Date
DELTA Pg No.

Datawaare house

A datawaare house is a powerfull database model that sig niffiantly enhance the user ability to quickly analyze larg multidimensional data sets. It cleanse and organize data to allow users to make buisness decision based o facts. Hence, the data in data-warehouse must hav strong analytical characterstic

Datawaare house is a field that has grown from the integration of a number of different technolog and experience over the past decade. These experienc have allowed the IT industries to identify the key problems that need to be solved

Data can be classified into 3 categories

- Reference and transaction data ⇒ Reference and transacti data orginate from operational system and it normally kept in conventional database system. At a predeter mined ~~period~~ periodicity, It will be loaded for refreshing the datawarehouse. once it put in Data warehouse then it can-not be modified

- Derived data -> Derived data on the other hand, is only derived from reference and transaction data, based on certain rules and computations. It can always be

derived again on fresh or modified basis of deviation

- Denormalize data => Denormalize data, which is the basis for online analytical processing (OLAP) tools is prepared periodically but is directly based on detailed reference (transaction) data

* Data-warehouse Characteristic

- It typically integrate several resources e.g. sales database from various region/ states/ year
- It require more historical data than generally maintained in operational database
- It must be optimized for access to very large amount of data
- It is mostly read-accessed and rarely write accessed
- Data may be more coarse grained than in operational database.
- Data ware house are maintained seperately from operational database

- It is based on client server architecture
- It provide multi-user support
- It is capable of handling dynamic sparse matrix
- It support unrestricted cross-dimensional operation
- It maintain transparency

Database vs Datawarehouse

Database	Data warehouse
It support operation process	It support analysis and performance reporting
capture and maintain data	Explore the data
current data	multiple layers of history
Data is balanced within the scope of this one system	Data must be integrated and balanced from multiple system
Data is updated when transaction occur	Data updated on scheduled process
Data verification occur when entry is done	Data verification occur after the fact
Size 100 mb to gb	100 Gb to Tb

Application oriented	subject oriented
Flat relational	multi-dimensional

The compelling need and purpose of Data Warehouse

The principle purpose of data warehouse is to provide Information to business users for strategic decision-making. These users interact with data-ware house using front end tools, or by getting the required information through the information delivery system

Need of data ware house

- Rapid growth of data volume and variety in modern organization has created a need for a centralized repository for storing and managing data from disparates sources
- Operational system are often designed for specific tasks and are not well suited for complex analysis. A data ware house provide a unified view of data from across the organization, making it easier to

perform Queries and analysis.

- The need of timely and accurate information for decision making is more important than ever before. A datawarehouse can provide users with access to the data they need to make informed decision

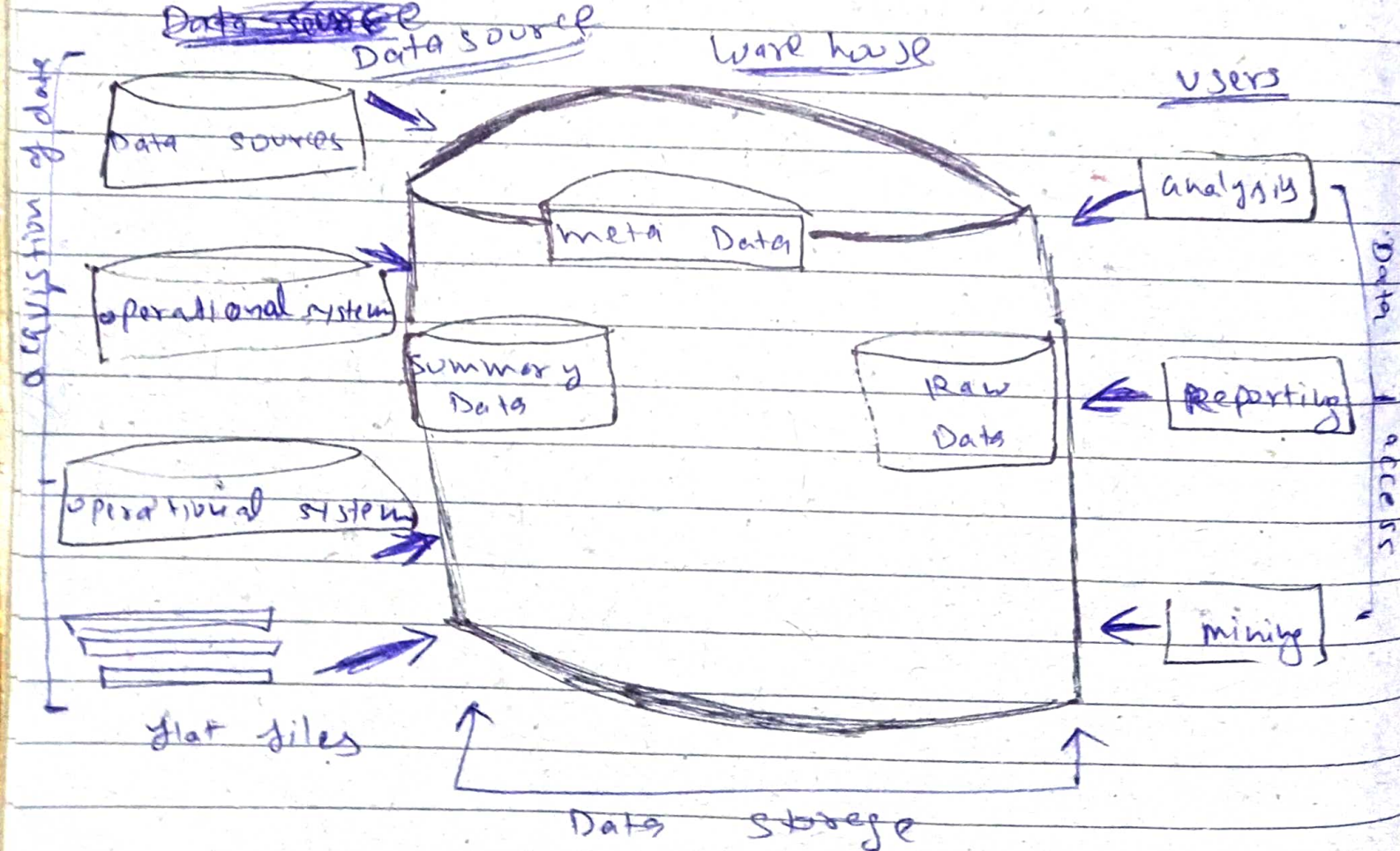
Purpose of data warehouse

- To provide a centralized repository for storing and managing the data from disparates sources
- To transform data into a consistent format that is easy to understand and analyze
- To support the development of decision support system (DSS) and business inteligence
- To enable data mining and machine learning initiative.

Data-Warehouse architecture

The hardware and software and data resources

required to construct the data warehouse depends upon the organization that went to construct it. The need and resource available, force the decision of organization regarding the architecture of particular data warehouse, there are many phases, that are common to all the data warehouse regardless of the organization or the design selected. The most common phases are acquisition of data, storage of data and data access



Date
DELTA Pg No.

- Acquisition of Data ⇒ All the data warehouse must have a source from where the data is acquired. Most of data in the warehouse is derived from the operational data of the organization. The required data is extracted, filtered, translated and integrated into data storage environment

- Storage of Data ⇒ The large amount of operational data that is historical in nature are defined, indexed and then partioned allow for economic and efficient access

- Data access ⇒ A number of data mining applications allow many users throughout the organization to retrieve, analyze, query and generate reports. The ability to access data is fundamental to the concept of data warehouse in the organization

Datawarehouse component

There are mainly six component

(i) summarized data (ii) Integration/Transformation program
(iii) Operational data store (iv) Detail data (v) meta data
(vi) Archives

(i) Summarized data => The raw data generated by a transaction - Processing system may be too large to store online. However, many queries can be answered by just maintaing the ~~Data warehouse~~ summary data obtain by aggregation on a relation, rather than maintaining the entire relation.

Summery data is classified into two catego-ry

(a) Light Summarized data => ~~obtained by aggregati-~~ This represent data distilled from current detailed data. It is summarized according to some unit of time and always resides on disk

(b) ~~Highly~~ summarized data => This represent data distilled from lightly summarized data. It is more compact and easily accesible and resides on disk.

(ii) operational data - store => operational data - base are source data for the data ware house. operational data store is repository of

operational data

(iii) Integration / transformation system → The integration and transformation program convert the operational data that is application specific into enterprise data. The major function performed by these program are follows

→ Reformating, re-evaluation or changing key structure
→ Adding time element
→ Default value identification
→ Providing logic to choice b/w multiple data store

(iv) Detailed data → Detailed is of two types older detail data or current detail data. The older detail data represent data whose not very rece may as old as ten year or longer

The current details data represent data of a recent nature and always has shorter time horizon than older detail data. It can be volatile minous.

