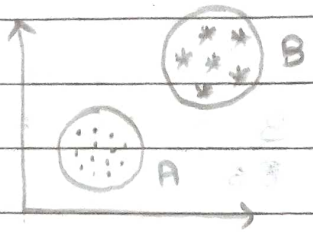
* Clustering

1) Partition Based
2) Hierarchy Based
3) Density Based

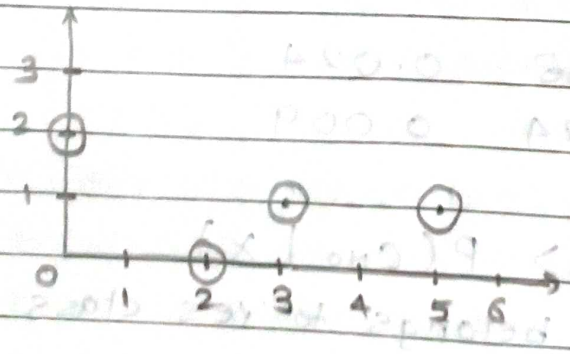


Distance Function

$$D(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$$ → Minkowski distance formula

$$D(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^1\right)^{1/1}$$ → Manhatton

$$D(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^2\right)^{1/2}$$ → Eucledian



Point	X	Y
A	0	2
B	2	0
C	3	1
D	5	1

$$D(P_i, P_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

$$D(A, B) = \sqrt{(0-2)^2 + (2-0)^2} = 2\sqrt{2} = 2.82$$

$$D(A, C) = \sqrt{(0-3)^2 + (2-1)^2} = \sqrt{10} = 3.162$$

$$D(A, D) = \sqrt{(0-5)^2 + (2-1)^2} = \sqrt{26} = 5.099$$

$$D(B, C) = \sqrt{(2-3)^2 + (0-1)^2} = \sqrt{2} = 1.414$$

$$D(B, D) = \sqrt{(2-5)^2 + (0-1)^2} = \sqrt{10} = 3.162$$

$$D(C, D) = \sqrt{(3-5)^2 + (1-1)^2} = \sqrt{4} = 2$$

for a given data point convert into Data Matrix

	A	B	C	D
A	0	2.82	3.16	5.09
B	2.82	0	1.41	3.16
C	3.16	1.41	0	2
D	5.09	3.16	2	0

* K-Means Clustering Algorithm (Partition Based)

For given value of K, K-Means is implemented in 4 steps:

1) Partition objects into K non-empty subsets.
2) Compute the seed points as centroids of the clusters of the current partition. (centroid is the mean point of the cluster)
3) Assign each object to cluster with the nearest seed point.

4) Go back to step ②, stop when no more new assignments are there.

Q. For given dataset create 3 clusters using k-means clustering algorithm.

Data = 22, 9, 12, 15, 10, 27, 35, 18, 36, 11

K = 3

→

1st Iteration :

K=1 {22} = {22, 27, 35, 36, 18} Mean = 27.6

K=2 {9} = {9, 10} Mean = 9.5

K=3 {12} = {12, 11, 15} Mean = 12.6

Calculate the mean of each set as the new centroid.

2nd Iteration :

K=1 {27.6} = {22, 27, 35, 36} = 30

K=2 {9.5} = {9, 10, 11} = 10

K=3 {12.6} = {12, 15, 18} = 15

3rd Iteration :

K=1 {30} = {27, 35, 36} = 32.6

K=2 {10} = {9, 10, 11, 12} = 10.5

K=3 {15} = {22, 15, 18} = 18.3

4th Iteration :

