Name: Sanket Chandrashekhar Harvande
Roll No : 19

Assignment No. 01

Date :

Q.1 What is meant by metadata in the context of a Data warehouse? Explain the different types of metadata stored in the data warehouse. Illustrate with a suitable example.

→ a) Metadata is simply defined as data about data. The data that is used to represent other data, is known as metadata.

b) Metadata are created for the data names & definitions of given warehouse.

c) Additional metadata are created & captured for timestamping any extracted data. the source of the extracted data and missing fields that have been added by data cleaning or integration process.

d) Metadata act as a directory. The directory helps the decision support system to locate the contents of a data warehouse.

e) Metadata is the road map to a data warehouse.

f) Metadata can be broadly categorized into three types :

I. Business Metadata :-

It is the data ownership information metadata are the data that remains warehouse objects.

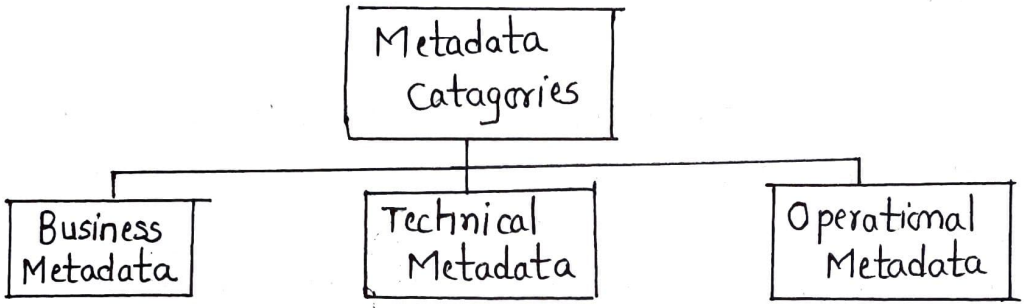
II. Technical Metadata :-

It includes database system names - table & column names and sizes, data types and allowed values. It also includes structural information, such as primary & foreign key attributes, and indices.

III. Operational Metadata :-

It includes currency of data & data and data lineage. Currency of data means whenever the data is active

archived or purged. Lineage of data means the history of data migrated & transformation.

