Don Bosco Institute of technology, Mumbai-400070
Department of Information Technology

Name: Shruti Chaube
Roll No.: 66

Assignment No.: 2

1. Consider the following transaction database:

TID	Items
01	A, B, C, D
02	A, B, C, D, E, G
03	A, C, G, H, K
04	B, C, D, E, K
05	D, E, F, H, L
06	A, B, C, D, L
07	B, I, E, K, L
08	A, B, D, E, K
09	A, E, F, H, L
10	B, C, D, F

Apply the Apriori algorithm with minimum support of 30% and minimum confidence of 70% and find all the association rules in the data set.

Solution:

Step 1: Scan D for count of each candidate. The candidate list is {A, B, C, D, E, F, G, H, I, K, L}

C1=

I=Itemsets	Support count
A	6
B	7
C	6
D	7
E	6
F	3
G	2
H	3
I	1
K	4
L	4

Step 2: Compare candidate support count with minimum support count (i.e., 3)

L1=

I=Itemsets	Support count
A	6
B	7
C	6
D	7
E	6
F	3
H	3
K	4
L	4

Step 3: Generate candidate C2 from L1 and find the support.

C2=

I=Itemsets	Support count
A,B	4
A,C	4
A,D	4
A,E	3
A,F	1
A,H	2
A,K	2
A,L	1
B,C	5
B,D	5
B,E	4
B,F	1
B,H	0
B,K	3
B,L	2
C,D	5
C,E	2
C,F	1
C,H	1
C,K	2
C,L	1
D,E	4
D,F	2

D,H	1
D,K	2
D,L	2
E,F	2
E,H	2
E,K	3
E,L	3
F,H	2
F,K	0
F,L	2
H,K	1
H,L	2
K,L	1

Step 4: Compare candidate (C2) support count with the minimum support count

L2=

I =Itemsets	Support count
A,B	4
A,C	4
A,D	4
A,E	3
B,C	5
B,D	5
B,E	4
B,K	3
C,D	5
D,E	4
E,K	3
E,L	3

Step 5: Generate candidate C3 from L2 and find the support.

C3=

I =Itemsets	Support count
A,B,C	3
A,B,D	4
A,B,E	2
A,B,K	1
A,C,D	3
A,C,E	1
A,D,E	2
B,C,D	
B,C,E	2
B,C,K	1
B,D,E	3
B,D,K	2
B,E,K	3
C,E,D	2
D,E,K	2
D,E,L	1

Step 6: Compare candidate (C3) support count with the minimum support count

L3=

I =Itemsets	Support count
A,B,C	3
A,B,D	4
A,C,D	3
B,C,D	5
B,D,E	3
B,E,K	3

Step 7: Generate candidate C4 from L3 and find the support.

C4=

I =Itemsets	Support count
A,B,C,D	3
A,B,D,E	2
B,C,D,E	2
B,D,E,K	2

Step 8: Compare candidate (C4) support count with the minimum support count

L4=

I =Itemsets	Support count
A,B,C,D	3

Step 9: So, data contains the frequent item sets: {A, B, C, D}

Generate the Association rule from frequent item sets with the support and confidence.

Association Rule	Support	Confidence	Confidence %
BCD => A	3	3/5	60%
ACD =>B	3	3/3	100%
ABD => C	3	3/4	75%
ABC => D	3	3/3	100%
CD =>AB	3	3/5	60%
BD => AC	3	3/5	60%
BC =>AD	3	3/5	60%
AD =>BC	3	3/4	75%
AC => BD	3	3/4	75%
AB =>CD	3	3/4	75%
D => ABC	3	3/7	43%
C =>ABD	3	3/6	50%
B =>ACD	3	3/7	43%
A => BCD	3	3/6	50%

Given minimum confidence threshold is 70%. So only ACD => B, ABD => C, ABC => D, AD =>BC, AC =>BD, AB =>CD rules are output.