K=1 {32.6} = {27, 35, 36} = 32.6

K=2 {10.5} = {9, 12, 10, 11} = 10.5

K=3 {18.3} = {22, 15, 18} = 18.3

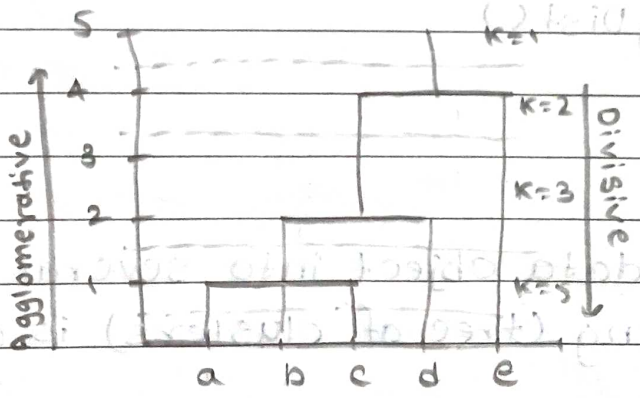
As the means in 3rd and 4th are same, which is stopping criteria for K-Means Algorithm, so clusters are-

$C_1 = \{27, 35, 36\}$
$C_2 = \{9, 10, 11, 12\}$
$C_3 = \{15, 18, 22\}$

Drawback of K-Means Algorithm :

① The value of k should be given.
② It is sensitive to outliers.

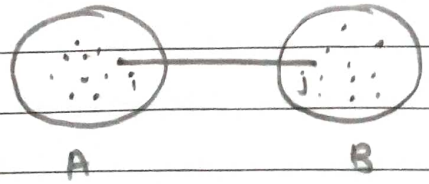
* Hierarchical Clustering :-



Dendrogram

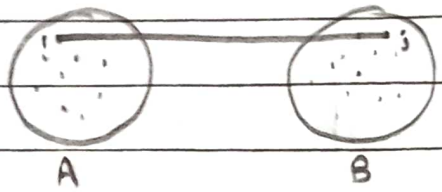
* Agglomerative

1) Single Linkage



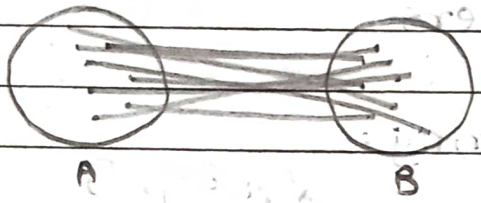
$D(A,B) = \min Dist(i, j)$

2). Complete Linkage



$D(A, B) = \text{Max Dist}(i, j)$

3) Average Linkage



$D(A, B) = \text{Avg Dist}()$

* Dendrogram :-

- Decomposing a data object into several levels of nested partitioning (tree of clusters) is called a Dendrogram.
- A clustering of data object is obtained by cutting the dendrogram depending upon the value of K. (at the derived level)
- Each connected components forms a cluster.