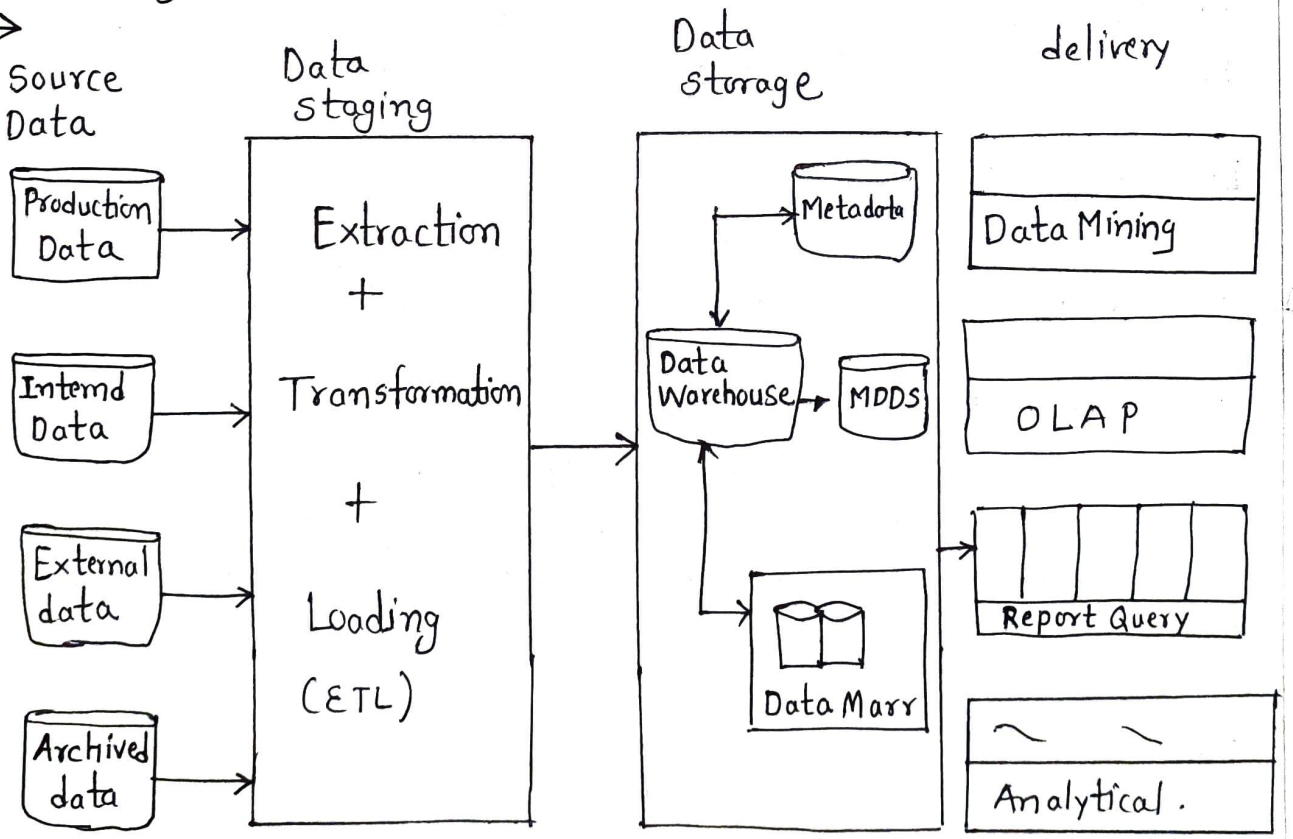


Example:-

The index of book service as a metadata for the contents in the book. Metadata is the summarized data that leads to detailed data.

Q.2 Explain the architecture of Data Warehouse with a suitable block diagram.

→



Date :

i) Data Source layer :-

The layer representing various data sources that data into the datawarehouse. The data into the datawarehouse. The data into the datawarehouse. The data can be in any of These formats : Plain tetx fire, relational database, Excel File & other types of data can act as data sources.

a. Production source - Represents sales data, HR data & Product data.

b. Internal data - Represents data of a deported of an organisation such as employ data

c. External data :- Represents data from outside the organisation of third party such as survey data of demographic.

d. Achieved data :- Represents logs of the web server along with the user's browsing data.

ii) Data staging layer :-

The storage area for data processing where data comes before being transformed into the data, that is entered in a data that entered in a datawarehouse.

a. Extractim - The process of extracting data from different source system and validating it against certain quality.

b. Data Warehouse - It is maintained by organisations ac central warehouse of data that can be equally accessed by all business experts & end users.

c. Loading :- The process of loading the data either from datawarehouse data mart.

(iii) Data Storage layer :-

The Layer in which the transformed data & cleaned data is stored.

a. Data Warehouse :- It is maintained by organisations as central warehouse of data that can be equally accessed by all business experts & end users.

b. Data Mart :- When data warehouse is created at the department level is known as data mart.

c. Metadata :- Details about the data is known as metadata In other words, it is a catalog of data warehouse.

d. MDDS - It is multi dimensional database that allows data to be molded & viewed in multiple dimensions. It is defined by dimensions & facts.

iv) Information Delivery:

It provides the information that reached to end users. The information that can be in any form such as tables, chart, graphs or histograms.

a. Data Mining :- The process of finding relevant & useful information, large amount of data.

b. OLAP :- Allows the navigation of data of different levels abstraction. such as down, rollup, slice, dice, & so on.

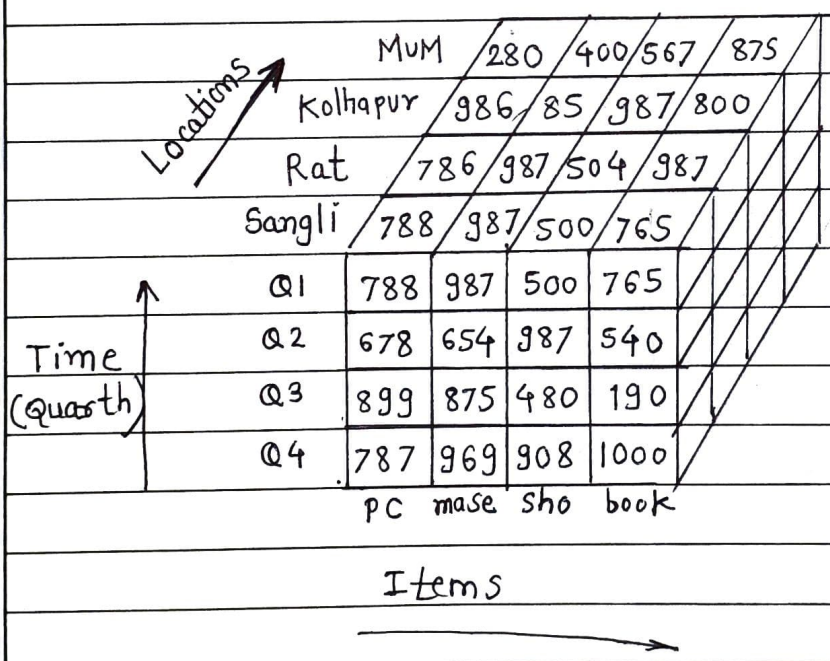
Date :

Q.3 We would like to view sales data of a company w.r.t. 3D's namely location, Item time. Represents sales data in the form of a 3D data cube for the above and perform Roll-up, Drill down, Slice, Dice & pivot. OLAP operations on the above data cube & illustrate.