Final rules are:

Rule 1: ACD => B

Rule 2: ABD => C

Rule 3: ABC => D

Rule 4: AD =>BC

Rule 5: AC =>BD

Rule 6: AB =>C

2. Explain multilevel and multidimensional association rule with example.

MULTILEVEL ASSOCIATION RULES:

- Association rules generated from mining data at multiple levels of abstraction are called multiple-level or multilevel association rules.
- Multilevel association rules can be mined efficiently using concept hierarchies under a support-confidence framework.
- Rules at high concept level may add to common sense while rules at low concept level may not be useful always.
    - Using uniform minimum support for all levels:
- When a uniform minimum support threshold is used, the search procedure is simplified.
- The method is also simple, in that users are required to specify only one minimum support threshold.
- The same minimum support threshold is used when mining at each level of abstraction.
- For example, in Figure, a minimum support threshold of 5% is used throughout.
- (e.g., for mining from “computer” down to “laptop computer”).
- Both “computer” and “laptop computer” are found to be frequent, while “desktop computer” is not.
- Using reduced minimum support at lower levels:
    - Each level of abstraction has its own minimum support threshold.
    - The deeper the level of abstraction, the smaller the corresponding threshold is.
    - For example, in Figure, the minimum support thresholds for levels 1 and 2 are 5% and 3%, respectively.
    - In this way, “computer,” “laptop computer,” and “desktop computer” are all considered frequent.

Multilevel Association rule consists of alternate search strategies and Controlled level cross filtering:

1. Alternate Search Strategies:

- Level by level independent:
    - Full breadth search.
    - No background knowledge in pruning.
    - Leads to examine lot of infrequent items.
- Level-cross filtering by single item:
    - Examine nodes at level i only if node at level (i-1) is frequent.
    - Misses frequent items at lower level abstractions (due to reduced support).
- Level-cross filtering by k-item set:
    - Examine k-item sets at level i only if k-item sets at level (i-1) is frequent.
    - Misses frequent k-item sets at lower level abstractions (due to reduced support).
- Controlled Level-cross filtering by single item:
    - A modified level-cross filtering by single item.
    - Sets a level passage threshold for every level.
- Allows the inspection of lower abstractions even if its ancestor fails to satisfy min_sup threshold.

MULTIDIMENSIONAL ASSOCIATION RULES:

1.In Multi-dimensional association:

- Attributes can be categorical or quantitative.
- Quantitative attributes are numeric and incorporates hierarchy.
- Numeric attributes must be discretized.
- Multi-dimensional association rule consists of more than one dimension:

E.g.: buys(X, ”IBM Laptop computer”)buys(X,”HP Inkjet Printer”)

2.Three approaches in mining multi-dimensional association rules:

1.Using static discretization of quantitative attributes.

- Discretization is static and occurs prior to mining.
- Discretised attributes are treated as categorical.
- Use apriori algorithm to find all k-frequent predicate sets (this requires k or k+1 table scans).
- Every subset of frequent predicate set must be frequent.
- E.g.: If in a data cube the 3D cuboid (age, income, buys) is frequent implies (age, income), (age, buys), (income, buys) are also frequent.
- Data cubes are well suited for mining since they make mining faster.
- The cells of an n-dimensional data cuboid correspond to the predicate cells.

2.Using dynamic discretization of quantitative attributes:

- Known as mining Quantitative Association Rules.
- Numeric attributes are dynamically discretized.
- E.g.: age(X, ”20..25”) Λ income(X,”30K..41K”)buys (X, ”Laptop Computer”)

	Age=20	Age=21	Age=22	Age=23	Age=24	Age=25
Income,38 to 41						
Income,34 to 37						
Income,30 to 33						

GRID FOR TUPLES

3.Using distance-based discretization with clustering.

This id dynamic discretization process that considers the distance between data points.

- It involves a two steps mining process:
  - Perform clustering to find the interval of attributes involved.
  - Obtain association rules by searching for groups of clusters that occur together.
- The resultant rules may satisfy:
  - Clusters in the rule antecedent are strongly associated with clusters of rules in the consequent.
  - Clusters in the antecedent occur together.
  - Clusters in the consequent occur together.

