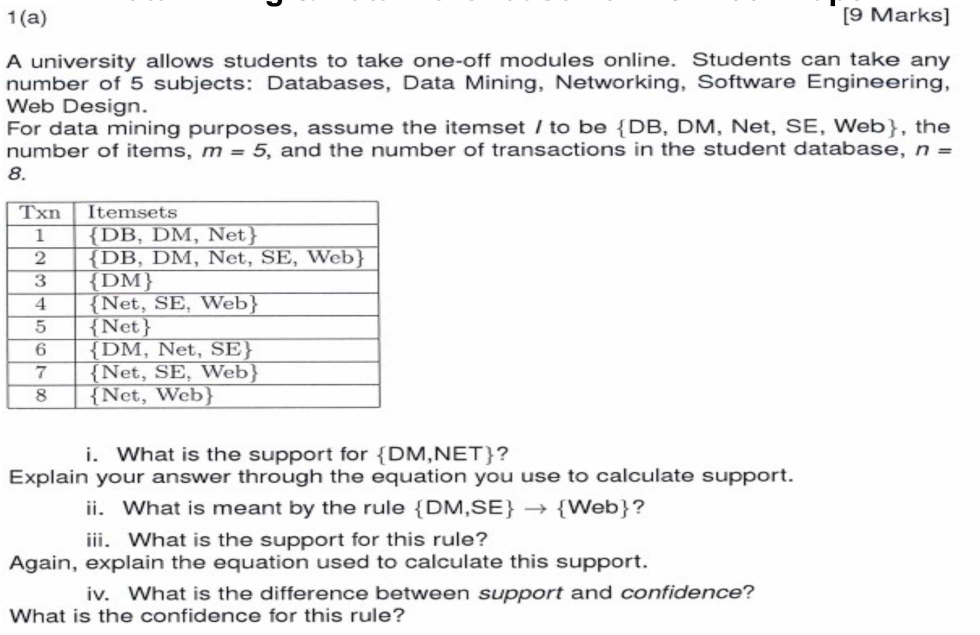
Data Mining & Data Warehouse 2014 CA4002 Paper



1(a) [9 Marks]

A university allows students to take one-off modules online. Students can take any number of 5 subjects: Databases, Data Mining, Networking, Software Engineering, Web Design.
For data mining purposes, assume the itemset *I* to be {DB, DM, Net, SE, Web}, the number of items, $m = 5$, and the number of transactions in the student database, $n = 8$.

Txn	Itemsets
1	{DB, DM, Net}
2	{DB, DM, Net, SE, Web}
3	{DM}
4	{Net, SE, Web}
5	{Net}
6	{DM, Net, SE}
7	{Net, SE, Web}
8	{Nct, Web}

i. What is the support for {DM,NET}?
Explain your answer through the equation you use to calculate support.

ii. What is meant by the rule {DM,SE} → {Web}?

iii. What is the support for this rule?
Again, explain the equation used to calculate this support.

iv. What is the difference between *support* and *confidence*?
What is the confidence for this rule?

I)

Support for an itemset is the proporition of transactions that contain all items in the itemset S the formula for this is:

Support(S) = Count(S)/n
where:
Count(S) is the number of transactions T matched by S, S matches T if S is a subset of T.
and n is the total number of transactions in the database

{DM,NET} matches the transactions 1,2,6 as all its items are in 1,2,6 (subset) so: Count({DM,NET}) = 3

Support({DM,NET}) = 3/8 = 0.375.

ii)

The rule {DM,SE} → {Web} means that if students choose the subjects DM and SE we can predict that they will also choose the subject Web.

iii)

Support for a rule L → R is the proporition of transactions in which the items in L and the items in R occur together L *union* R so support for a rule is calculated as:
Support(L → R) = Support(L *union R*) = Count(L *union* R)/n

{DM,SE,Web} matches the transaction 1 so:
Count({DM,SE,Web}) = 1

Support({DM,SE} → {Web}) = Support(L *union* R) = 1/8 = 0.125

iv)

Support of a rule is just how many transactions the rule applies to in the entire database where as the Confidence of a rule is defined as the proportion of transactions for which the rule is satisfied. This is calculated as the number of transactions matched by the left and right hand sides as a proportion of the number matched by the left on its own the formula is:

Confidence(L → R) = Count(L *union* R) /Count(L)

Count({DM,SE,Web}) = 1 *as shown above*
Count(L) = Count({DM,SE}) = 2

Confidence(L → R) = Count(L *union* R) / Count(L) = 1/2 = 0.5

1(b) [11 Marks]