(v) Meta data → meta data is the data about data. meta data for data warehouse users

are part of the data ~~itself~~ warehouse itself and controls access and analysis of data warehouse contentents. The meta data repository is the key data warehouse component. It contain both technical and buisness meta data. The technical data contains details about acquistion, processing, storage structure, data description, warehouse operation and maintance functionality. The buisness meta data cover the relevant buisness rule and organizational details

(vi) Archieves ⇒ These contain old or historical data of significant intereset and have value to enterprise. It is generally used for forecasting and trend analysis, thus these archieves store old and meta data that describe the characterstics of old data

Data mart

A data mart is a subset of a data warehouse that is designed to serve the specific data and analytical needs of a particular buisness unit, department, or group within an organization

Date
DELTA Pg No.

Data mart are smaller, more focused, and more specialized data storage or retrival system compared to the larger and more comprehensive data warehouse

Characterstic of Data mart

- Focused data → Data mart contain data that is highly relevant to a particular buisness or user group. This data is typicaly a subset of data found in the organization's data ware-house
- Department or bisness unit specific → Data marts are often created to cater needs of specific department or buisness unit within an organization such as marketing, finance, sales or human resource.
- Data consolidation → In some case, data marts may store data from multiple source but only related to particular buisness area. consolidation make it easier for user in that area to access data they need without the complexity of dealing with data from other part of organization.

- Optimized query Performance ⇒ Data mart are designed with query performance in mind. They are structured and indexed to facilitate quick retrieval of data, making them ideal for reporting and analysis

- user friendly ⇒ Data mart are typically user friendly and tailored to the need of specific user group

- Scalability ⇒ As data needs grow within a department or business unit, the data mart can be expanded and scaled accordingly

A 3-tier Data ware-house architecture

There are mainly 3 types of Data ware-house architecture

- Single-tier architecture ⇒ The objective of a single tier architecture is to minimize the amount the of data stored. This goal is to remove data redundancy. This architecture is not frequently used in practice

- Two-tier Architecture

two layer architecture seperates physically available sources and data-warehouse. This architecture is not expandable and also not supporting a large number of end users. It also has connectivity problem because of network limitation.

- Three-tier architecture

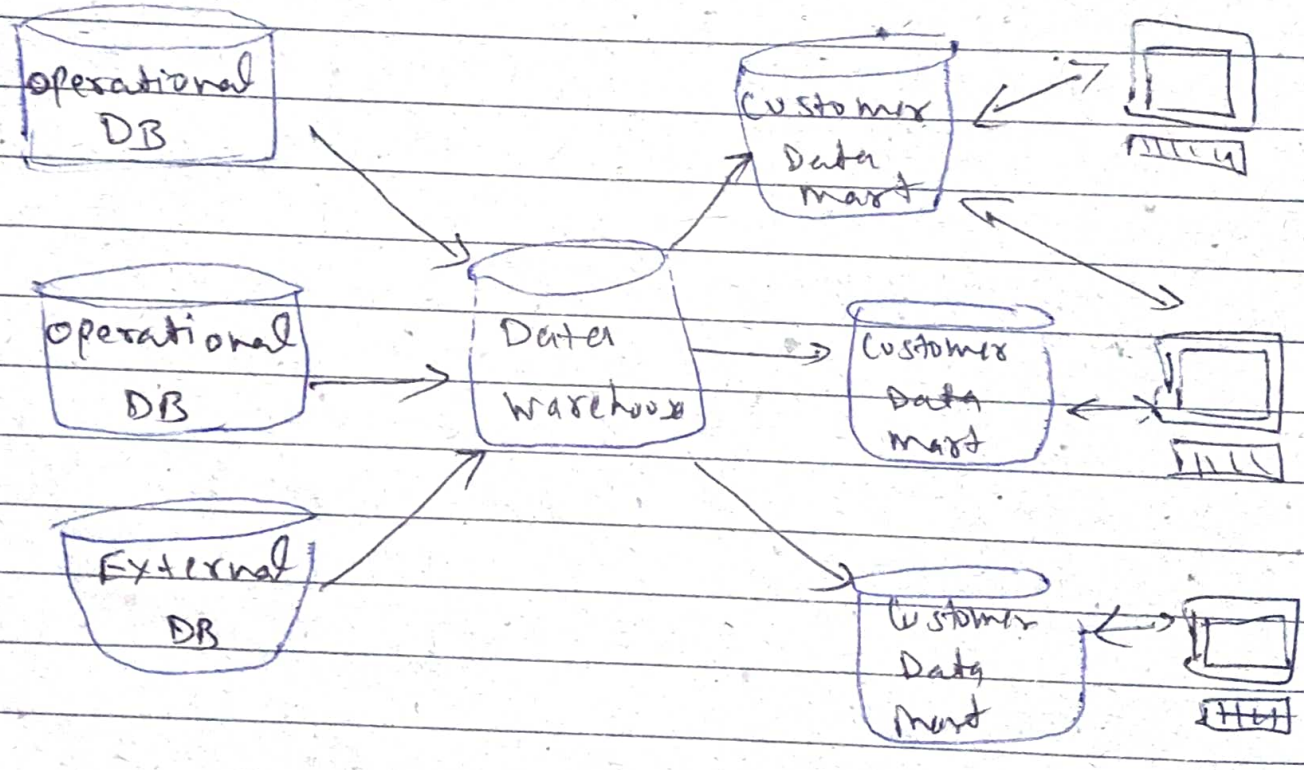
This is most widely used architecture, it consist 3 layer top, middle, bottom

1) Bottom tier => The database of the Datewarehouse servers as the bottom tier. It is usually a relational database system. Data is cleansed, transformed, and loaded into this layer using back-end tools

2) Middle tier => The middle tier in Data-Warehouse is an OLAP server which is implemented using either ROLAP or MOLAP model. For a user, this application tier present an abstracted use of database. This layer also acts as a mediator b/w the end-user and database

Date
DELTA Pg No.

3) Top tier -> The Top tier is a front end client layer. Top tier is the tool and API that you connect and get data out from the data warehouse. It could be query tools, reporting tools, managed query tools, Analysis tools and data mining tools



Data Pre-Processing

Date
DELTA Pg No.

Data cleaning

Data cleaning is also known as data scrubing. Data cleaning is a process which ensure the set of data is correct and accurate. Data accuracy and consistency, data integration is checked during data cleaning. Data cleaning can be applied for a set of records or multiple set of data which need to be merged.

Data cleaning is performed by reading all records in a set and verifying their accuracy. Typos and spelling error are rectified. mislabelled data if available is labelled and filed, missing or incomplete entry are completed unrecoverable records are purged, for not to take space and inefficient operation.

Data transformation

In data transformation process data are transformed from one format to another format, that is more appropriate for data mining.

Some Data transformation strategies

1) Smoothing ⇒ It it the process of removing noise from Data

2) Agregation ⇒ It is the process where summary or aggregation operations are applied to the Data

3) Generalization ⇒ In generalization low-lvl data are replaced with high-lvl data by using concept hierarchies climbing

4) Normalization ⇒ normalization scaled attribute data so as to yall within a small specified range such as 0.0 to 1.0

5) Attribute Construction ⇒ In attribute construction, new attribute are constructed from given set of attributt

ETL (extraction Transform Load)

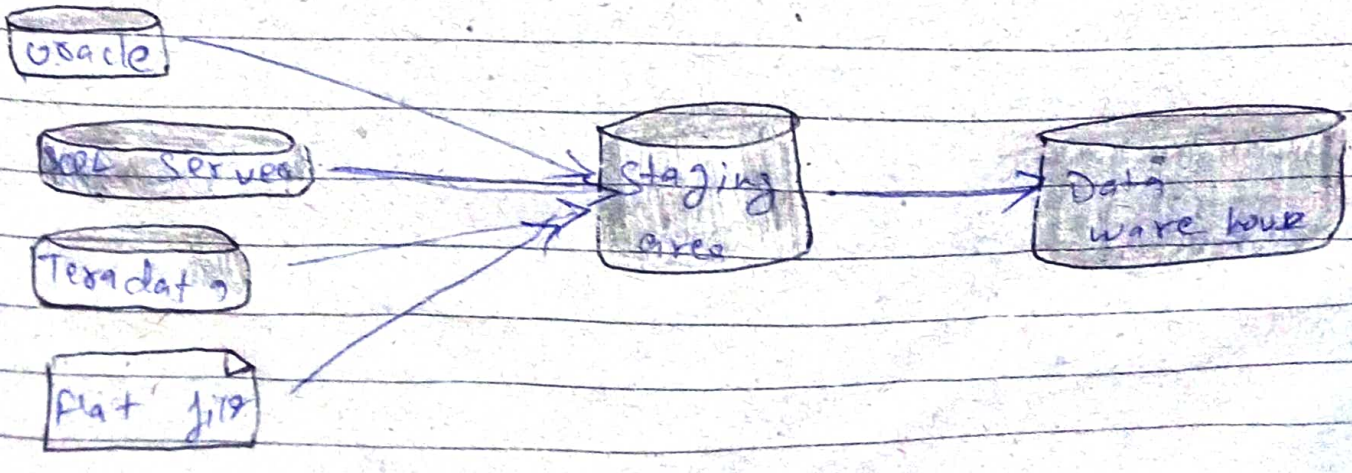
ETL is the process that defined as a process that extract the data from different RDBMS

source system, then transform the data (like apply calculation, concatenation, etc...) and finally load the data into Data warehouse system.