3. Construct FP-tree for the following database where minimum support is 3. {Answer should contain frequent items list, header table with fp-tree and F-list}

TID	Items
01	A, B, D, E, F
02	B, C, E
03	A, B, D, E
04	A, B, C, E
05	A, B, C, D, E, F
06	B, C, D
07	A, B, D, E

Solution:

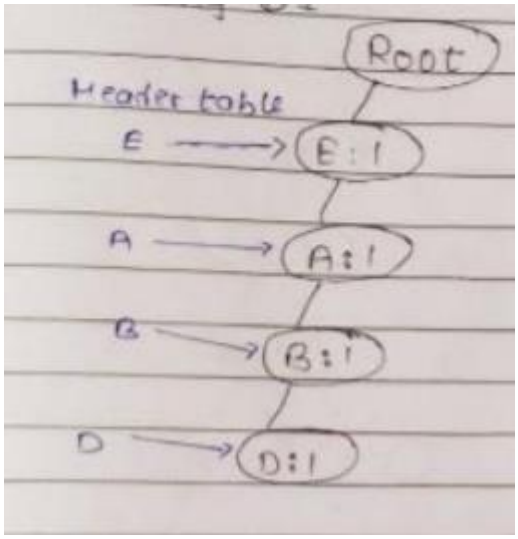
Item	**Support**
A	5
B	5
C	4
D	5
E	6
F	2

Don Bosco Institute of technology, Mumbai-400070
Department of Information Technology

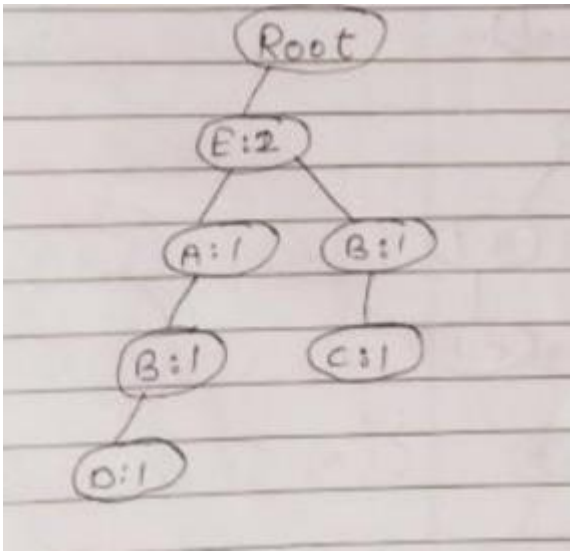
TID	Items
01	E, A, B, D
02	E, B, C
03	E, A, B, D
04	E, A, B, C
05	E, A, B, D, C
06	B, D, C
07	E, A, B, D