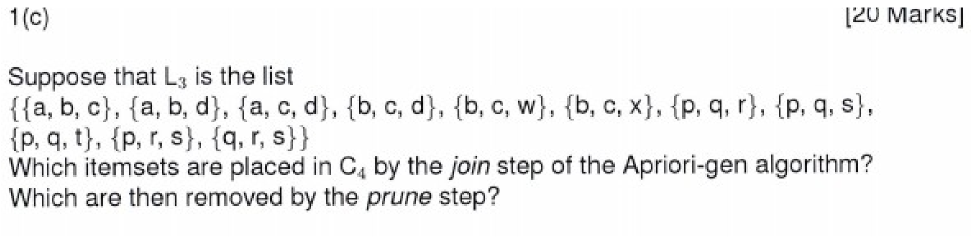
Describe, as a set of steps, the Apriori-gen algorithm that takes the L_{k-1} itemset and generates the new C_k itemset.

b)

The Apriori-gen algorithm uses two steps to take the L_{k-1} itemset and generate the new C_k itemset. These two steps are the **Join Step** and the **Prune Step**. In the Join step you compare each member A of L_{k-1} with every other member B in turn. If the first **k-2** items in A and B are the same, place the union of the two into C_k. The next step is the Prune step where you iterate all members c of C_k in turn and Examine all subsets of c with **k-1** elements and delete c from C_k if and of the subsets is not a member of L_{k-1}

Pesudo Code for Apori-gen:

```
$C_k$ = empty;
//join step
for( A : $L_{k-1}$)
{
   for(B : $L_{k-1}$)
   {
      if(A!=B)
      {
         if(A.sub(0,k-2) == B.sub(0,k-2) && !$C_k$.contains(union(A,B)))
              $C_k$ .add(union(A,B));
       }
   }
}
//prune step
for( c : $C_k$)
{
   for( sub : subset(c, k-1))
   {
     if( !$L_{k-1}$.includes(sub) )
     {
        $C_k$.remove(c);
        break;
     }
   }
}
return $C_k$;
```



1(c) [20 Marks]

Suppose that L_3 is the list
{{a, b, c}, {a, b, d}, {a, c, d}, {b, c, d}, {b, c, w}, {b, c, x}, {p, q, r}, {p, q, s}, {p, q, t}, {p, r, s}, {q, r, s}}
Which itemsets are placed in C_4 by the *join* step of the Apriori-gen algorithm?
Which are then removed by the *prune* step?

c)

Join step:
k = 4, k-2 = 3

First itemset	Secound itemset	Contribution to C_4
{a,b,c}	{a,b,d}	{a,b,c,d}
{b,c,d}	{b,c,w}	{b,c,d,w}
{b,c,d}	{b,c,x}	{b,c,d,x}
{p,q,r}	{p,q,s}	{p,q,r,s}
{p,q,r}	{p,q,t}	{p,q,r,t}

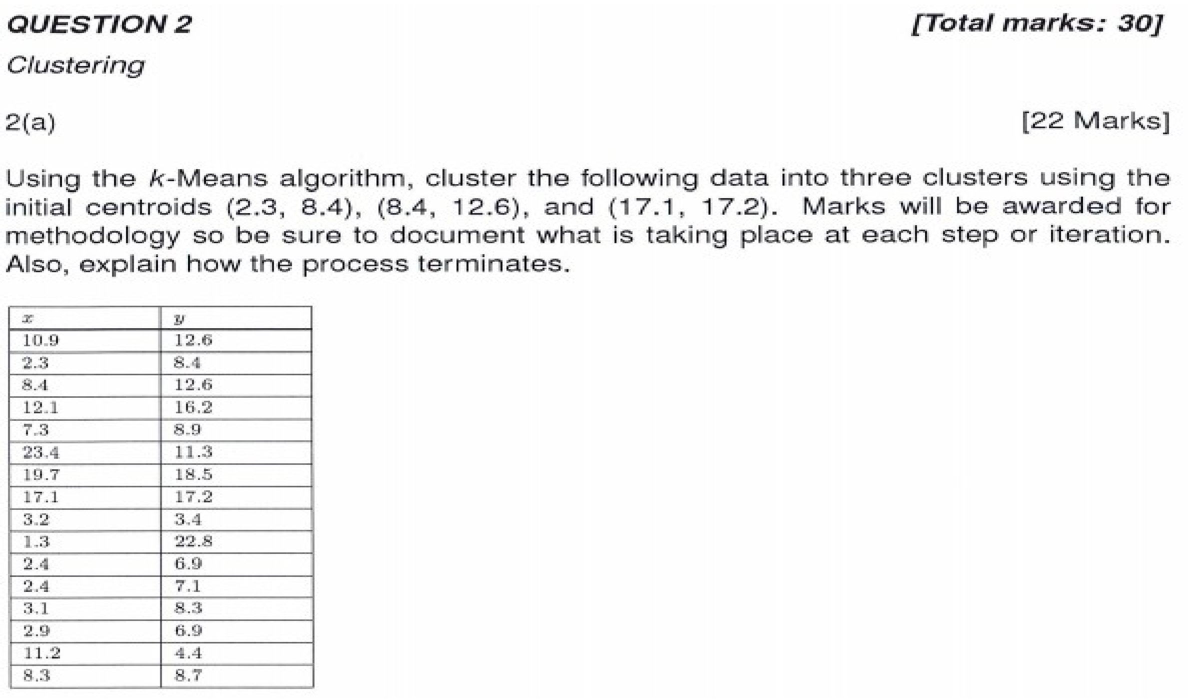
Items placed after join step:
C_4 = { {a,b,c,d} , {b,c,d,w} , {b,c,d,x} , {p,q,r,s} , {p,q,r,t} }

