Refrence: Arjun Singh Saud Sir's note. (Asst. Prof. CDCSIT)

[Around 8 to 10 marks]

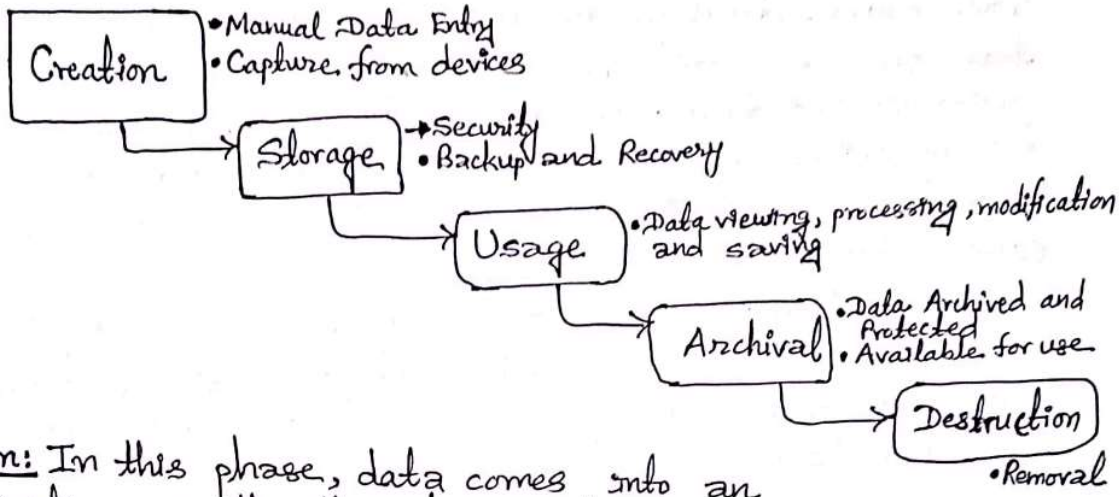
Used Colors and Usage:
- ■ → represents unit and main headings.
- ■ → represents sub-headings
- ■ → content on headings and sub-headings.

UNIT-1

Introduction to Data Warehousing

Lifecycle of data:

The data lifecycle represents all of the stages of data throughout its life from its creation for a study to its distribution and destruction.



Creation: In this phase, data comes into an organization usually through manual data entry and capture devices such as transmitted sensor data.

Storage: Once data has been created within the organization, it needs to be stored and protected. So in this phase security policies are applied and backup and recovery process are also implemented.

Usage: In this phase, data is used to support activities in the organization. Data can be viewed, modified and saved. Data may also be made available to share with others outside the organization.

Archival: Data Archival is the process of removing data from active production environment and keeping copy of data so that it can be used again in active production environment in future, if needed.

Destruction: Data destruction is the removal of every copy of data item from an organization. It is typically done from an archive storage location.

#Types of Data:

i) **Structured Data:** Structured data is the easiest to search and organize, because it is usually contained in rows and columns and its elements can be mapped into fixed pre-defined fields. Relational Databases and SQL is suitable for managing structured data.

ii) **Unstructured Data:** Data that cannot be contained in a row-column database is called unstructured data and doesn't have an associated data model. The lack of structure made unstructured data more difficult to search, manage and analyze, due to which companies have widely discarded it until the recent growth of AI and machine learning algorithms made it eaiser to process.

Examples of unstructured data include photos, video, audio files, text files etc. Instead of relational databases, unstructured data is usually stored in NoSQL databases and data warehouses.

iii) **Semi-Structured Data:** It contains characteristics of both structured data and unstructured data in a mixed way. There are some organizational properties such as semantic tags to make it eaiser to organize, but there's still variability in the data.

A good example of semi-structured data is Email message. In this Email message is unstructured content but there are some structured contents also like name of sender, name of receiver, time of message sent or received etc.

#Data Warehouse and Data Warehousing:

A data warehouse is a repository of information collected from multiple sources that stores historical data and provides support for decision-makers for data modeling and analysis. The data warehouse is the core of Business Intelligence system which is built for data analysis and reporting.

Data warehousing is the process of building data warehouse. It requires ETL operations and requires periodic data refreshing. ETL is a process that extracts the data from different source systems, then transforms the data and finally loads the data into the Data Warehouse system.

Features/Characteristics of Data Warehouse:

i) Subject Oriented: A data warehouse targets on modeling and analysis of data for decision-makers. Data warehouses provide a concise and straightforward view around a particular subject, such as customer, product, or sales, instead of the global organization's ongoing operations. This is done by excluding data that are not useful concerning the subject and including all data needed by the users to understand the subject.

ii) Time-Variant: Historical data is kept in a data warehouse. For example, one can retrive data from 3 months, 6 months, 12 months, or older from a data warehouse. This contrasts with a transactional database system, where only the most recent data is kept.

iii) Integrated: A data warehouse integrates various heterogeneous data sources like RDBMS, flat files, online transaction records etc. It requires performing data cleaning and transformation during data warehousing to ensure consistency in naming conventions, attribute types etc., among different data sources.

iv) Non-Volatile: The data in a warehouse is non-volatile. This ensures that previous data is not lost as new data is ~~upto~~ updated rather new version of the data is inserted. This separates data warehouse from operational databases which are subject to frequent changes.

