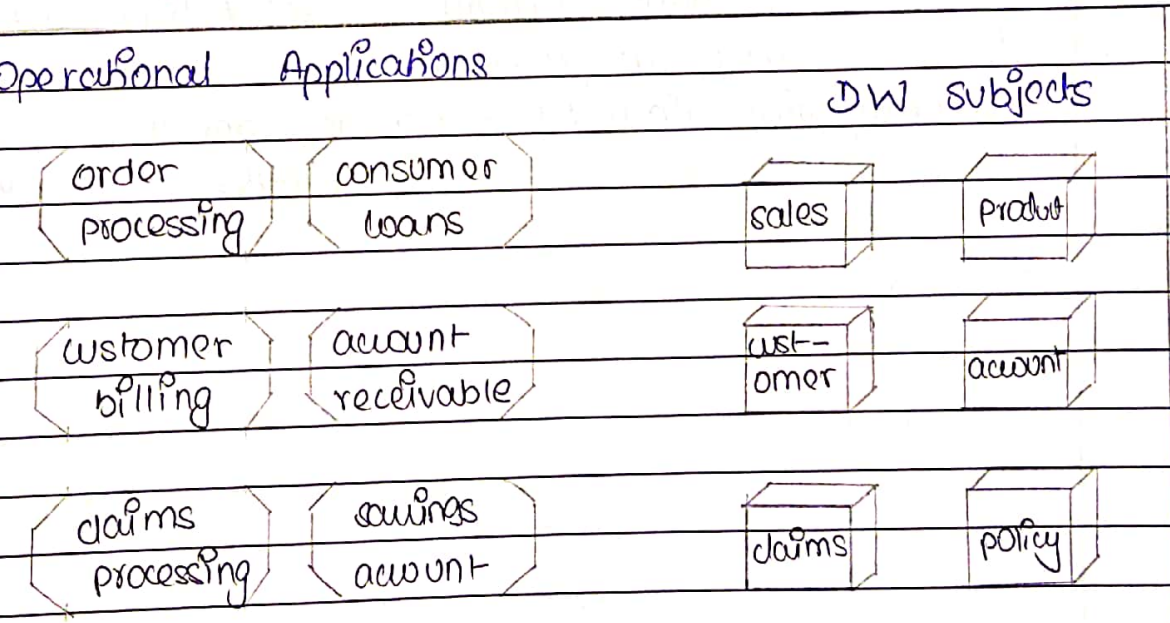
MCT
MANJARA CHARITABLE TRUST
RAJIV GANDHI INSTITUTE OF TECHNOLOGY, MUMBAI
JUHU VERSOVA LINK ROAD, VERSOVA, ANDHERI (WEST), MUMBAI - 53.

ASSIGNMENT - 01

Q1. Explain features of Datawarehouse in detail.

– Data warehouse is a subject oriented, integrated, non-volatile and time variant collection of data in support of managements decisions.

1. Subject oriented -
– In database, data is stored in the form of application whereas in datawarehouse, data is stored in the form of subjects. As every thing is available in the form of subjets, any complex query can be answered with less procesing time and it will provide 360° view of the organisation.



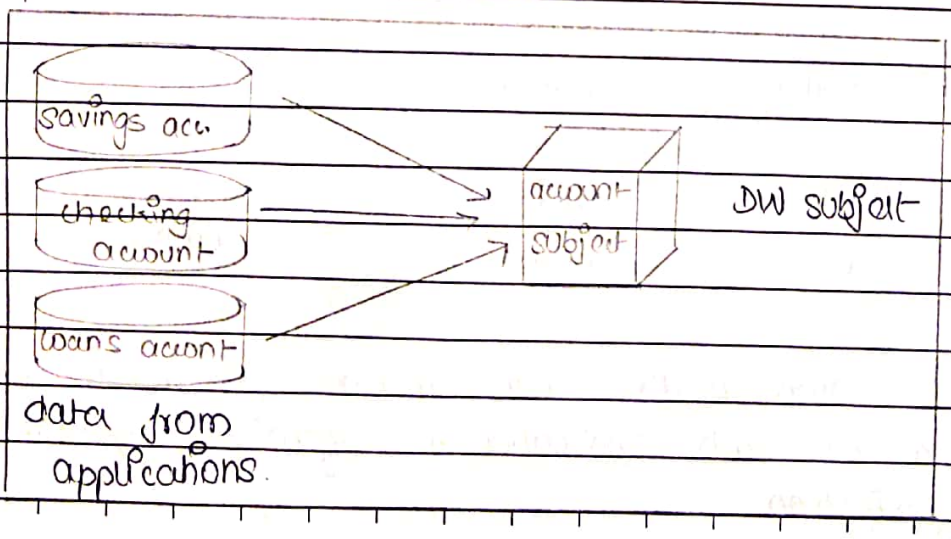
– As shown in the figure, in the operational systems shown data for each application is organized separately by application.

- whereas in data warehouse, data sets contain data needed for functioning of the particular applications as subjects.