FP-Tree

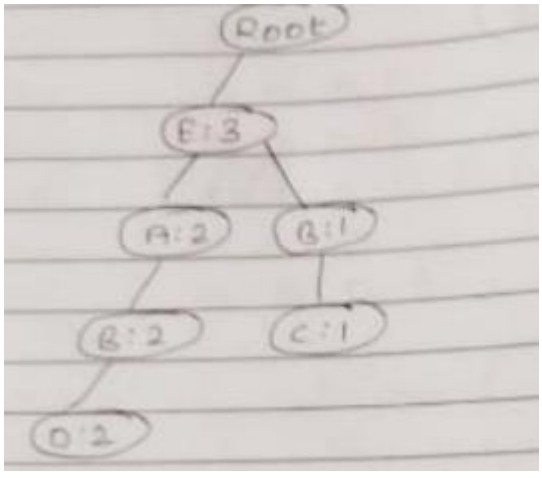
Inserting 01



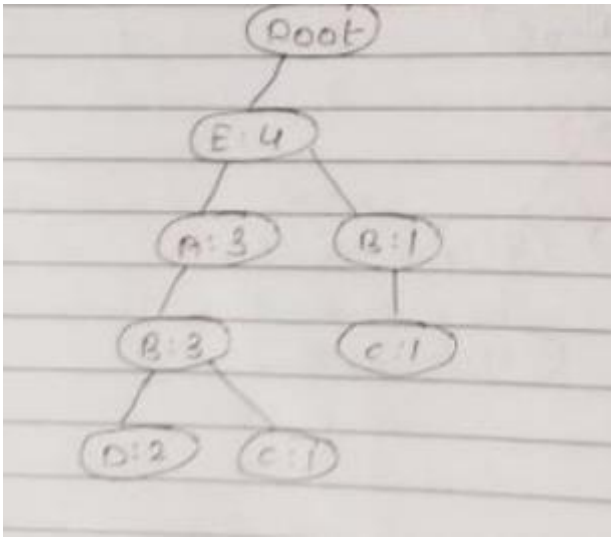
Inserting 02



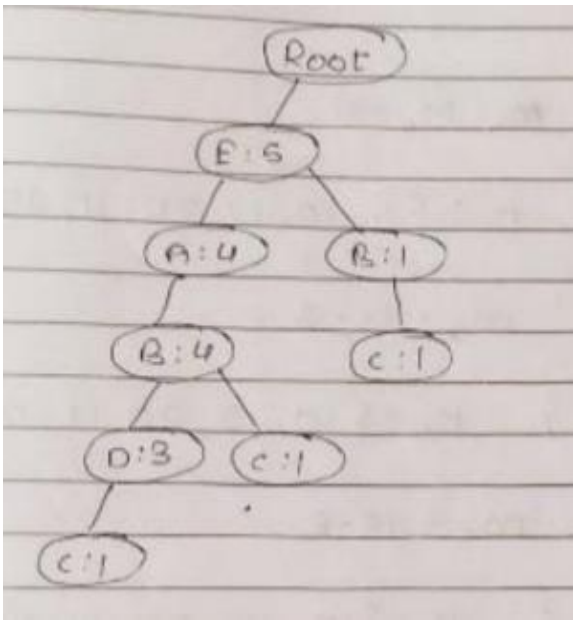
Inserting 03



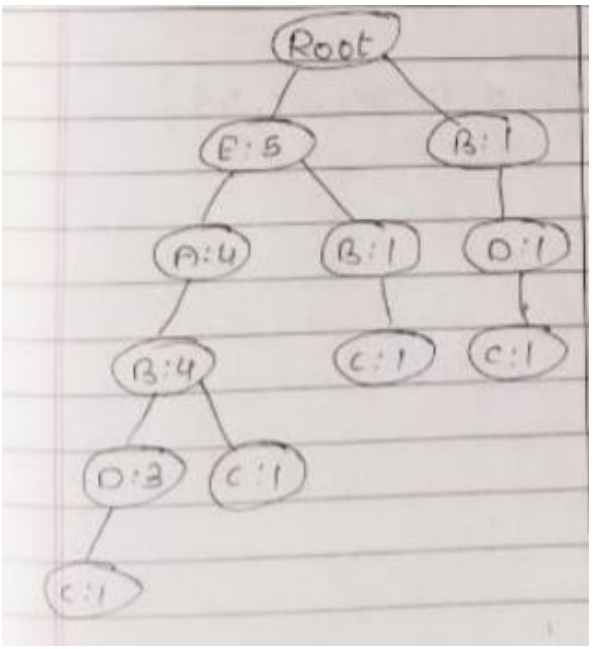
Inserting 04



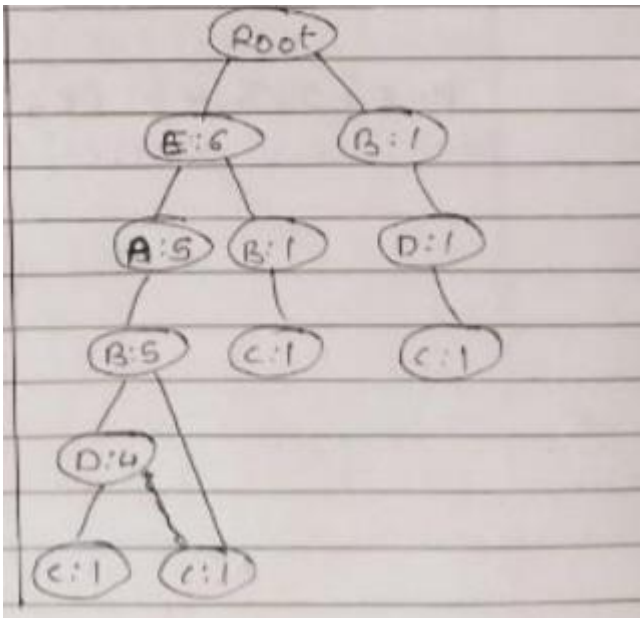
Inserting 05



Inserting 06



Inserting 07



Frequent Item List: E, A, B, D

4. What is clustering? Explain k-means clustering algorithm. Suppose the data for clustering is {2, 4, 10, 12, 3, 20 ,11, 25} Consider k=2, cluster the given data using K-means algorithm.

Solution:

- Clustering is a data mining technique used to place data elements into related groups without advance knowledge of the group definitions.
- Clustering is a process of partitioning a set of data in set of meaningful sub-classes, called as clusters.
- A cluster is therefore a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters.

Fig: Clustering graphical exampl

- In this case, we easily identify the 4 clusters into which the data can be divided.

k-means algorithm:

- K-means clustering is an algorithm to classify or to group the different object based on attributes or features into K number of group.
- K is positive integer number(which can be decided by user)
- Define K centroids for K clusters which are generally far away from each other.
- Then group the elements into clusters which are nearer to the centroid of that cluster.
- After this first step, again calculate the new centroid for each cluster based on the elements of that cluster.
- Follow the same method, and group the elements based on new centroid.