Prune step:

Itemset in C_4	Subsets of C_4 with 3 elements	Subsets all in L_3 ?
{a,b,c,d}	{{a,b,c} , {a,c,d} , {b,c,d} , {a,b,d}}	Yes
{b,c,d,w}	{{b,c,d} , {b,d,w} , {c,d,w} , {b,c,w}}	No
{b,c,d,x}	{{b,c,d} , {b,d,x} , {c,d,x} , {b,c,x}}	No
{p,q,r,s}	{{p,q,r} , {p,q,s} , {q,r,s} , {p,r,s}}	Yes
{p,q,r,t}	{{p,q,r} , {p,q,t} , {q,r,t} , {p,r,t}}	No

Items removed after prune step = {b,c,d,w} , {b,c,d,x} , {p,q,r,t}

C_4 = {{a,b,c,d} , {p,q,r,s}}



QUESTION 2 **[Total marks: 30]**

Clustering

2(a) [22 Marks]

Using the *k*-Means algorithm, cluster the following data into three clusters using the initial centroids (2.3, 8.4), (8.4, 12.6), and (17.1, 17.2). Marks will be awarded for methodology so be sure to document what is taking place at each step or iteration. Also, explain how the process terminates.

x	y
10.9	12.6
2.3	8.4
8.4	12.6
12.1	16.2
7.3	8.9
23.4	11.3
19.7	18.5
17.1	17.2
3.2	3.4
1.3	22.8
2.4	6.9
2.4	7.1
3.1	8.3
2.9	6.9
11.2	4.4
8.3	8.7

a)

First we decide how many clusters *k* we want to create after this we then select *k* points which will act as our initial centroids. In the question we are to create 3 clusters so *k* = 3 and our initial centroids are (2.3, 8.4) , (8.4, 12.6) and (17.1, 17.2).

Next for each point we calculate the distance to the centroid using a distance function. The distance function we will use is the euclidean distance:

$$d(p,q)=\sqrt{(p_1-q_1)^2+(p_2-q_2)^2+(p_n-q_n)^2}$$

$$d(p,q)=\sqrt{\sum_{i=1}^{n}(p_i-q_i)^2}$$

The centroid that the point is closest to is the cluster that point belongs to.

First iteration:

point	X	Y	D1	D2	D3	C
1	10.9	12.6	9.57	2.5	7.72	2
2	2.3	8.4	0			1
3	8.4	12.6	7.4	0		2
4	12.1	16.2	12.52	5.16	5.09	3
5	7.3	8.9	5.02	3.86	12.64	2
6	23.4	11.3	21.29	15.05	8.63	3
7	19.7	18.5	20.11	12.74	2.9	3
8	17.1	17.2			0	3
9	3.2	3.4	5.08	10.56	19.58	1
10	1.3	22.8	14.43	12.42	16.76	2
11	2.4	6.9	1.5	8.27	17.95	1
12	2.4	7.1	1.3	8.13	17.83	1
13	3.1	8.3	0.8	6.82	16.58	1
14	2.9	6.9	1.6	7.92	17.54	1

15	11.2	4.4	9.75	8.66	14.09	2
16	8.3	8.7	6	3.9	12.23	2

Cluster 1 = {2 , 9, 11, 12, 13, 14} $C_1 = 2$
Cluster 2 = {1,3,5,10,15,16} $C_2 = 3$
Cluster 3 = {4,6,7,8} $C_3 = 8$

we now have the clusters for the first iteration but the centroids are no longer the true centroids of the above clusters so we have to recalculate the centroids and repeat the above calculations with the new centroids. We repeat this process until the centroids no longer change. To calculate the centroids we just use:

$$C_k=(\frac{p_{11}+p_{12}+..p_{1n}}{n},\frac{p_{21}+p_{22}+..p_{2n}}{n},....\frac{p_{i1}+p_{i2}+..p_{in}}{n})$$