2. Data integration :

- For proper decision making, all the relevant data from various applications are to be pulled together.
- The data in the DW comes from various & operational systems. Source data are in different database files and data segments.
- Before data from various sources can be usefully stored in the DW, the inconsistencies are have to be removed. and data elements have to be standardized. Before moving data into the DW, a process of transformation, consolidation and integration of data source is done.
- Standardization includes - naming conventions, codes, data attributes, measurements.

example,



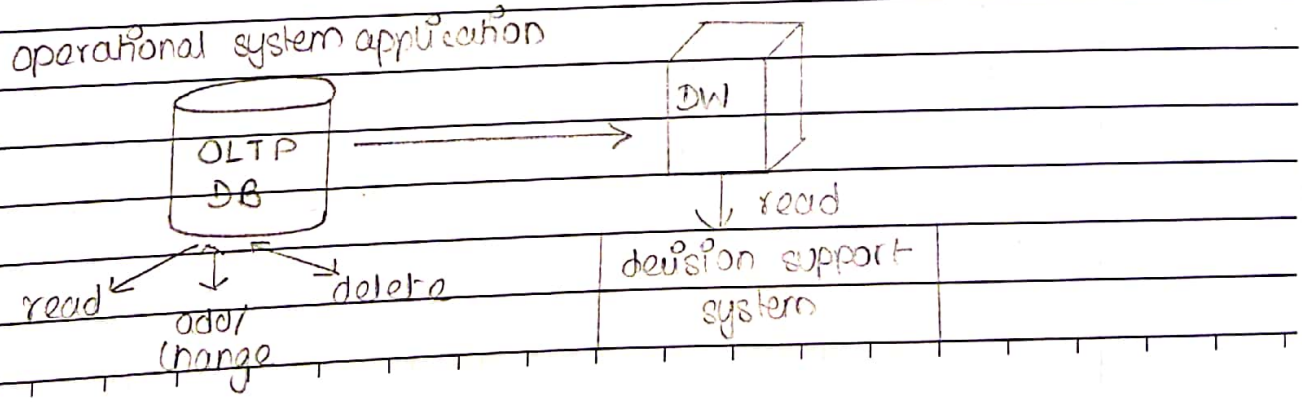
3. Time variant data :

- For an operational system, stored data contains the current values.
- A DW because of it's vary nature of its purpose has to contain historical data, not just current values.
Data is stored as snapshots over past and current periods. Every data set in DW contains the time elements.

- For example, in DW containing units of sales, the quantity stored in each file record relates to the specific time element. Depending upon the level of detail, the sales quantity in a record may relate to a specific date, week, month, or year.

4. Non-volatile data :

- The data in the DW is not intended to run the day to day business.
Data from operational systems are moved into the DW at specific intervals depending upon the requirements.
- In DW only a single operation i.e. read operation is performed and therefore data items are preserved for future analysis.

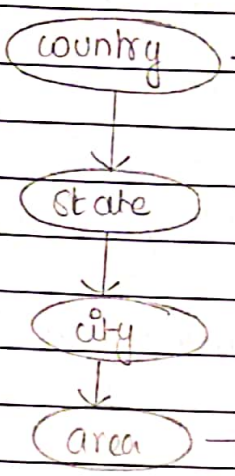


Page No.

5. Data granularity:

- Data granularity in DW refers to the level of detail. The lower the level of detail, finer the data granularity.

- example,

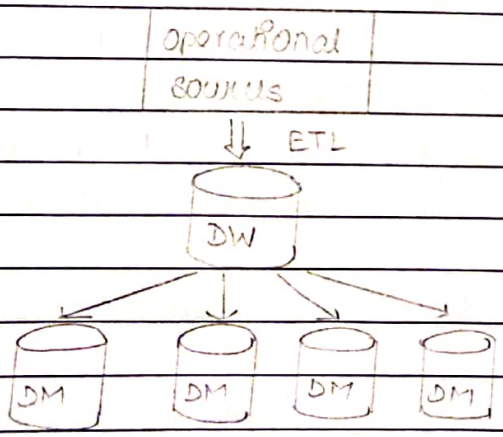


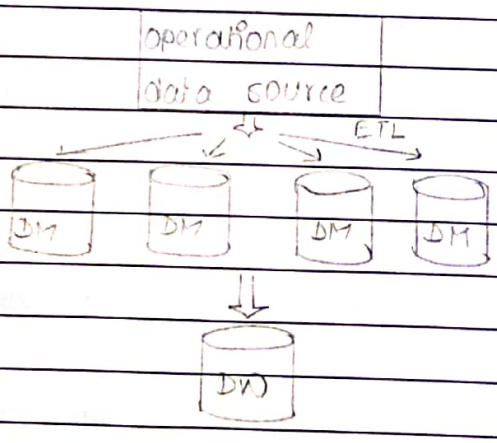
- The granularity levels has to be decided on the basis of data types and the expected system performance for queries.

Q2. Differentiate between :

1. Top down and bottom up approach.

Top down	Bottom up approach
– DW is created first and then individual data marts are created.	– Individual data marts are created first and then the wide view of DW is created.
– Answering the queries is easy as compared to bottom up approach.	– Answering the queries takes more time as compared to top down.
– As DW is created first, technical / professional skills are needed.	– Users with less cross - functional skills can build data marts.
– It is time consuming as iterations are needed to build DW.	– It is less time consuming as data marts are created first.
– Risk failure is higher	– Risk failure is lower.





2. ER modelling and dimensional modelling.

ER modelling	Dimensional modelling
- It is used for designing tables from database.	- It is used in order to create DW.
- It is used to specify relationship between entities.	- It is used to focus on business analysis.
- The main aim of ERM is to remove redundancies by normalising tables.	- Most of the tables of dimensional modelling are denormalised.
- It is well suited for answering queries at transactional level.	- It is used to answer queries at ~~drag~~ strategic level.