- In every step, the centroid changes and elements move from one cluster to another.
- Do the same process till no element is moving from one cluster to another.

Algorithm:

k: number of clusters

n :sample features vectors $x_1, x_2 \ldots\ldots\ldots x_n$x1,x2……xn

$m_i$mi: the mean of the vectors in cluster i

Assume k<n< p="">

- Make initial guesses for the mean m_1 , m_2…… ,m_k
- Until there is no changes in any mean
  - Use the estimated means to classify the samples into clusters.
  - For I from 1 to k

Replace m_i with the mean of all of the samples for cluster i

- End _for
- End _until

Suppose the data for clustering – 2, 4, 10, 12, 3, 20, 11, 25

1. Randomly assign means m_1m1=3 and m_2m2=4
2. The number which are close to mean m_1m1=3 is grouped into cluster k_1k1 and numbers which are close to mean m_2m2=4 are grouped into cluster k_2k2
3. Again, calculate the new mean for new cluster groups
4. k_1k1 = {2,3}, k_2k2= {4,10,12,20,30,11,25}, m_1=2.5, m_2m2=16
5. k_1k1 = {2,3,4}, k_2k2= {10,12,20,30,11,25}, m_1m1=3, m_2m2=18
6. k_1k1 = {2,3,4,10}, k_2= {12,20,30,11,25}, m_1m1=4.75, m_2m2=19.6
7. k_1k1 = {2,3,4,10,11,12}, k_2k2= {20,30,25}, m_1=7, m_2=25m1=7, m2=25
8. k_1k1 = {2,3,4,10,11,12}, k_2k2= {20,30,25}
9. Stop as clusters with these means in step 7 and 8 are same.
10. So, the final answer is k_1=2,3,4,10,11,12, k_2=20,30,25

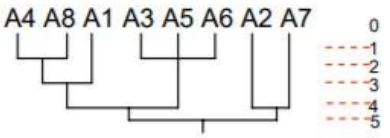
5. Use Hierarchical clustering algorithm to cluster the following 8 points into 3 clusters

A1 = (2,10), A2 = (2,5), A3 = (8,4), A4 = (5,8),

A5 = (7, 5), A6 = (6, 4), A7 = (1, 2), A8 = (4, 9).