Our new centroids after first iteration are:

$$C_1=(\frac{2.3+3.2+2.4+2.4+3.1+2.9}{6},\frac{8.4+3.4+6.9+7.1+8.3+6.9}{6})=(2.7,6.8)$$

$$C_2=(\frac{10.3+8.4+7.3+1.3+11.2+8.3}{6},\frac{12.6+12.6+16.2+8.9+22.8+4.4+8.7}{6})=(7.9,11.6)$$

$$C_3=(\frac{12.1+23.4+19.7+17.1}{4},\frac{16.2+11.3+18.5+17.2}{4})=(18.0,15.8)$$

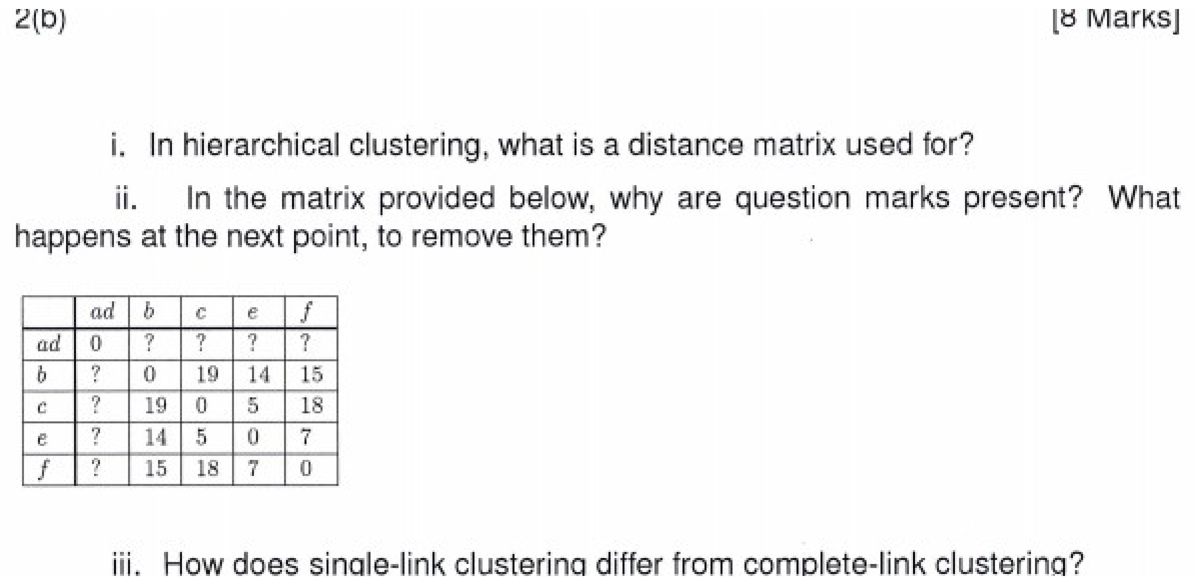
second iteration:

point	X	Y	D1	D2	D3	C
1	10.9	12.6	10.04	3.16	7.78	2
2	2.3	8.4	1.64	6.44	17.35	1
3	8.4	12.6	8.13	1.11	10.11	2
4	12.1	16.2	13.29	6.22	5.91	3
5	7.3	8.9	5.05	2.76	12.73	2
6	23.4	11.3	21.18	15.5	7.02	3
7	19.7	18.5	20.63	13.66	3.19	3
8	17.1	17.2	17.76	10.77	1.66	3
9	3.2	3.4	3.43	9.45	19.3	1
10	1.3	22.8	16.06	13	18.1	2
11	2.4	6.9	0.31	7.23	17.96	1
12	2.4	7.1	0.42	7.1	17.86	1
13	3.1	8.3	1.55	5.82	16.68	1
14	2.9	6.9	0.22	6.86	17.52	1
15	11.2	4.4	8.83	7.92	13.27	2
16	8.3	8.7	5.9	2.92	12.02	2

Cluster 1 = {2,9,11,12,13,14}
Cluster 2 = {1,3,5,10,15,16}
Cluster 3 = {4,6,7,8}

All points stay in their same cluster so the centroids have not changed so we can stop.

d1,d2,d3 calculated with Q2Dists.java. change the c values to the centroid values. (Will have to do them all on a calculator for the exam.......... going to take like an hour!!!!)