3. OLTP and OLAP :

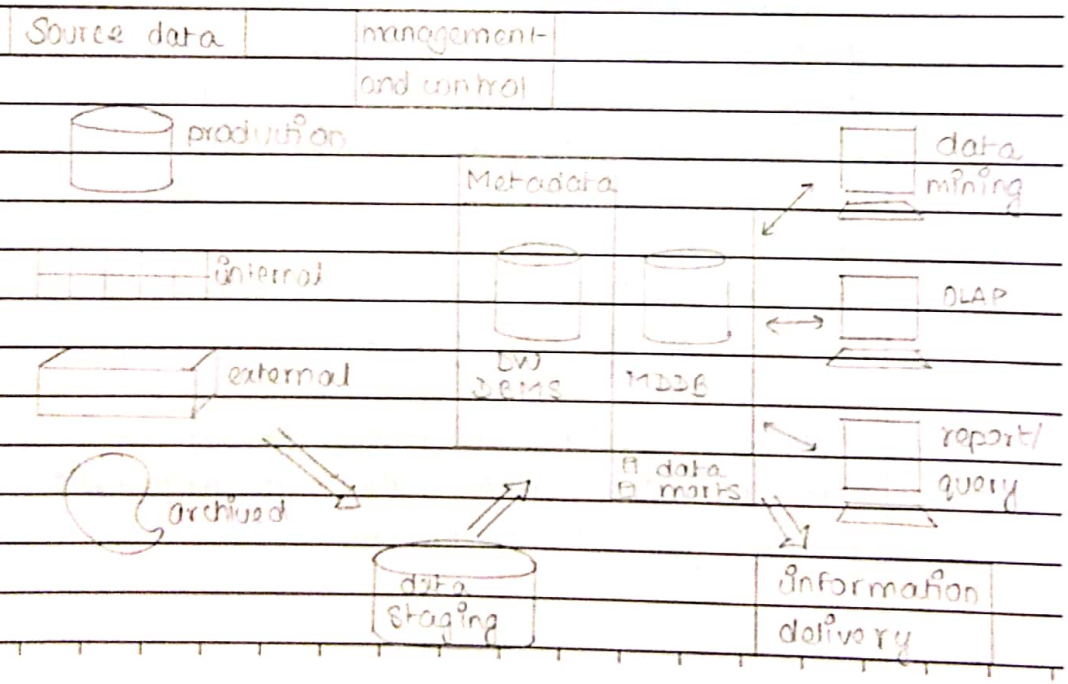
OLTP	OLAP
(online transactional processing)	(online analytical processing).
- It is used for performing day to day operations.	- It is used for analysis and decision purpose.
- It is application oriented.	- It is subject oriented.
- Detailed data is stored	- Summarized data is stored.

MCT MANJARA CHARITABLE TRUST

RAJIV GANDHI INSTITUTE OF TECHNOLOGY, MUMBAI
JUHU VERSOVA LINK ROAD, VERSOVA, ANDHERI (WEST), MUMBAI - 53.

OLTP	OLAP
- Read, write and update operations can be performed	- Only read operation can be performed
- It is volatile.	- It is non-volatile.
- Clerical users are involved in this.	- Executives and knowledged workers are involved.
- Data here is not time variant	- Here, data is time variant
- Tables are normalised.	- Tables are in denormalised form
- Size of OLTP is b/w 100MB to 100 GB.	- Size of OLAP ranges from 100 GB to few TBs
- Queries are simple.	- Queries involved in this are comparitively complex.

Q3. Describe architecture of Data warehouse with diagram.



- Source data: Input to the DW is taken from various source applications and categorised into 4 parts which are as follows:

1. Production data - In this, data is collected from various operational systems of organisation.
e.g. - employee table may be stored in account department, HR department and payroll department. For data to be stored in DW, it must be collected from all these systems. This is production data. Problem arises when collected data is not in standard form. So, standardisation is done before production data is stored in DW.

2. Internal data - It is available within the organisation. This is information which must be kept confidential. e.g. personal details of employees.

3. Archived data - Operational system is used to run daily transaction, So, old data from it is periodically taken out and stored in archived files. Thus, archived files are created for storing historical data.

4. External data - It is the data which is not available within the organisation and needs to be gathered from outside sources for analysis.

- Data staging component - Data staging area is the place where all the extracted data is temporarily stored and prepared for loading in DW.

Three major functions are performed:

1. Extraction – Source applications are extracted from and placed in data staging area.
2. Transformation – data cleansing and standardisation is done here.
3. Loading – Initial data is loaded in a data warehouse in a single run and then updated periodically.

– Data storage component: It is a separate repository. In this large amount of historical data is stored for analysis. It is designed specifically for analysis purpose and not for quick retrieval of information.

