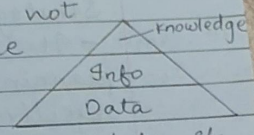
12/7/23

1. Data Warehousing Fundamentals

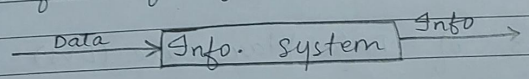
Data : Collection of raw facts which is not processed, not organised and available in its purest form.



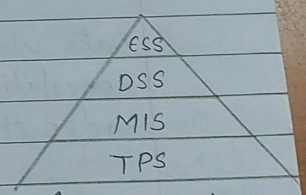
Information : If we apply certain rules on data it becomes an information.

Knowledge : What we understood from information.

Information System is responsible for transforming data in form of information.



TPS : Transactional Processing System
- It performs day to day applications.



MIS : Management Information System
- It takes input from TPS and manages the operations.

Info. system

DSS : Decision Support System
- It takes input from MIS and apply own knowledge to take decision.

ESS : Executive Support System
- There job is to plan for future.

DBMS also known as Online Transactional Processing (OLTP) is a part of TPS.
DMW also known as Online Analytical Processing (OLAP) is a part of ESS.

13/7/23

Strategic Management
- It is the process of setting an objective and planning policies in such a way that those objectives can be fullfilled.

Features of Strategic Management :-
(1) Integrated : Enterprize wide view of organization
(2) Data Integrity : Information must be accurate
(3) Credible : Exactly one value for an attribute
(4) Timely : Information must be available all time
(5) Accessible : Information must be available to all users.

IMP Features of Data Warehouse (DW)

Data Warehouse : It is subject oriented, integrated nonvolatile time-variant collection of data in order to take management decision

1. Subject Oriented
In database, data is stored in form of applications whereas in data warehouse, data is stored in form of subjects. As data is stored in form of subjects it provides centralized or 360° view of an organization.

2. Integrated
In these, source application are combined into one subject.
Integration is not easy task because of following
① Naming conventions
② Codes
③ Field lengths
④ Datatypes

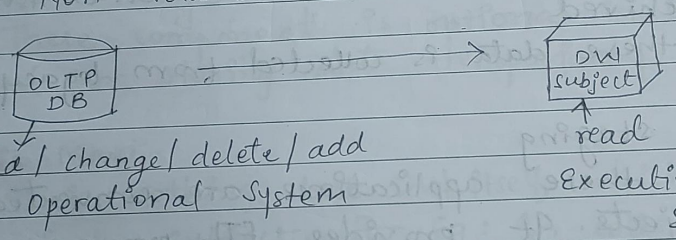
3. Time - variant

In datawarehouse, data varies over a time. It identifies the past values, compares those past value with current values and forecast future values based on current values. On the otherhand, in DB only current values are stored (time-invariant)

4. Data Granularity

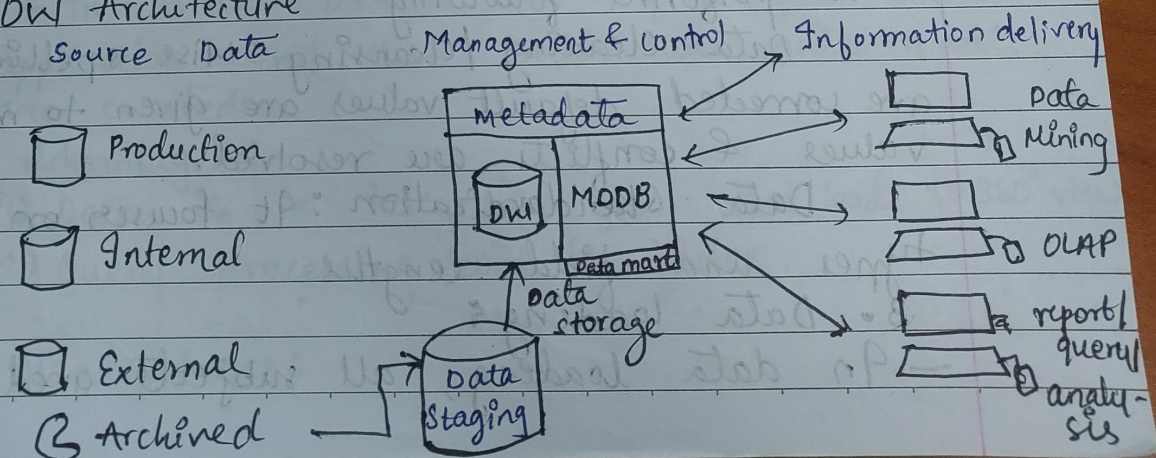
It refers to the level of details. Finer the data granularity, lower the level of detail.

5. Non - volatile



In datawarehouse only read operation is applicable and if we want to change / update any type of data present in datawarehouse then it must be updated in database first and then on periodic basic DW is reflected with the updations.

IMP DW Architecture



Source Data

In DW, data comes from hetergeneous sources, broadly divided into four parts :-

1. Production
- In there data comes from operational system (DBMS).

2. Internal
- There data is available within organization and is kept private. Ex:- Salary details of employees

3. External
- Such data is not available within organization and acquired from outside sources. Ex: list of vendors.

4. Archived
- In there data is collected from legacy system

Data staging

In there applications are converted in form of subjects. It provides ETL process

1. Data Extraction
- It focuses on extraction frequency (how frequently data is extracted) and extraction method (how data is extracted - manually or using readymade tools).

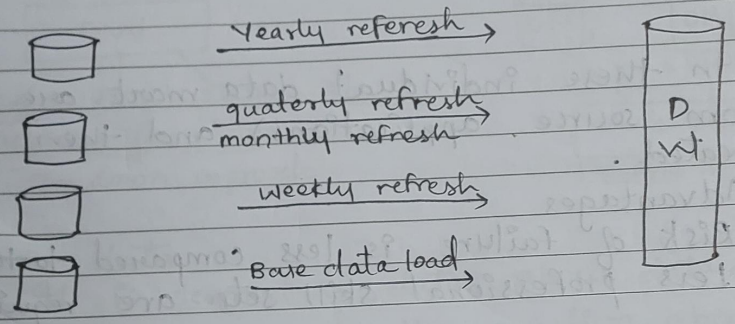
2. Data Transformation
- It performs following

Data Cleaning / Cleansing : Miss spellings are corrected, default values are given to missing values & conflicts are resolved.