2(b) [8 Marks]

i. In hierarchical clustering, what is a distance matrix used for?

ii. In the matrix provided below, why are question marks present? What happens at the next point, to remove them?

	ad	*b*	*c*	*e*	*f*
ad	0	?	?	?	?
b	?	0	19	14	15
c	?	19	0	5	18
e	?	14	5	0	7
f	?	15	18	7	0

iii. How does single-link clustering differ from complete-link clustering?

b)
I)

In hierarchical clustering a distance matrix is used to store the distances between each pair of clusters we use a distance matrix as it would be very inefficient to calculate the distance between each pair of clusters for each pass of the algorithm. Distance matrices allow us to only calculate and update distances for clusters involved in the recent merger eliminating any unnecessary recalculations.

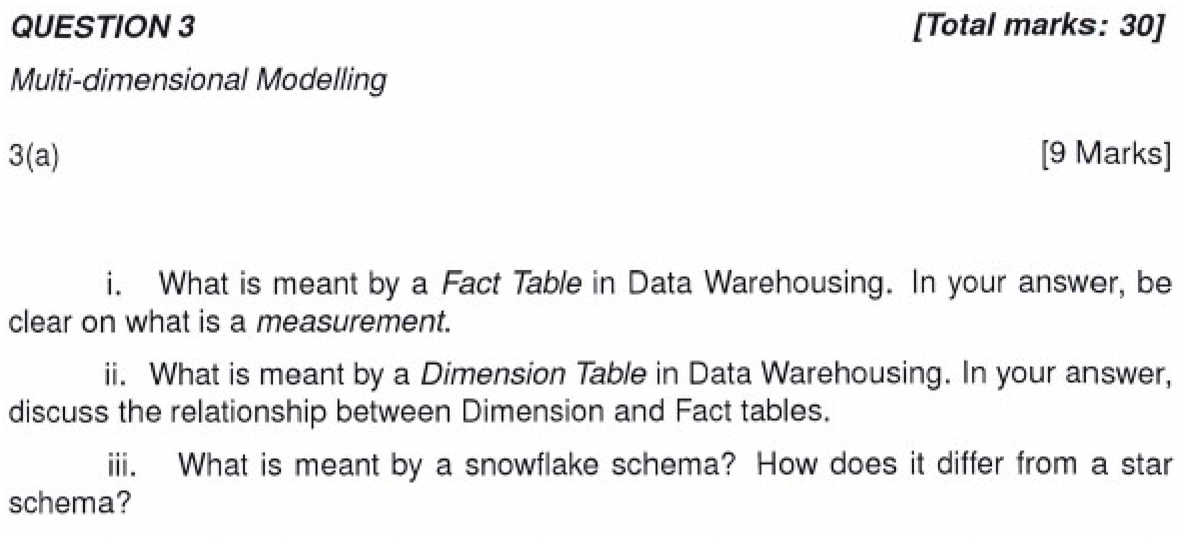
ii)

The question marks are present as we have just combined the clusters a and d as they were the closest and have yet to calculate distance between the new cluster and the previous ones. To remove them we have to calculate the distance between the new cluster and each of the others using either single-link or complete-link.

Iii)

In single-link clustering the distance between two clusters is taken as the shortest distance from any member of one to and member of the other.

In complete-link the distance is taken as the longest distance from any member of one to any member of the other.



QUESTION 3 ***[Total marks: 30]***

Multi-dimensional Modelling

3(a) [9 Marks]

i. What is meant by a *Fact Table* in Data Warehousing. In your answer, be clear on what is a *measurement*.

ii. What is meant by a *Dimension Table* in Data Warehousing. In your answer, discuss the relationship between Dimension and Fact tables.

iii. What is meant by a snowflake schema? How does it differ from a star schema?

a)
i)

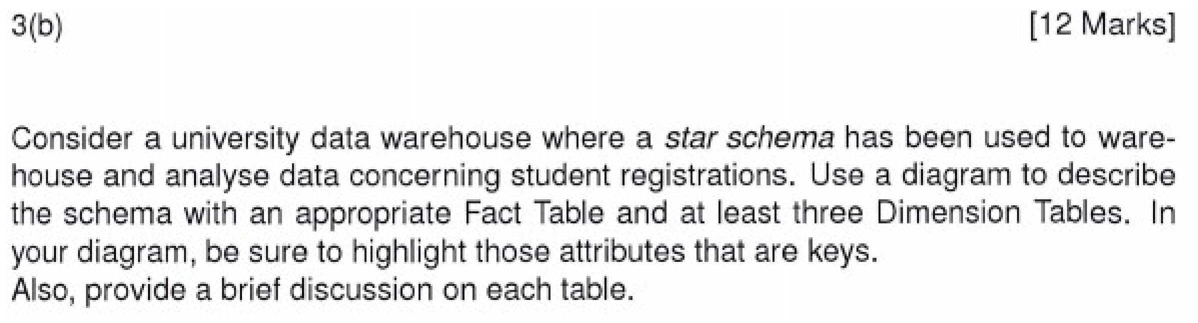
A fact table contains the bulk of the data it consists of measurements, metrics or facts (ie, sales, number of products, units sold etc.) The fact table is located in the centre of a star or a snowflake schema surrounded by dimension tables. When multiple fact tables are used they are arrange in a fact constellation schema. Fact tables typically have two types of columns those that contain facts and those that are foreign keys to dimension tables.