Q. **Single Linkage**

① Min Dist is 2
∴ Merge 1 & 2

	1	2	3	4	5
1	0				
2	(2)	0			
3	6	3	0		
4	10	9	7	0	
5	9	8	5	4	0

② Min Dist is 3
∴ Merge 1,2 & 3

	1,2	3	4	5
1,2	0			
3	(3)	0		
4	9	7	0	
5	8	5	4	0

③ Min Dist is 4
∴ Merge 4 & 5

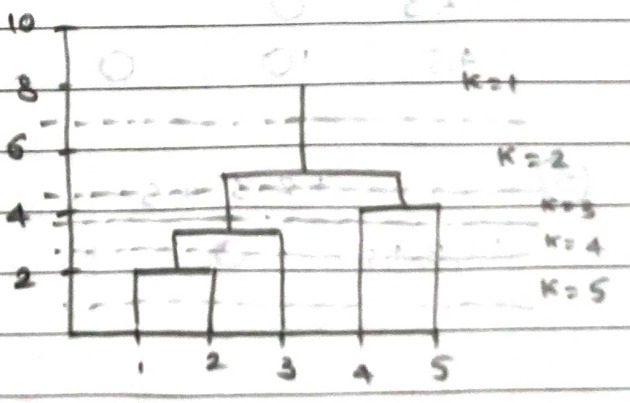
	1,2,3	4	5
1,2,3	0		
4	7	0	
5	5	(4)	0

④

	1,2,3	4,5
1,2,3	0	
4,5	5	0

⑤

	1,2,3,4,5
1,2,3,4,5	0



Dendrogram

Complete Linkage

①

	1	2	3	4	5
1	0				
2	(2)	0			
3	6	3	0		
4	10	9	7	0	
5	98	8	5	4	0

Min = 2
Merge 1 & 2

②

	1,2	3	4	5
1,2	0			
3	6	0		
4	10	7	0	
5	9	5	(4)	0

Min = 4
Merge 4 & 5

③

	1,2	3	4,5
1,2	0		
3	(6)	0	
4,5	10	7	0

Min = 6
Merge 1,2 & 3

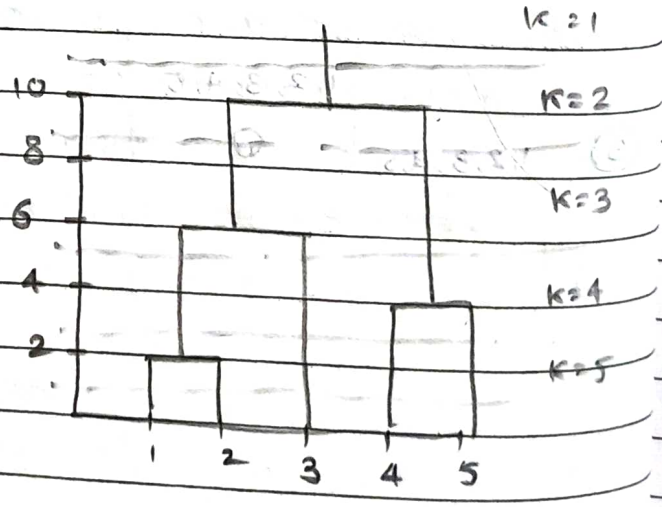
④

	1,2,3	4,5
1,2,3	0	
4,5	10	0

Merge 1,2,3,4,5

⑤

	1,2,3,4,5
1,2,3,4,5	0



Average Linkage

①

	1	2	3	4	5
1	0				
2	(2)	0			
3	6	3	0		
4	10	9	7	0	
5	9	8	5	4	0

Min = 2
Merge 1, 2

②

	1,2	3	4	5
1,2	0			
3	4.5	0		
4	9.5	7	0	
5	8.5	5	(4)	0

Min = 4
Merge 4 & 5

③

	1,2	3	4,5
1,2	0		
3	(4.5)	0	
4,5	9	6	0

Min = 4.5
Merge 1,2 & 3

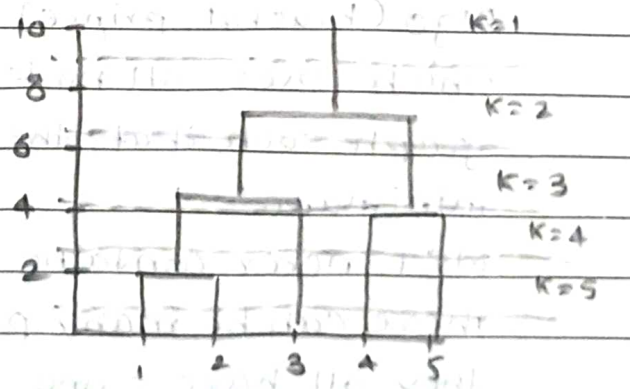
④

	1,2,3	4,5
1,2,3	0	
4,5	7.5	0

Min = 7.5
Merge 1, 2, 3, 4, 5

⑤

	1,2,3,4,5
1,2,3,4,5	0



H.W.

Q.

	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0

By using Hierarchical Clustering method draw Dendrogram using single linkage and average linkage method.