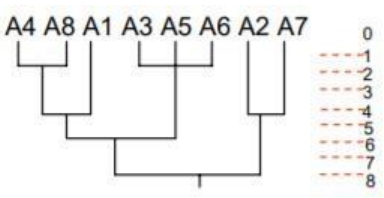
Solution: Single Link:

d	k	K
0	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
1	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
2	5	{A4, A8}, {A1}, {A3, A5, A6}, {A2}, {A7}
3	4	{A4, A8, A1}, {A3, A5, A6}, {A2}, {A7}
4	2	{A1, A3, A4, A5, A6, A8}, {A2, A7}
5	1	{A1, A3, A4, A5, A6, A8, A2, A7}



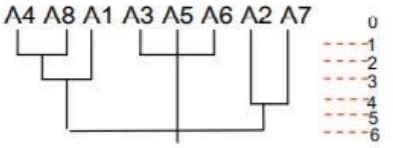
Complete Link:

d	k	K
0	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
1	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
2	5	{A4, A8}, {A1}, {A3, A5, A6}, {A2}, {A7}
3	5	{A4, A8}, {A1}, {A3, A5, A6}, {A2}, {A7}
4	3	{A4, A8, A1}, {A3, A5, A6}, {A2, A7}
5	3	{A4, A8, A1}, {A3, A5, A6}, {A2, A7}
6	2	{A4, A8, A1, A3, A5, A6}, {A2, A7}
7	2	{A4, A8, A1, A3, A5, A6}, {A2, A7}
8	1	{A4, A8, A1, A3, A5, A6, A2, A7}



Average Link:

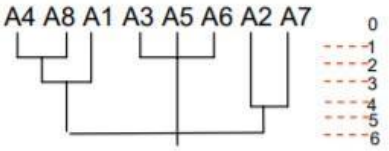
d	k	K
0	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
1	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
2	5	{A4, A8}, {A1}, {A3, A5, A6}, {A2}, {A7}
3	4	{A4, A8, A1}, {A3, A5, A6}, {A2}, {A7}
4	3	{A4, A8, A1}, {A3, A5, A6}, {A2, A7}
5	3	{A4, A8, A1}, {A3, A5, A6}, {A2, A7}
6	1	{A4, A8, A1, A3, A5, A6, A2, A7}



Average distance from {A3, A5, A6} to {A1, A4, A8} is 5.53 and is 5.75 to {A2, A7}

Centroid:

D	k	K
0	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
1	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
2	5	{A4, A8}, {A1}, {A3, A5, A6}, {A2}, {A7}
3	5	{A4, A8}, {A1}, {A3, A5, A6}, {A2}, {A7}
4	3	{A4, A8, A1}, {A3, A5, A6}, {A2, A7}
5	3	{A4, A8, A1}, {A3, A5, A6}, {A2, A7}
6	1	{A4, A8, A1, A3, A5, A6, A2, A7}

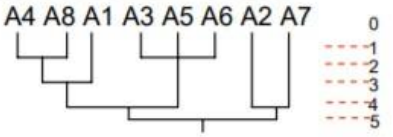


Centroid of {A4, A8} is B = (4.5, 8.5) and centroid of {A3, A5, A6} is C = (7, 4.33) distance(A1, B) = 2.91 Centroid of {A1, A4, A8} is D=(3.66, 9) and of {A2, A7} is E=(1.5, 3.5) distance(D,C)= 5.74 distance(D,E)= 5.90

Medoid:

This is not deterministic. It can be different depending upon which medoid in a cluster we chose.