ii)

a dimension table is a table that stores attributes that describe the objects in the fact table. Each table has a primary key that relates back to foreign key stored in the fact table. Dimension tables categorise and describe data warehouse facts and measures in a way to support meaningful answers to business questions.

Iii)

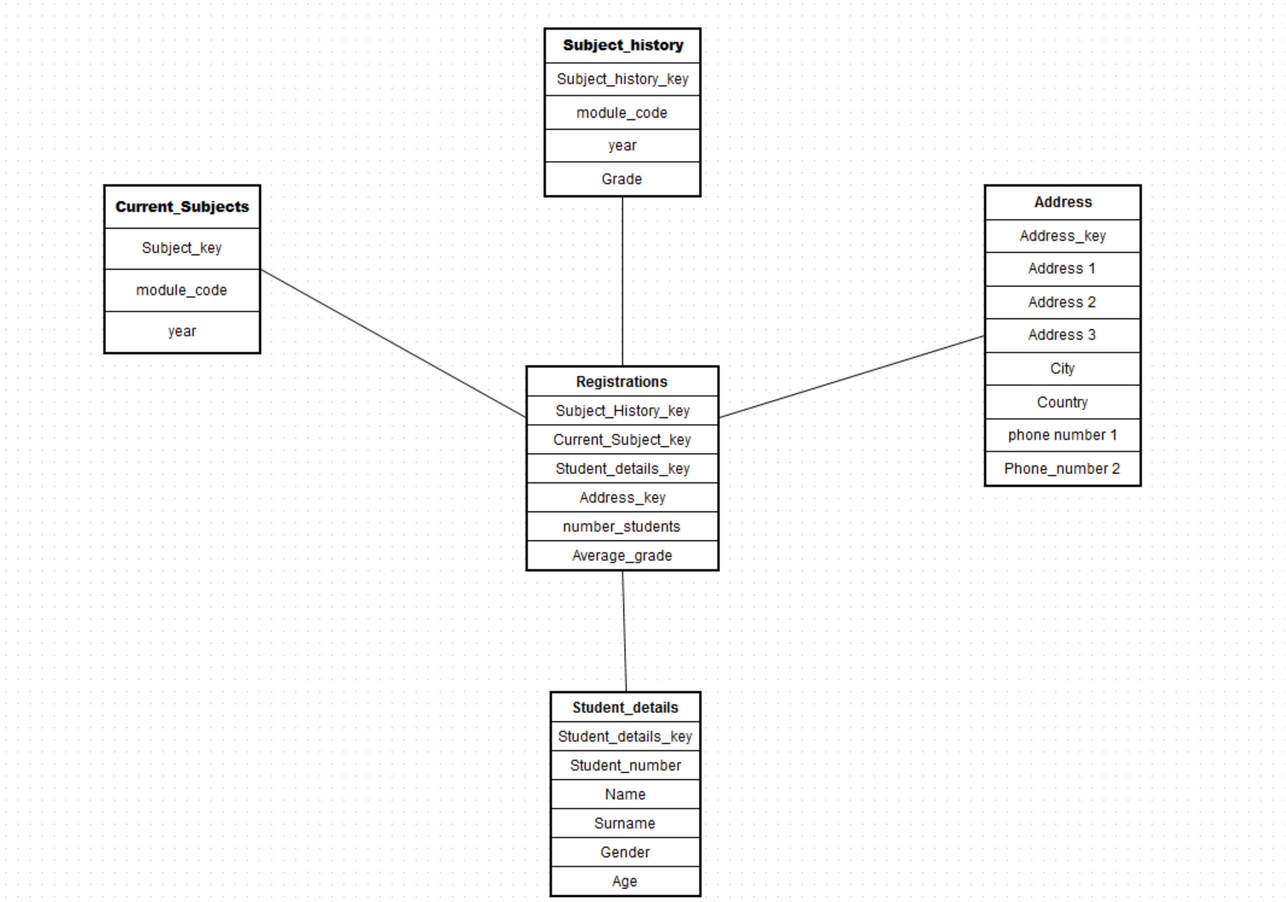
A snowflake schema is a variant of a star schema where some dimension tables are normalized thereby further spiting the data into additional tables.