It's tempting to ~~thing~~ think a creating a Data warehouse is simply extracting data from multiple source and loading into Database of Datawarehouse. This is far from the truth and require a complex ETL process. The ETL process require active i/p from various stakeholder including developer, analyst, tester etc...

ETL Process

ETL is ~~3~~ step process



Date
DELTA PG No.

Step 1 Extraction

In this step data is extracted from source system into staging area. Transformation if any done in staging area so that performance of source system is not degraded. Also if corrupted data is copied directly from source into data-warehouse database, rollback will be challenge. Staging area give option to validate extracted data before it move to Data warehouse

3 Data extraction method

1) Full extraction
2) Partial " without update notification
3) ——— with ———

Some validation are done during extraction)

- Reconcile records with source data
- make sure that no spam/unwanted data loaded
- Data type check
- Remove all type of duplicate and fragmented data
- check weather all key are placed or not

Step 2 Transformation

Data extracted from source server is raw and not usable in its original form. Therefore it

needs to be cleansed, mapped and transformed. In fact this is the key step where ETL adds value and change data such as that insightful BI reports can be generated.

In this step, a set of function applied on extracted data. Data that does not require any transformation is called as direct move or pass through data

Following are Data Integrity problem

1) Different spelling of same Person like John, Jonh, Jon.
2) There are multiple way to denote company name like Google, Google Inc
3) In some Data required file was blank

validation done during this stage

- filtering - select only certain column to load
- using rules and look-up table for Data standardi zation
- character set conversation
- Encoding handling
- Data threesold validation check

- Transposing Row and colourns
- use look-up to merge data

Step 3 Loading

Loading data into target data warehouse database is the last step of ETL. In a typical Data warehouse, huge volume of data need to be loaded in a relatively short period. Hence load process should be optimized for Performance

In case of load failure, recover mechanism should be configured to restart from the point of failure without data integrity loss. Data warehouse admin need to monitor, resume, cancel, load as per prevailing server performance

Types of loading

- Initial load -> Populating all the Data warehouse table
- Incremental load -> applying ongoing chang as we need Periodically
- full fresh -> erasing the content of one more table and reloading with fresh data

Date
DELTA Pg No.

Load verification

- Ensure key field is never missing or null
- Test modeling view based on target table
- Check combined value and calculated measure
- Data checks in dimension table as well as history table
- Check BI report on the loaded fact and dimension table

ETL Tools

There are many ETL tools available in market
here are some most prominent one

marklogic

marklogic is data warehousing solution which make data integration is easier and faster using a array of enterprise feature. It can query different type data like document, relationship and metadata

oracle

oracle is the industry leading database. It offer a wide range of choice of Datawarehouse solution for both on-premises and in cloud.

Amazon redshift

amazon redshift is Datawarehouse tool, It is

Date
DELTA Pg No.

simple and cost- effective to analyze all type of data using standard SQL and existing BI tools. It allows running complex queries against petabyte of structured data

Defining the Business requirement

* Dimensional analysis

- capability to view their information over time

so the concept dimensional analysis became a method for defining data warehouse

Dimensional analysis is a technique used in data mining to identify pattern and relationship in data by analyzing the dimension of data This is done by representing the data in a multidimensional space, where each dimension represent a different attribute of the data. The dimension are then analyzed to identify pattern and relation ship that may not be apparent from the data in its original form

Date
DELTA Pg No.

Dimensional data can be ~~explore~~ used to explore data from a variety of sources, including

- Transactional data => This type of data record events that occur over time, such as sale transaction, customer interaction and website visit

- Sensor data => This type of data is collected ~~by~~ from sensor that measure physical phenomena such as temprature, pressure, and humidity

- Text data => this type of data consist of written text, such as email, document and social media post

Dimensional analysis can be used to perform a variety of task including,

- clustering => This task involve grouping similar data point together

- classification => This task involve assigning data point to a predefined set of

Date
DELTA Pg No.

Categories

- Anomly detection ⇒ This task involve identifying data points that are outliners or deviate from norm

★ Multidimensional analysis

It is a data analysis ~~tech~~ technique that group data into multiple dimension, allowing for the exploration of complex relationship and pattern within data. It is often used in buisness intelligen (BI) and data mining application to gain insights into customer behaviours, sales, trending, market dynamic, and operational performance

key cherarcterstic of multidemensional analysis

1) Data Organization ⇒ MDA organize data into a multiple a multiple dimension array, where each dimension represent distinct attribute or category of the data. This structure allows for effitient analysis

Date
DELTA Pg No.

2) Hierarchies → Dimension can be organized into hierarchical relationship, reflecting parent-child relationship b/w different lvl of granularity. This enable users to drill down or roll up through data

3) measure → Numeric value associated with data points are called measure. These measure represent the quantitative aspects of the data and are used to analyze trends, performance and relationship

4) visualization technique → Multidimensional data is often visualized using techniques like Pivot table, charts and graph.

Application of MDA
used in..
- sales analysis
- customer segmentation
- Risk assessment
- operational efficiency

Date
DELTA Pg No.

* Information Package

Information Package are structured representation of specific buriness requirement or information needs in data mining. They serve as a bridge b/w abstracted buisness goals and concrete data mining tasks, ensuring that data mining efforts are aligned with overall buisness objectives. Information Package are typically include a business question, buisness requirement, data source, data mining task and expected outcome. They offer several benefit, including enhanced clarity, improved communication, focused data mining efforts, and measurable out comes.

* Requirement gathering

Good requirement start with good source. finding those quality source is an important task and, fortunately done that takes yew resource.

example

- Customer
- Administrator
- Domain expert
- user
- patner
- Industry analyst

Date
DELTA Pg No.

Requirement gathering technique

After getting and identified these sources, there are number of technique that may be used to gather requirement. The following will describe various technique

To go the requirement following technique are used

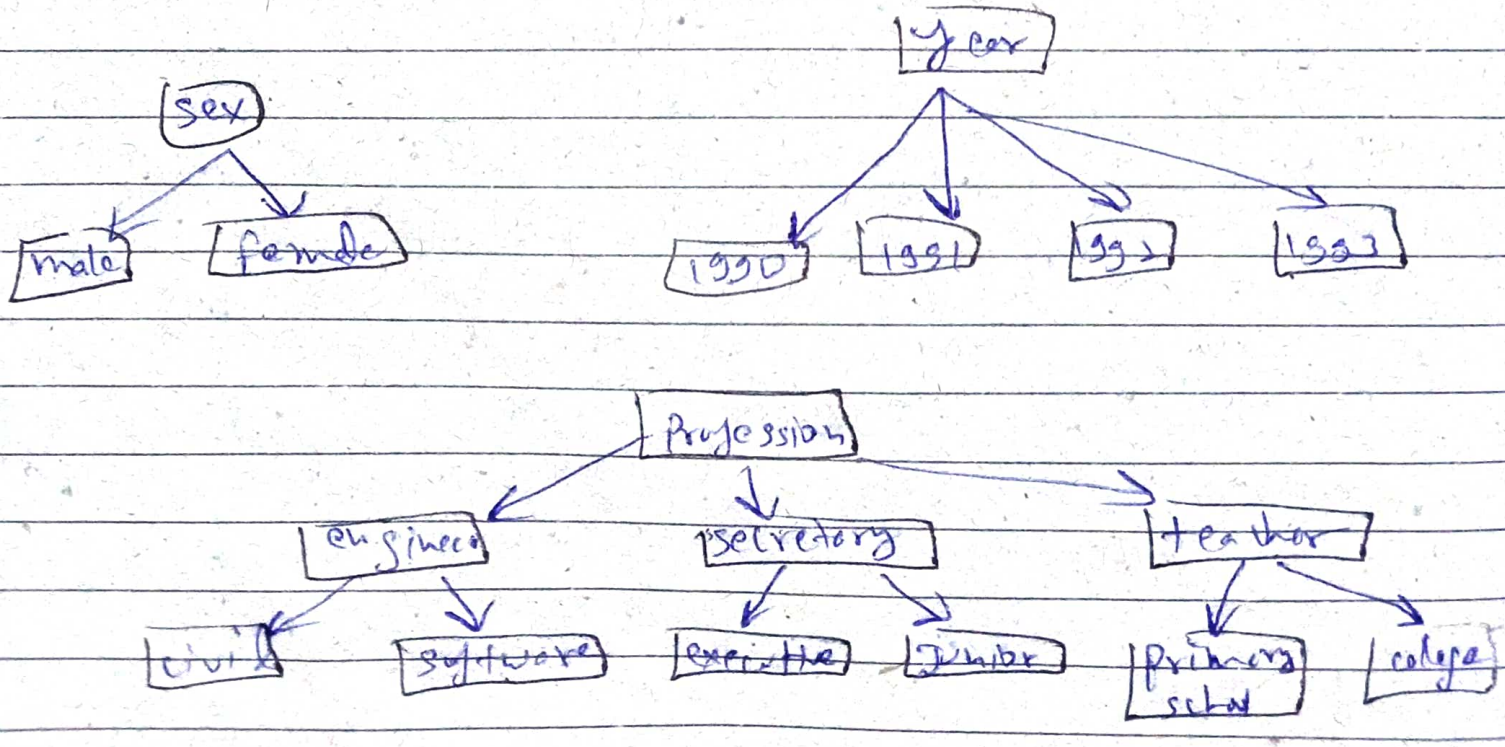
- Conduct a brainstorming session
- Interview user
- Send questionaries
- work in target envirment
- Examine suggestion and problem report
- Talk to support team.
- Conduct workshop
- Demonstrate prototy to stakeholder

UNIT 2

Date
DELTA Pg No.