Data Standardization : It focuses on data-types and field lengths

3. Data loading :
- In data loading, all subjects are loaded

in single run into DW and then on periodic basis, DW are updated.



Data storage Area
- These component is optional in DW and only used for storing data temporarly.

Data Mart
It is nothing but DW itself which focuses only on single subject. For Eg: In college DW, computer department data mart can be created which focuses on computer related activities only.

Two approaches for building DW and data mart:

① Top down approach
- In there DW is created first (from source application) and then individual data marts are created.

Advantages :-
(1) Data is in integrated form that leads to 360° view
(2) Query processing will be faster.

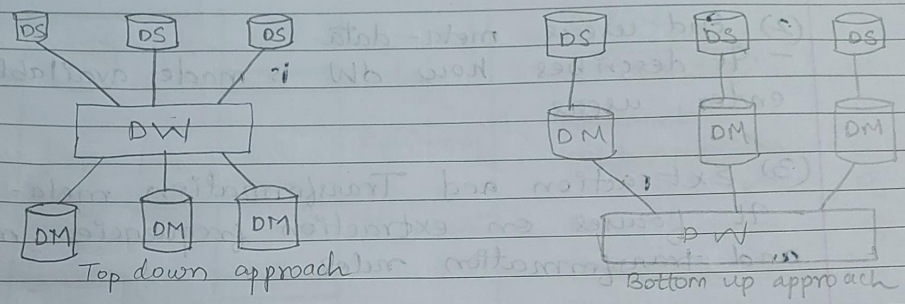
Disadvantages:

(1) Risk of Failure is high
(2) More professional skill sets are needed.

(2)
- In these individual data mart are created (from source applications) and then DW is created.
- Advantages
(1) Risk of failure is less compared to top down
(2) less professional skill sets are required.

- Disadvantages
(1) Data redundancy ~~makes~~ take place.
(2) Query processing will be slower.

DW	Datamart
1. It focuses on multiple subjects	It focuses on single subject
2. It is centralised wide view of an organisation	It is decentralised by specific subjects.
3. It is highly flexible	It is restrictive
4. DW is well-planned	It is possibly not planned.
5. Implementation takes months to years	Implementation is usually done in months
6. Size of DW is 100GB to few TB.	Size is less than 100GB.



Meta - data

It is termed as data about data, which provides answers to the question about data in DW. It is also termed as data dictionary or data catalogue.

Eg:- entity name : customer
Alias name : client, purchaser
Def^n :- A client, person or an organization that purchases goods or services
Remarks :- customer entity includes regular, current & past customers.
Source system :- goods, maintenance, online sale.
Create date :- 26th April 2015
Last update date :- 17th Sep, 2015
Last : Refresh cycle :- 20th Oct, 2015
Data Quality reviewed :- 18th Sep, 2015
Planned archival :- Quaterly
Responsible user :- John.

Format can be changed
No fix format

Types of meta data

(1) Operational meta-data
- It focuses on source data applications

(2) End users meta-data
- It describes how dW is made available to ender users

(3) Extraction and Transformation meta-data
- It focuses on extraction frequencies, method and transformation rules.

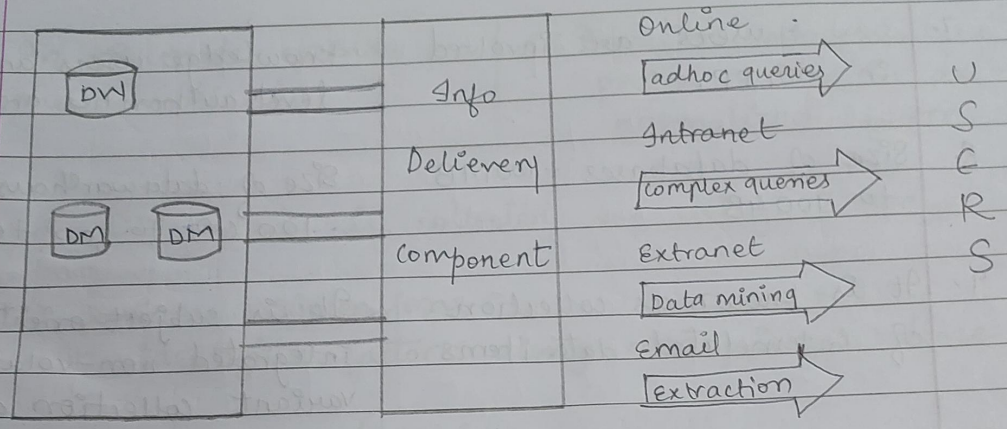
(4) Business meta-data
- It is created for stakeholders in DW.

(5) Management and control meta-data
- It describes management and control policies of DW.

Management and control component
It sets on the top of all components and manages and controls all the activities of data - staging area, meta data and data mart.

Multi-dimensional database (MDDB)
In DB, data is represented by using two dimensional tables (rows and columns). But in DW, as data is large in size 2-dimensional DB is not sufficient and therefore more than 2-dimensional data is needed for DW.

Information Delievery component
It plays an important role in making information available to the wide community of users.



OLTP	OLAP
1. It stands for online transactional process (DBMS)	It stands for online analytical processing (DW)
2. It is application oriented	It is subject oriented
3. It is detailed in nature	It is summarized in nature.
4. It is volatile in nature where all the operations are performed on databases	It is non-volatile as only read operation on DW.
5. It performs day to day activities needed for running the applications	DW is centralized wide view which is used to analyze data.
6. It stores only current values (time-invariant)	It stores current, past to predict future (time-variant).

7.	users are involved in OLTP.	knowledge users and higher level authorities are involved
8.	Size of databases - 100MB to 100GB	Size of data warehouses is 100 GB to few terabytes.
9.	It is known as collection of interrelated data items	It is subject oriented, integrated non-volatile time variant collection data to take management decisions.

Dimensional Modelling

Just like E-R modelling in DB, we have dimensional models in DWH to provide structural view of an organization.