Diff. betw OLTP and OLAP मानो पहिला 2nd point बाट शुरु लेख्ने

#Differences between operational database and data warehouse: [Imp]

Operational Database	Data Warehouse
i) Databases use Online Transactional Processing (OLTP).	i) Data warehouses use Online Analytical Processing (OLAP).
ii) Databases store current data only.	ii) Data warehouses store historical data.
iii) The data in databases are normalized to reduce or eliminate data redundancy.	iii) The data in data warehouses are denormalized so that data can be accessed faster.
iv) It is optimized for performing write operations.	iv) It is optimized for performing read operations.
v) It usually adopts an ER data model and an application-oriented database design.	v) It usually adopts star or snowflake model and subject oriented database design.
vi) ~~Dat~~ The transactions are usually executed in an ACID compliant manner.	vi) ACID compliance is less strictly enforced since data warehouses focus on reading, rather than modifying historical data.

#Multidimensional Data Model:

May not be asked in exam but concept imp for later use. Understand from tutorials on AI sthg youtube channel in detail. channel.

→ Data warehouses and OLAP tools are based on a multidimensional data model. This is the data model that views data in the form of a data cube.

→ A data cube allows data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts.

→ Dimensions are the entities with respect to which an organization wants to keep records. For example, an organization may create a sales data warehouse in order to keep records of the store's sales with respect to the dimensions time, item, branch, and location.

→ Each dimension may have table associated with it, called a dimension table.

→ A multidimensional data model is typically organized around a central theme, like sales. This theme is represented by fact table. Facts are numerical measures. Examples: sales_amount, units_sold etc.

#OLAP operations in multidimensional data model: [Imp]

1) Roll-up: Roll-up is also known as "consolidation" or "aggregation". It helps in generating summary from lower level to higher level. Roll-up operation can be performed in two ways:

→ Reducing dimensions

→ Climbing up concept hierarchy.

Concept hierarchy is a system of grouping things based on their order level. For Example in concept hierarchy from daily summary, weekly summary ~~, from weekly s~~ can be generated, again from weekly summary monthly summary can be generated and so on.

In the diagram below we are performing roll-up operation on the basis of location from cities to country.