Dimension modeling

The notion of a dimension provide a lot of semantic information, especially about the hierarchical relationship b/w its elements. It is important to note that dimension modeling is a special technique for structuring data around business concept. Unlike ER modeling, which describe entities and relationship, dimension modeling structures the numeric measure and the dimension. The dimension schema can represent the details of the dimensional modeling. The following figure show the dimension modeling



The Dimension hierarchy helps us view the multi dimensional data in several different data cube representation

Steps to create dimension modeling

- Step 1 => Identifying the buisness objective

The first step is to identifying the buisness objective. Sales, marketing, etc are some examples as per need of the organization. since it is the most important step of Data modeling the selection of buisness objective also depend on quality of data available for that proces

- step 2 => Identifying granuality.

Granuality is the lowest level of information stored in the table. The label of details for buisness problem and its solution is described by Grain

- Step 3 => Identifying Dimension and its attribute

Dimension are object things. Dimension

categorize and describe data warehouse facts and measure in a way that support meaning-full answer to business question. A data-warehouse organizes descriptive attribute as column in dimensional table for example the Date dimension may contain data like a year, a month, a date

Step 4

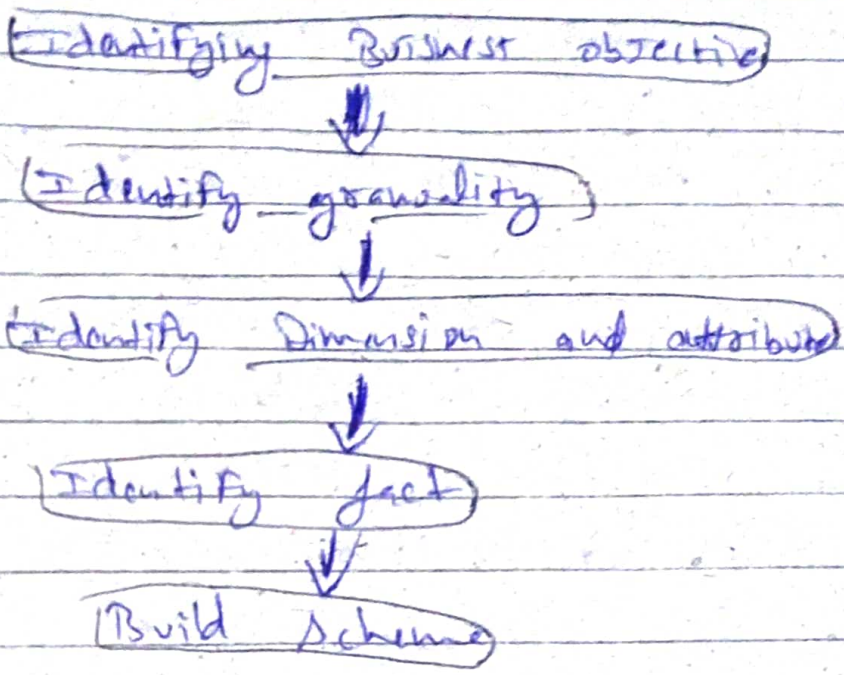
- Identifying fact building Schema

The measurable data is hold by fact table. Most of the fact table rows are numerical value like price or cost ... etc

- Step 5 Building of Schema

we Implement the Dimension model in this step. A Schema is a database structure. There are two popular schemas

→ star Schema
→ Snow-flake scheme



From Requirement to Data design effects

Good "Business intelligence (BI)", allows your organization to query data obtained from trusted source and use it gain a competitive edge in your industry. The first step to achieving effective BI is a well designed warehouse

- Require Gathering → Identifying long-term business strategy alongside current and technical requirement

Date
DELTA Pg No.

=> Analyze ~~long-term~~ user needs, reporting requirement, and plan for hardware, development, testing, implementation, and user training

=> Develop a disaster recovery plan for quick response to threat

- ~~Physical Enviornment Setup~~ => Establish separate physical environment for development testing, and production to ensure smooth transitions and data integrity

- Data modeling => Define how data structure will be accessed, connected, processed and stored

=> Identify data sources, ensuring knowledge of data location and availability for successfull project execution

- ETL (extract, transform, load) => identify data source during data modeling to reduce ETL development time.

Date
DELTA Pg No.

=> prioritize optimized load speeds without sacrificing quality to prevent poor performance

- OLAP cube design => obtain specification and measure during the design process

=> optimize the OLAP cube generation path to prevent performance issues after the data warehouse goes live

- Front end development => Ensure secure access to data from any device

=> Choose a tool that allows backend modification and provides a user-friendly GUI for report customization

Date
DELTA Pg No.

Multidimensional Data Model

multi-dimensional data model stores data in the data cube. mostly, data warehousing supports two or three-dimensional cubes.

A data cube allows data to be viewed in multiple dimension. Dimensions are entities with respect to which an organization wants to keep records. for example. in store sales record, dimensions allows the store to keep track thing like monthly sales of items and the branches and location.

A multidimensional database helps to provide data-related answer to complex business queries quickly and accurately.

Data warehouse and online analytical processing (OLAP) tools are based on a multidimensional data model. OLAP in data warehousing enable users to view data from different angle and dimen-

Date
DELTA Pg No.

What is multidimensional schemas ?

- Multidimensional schema is espicially designed to model data warehouse systems. The schemas are designed to address the unique needs of very large data-base designed for the analytic purpose (OLAP)

Types of Data warehouse schema

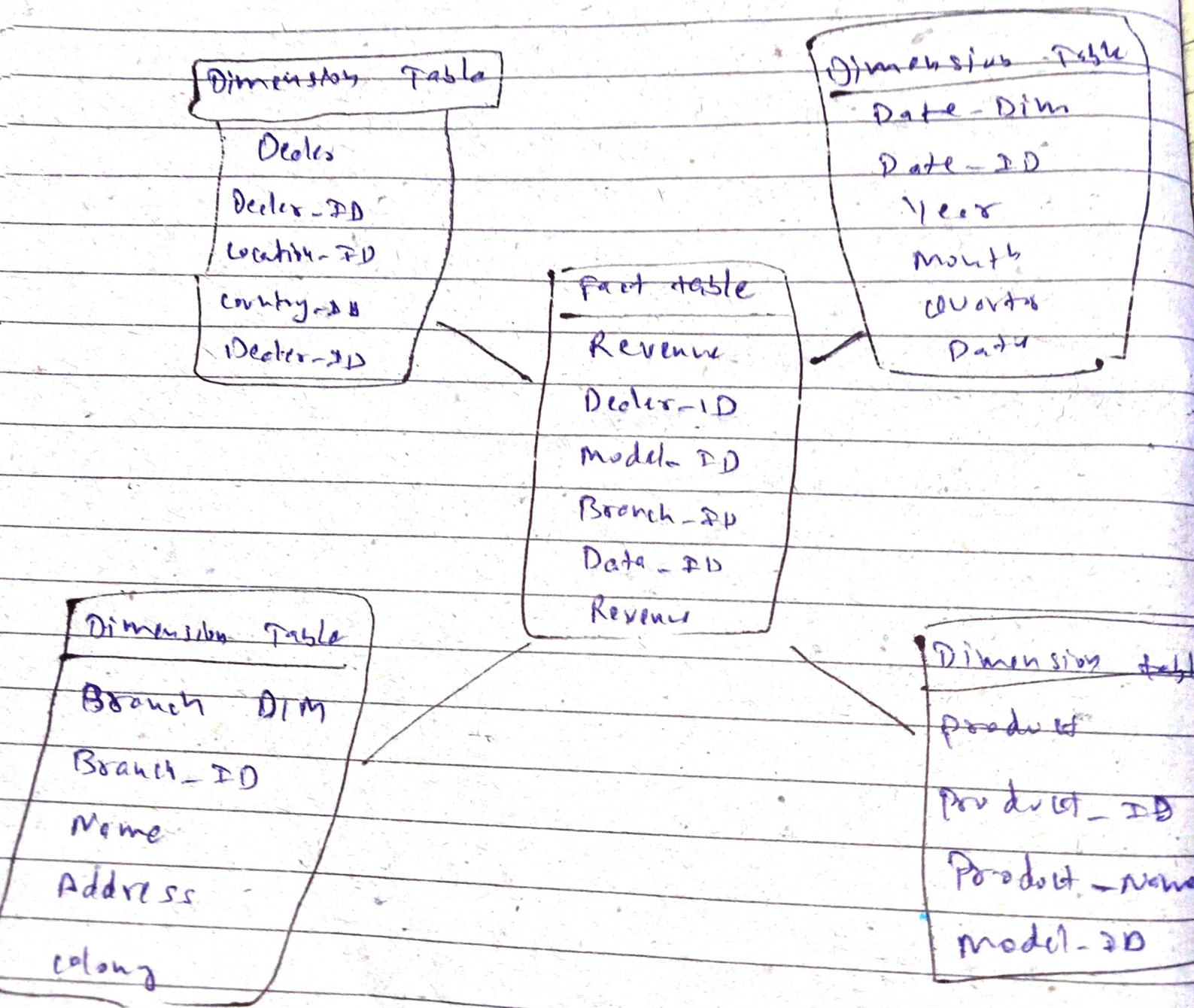
- Star schema
- Snow-flake schema
- Galaxy schema