* Divisive

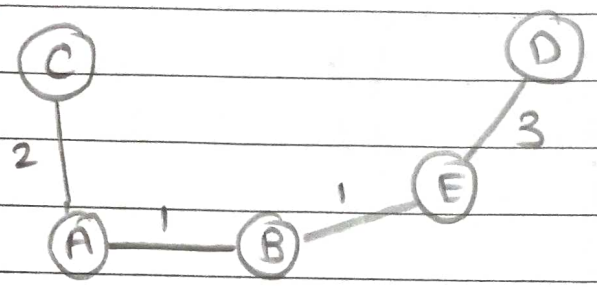
	A	B	C	D	E
A	0				
B	1	0			
C	2	3	0		
D	5	4	6	0	
E	4	1	2	3	0

- ~~Hiera~~ Divisive algorithm uses Minimum Spanning Algo (Kruskal, prince) to create tree like struc which covers all nodes of connected undirected graph such that the sum of cost of edges is minimum.
- MST never contains a cycle.

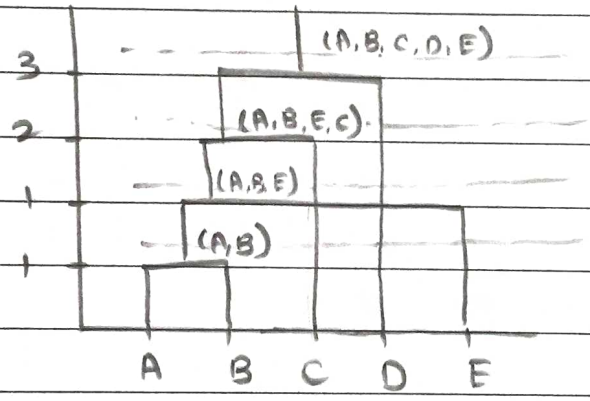
there can be many possible MSTs matrix, but they all have same minimum cost.

PAGE No.
DATE / /

MST :-



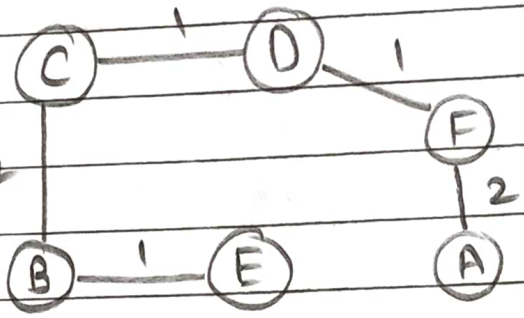
Dendrogram :-



Q.1

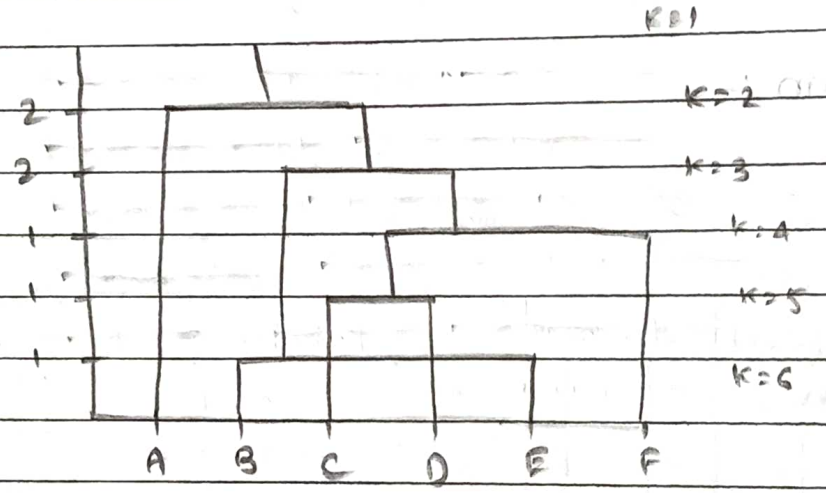
	A	B	C	D	E	F
A	30					
B	3	0				
C	4	2	0			
D	7	3	1	0		
E	5	1	2	3	0	
F	2	6	7	1	4	0