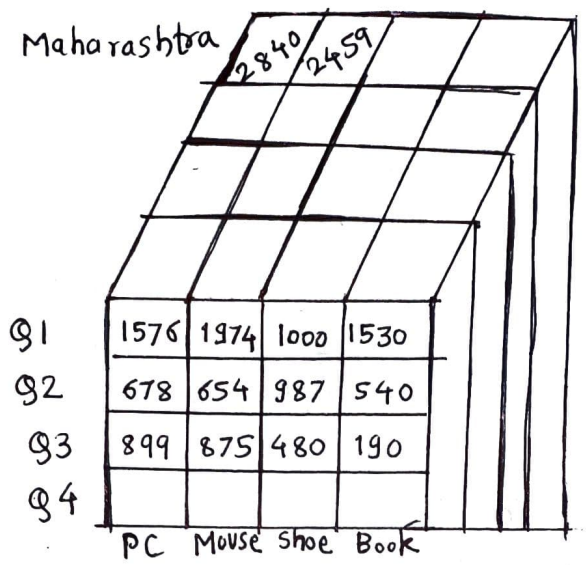
→

Dimensions → Location, Time

Facts → Sales, Data.

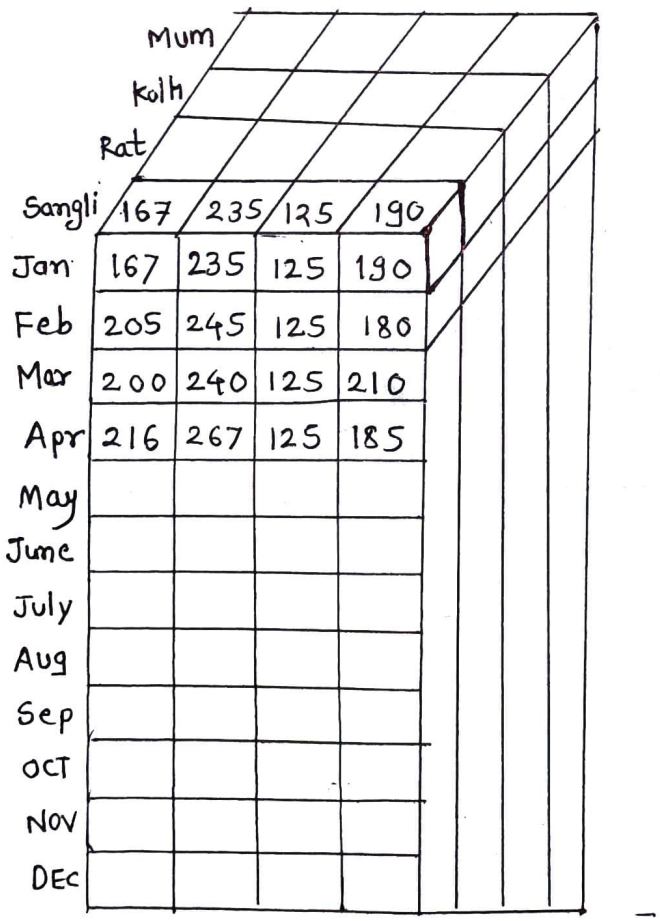


i) Roll-up



(Roll-up Location)

ii) Roll-down



Date :

iii) Slice

Mum				
Kolh				
Rat				
Sangli	788	987	500	765
	PC	Mouse	shoe	Book

iv) Dice

	Ratnagiri	786	987
	Sangli	788	987
Q1 :	788	987	
Q2 :	678	697	
	PC	Mouse	

v) Divot

			788	PC
			987	Mouse
			500	shoe
			765	Book
Mum	kolh	Ratn	Sangli	

Q.4 Consider a data warehouse for a hospital where there are three dimensions.

a. Doctor b. Patient c. Time

Consider two measures

i) Count

ii) charge where charge is the ~~same~~ Fee that the doctor charges a patient for a visit for the above example, create a cube & illustrate the OLAP operations.