- Star Scheme

The Star scheme is the simplest type of Data warehouse scheme. It is known as Star scheme as its structure resembles a star. In the Star scheme, the center of the

Date
DELTA Pg No.

Star can have one fact tables and and numbers of associated dimension tables. It is also known as Star Join schema and is optimized for querying large data sets.



Date
DELTA Pg No.

for example, as you can see in the above - given image the fact table is at the center which contains key to every dimension table like Deal-ID, Model-ID, product-ID, Branch-ID and other attributes like units sold and revenue

characteristic of star schema

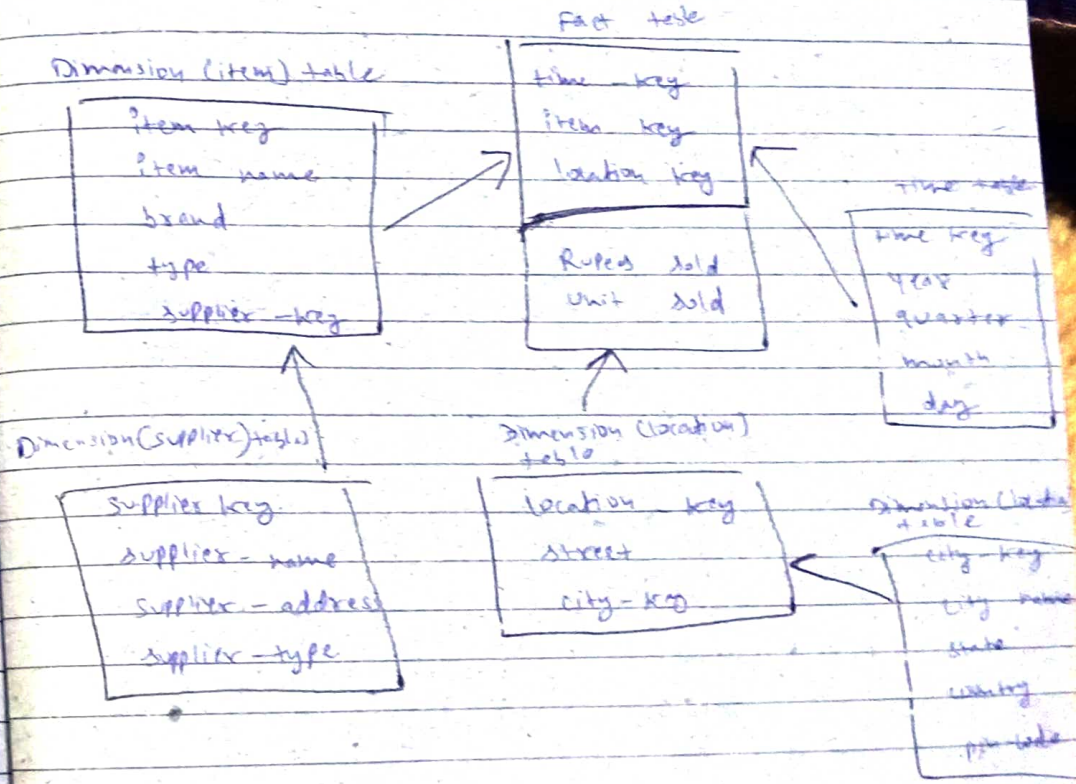
- Every dimension in a star scheme is represented with the only one-dimension table
- The dimension table should contain the set of attribute
- The dimension table is joined to the the fact table using a foreign key
- The dimension table are not joined to the fact table using a foreign key
- The dimension table are not ~~other~~

Joined to each other.

- Fact table would contain key and measure
- The schema is widely supported by BI Tools

- Snowflake Schema

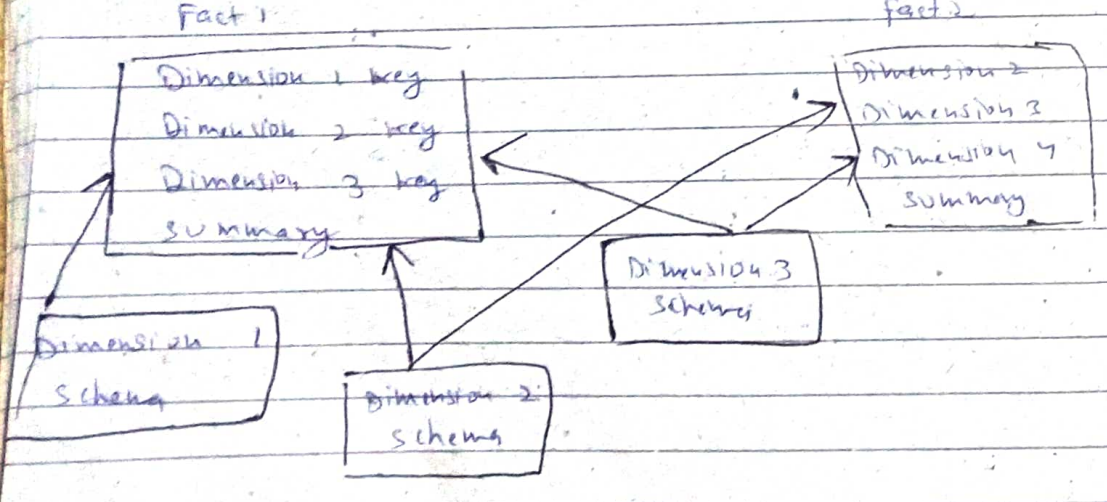
We notice that star schema consist of a single fact table and a single denormalized dimension table for each dimension of the multi-dimensional data model. To support attribute hierarchies, the dimension table can be normalized to create snowflake schema. A snowflake schema consist of single fact table and multiple dimension table. like star schema, each tuple of fact table consist of a (foreign key) pointing to each of the dimension table that provide its multidimensional co-ordinate. It also store numerical values (non-dimensional attribute) for those co-ordinate. Dimension table in a star schema is de-normalized while those in a snowflake schema are normalized.



- fact constellation

Most often, there may be a need to have more than one fact table and these are called fact constellation.

A fact constellation is a kind of schema where we have more than one fact table sharing among them some dimension tables. It is also called Galaxy schema for example let us assume that Deccan Electronic would like to have another fact table for supply and delivery. It may contain five dimension or key, time, item, delivery-agent, origin, destination along with the numeric measure as the number of unit supplied and the cost of delivery. It can be seen that both fact table can share the same item Dimension table



OLAP (online analytical processing)

OLAP is a category of software tools that allows users to interactively analyze multidimensional data from multiple perspective. It enable users to explore and analyze data to gain insights into various dimensions, such as time, geography, and Product categories. OLAP system are designed to provide a fast and efficient way to perform complex queries and analyses on large datasets

~~Key characteristic~~ and operation

1) Multidimensional Data model => OLAP system organizes data into multidimensional cubes. Each dimension represent a specific aspects of data, and the intersection of dimension contain aggregated data point

2) Cube and measure => The core components of OLAP, are data cube. Measure represent the numeric data being analyze and dimension categorize and organize

these measure

3) Slicing, Dicing and pivoting → OLAP allow user to "slice" and "dice" through the data cubes to view subset of inform ation. Slicing involve selecting a specific dimension to view a single layer of cube, while dicing involve selecting specific element from multiple dimension. pivoting involve rotating the cube to view it from a differ ent prespative.

4) Drill down and Roll up → user can Drill down in to more detailed data or roll up to view higher lvl summ aria. This flexibility allow for detailed analysis at a granual level or broder overview.