3(b) [12 Marks]

Consider a university data warehouse where a *star schema* has been used to warehouse and analyse data concerning student registrations. Use a diagram to describe the schema with an appropriate Fact Table and at least three Dimension Tables. In your diagram, be sure to highlight those attributes that are keys.
Also, provide a brief discussion on each table.

b)



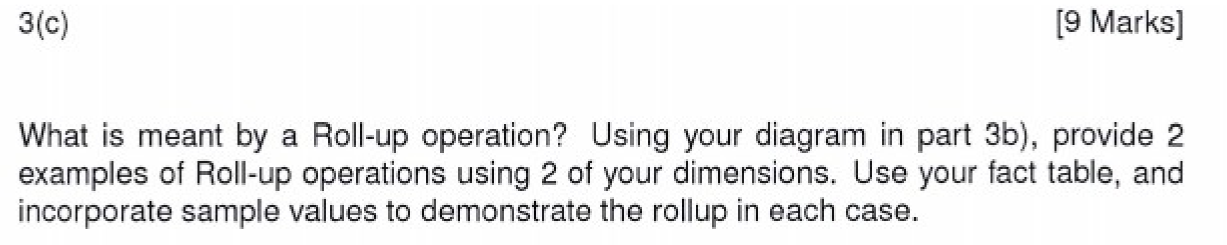
Registrations: the fact table situated in the centre of the the schema holds the foreign keys to each dimensions and a measurement on the total students registered for the subject a nd the students average grade.

Address: dimensions table containing the primary key address_key along with attributes on the students location

Student_details: dimension table holding attributes used to describe the student.

Subject_History : Holds data detailing the students previous subjects.

Current_subjects: Holds data on what subjects the student has chosen on registration.

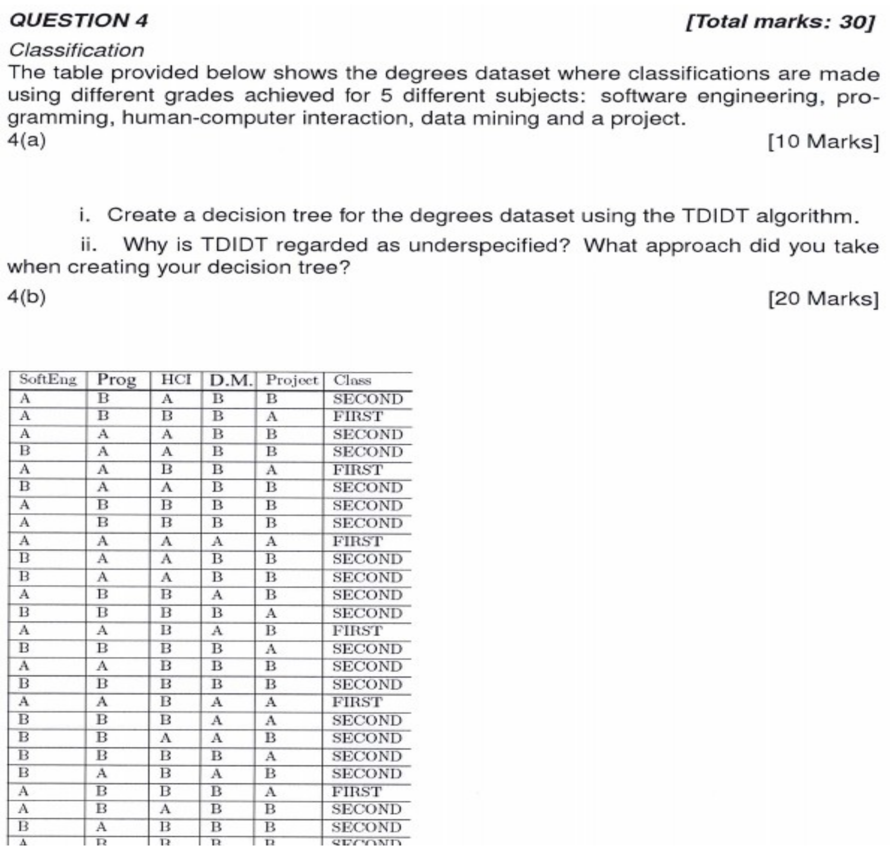


3(c) [9 Marks]

What is meant by a Roll-up operation? Using your diagram in part 3b), provide 2 examples of Roll-up operations using 2 of your dimensions. Use your fact table, and incorporate sample values to demonstrate the rollup in each case.

C)

A Roll-Up operation is to summarize data by climbing up hierarchy or by dimension reduction.