→ There are four tables out of 3 dimensions tables & 1 fact tables

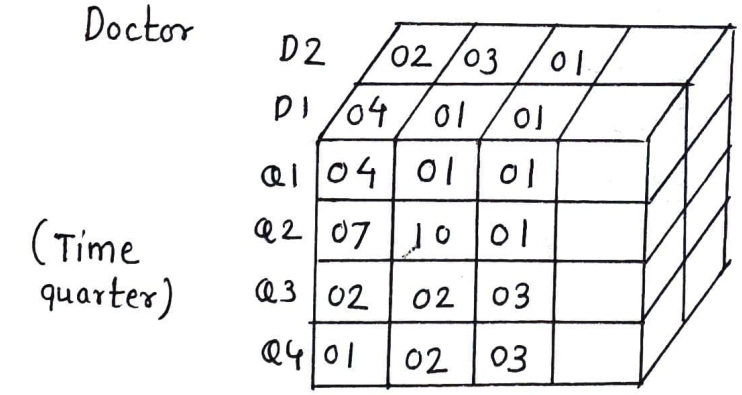
Dimension tables :-

i) Doctor (DID, name, phone, number, Location, pin, specialization)

ii) Patient (PID, name, phoneno, state, city, Location, pin)

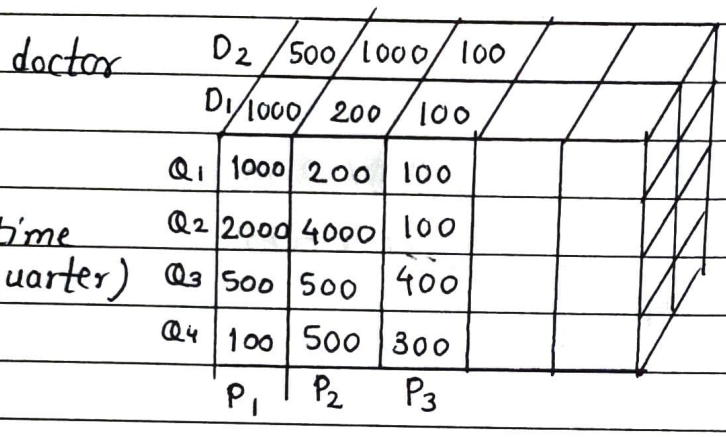
iii) Time (TID, Day, month, quarter, year)

Fact table :- Fact table (D_ID, PID, TID, count, charge)



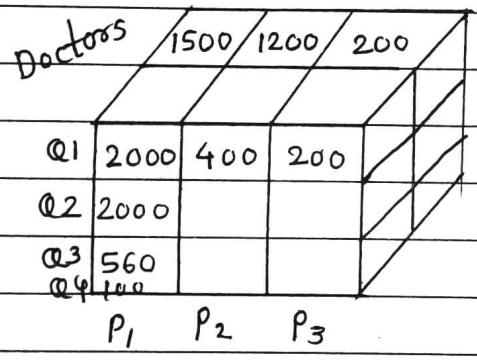
Fact - Count (No. of times patient visited doctor)

Date :



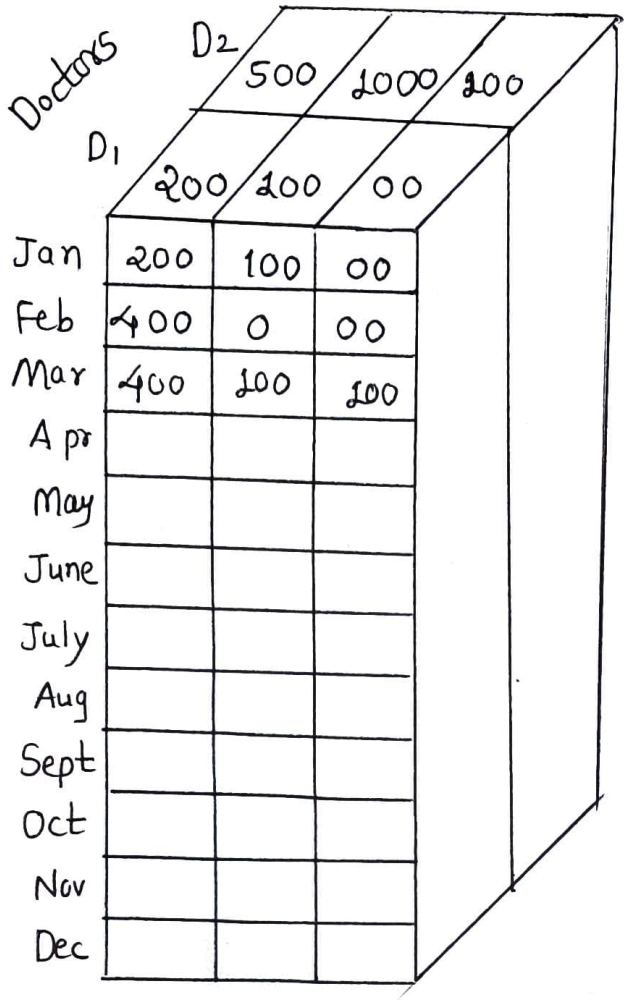
fact - sum or fees paid by patient.