There is 2 main type of OLAP

1) MOLAP (multidimensional OLAP)

molap system store data in a multidimensional cube format. This mean that data is pre-aggregated and organized into a structure where each cell in the cube represent the unique intersection of various dimension. and data store in these intersection. It is very efficient in data retrival and its fast and perform complex queries very well and it well suited for analytical process

2) ROLAP (Relational OLAP)

ROLAP system on the other hand, store the data in relational database. Instead of pre-aggregating data into a multidimension al cube. ROLAP dynamically generate multidimen sional view on demand by using SQL queries against the relational database. This approach leverage the relational database flexibility and scalability, as they can optimize and handle complex query. generaly data store in tabular method

Need of OLAP

Solving modern business problems such as market analysis and financial forcasting require query centric data-base schema

that are array-oriented and multidimensional in nature. These business problem are characterized by the need to retrieve large number of records from very large data sets (hundread of gigabyte and even terabytes) and summerize them on the fly. The multidimensional nature of the problems it is designed to address is the key driver of OLAP

2) Fast Query Response

OLAP system in particular, store pre-aggregated data in a cube format, leading to faster query response times. This responsiveness is crucial for interactive and ad-hoc quring

3) User friendly interface

OLAP system provide a user friendly interface with feature like drag and drop, point and click, and visualisation tool making it easier for non technical user

4) Complex business reporting

OLAP facilates complex reporting by allowing user to slice, dice, drill down and pivot through data cubes enabling the creation of detailed and comprehensive repor

5) Flexibility

6) Strategic planning

7) Scalability

OLAP feature

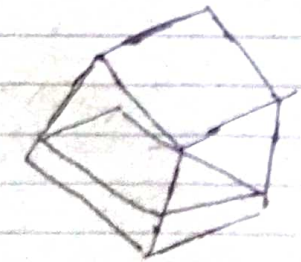
1) multidimensional conceptual view
2) Transparency
3) Accessibility
4) consistent Reporting performance
5) Client server architecture
6) Generic dimension
7) Sparse matrix handeling
8) multi-user support
9) flexible reporting
10) friendly user interface
11) scalable

Hyper cube

It is refer to multidimensional cube in with more than 3 dimension. In traditional three-dimensional OLAP cubes data is organized along three axes representing different dimension. A hyper-cube extends this concept to include additional dimension, creating a structure that allow for even more complex and detailed analysis

key parts

- Dimension => Each axis in a hypercube represent a different dimension.
- Cells => The intersection of dimension in a hypercube contain data points or cell each cells represent a unique combination of values from dimension
- Analysis => Hypercube enable users to analyze data across multiple dimension simultaneously
- flexibility => The concept of hypercube adds flexblity to OLAP by accomodating greater number of cubes

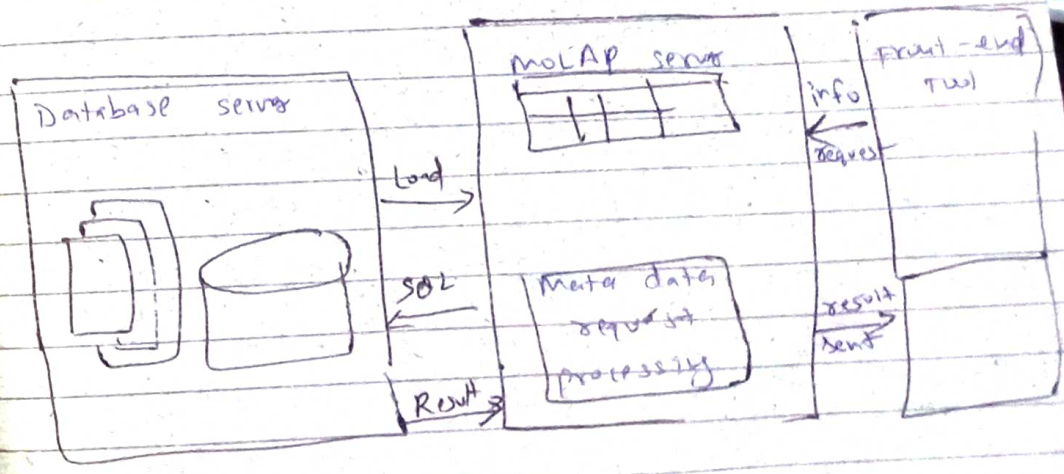


OLAP server / engine

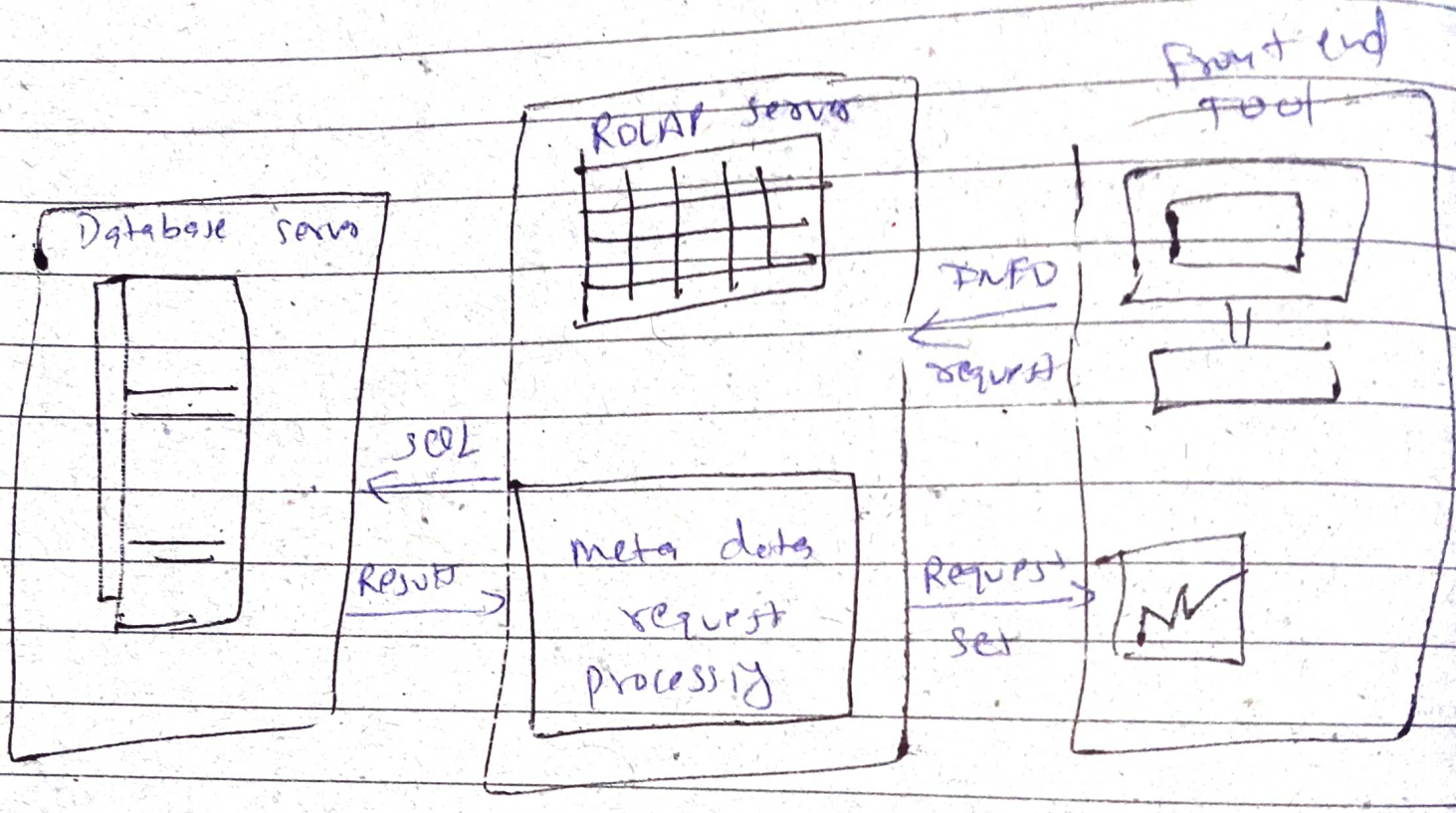
An OLAP server or OLAP engine is a software component responsible for managing and facilitating online analytical processing operations. It serve as the backend system that handles the storage, retrieval, and processing of multidimensional data to support analytical queries. The OLAP server play crucial role in providing a fast and efficient environment for user to interact with and analyze large datasets.