	ER Modelling	Dimensional modelling
1.	ER models are used to design tables for databases	Dimensional models are used to design DWH.
2.	It is well suited to answer queries at transactional level	It is well suited to answer queries at business level.
3.	ER models are used to express relationship, between entities	Dimensional models are used to express facts, dimensions.
4.	It has 2-dimensional view.	It has multi-dimensional view.

5.	In ER model tables are normalised.	In this we usually prefer tables in de-normalized form.

In DM, data is collected by using 3-ways

(1) Business Dimensions
- It identifies factors along analysis is done

(2) Facts / Measures / Metrics
- It defines what is being analyzed or calculated

(3) Business Hierarchies / Categories
- Every dimension of DW is represented with a level of detail termed as business hiearchies.

IPD

It stands for Information Package Diagram. IPD's are used to combine business dimensions, facts & categories of DW inorder to gather info which is needed for analysis of DW.

- IPD for student Monitoring system

Subject: Student monitoring system						
Facts : Agg_marks, attendance, no. of students placed						
stud_dim:	stud_id	name	div	age	DOB	marks
prof_dim:	P_id	name	sal	qual	sub	year of Exp
branch_dim:	b_id	name	Establishment			
sub_dim:	sub_id	name	type	duration		
Time_dim:	T_id	Day	week	sem	Years	month

fig. IPD for student monitoring system

- IPD for Railway Reservation System

Subject : Railway Reservation system
Facts : no. of passengers, no. of seats, no. of coaches, revenue generated, no. of staff, no. of stations.

train :	t_id	t-name	source	destination	Arrival	Dep
Passengers :	p-id	p-name	Aadhar no.	DOB	Address	
tickets :	ti-id	time	B-date	Money		
seats :	s-id	type				

Fig. Railway reservation system

Fact Table and Dimension Table

- Dimension table contains its own primary key, which will identify each row of D.T uniquely. It also has categories of the table
- Fact table includes primary key of all D.T (concatinated key) along with facts.

Eg : Student Monitoring system

Stud_dim
s_id
s_name
div
age
DOB
marks

prof_dim
p_id
name
sal
Qual
subject

sub_dim
sub_id
name
duration
type.

branch_dim
b_id
name
YOF

Time_dim
Tid
Day
Week

stud-monitoring fact-table	
s_id	Primary Key
p_id	
stud_id	
b_id	
T_id	
no. of students placed	facts
attendance	
avg_marks	

Characteristics of DT
- Dimension table Key
- Table is wide → Has more no. of columns
- Textual attributes
- Attributes are not directly related
- Not normalized
- Multiple hiearchies
- Fewer no. of records

Dimension table Key
- Each dt is represented with its primary key to identify each row uniquely.

Table is wide
- DT may have more than 50 attributes in it which makes it wide

Textual attributes
- Most of attributes of DT are in text format

- In DT we have categories which are not directly related to each other. Ex: In professor DT there is no relation between id of professor and salary of professor.

Not Normalized
- DT are not normalized to reduce complexity of primary key - foriegn key relationship.

Multiple Hiearchies
- Depending upon user's persecpective, hiearchies are formed.

Fewer no. of records
- As DT is wide, it has less no. of records as compared to fact table where in FT cross-join is performed.

Characteristics of Fact

Concatenated Key
- As FT includes primary key of all DT it acts as concatenated key for FT (primary key)

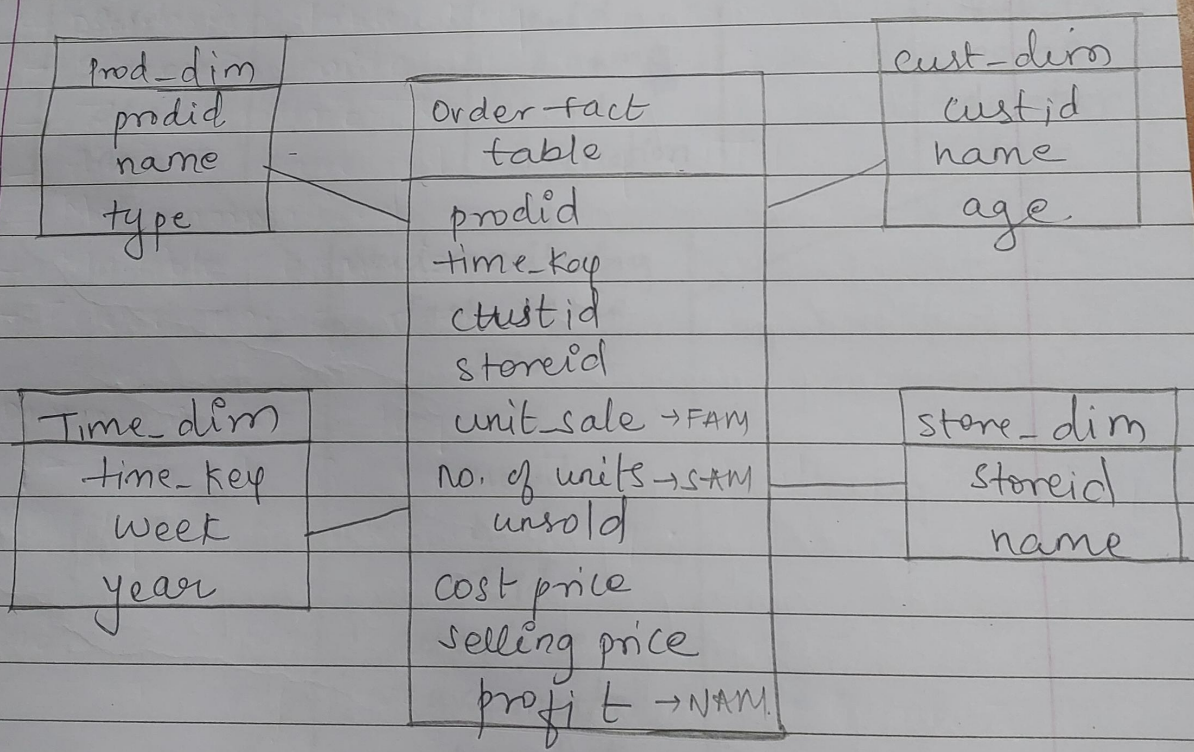
Sparse data
- In FT few gaps are created which are known as sparse data.