QUESTION 4 **[Total marks: 30]**

Classification

The table provided below shows the degrees dataset where classifications are made using different grades achieved for 5 different subjects: software engineering, programming, human-computer interaction, data mining and a project.

4(a) [10 Marks]

i. Create a decision tree for the degrees dataset using the TDIDT algorithm.

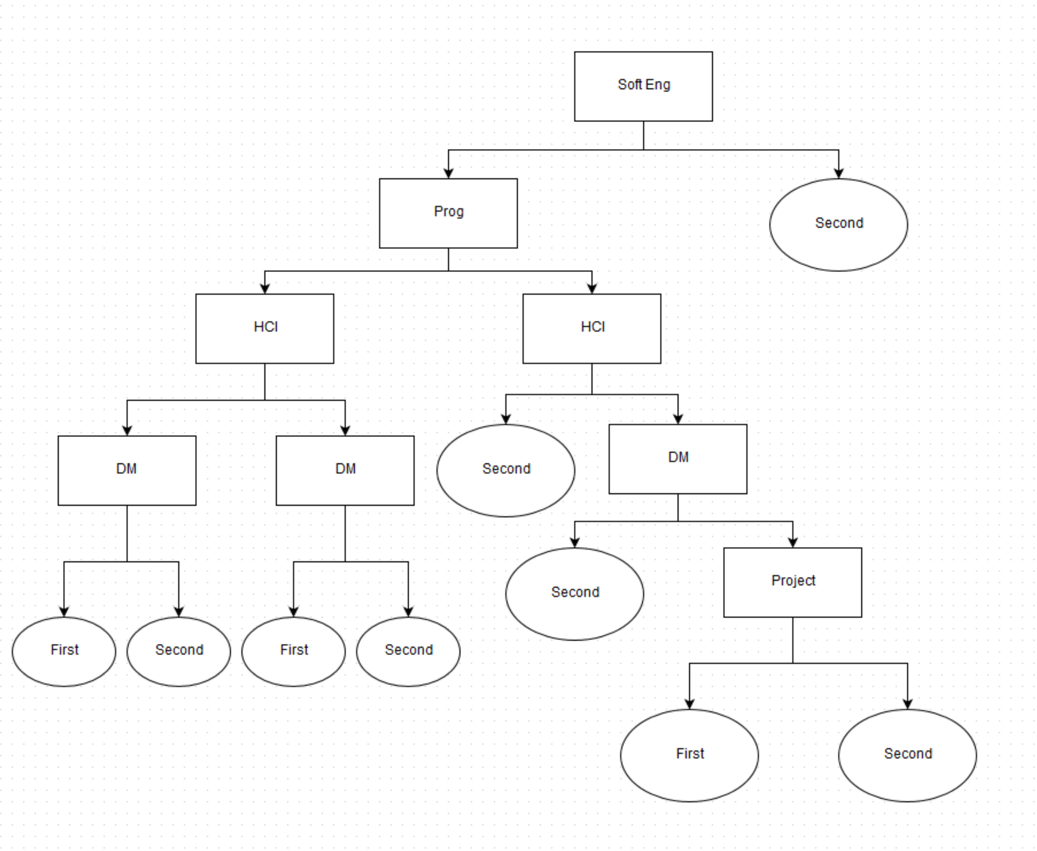
ii. Why is TDIDT regarded as underspecified? What approach did you take when creating your decision tree?

4(b) [20 Marks]

SoftEng	Prog	HCI	D.M.	Project	Class
A	B	A	B	B	SECOND
A	B	B	B	A	FIRST
A	A	A	B	B	SECOND
B	A	A	B	B	SECOND
A	A	B	B	A	FIRST
B	A	A	B	B	SECOND
A	B	B	B	B	SECOND
A	B	B	B	B	SECOND
A	A	A	A	A	FIRST
B	A	A	B	B	SECOND
B	A	A	B	B	SECOND
A	B	B	A	B	SECOND
B	B	B	B	A	SECOND
A	A	B	A	B	FIRST
B	B	B	B	A	SECOND
A	A	B	B	B	SECOND
B	B	B	B	B	SECOND
A	A	B	A	A	FIRST
B	B	B	A	A	SECOND
B	B	A	A	B	SECOND
B	B	B	B	A	SECOND
B	A	B	A	B	SECOND
A	B	B	B	A	FIRST
A	B	A	B	B	SECOND
B	A	B	B	B	SECOND
A	B	B	B	B	SECOND

a)

I)



ii)

TDIDT is under specified as it does not give a method in which to select attributes.

The approach I had to use to select attributes to generate the above tree was the TakeFirst approach.