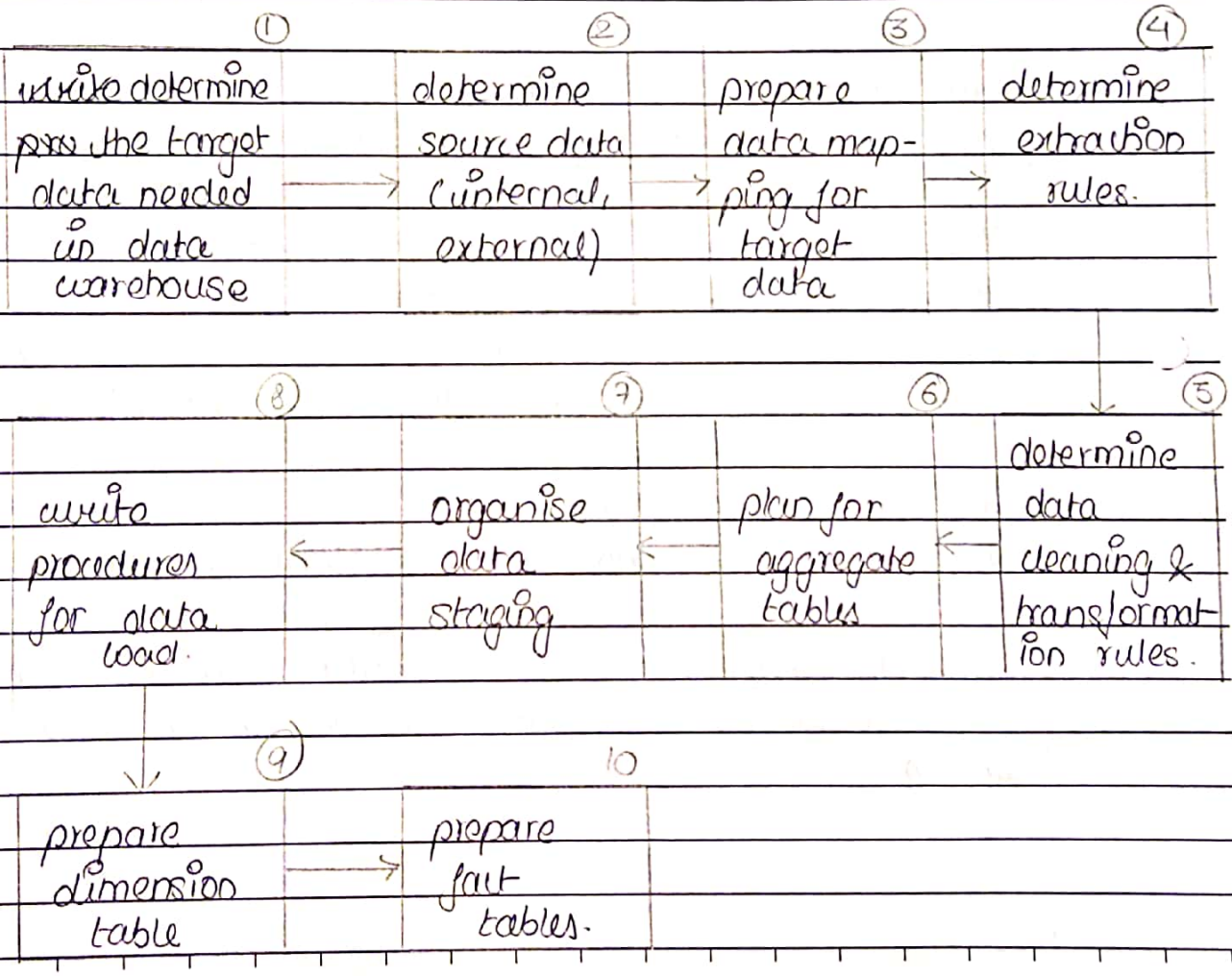
– Data marts: Data mart is a decision support system that stores number of subject areas based on the needs of the users in that department.
Thus, data mart can be thought of as a subset of the enterprise wide view DW.
Eg: Finance has its own department, marketing has its own dept. and so on. Thus, each dept. will have its own Data mart.
There are 2 approaches –
a) top down approach b) bottom up approach.

– Metadata component: It is data about data in DW. It is similar to the data dictionary or data catalog in DBMS.

- Management and control component: It sites on the top of all components and manages all the activities of DW. It helps data staging area to perform ETL activities.

- Information delivery component: It is used to provide info. to the wide variety of users with the help of ISC, info of DW can be used for various purposes like solving complex queries to do prediction and perform analysis.

Q4. Explain ETL process.



- The ETL (Extraction transformation and loading) process plays an important role in DW architecture.
Without ETL process there would be no strategic information present in the DW for future analysis.

- ETL process also starts with determing the target data (what output we want for future analysis).
- Once, the target data is set, source identification takes place, in which internal & external sources are identified from where input must be taken.
- The input is then mapped with the output.
- Once, mapping is done, extraction methods are applied.
- After extraction, data is cleansed and transformed into standardized format.
- In next step we decide whether to keep single fact table or plan for aggregate fact tables.
- After planning dimensional modelling, data staging area is organized
- Different procedures are written to load data in DW.
- All the subjects are defined in the terms of dimensional tables and fact tables.