Table deep not wide
- As FT involves cross-join, it has more no. of rows compared to DT

Fully additive measure (FAM)
- When all DT are involved in calculation of facts known as fully additive measure.

Semi additive measure. (SAM)
- when some DT are involved in calculation of facts it becomes semi additive measure

No additive measure (NAM)
- when no dimension is needed to calculate a fact we have no additive measure

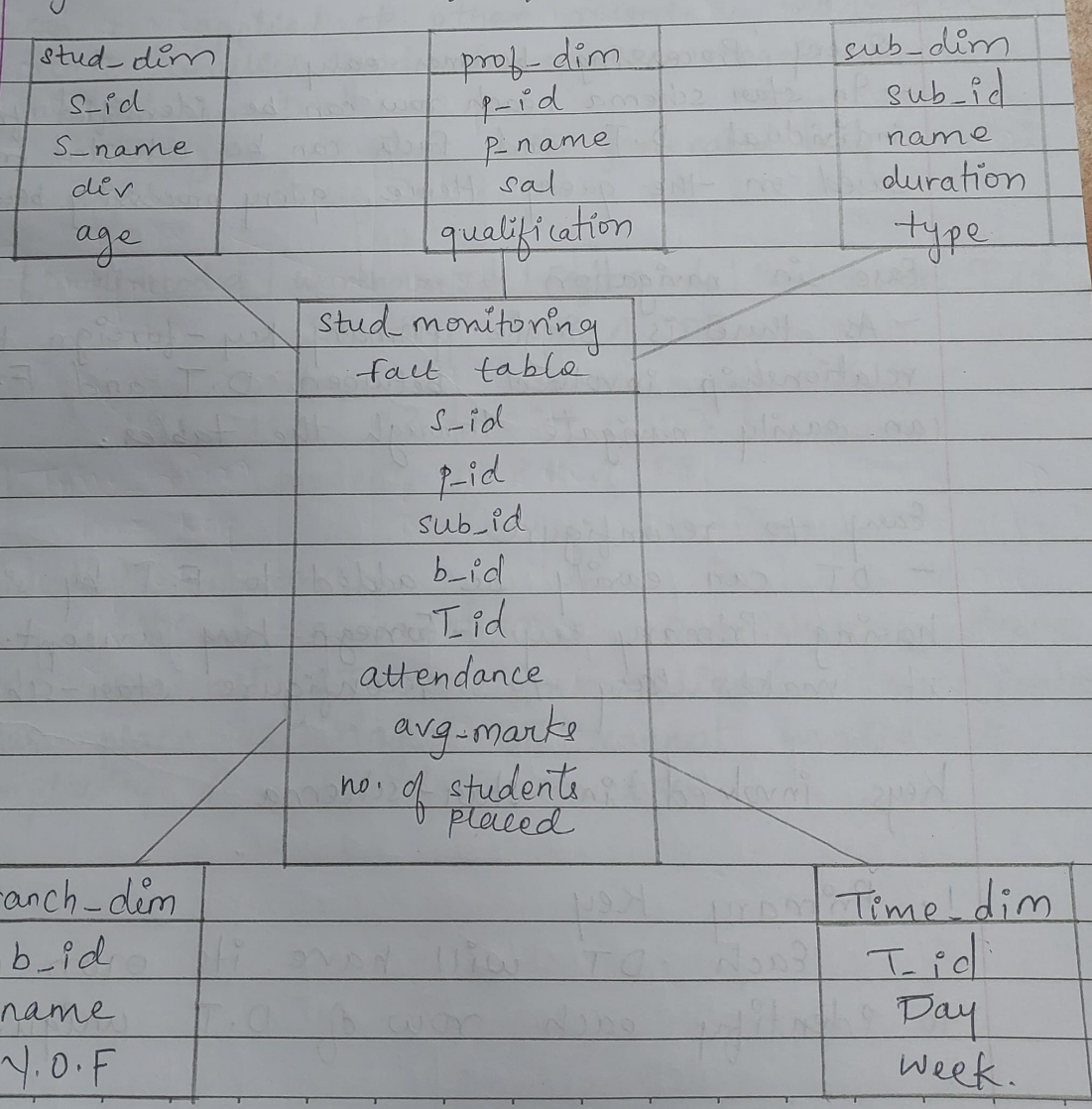


Star Schema

It is an arrangement where F.T is kept in middle to which all dimensions tables are directly connected. There is primary key & foreign key relationship between D.T & F.T.

As all dimensions tables are connected to the fact table each dimension table will have equal opportunity in query processing.

Eg: Student monitoring system



Advantages of star schema:
① Easy to understand & simplicity
② Query processing becomes easy.
③ Ease in navigation
④ Easy to reconfigure.

Easy to understand and simplicity.
- As all dimension tables are directly connected to F.T it makes design of star schema simple and easy to understand.

Query Processing becomes easy.
- In star schema, each row can be identified from its individual D.T and facts can be extracted from F.T based on the query. Hence, query processing becomes easy.

Ease in navigation
- As there is single primary key - foreign key relationship involved between D.T and F.T, user can easily navigate through the tables.

Easy to reconfigure
- DT can easily be added to F.T by simply having Primary key - Foriegn key concept. Hence it make easy to reconfigure star-schema.

Keys involved in star-schema

1. Primary Key
Each DT will have its own primary key to identify each row of D.T uniquely.

2. Foreign Key

Primary key of all D.T are added in fact table working as foreign key.

3. Surrogate Keys

Operational system primary key cannot be applied to DW directly because of two reasons:

(1) As data in DB, keep on changing frequently, PK is also going to be change

(2) As updations are done in DB's, same id can be applied to other record in DB which may create conflict and therefore surrogate key are involved in DW

Surrogate key is auto generated key created by system itself.