d	k	K
0	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
1	8	{A1}, {A2}, {A3}, {A4}, {A5}, {A6}, {A7}, {A8}
2	5	{A4, A8}, {A1}, {A3, A5, A6}, {A2}, {A7}
3	4	{A4, A8, A1}, {A3, A5, A6}, {A2}, {A7}
4	2	{A1, A3, A4, A5, A6, A8}, {A2, A7}
5	1	{A1, A3, A4, A5, A6, A8, A2, A7}

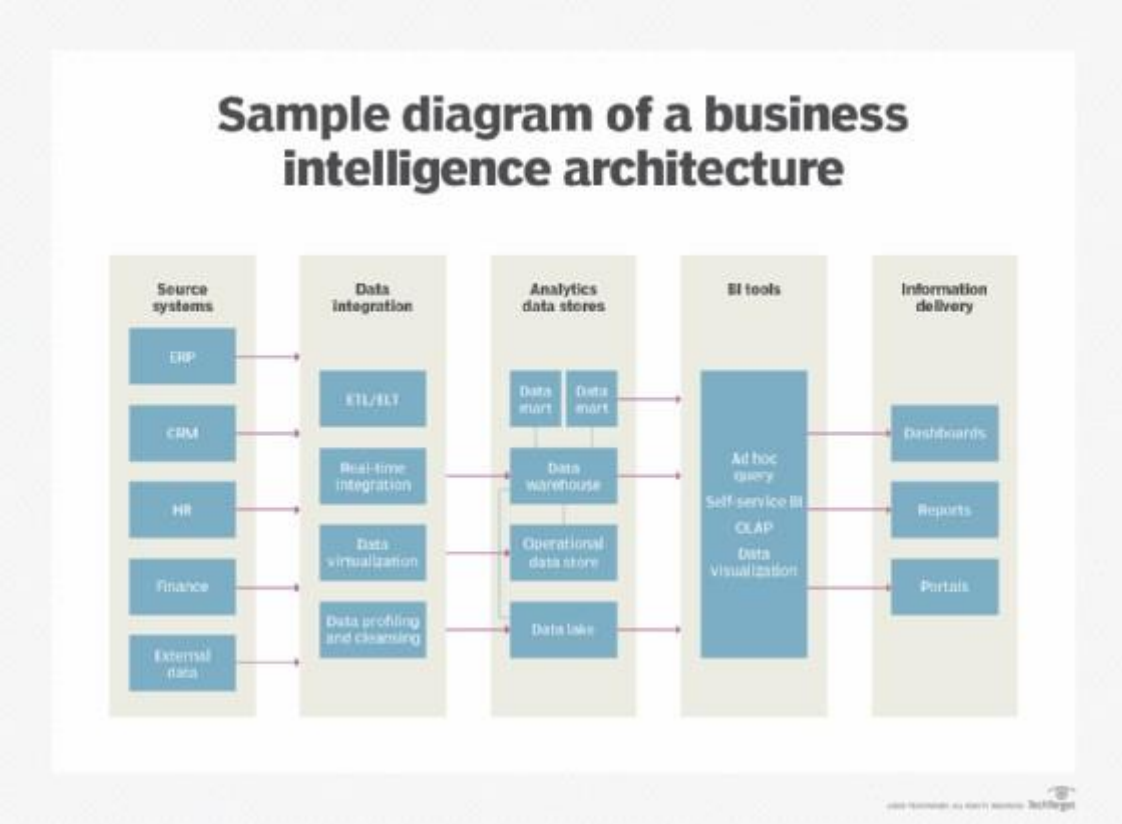


6. Define BI and give its architecture. Explain any business application where data mining can be used.

Solution:

A business intelligence architecture is the framework for the various technologies an organization deploys to run business intelligence and analytics applications. It includes the IT systems and software tools that are used to collect, integrate, store and analyse BI data and then present information on business operations and trends to corporate executives and other business users.

The underlying BI architecture is a key element in the implementation of a successful business intelligence program that uses data analysis and reporting to help an organization track business performance, optimize business processes, identify new revenue opportunities, improve strategic planning and make more informed decisions overall.



