will be tested and accumulate [2]. Furthermore, to detect frequency pattern in size 100 (eg) $V_1, V_2, \ldots, V_{100}$, it have to generate 2^{100} candidate itemset [1] that yield on costly and wasting of time of candidate generation. So, it will check from many sets from candidate itemset, also it will scan database many times repetedly for finding candidate itemsets. Apriori will be

Sec-A

My NoteBook Sign.

DELTA Pg No.
Date / /

Very low and ineffecient when memory capacity is limited with large no. of transaction.

Sec - B:-

Ans. 1:-

Sec-B

Min. support = 3

Step 1:- Creating a table for support count of each itemset.

Itemset	Support count
f	4
A	3
C	4
D	1
G	1
I	1
M	3
P	3
L	2
O	1
B	3
H	1
J	1
W	1
K	1
S	1
E	1
N	1

Step-2:- Now removing all itemset having support count less than 3

Item set	Sup. count
f	4
A	3
C	4
M	3
P	3
B	3

Step-3:- Arranging in order to priority of highest Sup. count.

Item set	Sup. count
f	4
C	4
A	3
M	3
P	3
B	3

DELTA Pg No. Date / /

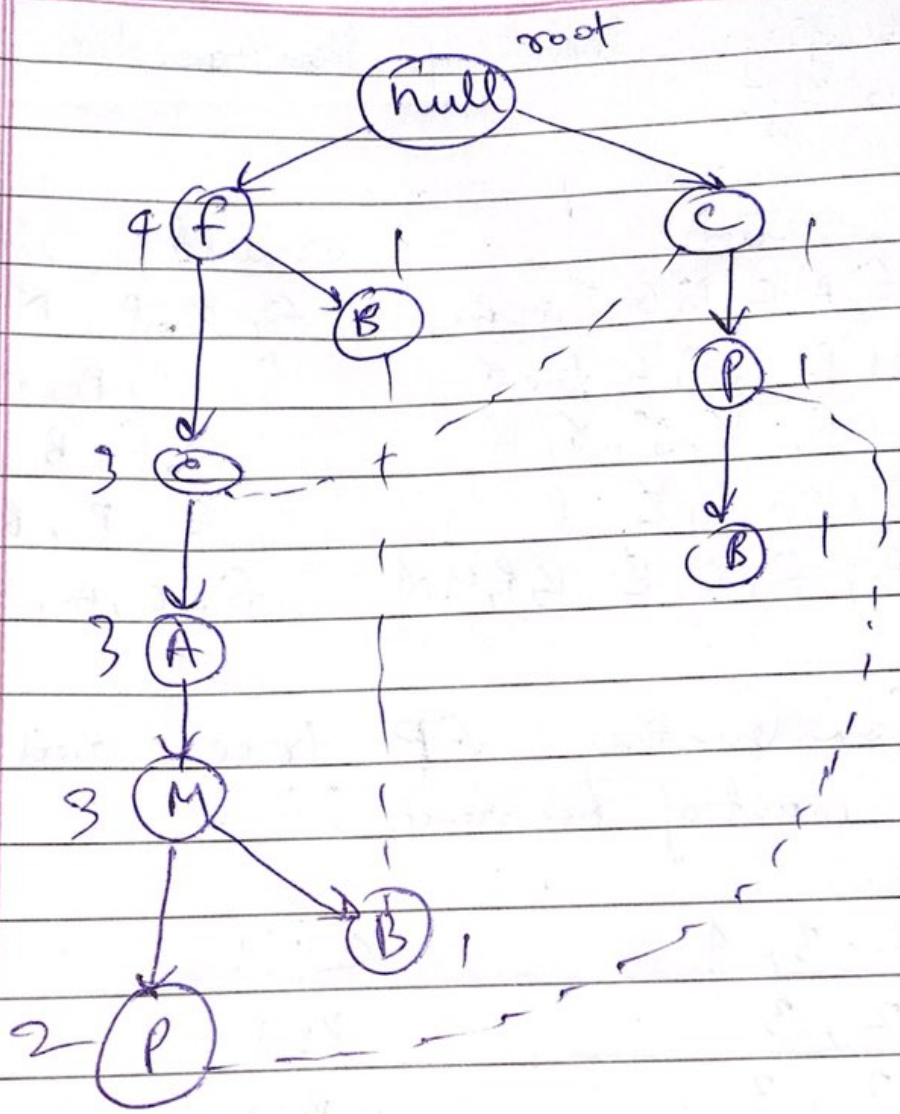
Step -4 Arranging list of item on the basis of priority.

Tr_id	items	ordered items
100	f, A, C, ~~B~~, ~~G~~, ~~I~~, M, P	f, C, A, M, P
101	A, B, C, F, ~~G~~, M, ~~O~~	f, C, A, M, B
102	B, f, ~~A~~, ~~J~~, ~~O~~, ~~W~~	f, B
103	B, C, K, ~~S~~, P	C, P, B
104	A, f, C, ~~E~~, ~~G~~ P, M, ~~N~~	f, C, A, M, P

Step -5. Constructing ~~fp~~ tree and noting their count of occurence.

f : 1, 2, 3, 4
C : 1, 2, 3
A : 1, 2, 3
M : 1, 2, 3
P : 1, 2
B : 1
B : 1

C : 1
P : 1
B : 1


root
null
4
F
C
B
P
3
A
B
3
A
3
M
B
2
P

Ans3:- Sec-B The, Apriori property is based on the following observation, By definition, If an itemset I dosenot satisfy the minimum support threshold, then I is not frequent, that is P(I) < min. support. If an item A is added to the itemset I, then the resulting itemset (i.e I ∪ A) cannot occur more frequently than I. Therefore, I ∪ A is not frequent either

Then is, P(I ∪ A) < min. support

This property belongs to a special category of properties called antimono-tonicity. In the sense that if a set can't pass a test, all of its superset will fail the same test as well.