Q5. Write short note on the following-

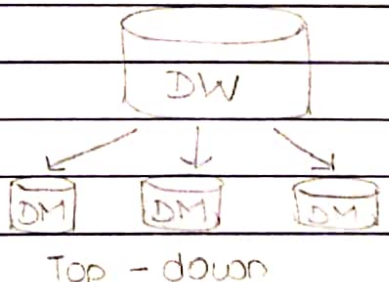
a) Data mart :

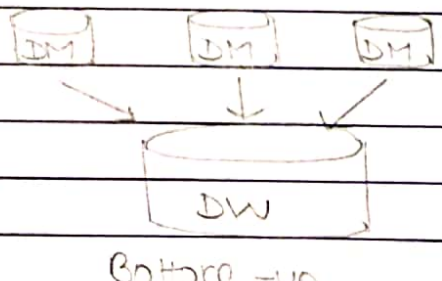
- Data mart is a decision support system that stores number of subject areas based on the needs of the users in their department.
- Data mart can be thought of as a subject of the enterprise wide data warehouse.
- e.g. the finance department has their own data mart, likewise every department will have its own data mart.

- There are two approaches for building data mart -
  a) top down approach
  b) bottom up approach

a) Top down approach- DW project team looks at the larger picture of the organization and builds huge DW first that will feed the individual data marts.

b) Bottom up approach- DW project team provides the requirements of individual department and builds data mart first that will feed data to the data warehouse so thus data marts are created first.

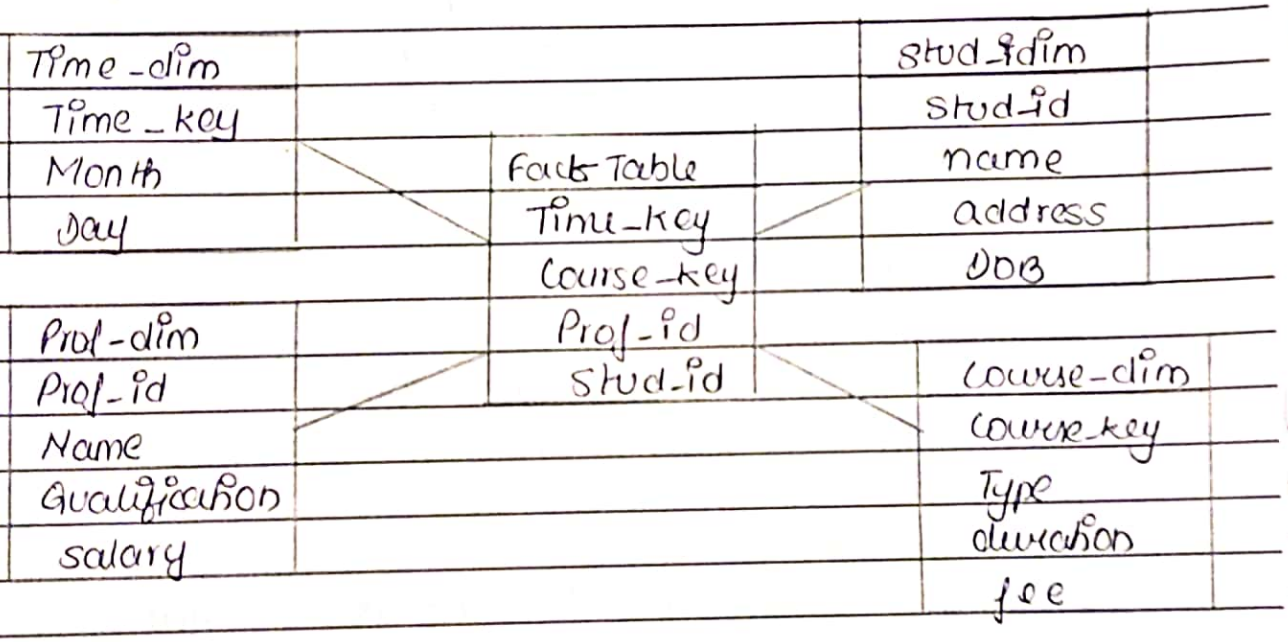




b) Fact less Fact table -

- In factless fact tables, no facts are present in the fact table. Only the primary keys of all the dimension tables are included.
- The job of factless fact table is to record event occurances. or describe conditions.

- example,



c) Star schema and keys included in star schema:

- Star schema is an arrangement where fact table is placed in the middle and all the dimension tables are directly connected to the fact table.
- In this schema all the dimension tables will have equal chance in query analysis as no intermediate tables are created.

example, DTs → ~~prod~~ prod_dim, time_dim, cust_dim, store_dim
FT → order_fact

Prod_dim
Prod_id
Name
Quantity
Brand

Time_dim
Time_id
Year
Month
Day

Order_fact
Prod_id
Store_id
cust_id
Time_id
Total_sales
Profit_dollars

Store_dim
store_id
Name
address

Cust_dim
Cust_id
Name
Address.

* Keys involved in star schema -

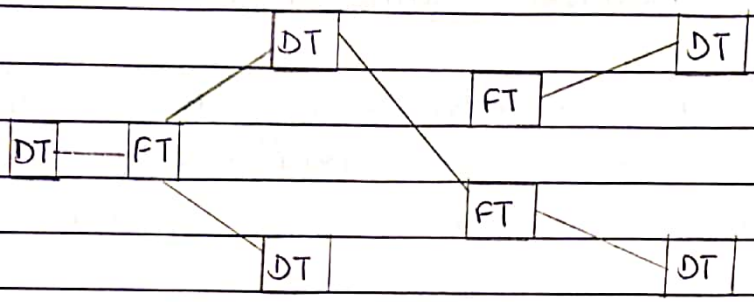
1. Primary key:
In dimension table, unique keys are used to identify each record separately. In FTs, we have concatenated keys which acts as primary keys for FTs.