Country code	Warehouse Id	Region code	Product code	SP Id

Primary key of DB

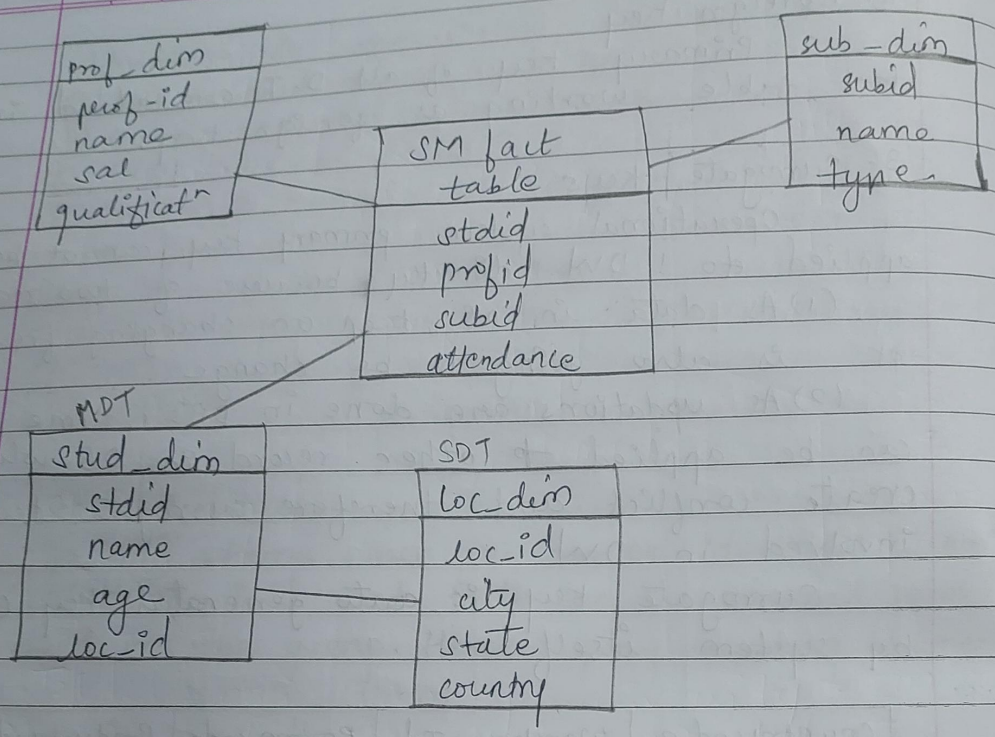
91	101	R2	1234	S02

contents of P.K of DB.

31/7/23

IMP Snowflake Schema

It is a variation of star schema, where some or all DT are normalized. Normalization takes place as single DT cannot handle of complexity of its data-structure



Advantages :

① As normalization is involved, it requires less storage space

Disadvantages :

① Complexity is more as compared to star schema (As more primary key - foreign key joins are involved)

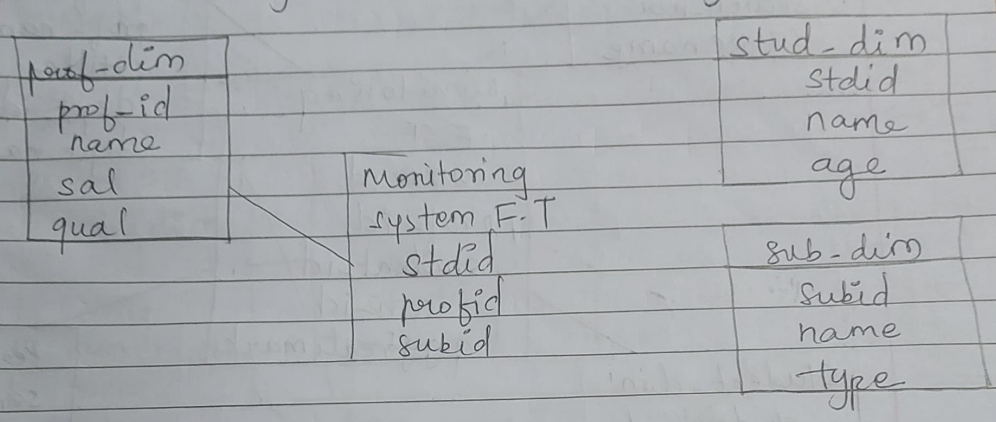
② Query processing takes time as more sub-DT are added.

③ Navigation will be slower.

Factless Fact table

As the name indicates, no fact are present in fact table referred as factless fact table.

It is used mainly in automated system for event tracking and monitoring.



V.Q. The Mumbai University wants you to design a star to records grades for the course completed by students. There are 4 DT namely professor, student, periods with attribute

10 M

Solution

course_section: course_id, section_no, course_name. During a given sem college offers an avg of 500 course sections.

professor: prof_id, prof_name, title, dept_id, dept_name

student: student_id, stud_name

Each course section has an avg of 60 students.

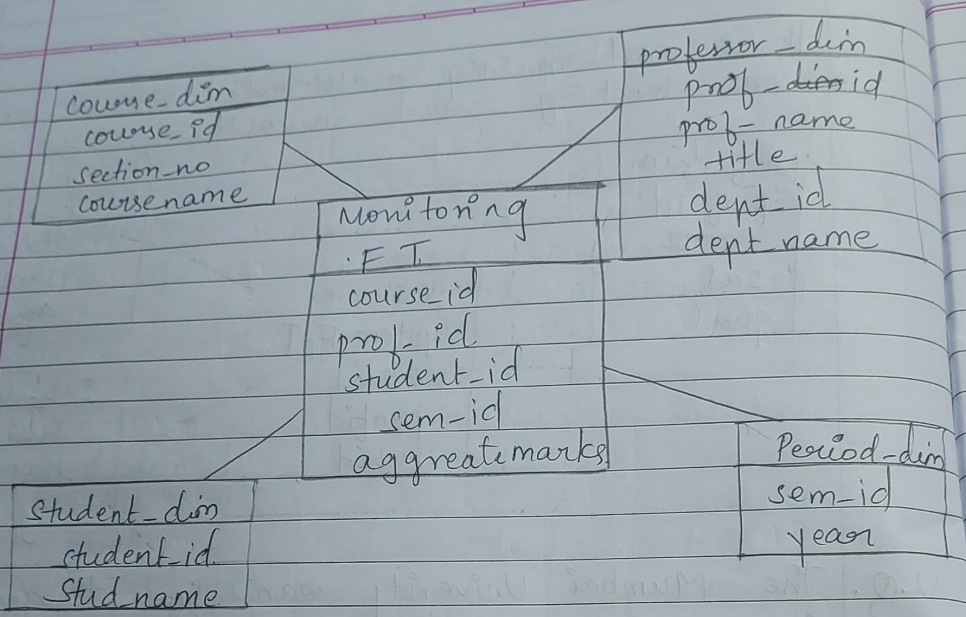
period: sem_id, year.