~~MST :-~~ Compare Ate Aglomerative & Divisive and create a Dedndrogram using single & complete linkage method.



cost = 7

Dendrogram :-



* Association Mining :-

Frequent Patterns :-

- ** A pattern is a set of items that occur frequently in a dataset. frequent
- A typical example of item set mining is market ~~Appl~~ basket analysis (MBA).
- Cross marketing, catalogue design, sale campaign analysis, web log analysis, DNA sequence analysis

Applications

* Apriori Algorithm

Two steps :

1) Join $L_1 \bowtie L_1$
2) Prune

Two Lists :

1) Candidate List → C_1
2) Item List → L_1

Eg.-

Tid	Items
10	I1 I3 I4
20	I2 I3 I5
30	I1 I2 I3 I5
40	I2 I5

SupMin = 2

1st scan → C_1

Itemset	Support
I1	2
I2	3
I3	3
~~I4~~	~~1~~
I5	3

C_1 → L_1

Item set
I1
I2
I3
I5

L_1

$L_1 \bowtie L_1$ →

Itemset
I1 I2
I1 I3
I1 I5
I2 I3
I2 I5
I3 I5

$L_1 \bowtie L_1$

2nd scan → C_2

Itemset	support
~~I1 I2~~	~~1~~
I1 I3	2
~~I1 I5~~	~~1~~
I2 I3	2
I2 I5	3
I3 I5	2

C_2 →

Item set
I1 I3
I2 I3
I2 I5
I3 I5

L_2

$L_2 \bowtie L_2$ →

Itemset
I1 I2 I5
I1 I2 I3 I5
I1 I3 I5
I2 I3 I5

$L_2 \bowtie L_2$

3rd Scan → C_3

Itemset	support
I1 I2 I3	1
I1 I2 I3 I5	1
I1 I3 I5	1
I2 I3 I5	2

C_3 → I2 I3 I5

Frequent Pattern

Q.1

Tid	Items
1	I1 I2 I5
2	I2 I4
3	I2 I3
4	I1 I2 I4
5	I1 I3
6	I2 I3
7	I1 I3
8	I1 I2 I3 I5
9	I1 I2 I3

SupMin = 2

→

Itemset	Support
I1	6
I2	7
I3	6
I4	2
I5	2

⟹

Itemset
I1 I2

PAGE No.
DATE / /

* Association Table Rules

Pattern : I2 I3 I5

I2 ⇒ I3 I5 = 2/3 = 66 %
I3 ⇒ I2 I5 = 2/3 = 66 %
I5 ⇒ I3 I2 = 2/3 = 66 %
I2 ⇒ I5 = 3/3 = 100 %
I2 ⇒ I3 = 2/3 = 66 %
I3 ⇒ I5 = 2/3 = 66 %
I3 I5 ⇒ I2 = 2/2
I2 I5 ⇒ I3 = 2/3
I3 I2 ⇒ I5 = 2/3

The association rules who's confidence & support is greater than given threshold value, it is called as strong association rule.

Drawback of Aprion :-

① Too many scans.

Q. A database has 5 transitions. let Min support be 60% & min confidence is 80%. Find all frequent item sets.

T100	'M', 'O', 'N', 'K', 'E', 'Y'
T200	'D', 'O', 'N', 'K', 'E', 'Y'
T300	'M', 'A', 'K', 'E'
T400	'M', 'U', 'C', 'K', 'Y'
T500	'C', 'O', 'O', 'K', 'I', 'E'