2. Foreign key:
Primary key of DTs act as normal columns in FT termed as foreign key for FTs.

3. Surrogate key:
These keys are generated automatically by the software. Primary keys of database and data warehouse cannot be same because,
if primary key of DB keeps on changing rapidly
if conflicts may be present in the DB.

d) Fact constellation schema -

- In fact constellation schema, more than one fact tables are present in the organization.
- No two fact tables are directly connected to each other although they are connected using dimension tables.



- There are four ways by which fact constellation schema can be created.

10 Aggregate FT and derived DT:

In aggregate FT, pre calculated summaries of FTs are stored.
The dimensions which are shared between base FT and aggregate FT are termed as derived DTs.

20 Core and custom table:

Core FT is basically used to store values which are similar in nature whereas custom table is used to store the values which are dissimilar in nature.

5. Snapshot and transactional FT:

In snapshot FT, information based on particular point of time is stored whereas in transactional FT, fullest view of the organisation is stored

4. Supporting enterprise value chain:

As product goes from raw facts to finished product, it has to go through various stages. In each stage an individual star schema is maintained.

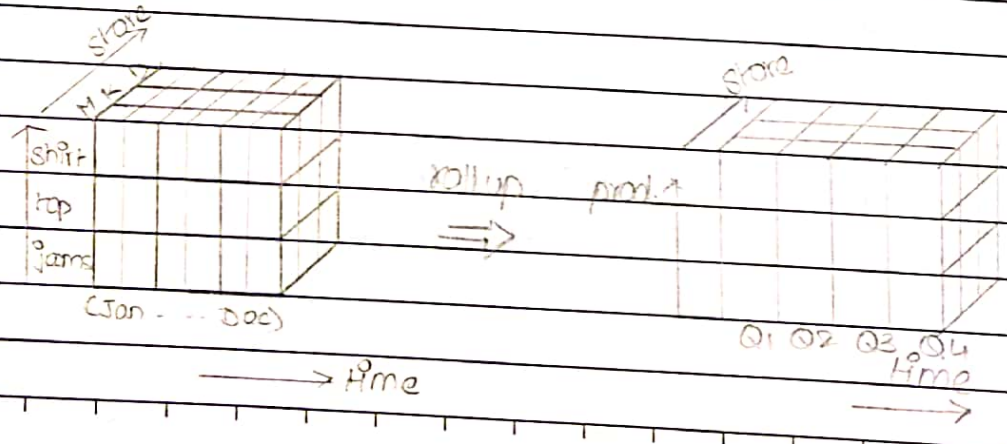
Q6. Explain OLAP operations in detail.

– There are five types of OLAP operations-

1. roll up
2. drill down
3. slice
4. dice
5. pivot / rotation.

1. Roll up: In roll up operation (also known as drill up), we move upwards in the hierarchy from less detailed to more summarised data.

Prob.

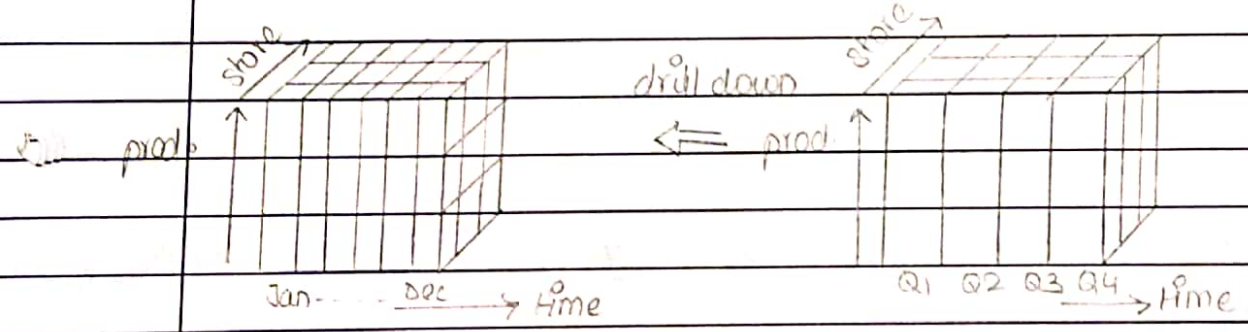


eg store Mumbai.

shirt	900					shirt					
top	450			⇒		top					
jeans	200					jeans	1300				
	Jan ---- Dec						Q1	Q2	Q3	Q4	

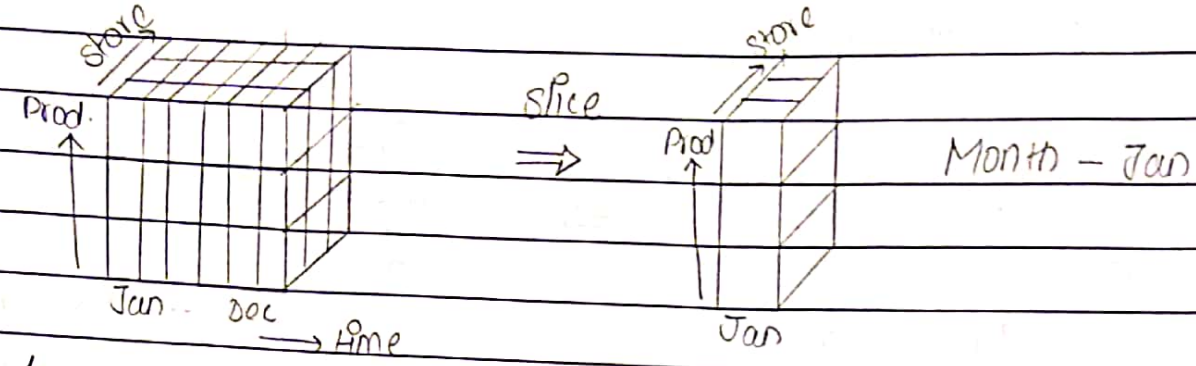
- In this example, time dimension is represented using roll up operation, where time dimension is represented in quater using the hierarchy as,
day < month < quater < half year < year.