- Roll-up



Roll Up for fees.

- Drill down :-

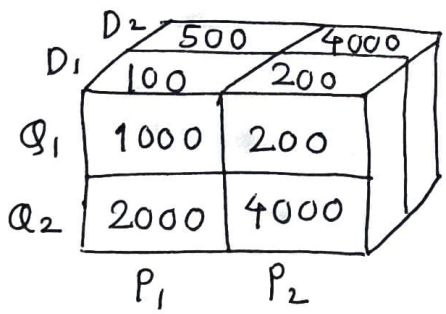


Drill Down for fees

- Slice

D_2	500	1000	100
D_1	1000	200	100
	P_1	P_2	P_3

Dice



- Pivot :-

D_2	500	1000	100
D_1	1000	200	100

↓

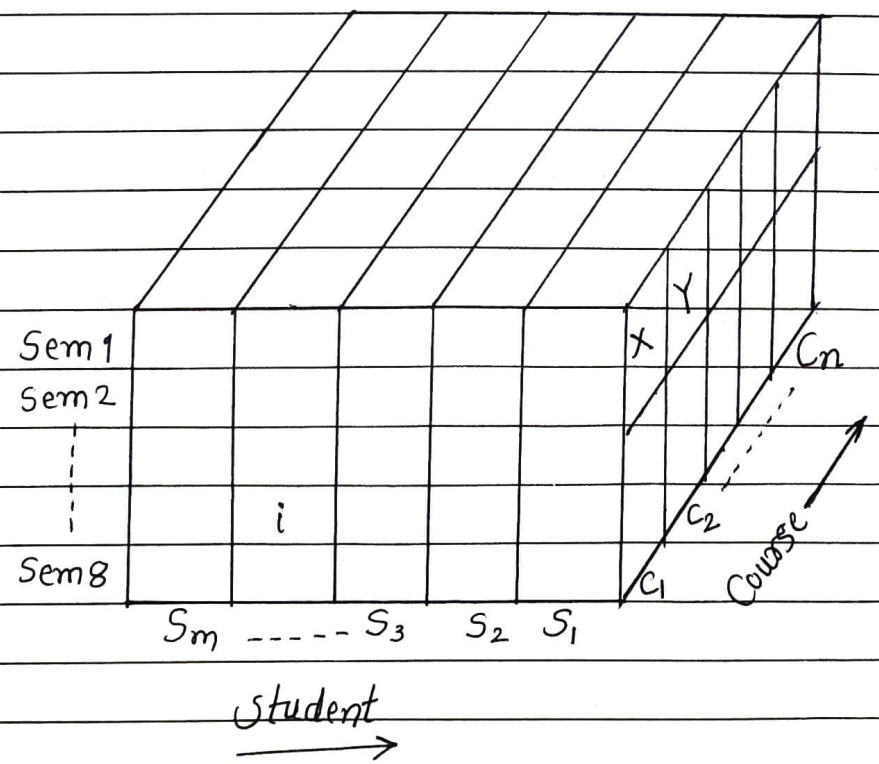
1000	500	P_1
200	1000	P_2
100	100	P_3

Date :

Q.5 The college wants to record the marks for the course completed by students using the dimensions
a) course b) students c) Time & measure Aggregate marks Create a cube & apply the OLAP operations on the cube for illustration.