- In the computation of the itemset in L_k using L_{k-1}

My NoteBook

Sign.

- It is done in two steps
  - → Join
  - → Prune

→ Join step :-

- The set of candidate k itemset (elements of L_k), C_k, is generated by joining L_{k-1} with itself

$$L_{k-1} \bowtie L_{k-1}$$

- Given I_1 & I_2 of L_{k-1}

$L_i = I_{i_1}, I_{i_2}, \ldots\ldots I_{i(k-2)}, I_{(k-1)}$

$L_j = I_{j_1}, I_{j_2}, \ldots\ldots I_{j(k-2)}, I_{j(k-1)}$

where L_i & L_j are sorted

- L_i & L_j are joined if there are different (no. duplicate generation). Assume the following.

$I_{i1} = I_{j1}, I_{i2} = I_{j2}, \ldots\ldots, I_{i(k-2)} = I_{j(k-2)}$

& $I_{i(k-1)} < I_{j(k-1)}$

- The resulting itemset is

$I_{i1}, I_{i2}, \ldots \ldots I_{i(k-1)} I_{j(k-1)}$

- Example of candidate generation

$L_3 = \{ abc, abd, acd, ace, bcd \}$

self-joining $L_3 \infty L_3$

abcd from abc & abd

acde from acd & ace

→ Pruning step :-

* C_k is superset of L_k → some itemset in C_k may or may not be frequent.

• L_k. Test each generated itemset against the database.

- Scan the database to determine the count of each generated itemset & include those that have a count no. less than the minimum support count.

- This may require intensive computation.

- Use apriori ~~algo~~ property to reduce the search space

• Any $(k-1)$ itemset is not frequent cannot be a subset of a frequent k itemset.

- Remove from C_k any k-itemset that has a (k-1) subset not in L_{k-1} (itemset that are not frequent)
- Efficiently implemented : maintain a hash table of all frequent itemset.

⇒ Example :-

Database of TDB

T.id	item
10	A C D
20	B C E
30	A B C E
40	B E

1st Scan → C_1

item set	Support
A	2
B	3
C	3
D	1
E	3

→ L_1

item	support
A	3
B	3
C	3
E	3

→

itemset
A, B
A, C
A, E
B, C
B, E
C, E

→ C_2 → 2nd Scan →

itemset	support
A, B	1
A, C	2
A, E	1
B, C	2
B, E	3
C, E	2

→ L_2

itemset	support
A, C	2
B, C	2
B, E	3
C, E	2

→ C_3 → 3rd Scan →

itemset	support
B, C, E	2

→ L_3

itemset
B, C, E

My NoteBook Sign.

Ans 4:- There are four types of data mining architecture -

Sec-B

→ No coupling data mining:- In this architecture, a mining system does not use any functionality of a database. A no coupling data mining system retrieves data from a particular data sources. The no-coupling data mining system architecture does not take advantage of a database That is already very effecient in organizing, storing, accessing & retrieving data

→ Loose coupling data mining :- In this architecture, data mining uses a database or data retrieves. In ~~loose~~ loose coupling, data mining architecture data mining system retrieves data from a database and it stores the result in those

systems. Data mining architecture is for memory based data mining system.

→ Semi-Tight coupling data mining:-

In semi tight coupling data mining system uses features of data warehouse system. That is to perform some data mining takes. This includes sorting, indexing, aggregation. In this some intermediate result can be stored in a database for better performance.

→ Tight coupling data mining:-

In tight coupling data warehouse is treated as an information retrieval component. All the features of database or data warehouses all used to perform data mining tasks.

There are three tiers in the tight coupling data mining architecture —

i) Data layer :- We can define data layer as database. This layer is a interface for all data sources. Data mining results are stored in data layer.

ii) Data mining application layer :- It is to retreive data from a database. Some transformation routine has to perform here that is to transform data into the desired format. Then we have to process data using various data mining algorithms.

iii) front end layer :- It provides the intuitive and friendly user interface for end user i.e to interact with data mining system. Data mining result presented in visualization from to the user in the front end layer.

- The major component of any data mining system are data source, data warehouse server, data mining engine, pattern evaluation module, graphical user interface and knowledge base.

- Data sources :- Data bases, data warehouse, world wide web, text files and other documents are the actual source of data. You need large volume of historical data for data mining to be successful. Organizations usually store data in database or data warehouses.

- Database or data warehouse servers :- The database or data warehouse servers contains the actual data that is ready to be processed. Hence, the server is responsible for retrieving the relevant data based on the data mining request of the user.

→ **Data Mining Engine:-** The data mining engine is the core component of any data mining system. It consist of a number of modules for performing data mining tasks including association, clustering, characterization, classification, prediction, time series analysis etc.

→ **Pattern Evaluation Modules:-** The pattern evaluation module is mainly responsible for the measure of interestingness of the pattern by using a threshold value. It interacts with the data mining engine to focus the search towards interesting pattern.

→ **Graphical User Interface:-** The graphical user interface module communicates between the user and the data mining system. This module helps the user use the system easily and efficiently without knowing the real complexity behind the process.