- Multidimensional storage → The OLAP server store data in a multi-dimensional format / typically organized in cube.
- Data aggregation → OLAP server pre-aggregate data at different level of granularity to optimize query performance
- Query processing → OLAP server process complex analytical query such as slicing, dicing, drill down and rolling up.
- Integration with BI tools → OLAP server are commonly integrate with Business intelligence tool. BI provide user interface to interact with OLAP server.
- Scalability → OLAP server need to scale efficiently to handle increasing volume of data and user queries

MOLAP architecture



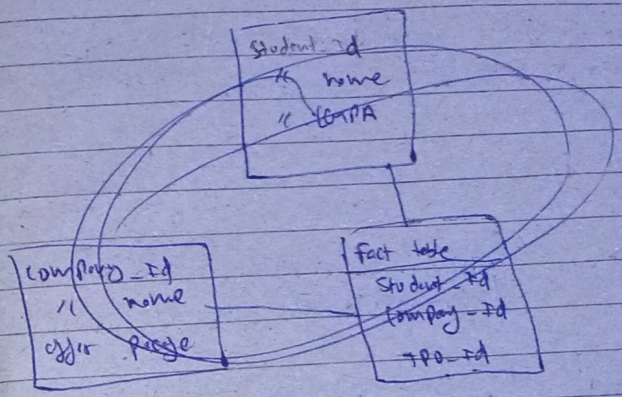
ROLAP architecture

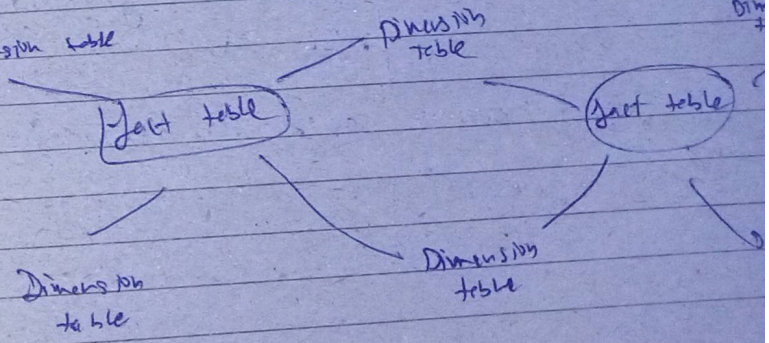


star schema vs Snowflake

• fact table and dimension table are contained	• fact table and dimension table and sub dimension are contained
• star schema is top-down model	• It is bottom-up model
• star schema use more space	• It use less space
• It take less time for execution of query	• It take more time for execution of query
• It design is very simple	• It design is complex.
• It is simple to understand	• difficult to understand
• less number of foreign key	• high number of foreign key
high data redundancy	low redundancy

Example of Galaxy or fact constellation sch

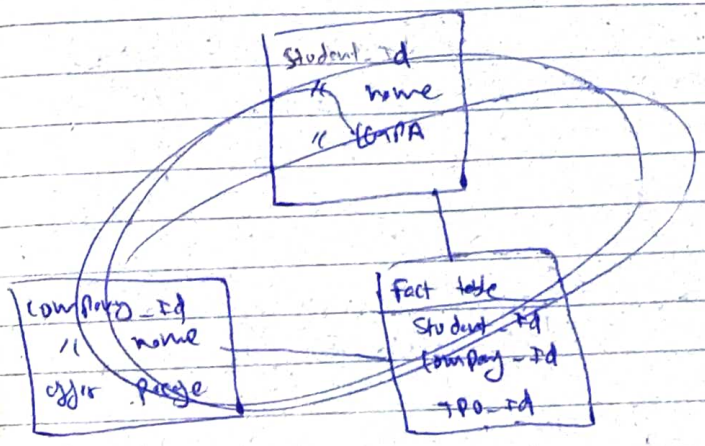


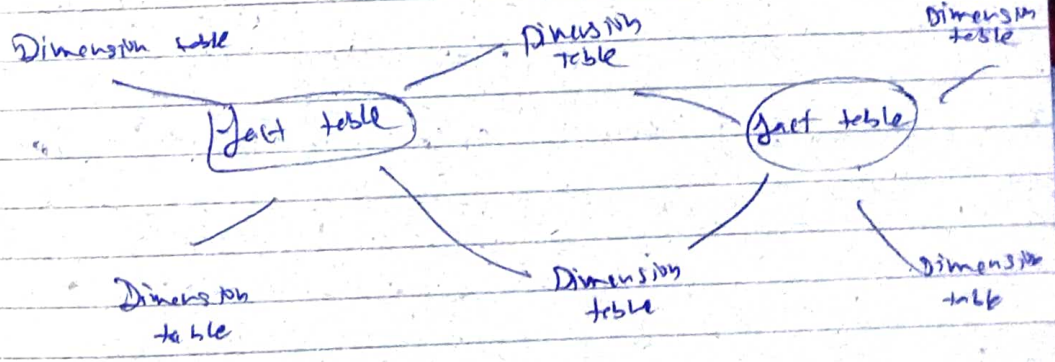


Star Schema vs Snowflake

Fact table and dimension tables are contained	Fact table and dimension table and sub dimension are contained
Star schema is top-down model	It is bottom-up model
Star schema use more space	It use less space
It take less time for execution of query	It take more time for execution of query
Its design is very simple	Its design is complex
It is simple to understand	Difficult to understand
less number of foreign key	high number of foreign key
high data redundancy	low redundancy

Example of Galaxy or fact constellation sch





MOLAP vs ROLAP

ROLAP	MOLAP
full form	full form
used for large data volume	used for limited data volume
access of ROLAP slow	access is fast
Data store in relational table	Data store in multi-dimensional array
Data is fetched from data-warehouse	Data is fetched from MDB database
complicated SQL queries are used	sparse matrix is used

EIS

An executive Information System (EIS) is a decision support system (dss) it used to assist senior executive in the decision making process. It does this by providing easy access to important data needed to achieve strategic goal in an organization. An EIS normally feature graphically display on an easy to use interface

EIS can be used many different types of organization to monitor enterprise performance as well as to identify opportunities and problem

DELTA

Codds rule for OLAP

In 1993, Dr. E.F Codds originate 12 rule as the basis for selecting OLAP tool.

These 12 rules are :

1) Multidimensional conceptual view of database
2) Concept of transparency
3) Concept of accessbility
4) Consistent reporting performance
5) client server architecture
6) Generic dimensionality
7) multi-user support
8) Dynamic ~~matrix~~ sparse matrix handeling
9) Unrestricted cross dimensional operation
10) Intitutive data manipulation
11) flexible reporting
12) unlimited dimension and aggregation lvl

UNIT 3

Data mining

Data mining is looking for hidden, valid and potentially useful pattern in huge data sets. Data mining is all about discovering unsuspected / previously unknown relationship among the data.

It is a multi - disciplinary skill that uses machine learning, statistic, AI and database technology.

The insight derived via Data-mining can be used for marketing, fraud detection and scientific discovery, etc.

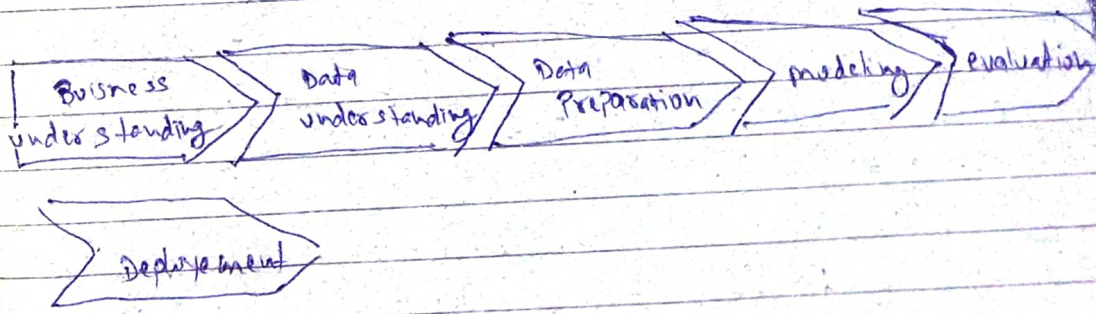
Data mining is also called as knowledge discovery, knowledge extraction, data/pattern analysis, information harvesting, etc.

Types of DATA

Data mining can be performed on following types of data

- Relational database
- Data warehouse
- Advance DB and information repositories
- object - oriented and object - relational database
- Transactional and spatial database.
- Heterogenous and legacy database
- Multimedia and streaming database
- Text database
- Text mining and web mining

Data mining Implementation process



1) Business understanding

In this phase, business and data-mining goals are established

- First need to understand business and client objective. You need to define what your client want
- Take stock of the current data-mining scenario, factor in resource, assumption, constraints and other significant factor into your assessment.
- Using business objective and current scenario, define your data mining goals
- A good data mining plan is very detailed and should be developed to accomplish both business and data mining goals

2) Data understanding

In this phase, sanity check on data is performed to check whether its appropriate for data mining goals

- First data is collected from multiple data source available in the organization.
- These data sources may include multiple databases, flat either or data cubes.
- next step is to search for properties of acquired data. A good way to explain the data is to answer data mining questions using query, reporting and visualization tools
- Based on result of query, the data quality should be ascertained.

3) Data Preparation

In this phase, data is made production ready

- The data preparation consume about 90% of the time of the project.
- The data from different source should be selected, cleaned, transformed, formatted, anonymize and constructed
- Data cleaning is process to "clean" the

data by smoothing noisy data and filling in missing values

4) Data Transformation

Data transformation operation would contribute towards the success of mining process.