Database will contain data for 30 months.

The only fact that is recorded from system is aggregate marks of students

(a) Design star schema

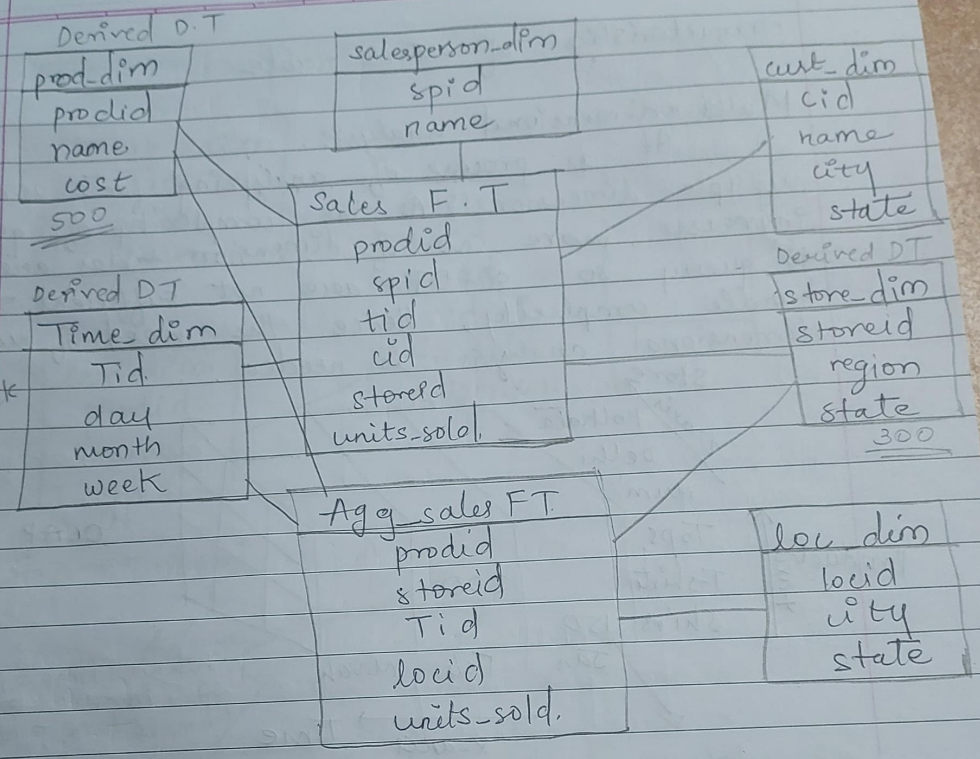
(b) Estimate no. of rows in FT using the assumptions stated above. Estimate total size of FT in bytes assuming that each field is of 5 bytes



Total courses given by university = 500 courses
No. of students enrolled = 60 students.
As period is represented in semid, 30 months are converted into sem, where 1 sem = 6 month
∴ 30 months = 5 sems.
Therefore total no. of rows fact table will have = 500 × 60 × 5 = 1,50,000

As each field is of 5 bytes
1,50,000 rows will have the size as 150000 × 5 = 7,50,000.

Aggregate fact table
It is used to store precalculated summaries of data set. As complexity of query increases more no. of rows are added to the fact table. As single F.T cannot handle complexity, more than one F.T is needed in DW which results in fact —



Constellation

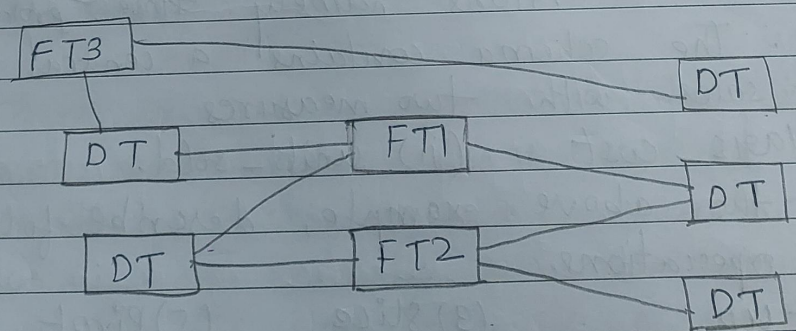
(1) 1 prod, 1 store, 1 week = 1 row.

(2) 1 prod, all store, 1 week = 1 X 300 X 1

(3) all prod, all store, 1 week = 500 X 300 X 1

(4) all prod, all store, 1 year = 500 X 300 X 52

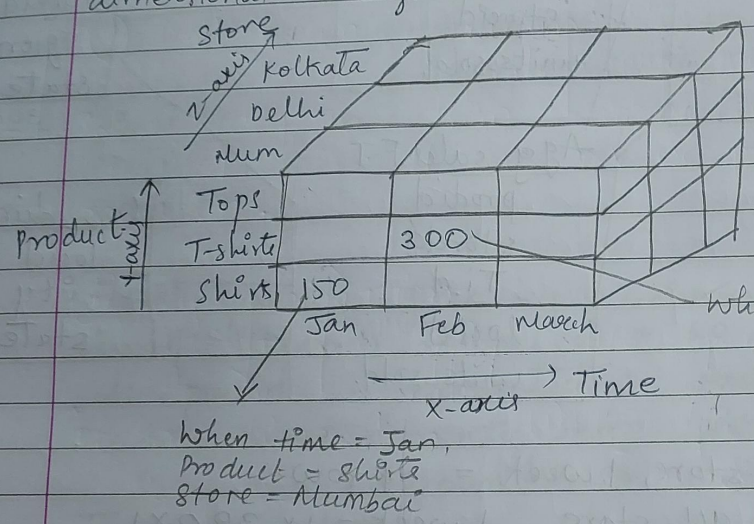
(5) all prod, all store, 5 years = 500 X 300 X 52 X 5



constellation schema (Family of Stars)

characteristics of Fact table

Multi-dimensional analysis.
It is process of analyzing facts across multiple dimensions. As the complexity of query increases, more no. of dimension tables are added to query so 2-DT are not sufficient to handle complexity. Hence, the need of multi-dimensional analysis is needed.