→

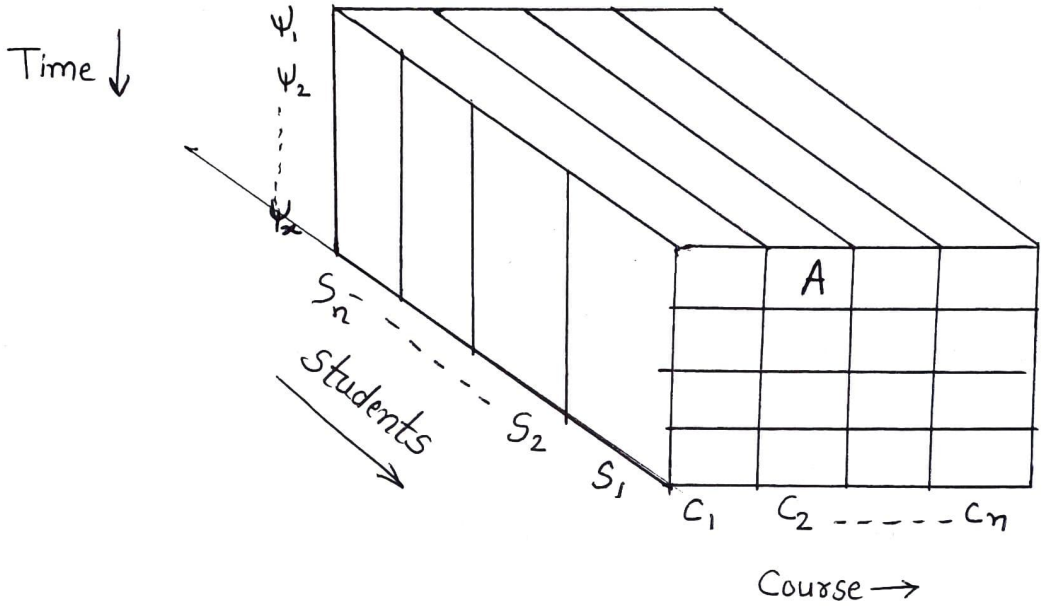
Marks obtained ⇒ fact
Dimensions - course, student & time
Let the current level of gram is each cell gives aggregate marks of each student for an individual duster of course for each semester.
"X" is the total aggregate marks obtained by students 'Si' in semester 1 for all the subjects that belongs to cluster C1.



Roll-Up:-

Obtaining aggregated/generalised level of information from the current level is called roll-up.

Here we obtain year wise aggregate marks from semesterwise aggregate marks.



Here "A" is aggregate total marks of student S_1 in all the courses of cluster C_1 in the year "Ψ" i.e.

$$A = x + z.$$

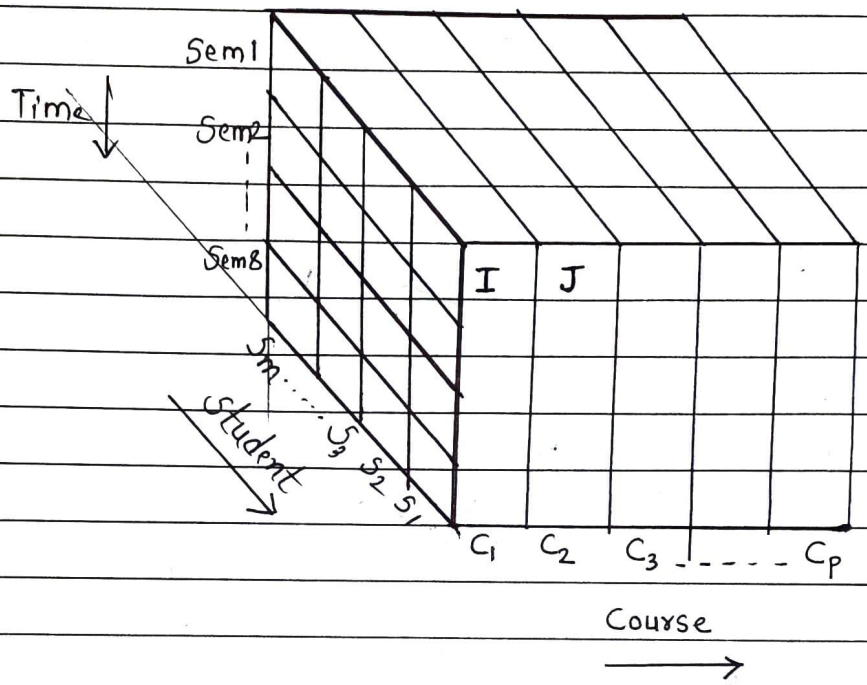
Drill Down:-

Obtaining detailed level of information from the current level is called down.

For example, getting weekly profit from montly profit. It is done by decomposing the cells along one or more dimensions.

In this example, obtaining aggregate marks for individual courses from cluster level.