2. Drill down: In drill down operation (also known as roll down), we move downwards in the hierarchy that is from summarised data to detailed data.



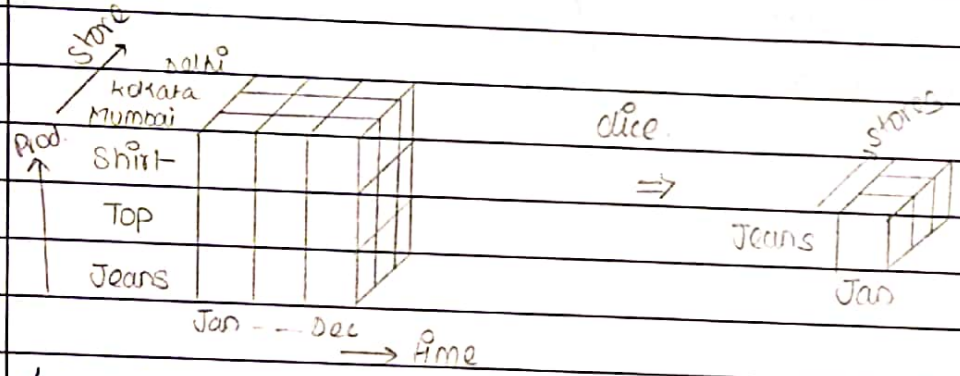
3. Slice operation:
If we apply selection operation on only one dimension, it will result in a sub cube. This operation is termed as slice operation.

MCT
MANJARA CHARITABLE TRUST
RAJIV GANDHI INSTITUTE OF TECHNOLOGY, MUMBAI
JUHU VERSOVA LINK ROAD, VERSOVA, ANDHERI (WEST), MUMBAI - 53.



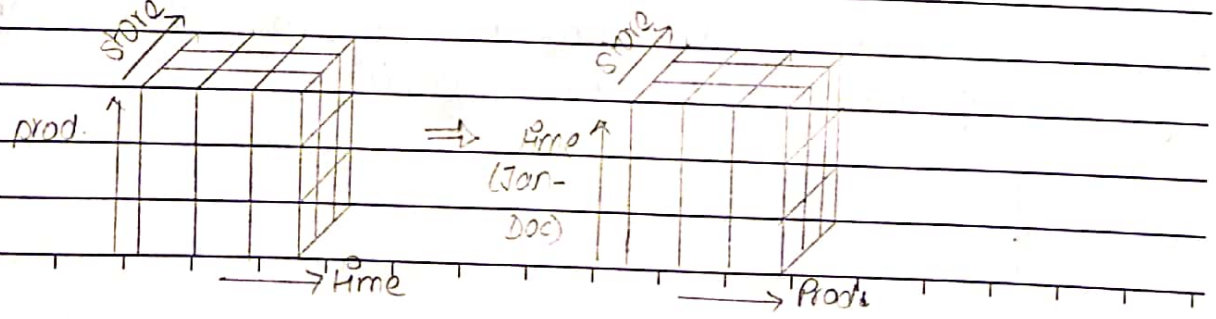
In above example, product from all the stores for the month of January are selected.

4. Dice operation : If we apply selection operation on more than one dimension, it results in a subcube and this operation is termed as dice operation.



- In the above example, selection of Jeans for the month of January is shown.

5. Pivot operation : It is obtained by interchanging the dimensions.



RAJIV GANDHI INSTITUTE OF TECHNOLOGY, MUMBAI
JUHU VERSOVA LINK ROAD, VERSOVA, ANDHERI (WEST), MUMBAI - 53.

Q7. Explain metadata, types of metadata and example describing metadata in detail.

- Metadata is termed as data dictionary or data catalog as it describes all the entities of a data warehouse briefly.
It is also called as data about data.

- In most of the organisations, data warehouse is created externally and therefore the users of DW are not aware about the contents of it.
And therefore, along with DW, data dictionary is given which not only describes the entities but also describes the syntaxes and semantics of DW.

- Types of metadata -

1. Operational metadata - It is created for source application and details of input of the DW.

2. Extraction and transformation metadata -
It gives details about the extraction method, frequency and transformation rules.

3. End users metadata -
It is created for the stakeholders using DW for analysis.

4. Business metadata -
It is created for the managers and the analysts.

5. Technical metadata :
It gives the technical details required by the developers.

example,

Sample metadata for customer.