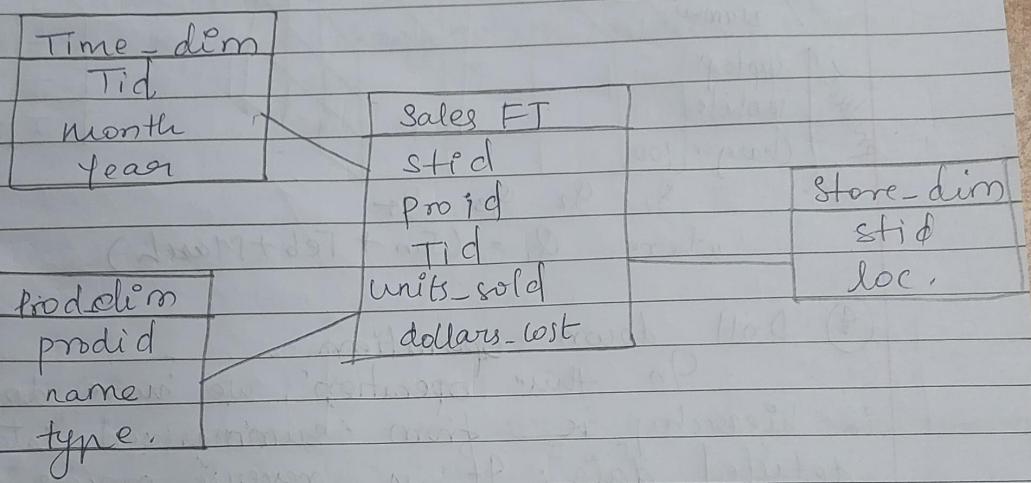
OLAP cube.

V.V.IMP OLAP operations

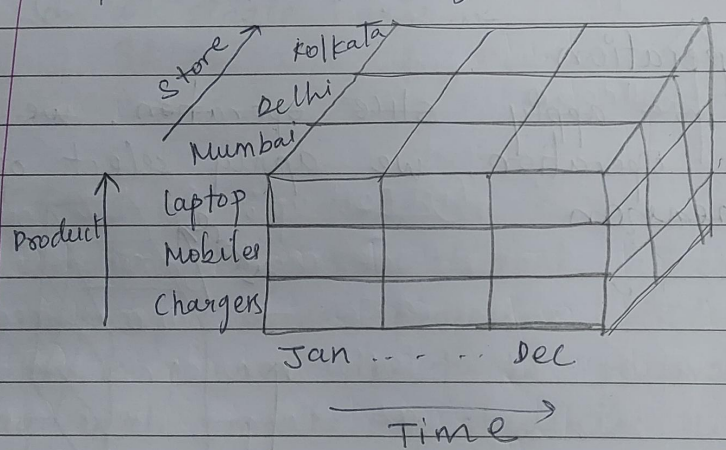
Q. All electronics company have sales department. Consider 3-dimensions namely time, product, store. The schema contains a central fact table :- sales with two measures
(i) dollars_cost (ii) units_sold.
Using the above example, describe following OLAP operations.
(1) Roll-up (3) Slice (5) Pivot
(2) Drill down (4) Dice

Can be asked in 2M

→ Step 1: Creation of Star schema.

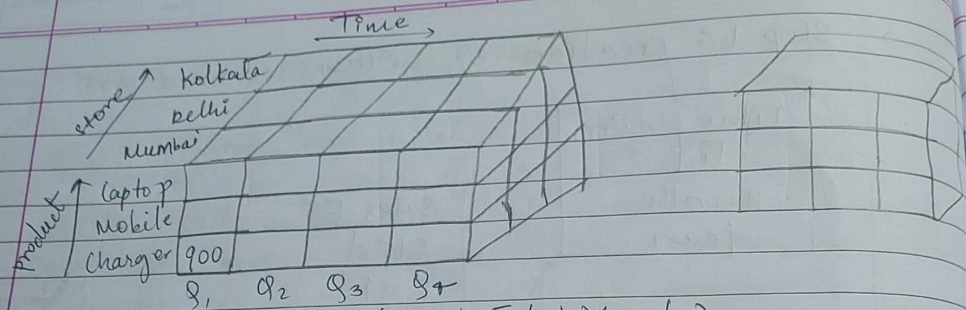


Step 2: Creation of cube.



③ Rollup Operation

It is also known as Drill up operation where we move upward in heiarchy that is from detailed data to summarized data. lets consider hiearchy as week → month → quarter → year then in an example, as data is represented monthly by applying roll up operation, data can be represented quaterly.



where Q_1 = (Jan + Feb + March)

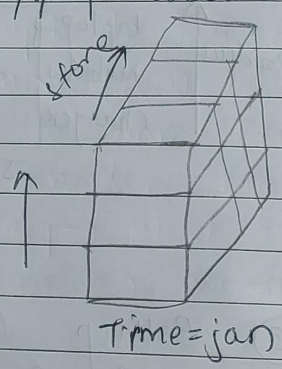
④ Drill down operation.

In this operation, we move downward in hierarchy ie from summarized data to detailed data. It is reverse process of rollup operation.

⑤ Slice operation

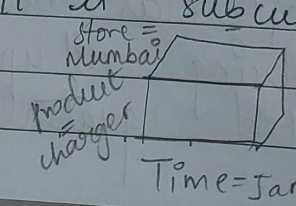
~~When we apply slice operation, we apy~~

In this operation, we apply select operation on one dimension



⑥ Dice operation

If we apply selection operation on more than one dimension, it results in dice operation. It is subcube of main cube



(7) Pivot operation.