Date :

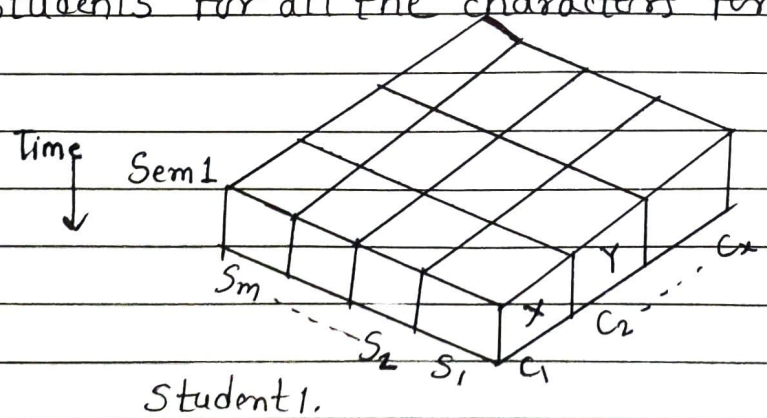


Here, "I" is the aggregate marks obtained by student S1 in semester for course that belong to cluster C1.

Slice & Dice :-

obtaining one specific part of cube for detailed investigation is called slice & dice. for example getting profit or all products for all the years for a specific region.

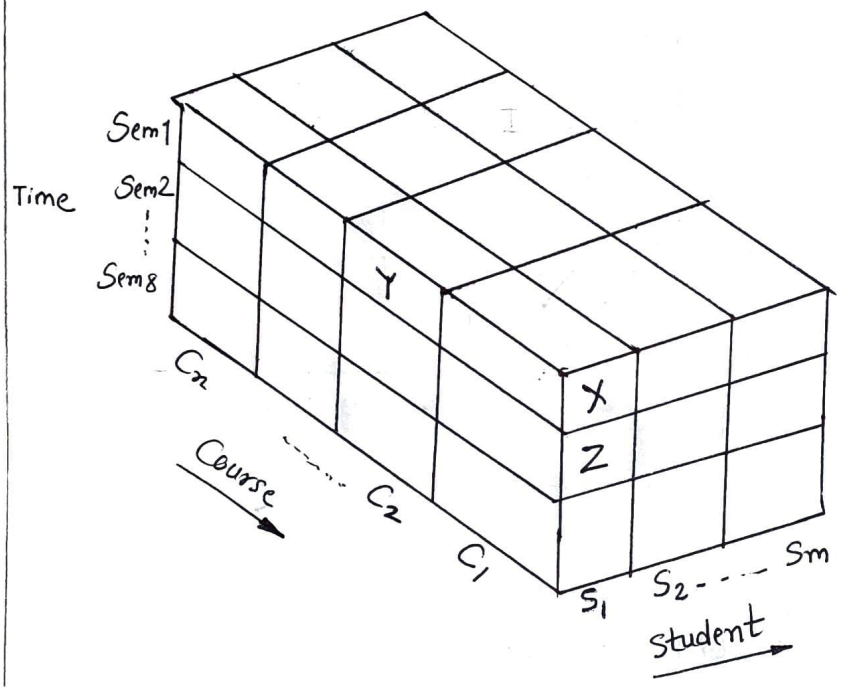
In this example, obtaining aggregate marks of all the students for all the characters for sem 1.



Pivot :-

It is rotation of the cube, it is used to analyze the same data from different perspective, for example, anlyzing profit region wise & then product wise.

In this example, analysing aggregate marks obtained by students character wise & then student wise.