Entity name : Customer
Definition : A client, a person or an organisation that purchases goods or services from company.
Remark : It includes current and past customers
Source systems : Finished good orders, online sales, maintainance contracts.
Create date : 25th April 2015
last update date : 22nd September 2016
update cycle : yearly
Data quality recieved : 20th September 2016
Responsible user : John M.

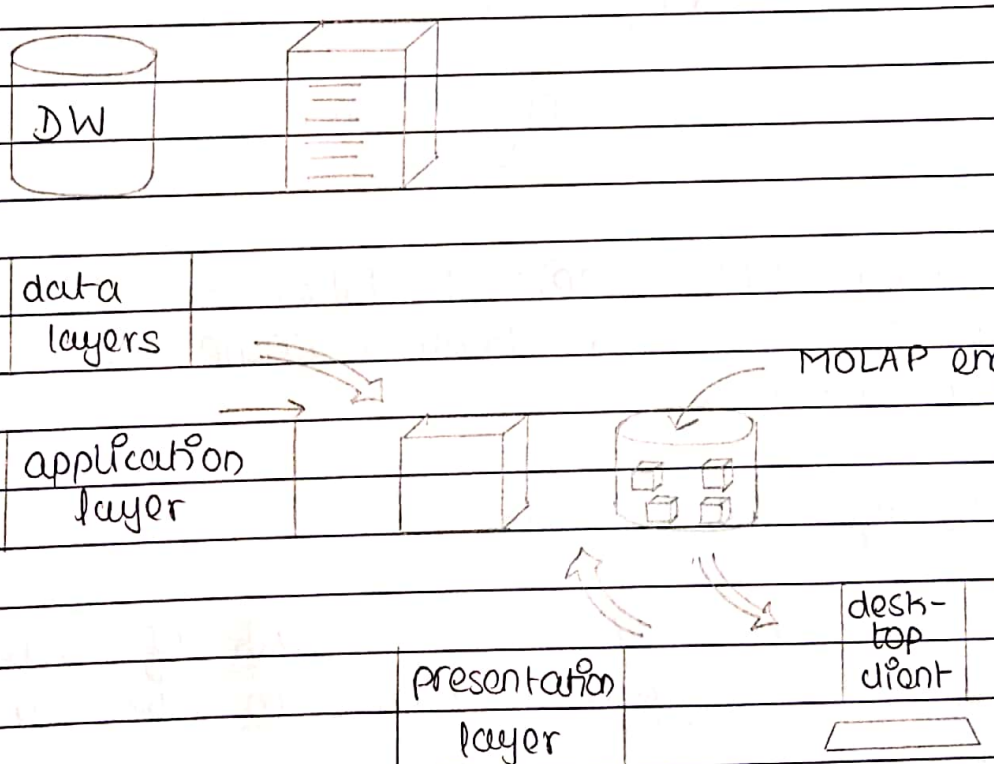
Q8. Explain different OLAP models.

There are four OLAP models -

1. MOLAP
2. ROLAP
3. HOLAP
4. DOLAP

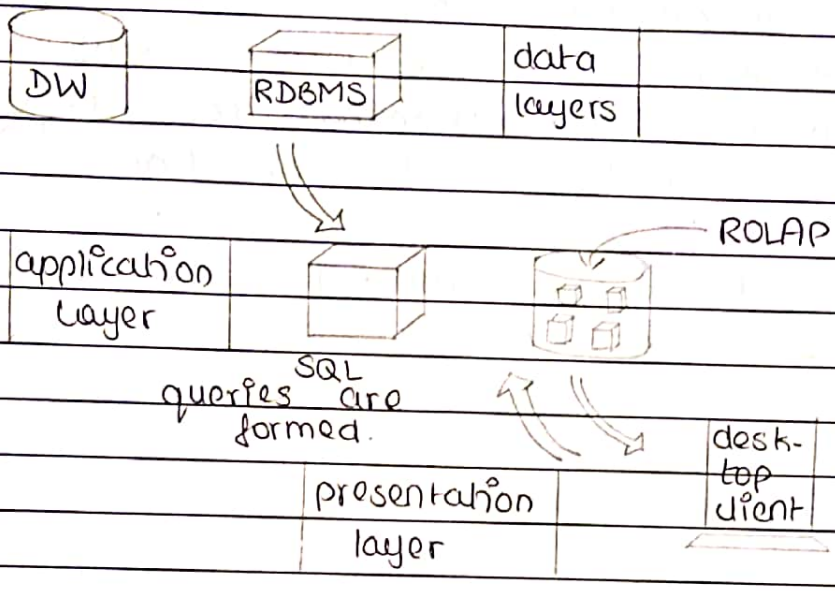
1. Multidimensional OLAP (MOLAP):

- In MOLAP, pre calculated cubes are created by the MOLAP engine depending upon the data present in DW. MOLAP engine creates as many cubes as possible and therefore retrieval of information is rapid.
- As many outcomes are created, complexity of MOLAP is more.



2. Relational OLAP (ROLAP):

- In ROLAP, pre calculated cubes are not created and cubes are created depending upon the requirements of the desktop client.
- SQL queries are formed in order to get information from DW. ROLAP engine will create cubes based on the requirements, therefore, information retrieval is slower.



3. Hybrid OLAP (HOLAP): It takes the advantage of both the OLAP models i.e HOLAP & ROLAP and removes the disadvantages.

4. Desktop OLAP (DOLAP):

It programs performs the task of sending the cubes of requested data to the client.