This operation is used for visualization purpose by ~~apt~~ changing the access of dimensions

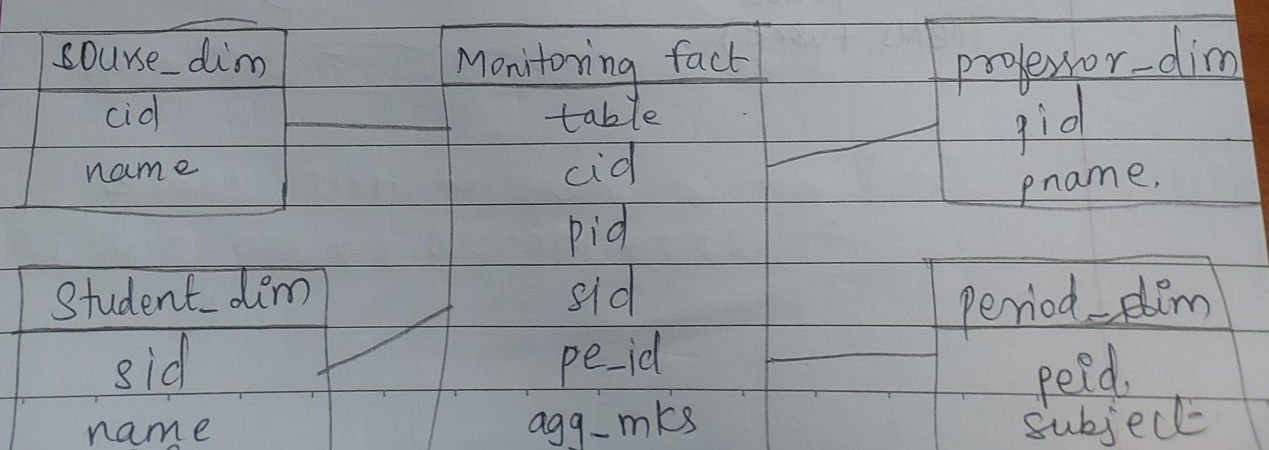
Hyper Cube

In Hyper cube, we have more than 3 dimensions tables needed to be describe with the help of 3-D cube, where any two DT can be merged on a single axis.

Hyper cube is the generalization method from 3Dimensional cube to N- dimensional cube.

U.Q. IMP The college wants to record grades for the courses completed by the student. There are 4 dimensions namely course, professor, student and periods. The only fact ie to be recorded in table is aggregate marks of students. Using the above example, perform following operations.
(i) Slice (ii) Dice (iii) Drill - down (iv) pivot.

→



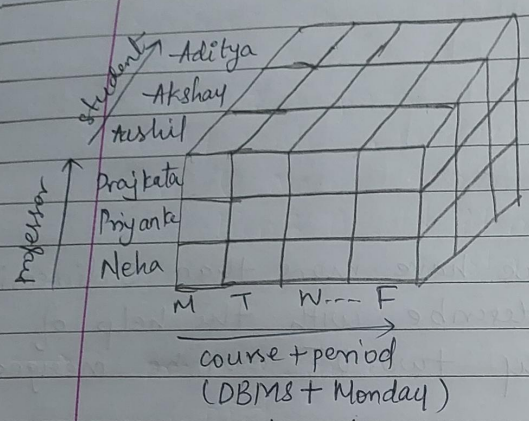
Step 2: Creation of cube

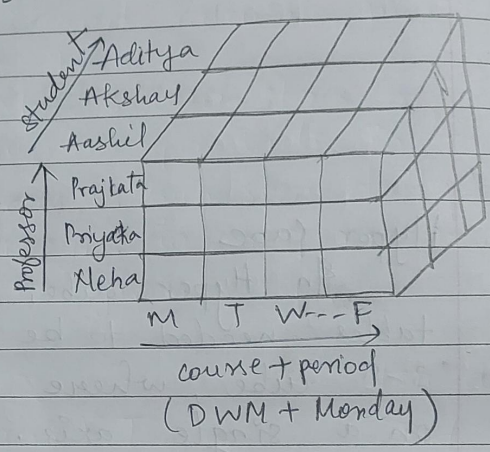
Student: Aditya, Akshay, Aashil

Neha, Priyanka, Prajakta

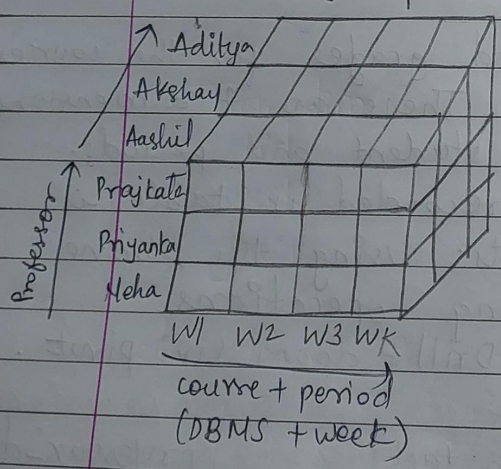
DBMS, DWM

Monday, Tuesday, ..., Friday.





Rollup ↓ ↑ Drill down



① Roll up: In this operation we move from detailed data to summarized data. let's consider hiearchy as day → week → month then in an example.

U.Q For a supermarket chain, consider following dimensions namely product, store, time and promotion. The schema contains central fact table 'sales' with 3 major measures. (1) units_sales (2) dollars_sales (3) dollar_cost design star schema and calculate maximum no. of rows in fact table
Time : 5 years, store : 300 stores reporting daily sents, product : 40,000 products in each store (About 4000 sells in each store daily).
promotion : A sold item maybe in only one promotion in a store on given day.

→

product_dim	Sales F.T	store_dim.
pid	pid	sid.
name	sid	loc
type	tid	
	prid	
	unit_sales	
	dollar_sales	
	dollar_cost	

Time_dim	promotion_dim
tid	prid
week	~~name~~type
year	discount

As time is given in years it has to be calculated in days
∴ 5 years = 365 × 5 = 1825 days

∴ Total no. of rows in fact table
= 1825 × 4000 × 1 × 300 = 2,19,00,00,000.