- Smoothing: => It help to remove noise from data.
- Aggregation: => aggregation operation are applied to the data example -> The weekly sales data is aggregate to calculated the monthly and yearly total
- Generalization: => In this step, low-lvl data is replaced by high-lvl ~~concepts~~ concept with help of concept hierarchies
- Normalization: => Normalization is perform when the attribute data are scaled up or scaled down

5) Modeling

In this phase, mathematical models are used to determine Patterns

- Based on buisness objective, suitable modeling techniques should be selected for prepared data set
- Create a scenario to test check the quality and validity of model.
- Run the model on prepared dataset

6) Evaluation

In this phase, Patterns are evaluated against the buisness objective

- Result generated by the data mining model should evaluated against buisness objective
- Gaining buisness understanding is an iterative process. In fact, while understanding, how buisness requirement may be raised beause of data mining

7) Deployement: In phase, You ship your data-mining discoveries to everyday buisness operation.

- The knowledge or information discovered during data mining process should be made easy to understand for non-technical stakeholders
- A detailed deployement plan, for shipping, maintenance and monitoring of data mining discoveries is created

Data mining techniques;

1) classification => This analysis is used to retrieve important and relevant information about data, and metadata. This data mining method help to classify data in different classes.

2) clustering => This analysis is a data mining technique to identify data that are like each other. This process help to understand the difference and similarities b/w data

3) Regression => Regression method is the data mining method of identifying and analyzing the relationship b/w variable. It is used to identified the likelihood of a specific variable, given the presence of other variable

4) Association rule => This data mining technique help to find the association b/w two or more item. It discover a hidden pattern in the data set.

5) outer detection technique => This type of technique refer to observation of data item in the data set which do not match an expected pattern or expected behaviour. This technique can be used in variety of domain, such as intrusion, detection, fraud or fault detection etc.

6) sequential pattern => This data mining technique help to discover or identify similar patterns or trend in transaction data for certain period

7) Prediction => It has used a combination of the other data mining techinque like trend, sequential pattern, clustering, classification etc. It analyze past event or instance in a right sequence of predicting future event

challenges for implementation of Data mine

- Skilled experts are needed to formulate the data mining queries
- Overfitting, due to small size of (really) database, ~~is~~ ~~sometime are difficult to manage~~ a model may not fit in future State.
- Data-mining need large database which sometime are difficult to manage
- Business practice may need to be modified to determine to use the information uncovered

- If the data set is not diverse, data mining result may not be accurate.

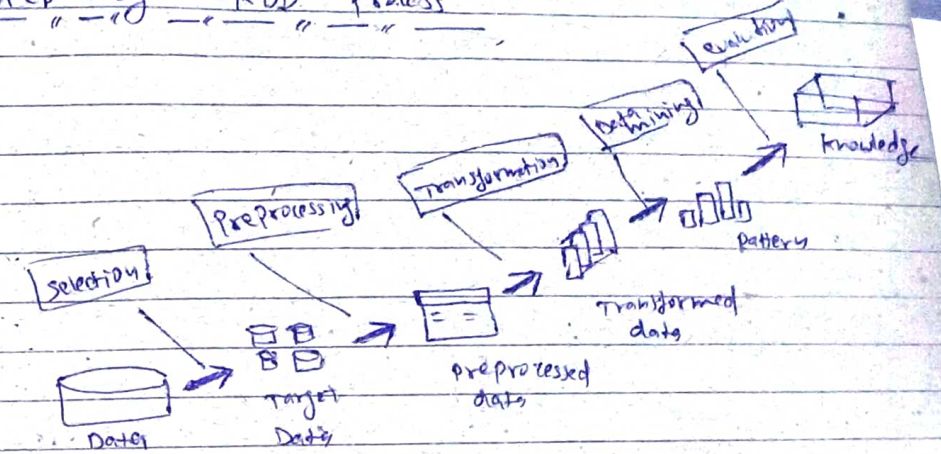
KDD Process

The term "knowledge discovery in database" (KDD) refer to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods. It is of interest to researcher in machine learning, pattern recognizing, database statistic, artificial intelligence, knowledge acquisition for expert system and data visualization

The unifying goal of KDD Process is to extract knowledge from data in context of large database

It does this by using data mining method to extract what is deemed knowledge, according to the specification of measure and threshold, using a database along with any required preprocessing, subsampling and transformation of database

- Step of KDD Process



The overall process of finding and interpreting patterns from data involve the repeated application of following steps

1) Developing an understanding of :
   - the application domain
   - the relevant prior knowledge
   - the goal of end-user

2) creating a target data set : selecting a data set, or focusing on a subset of variable or data samples, on which discovery is to be performed.

3) Data cleaning and Preprocessing :

- Removing of noise and outliers
- collecting necessary information to model or account for noise
- strategies for handeling missing data field
- Accounting for time sequence information and known change

4) Data reduction and Projection :

- Finding useful feature to represent the data depending on goal of task
- using dimensionality reduction or transformation method to reduce the effective number of variables under consideration or to find invariant representation for data

5) choosing data mining task

- Decide whether the goal of KDD Process is classification, regression, clustering etc.

6) choosing data mining algorithm

- selecting method to be used for searching pattern in data.
- Deciding which model and parameters may be appropriate.

- matching a particular data mining method with the overall criteria of KDD Process

7) Data mining

- Searching for Pattern of interest in a particular representational form or a set of such representation as classification rules or trees, regression, clustring and so on

8) Interpreting mined patters
9) Consolidating discovered knowledge

The Business context of datamining

Why does an organisation have to practise data mining when it does not bring impact to their business? In Product marketing, the marketing manager should identify the segment of the population who is most likely to respond to your product. Identify these segment of population involve understanding the overall population and deploying the right technique to classify the population. like we in Predictive modeling, there

are several way to interact with customer using different channel. These include direct marketing, print advertisement, telemarketing, radio, and TV ads. It is only through data mining, that an analyst would conclude which is optimal channel for sending communication for customer

Process improvement through Data mining

Process improvement through data mining involve using data analysis technique to discover pattern, trends, and insights within business processes By identifying areas for improvement, organization can make inform-med decision to enhance efficiency, reduce cost, and optimize performance

How data mining can applied to process improvement.

1) Define objective and scope => clearly articulate the objective of the process improvement initiative, and define scope by identifying the specific processes or areas to be analyzed.

2) Data collection => Gather relevant data related to the targeted processes.

3) Data integration => Integrate data from various resou-rces to create a comprehsive data set

4) Data cleaning and Preprocessing => Identifying and handle missing or erroneous data and nor-malize data to ensure consistency

Data mining as research tool

Data mining serve as Powerfull research tool across various domain, enabling researcher to extract valuable patterns, trends and insights from large data set.

Here how data mining can be applied as research tool.

1) Hypothesis Generation => Researcher can use data mining to explore large dataset and generate hypothesis or research questions base on observed patterh and trends

2) Data collection and integration => Gather relevant data from various source to create a comp-rehsive dataset for analysis.

Integrate and Preprocess the data to ensure its quality and consistency

3) Regression analysis => employ regression analysis to model relationship b/w variable, allowing researcher to predict and understanding the impact of one variable on another

4) clustering => use clustering algorithm to group similar data points together, providing insights into the inherent structure of data

Data mining in marketing

Data mining play crucial role in marketing by helping businesses analyze large volumes of data to extract valuable insights, pattern and trends

Data mining technology allow to learn more about their customer and make smart marketing decision

- Data mining information can help to
  - => increase return on investment
  - -> improve CRM and market analysis
  - => reduce marke campaign cost

- There is multiple benifit y using data mng in marketing
  - => Predict future trend
  - -> understand customer Purchase habbit
  - => Help with decision making
  - => improve company revenue and lower cost
  - => market basket analysis

* Data mining technique for marketing

1) knowledge base marketing

- It is marketing which make use of the macro and micro-environmental knowledge that is available to the marketing functional unit in an organization.

- There are 3 major area of application

Data
DELTA Pg No.

- ~~of data mining for knowledge~~ - based marketing are customer profiling, deviation analysis, and trend analysis

- The customer profiling system can analyse the frequency of purchase, companies can know how many time the customer can buy this product or visit the store

- The deviation analysis gives marketer a good capability to query changes that occured as a result of recent price changes or promotion

- The trend analysis can determine ~~the~~ trend in sales, cost and profits by products or market in order to achieve the highest amount of sale

2) Market Basket marketing

- most common and useful type of data analysis for marketing and retailing

Data
DELTA Pg No.

- Determine what customer purchase together
- Improve the effectiveness of marketing and sales tactic using customer data already available to company

3) Social media marketing

- SMM is a form marketing that implement various media network in order to achieve marketing communication and branding goals

- SMM primarly covers activities involving social sharing of content, videos and images for marketing purposes, as well as paid social media advertising

- Data mining tools for marketing

-> weka
-> Rapid miner
-> R- programming