PAGE No.
DATE / /

* FP growth

Minsupport = 2

TID	Items
T1	E, A, D, B
T2	D, A, C, E, B
T3	C A B E
T4	B A D
T5	D
T6	D B
T7	A D E
T8	B C

scan ⟹

Itemset	support
A	5
B	6
C	3
D	6
E	4

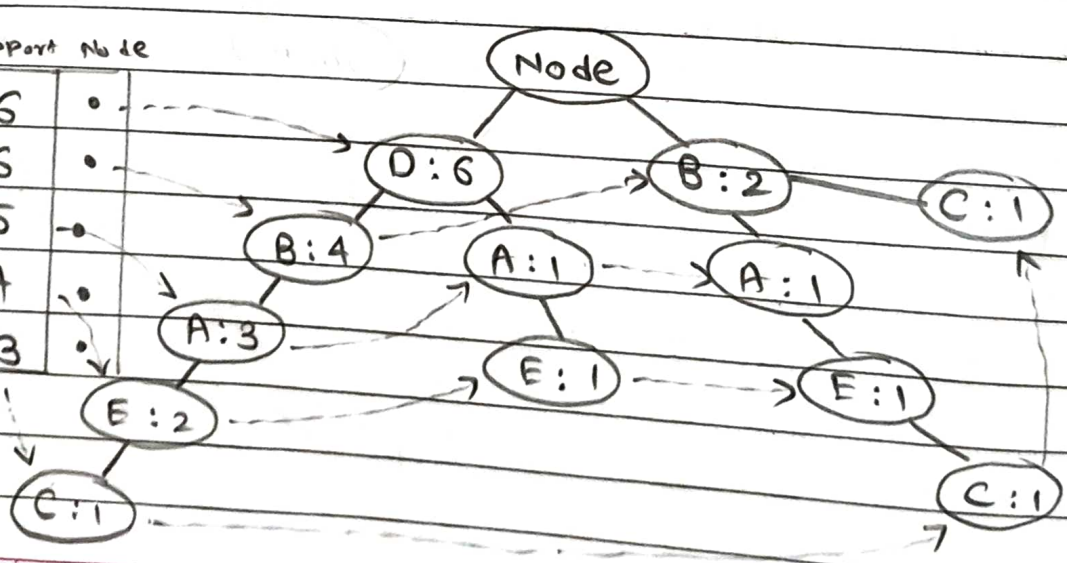
Decreasing order ⟹

TID	Items
T1	D, B, A, E
T2	D, B, A, E, C
T3	B, A, E, C
T4	D, B, A
T5	D
T6	D, B
T7	D, A, E
T8	B, C

⟸ Rearrange ~~Transactions~~

Itemset	support
D	6
B	6
A	5
E	4
C	3

Item	Support	Node
D	6	•
B	6	•
A	5	•
E	4	•
C	3	•



FP growth Tree

	Condition Frequent Base	FP
C	<B:1> <EAB:1> <EABD:1>	<B:3> <EAB:2> <A:2> <E:2> <AB:2> <EB:2> <EA:2>
E	<ABD:2> <AD:1> <AB:1>	<ABD:2> <A:4> <AB:3> <AD:3> <B:3> <D:3> <BD:2>
A	<BD:3> <D:1> <B:1>	<BD:3> <B:4> <D:4>
B	<D:4>	<D:4>
D		