As shown in the accompanying business intelligence architecture diagram, the core components include the following items.

Source systems. These are all of the systems that capture and hold the transactional and operational data identified as essential for the enterprise BI program -- for example, ERP, CRM, finance, manufacturing and supply chain management systems.

Data integration and cleansing tools. To effectively analyse the data collected for a BI program, an organization must integrate and consolidate different data sets to create unified views of them. The most widely used data integration technology for BI applications is extract, transform and load (ETL) software, which pulls data from source systems in batch processes.

Analytics data stores: This encompasses the various repositories where BI data is stored and managed. The primary one is a data warehouse, which usually stores structured data in a

relational, columnar or multidimensional database and makes it available for querying and analysis. An enterprise data warehouse can also be tied to smaller data marts set up for

BI and data visualization tools. The tools used to analyze data and present information to business users include a suite of technologies that can be built into a BI architecture -- for example, ad hoc query, data mining and online analytical processing, or OLAP, software

Dashboards, portals and reports. These information delivery tools give business users visibility into the results of BI and analytics applications, with built-in data visualizations and, often, self-service capabilities to do additional data analysis. For example, BI dashboards and online portals can both be designed to provide real-time data access with configurable views and the ability to drill down into data. Reports tend to present data in a more static format.

Business Application where Data Mining can be used:

Future Healthcare: Data mining holds great potential to improve health systems. It uses data and analytics to identify best practices that improve care and reduce costs. Researchers use data mining approaches like multi-dimensional databases, machine learning, soft computing, data visualization and statistics. Mining can be used to predict the volume of patients in every category. Processes are developed that make sure that the patients receive appropriate care at the right place and at the right time. Data mining can also help healthcare insurers to detect fraud and abuse.

Market Basket Analysis: Market basket analysis is a modelling technique based upon a theory that if you buy a certain group of items, you are more likely to buy another group of items. This technique may allow the retailer to understand the purchase behaviour of a buyer. This information may help the retailer to know the buyer's needs and change the store's layout accordingly. Using differential analysis comparison of results between different stores, between customers in different demographic groups can be done.

Education: There is a new emerging field, called Educational Data Mining, concerns with developing methods that discover knowledge from data originating from educational Environments. The goals of EDM are identified as predicting students' future learning behaviour, studying the effects of educational support, and advancing scientific knowledge about learning. Data mining can be used by an institution to take accurate decisions and also to predict the results of the student. With the results the institution can focus on what to teach and how to teach. Learning pattern of the students can be captured and used to develop techniques to teach them.

Manufacturing Engineering: Knowledge is the best asset a manufacturing enterprise would possess. Data mining tools can be very useful to discover patterns in complex manufacturing process. Data mining can be used in system-level designing to extract the relationships between

product architecture, product portfolio, and customer needs data. It can also be used to predict the product development span time, cost, and dependencies among other tasks.

CRM: Customer Relationship Management is all about acquiring and retaining customers, also improving customers' loyalty and implementing customer focused strategies. To maintain a proper relationship with a customer a business need to collect data and analyse the information. This is where data mining plays its part. With data mining technologies the collected data can be used for analysis. Instead of being confused where to focus to retain customer, the seekers for the solution get filtered results.

7. Explain Business Intelligence issues.

Solution:

There are three major issues in Business Intelligence:

1. Data mining methodology:

- Mining different kinds of knowledge from diverse data types.
- Performance: efficiency, effectiveness and scalability.
- Pattern evaluation: The interestingness pattern.
- Incorporation of background pattern.
- Handling noise and incomplete data.
- Parallel, distributed and incremental mining methods.
- Integration of the discovered knowledge with existing one: Knowledge fusion.

2. User interaction:

- Data mining query languages and ad-hoc mining.
- Expression and visualization of resultant knowledge.
- Interactive mining of knowledge at multiple levels of abstraction.

3. Applications and social impact:

- Domain specific data mining and invisible data mining.
- Protection of data security, integrity and privacy.