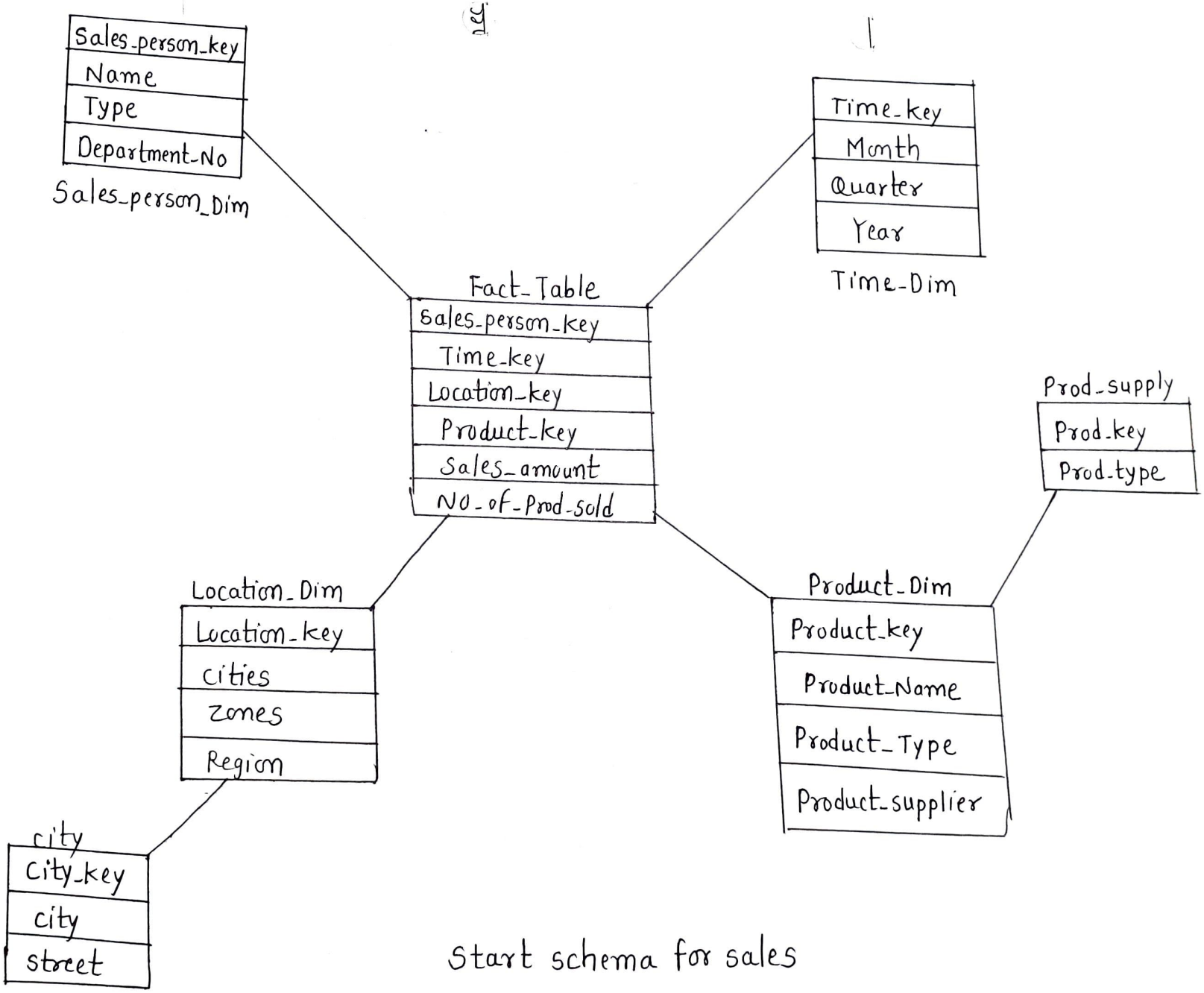


Date :

Q.6 A manufacturing company has a ~~large~~ huge sales network. to control the sales, it is divided into regions, each region has multiple zones. Each zone has different cities, each sales person is allowed to different cities. The objective is to track sales figures at different granularity levels for region, sales person & time. convert the star schema to snoflake schema.

→

Each zone has different cities. such sales person is allocated different cities. The objective is to track sales fig. at different granularity levels of region & to count no. of products sold. Design a star schema by considering granularity levels for region, salesperson & time convert the star schema to snowflake schema.



Start schema for sales

Date :

Q.7 Write down differences between the following.

a. Star schema vs snoflake schema.

Star Schema	Snoflake Schema
1. In star schema, the fact tables & the dimension tables are contained.	1. In snowflake schema, the fact tables dimensions, tables as well as sub dimension tables are connected.
2. Uses more space	It uses less space.
3. Star schema uses ~~m~~ top-down model.	It is bottom-up model.
4. It takes less time for the execution of queries.	while it takes more time than star schema for execution of queries.
5. In star schema, Normalization is not used.	In this, Both normalization & denormalization are used.

b. Data Warehouse Vs Data Mart.

	Data Warehouse	Data Mart
1.	Data Warehouse is a centralised system.	While it is a decentralized System.
2.	In data warehouse, lightly denormalization takes place.	While in data mart, highly denormalized takes place.
3.	Data-warehouse is top-down model.	While it is a bottom up Model.
4.	To built a warehouse is difficult	While to built Mart is easy.
5.	In Data Warehouse, Fact constellation schema is used.	In data mart, star schema & snowflake schema are used.

Date :

C. OLTP & OLAP.

OLTP	OLAP
1. OLTP consists only operational current data.	OLAP consists of historical data from various Databases.
2. It is application oriented. Used for business tasks.	It is subject oriented Used for Data Mining, Analytics, Decision making etc.
3. It provides a multidimensional view of different bussiness tasks.	It reveals a snapshot of present business tasks.
4. The size of the data is relatively small as the historical data is archived. For eg. MB, GB.	Large amount of data is stored typically in TB PB.