Drawback :-

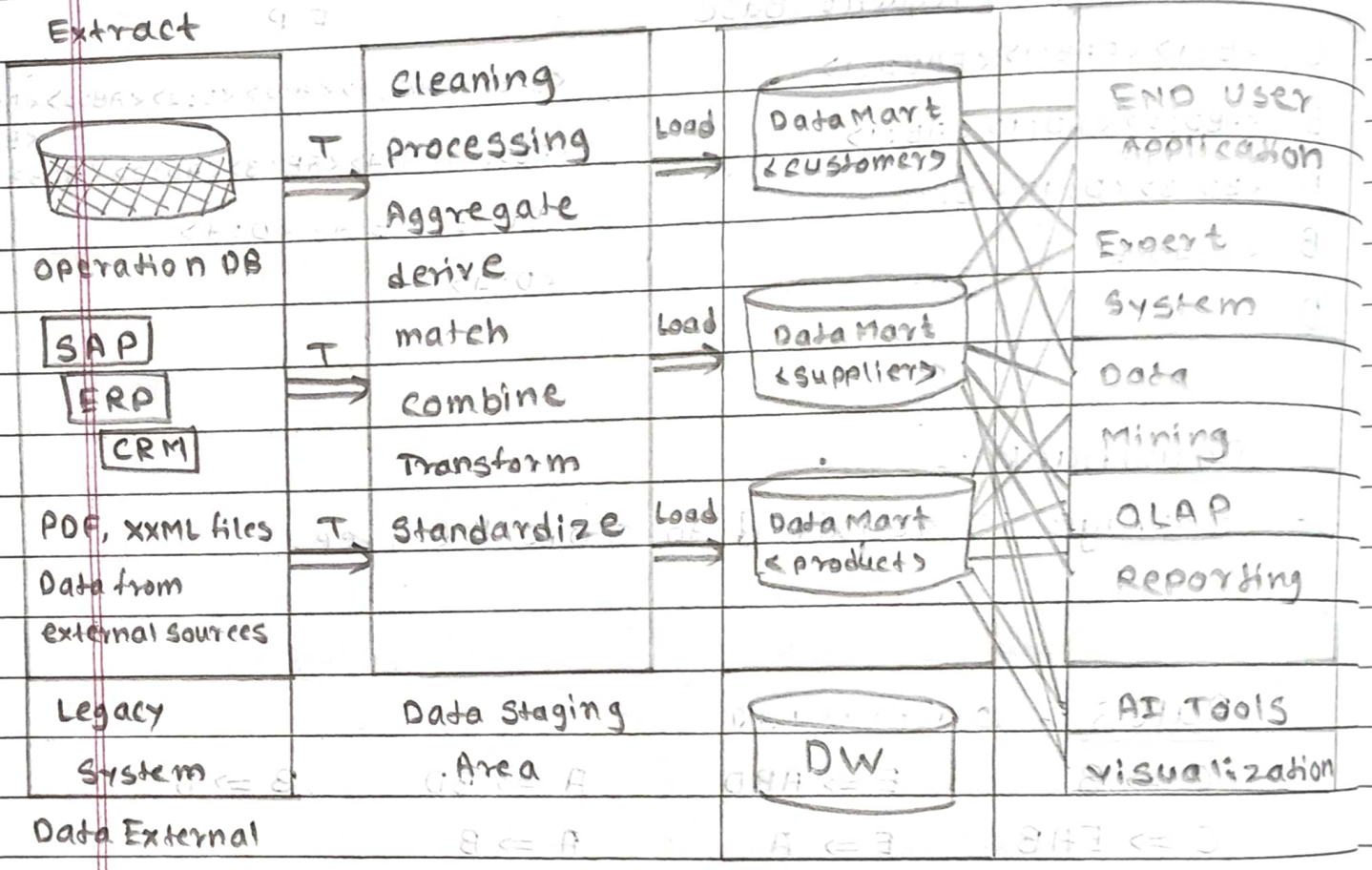
- Storage is required to store the tree.

Rules :-

C ⇒ B	E ⇒ ABD	A ⇒ BD	B ⇒ D
C ⇒ EAB	E ⇒ A	A ⇒ B	
C ⇒ A	E ⇒ AB	A ⇒ D	
C ⇒ AB	E ⇒ AD		
C ⇒ E	E ⇒ B		
C ⇒ EB	E ⇒ BD		
C ⇒ EA	E ⇒ D		

Q.

* Generic Architecture of Data warehouse :-



Data warehouse :-

1. Subject oriented
2. Integrated
3. Time variant
4. Non Updatable (read only)

Data warehouse :-

A subj collection of data used in support of management decision making process.

Data warehousing :-

Process of constructing and using datawarehouse