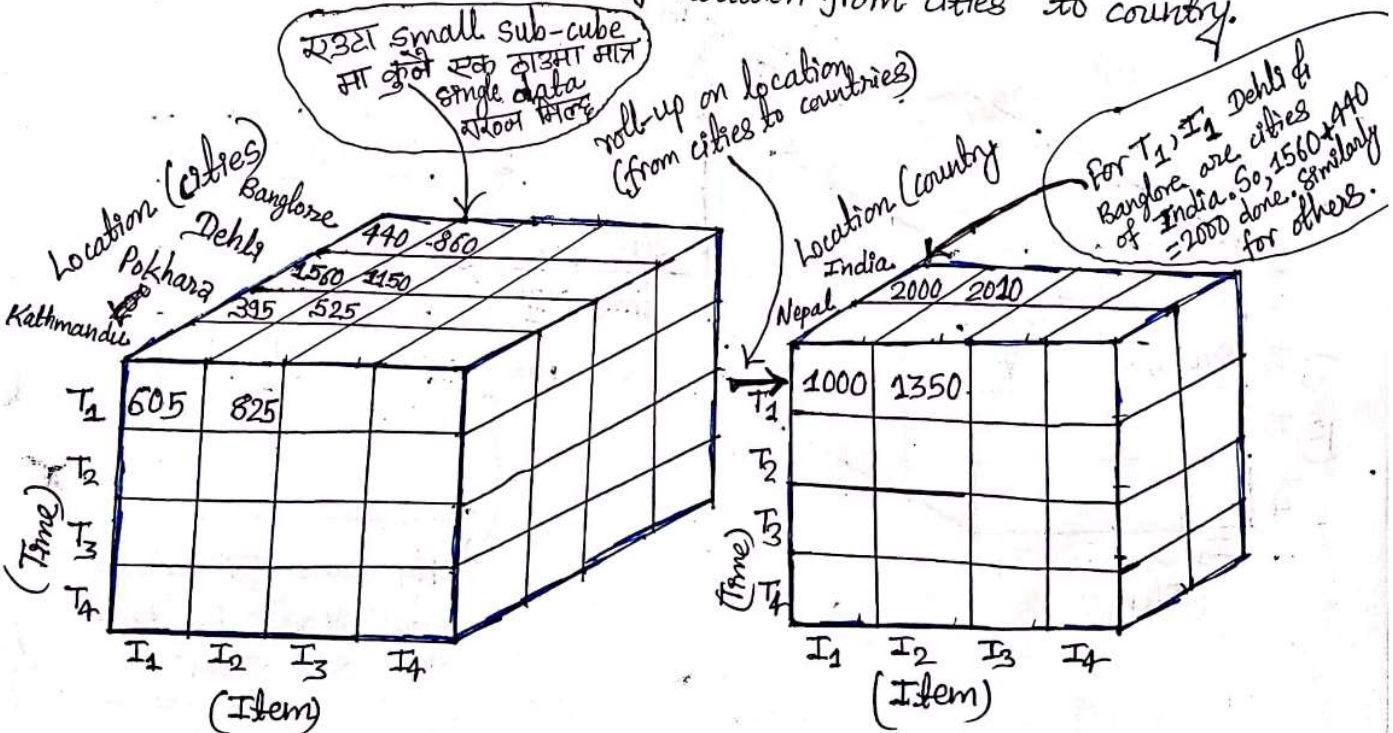


Fig: Cube 1 Fig: Cube 2

2) **Drill-down:** Drill-down is the reverse operation of roll-up. It helps in generating summary from higher level to lower level. It can also be performed by two ways:

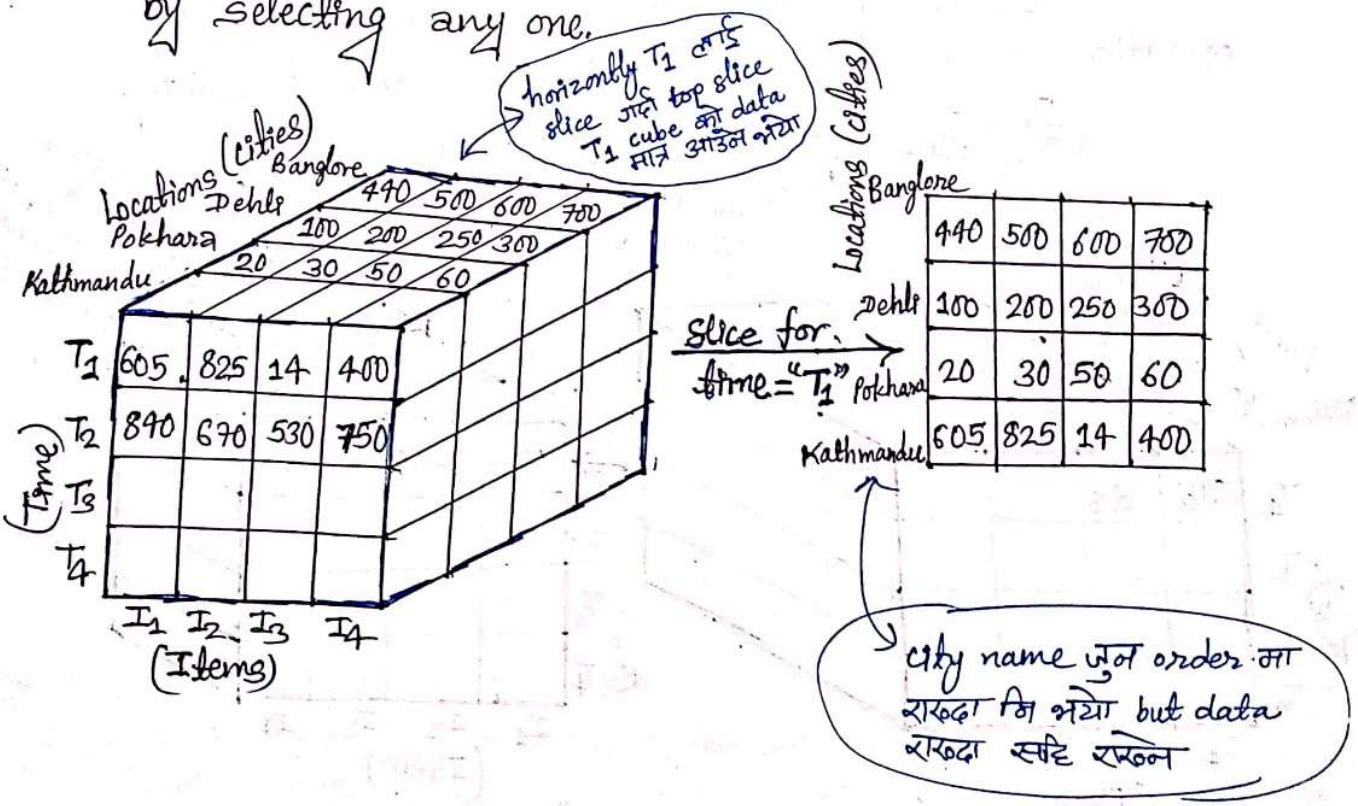
→ By stepping down a concept hierarchy for a dimension.

→ By introducing a new dimension.

For example in drill-down operation from monthly summary weekly summary can be generated, again from weekly summary daily summary can also be generated and so on.

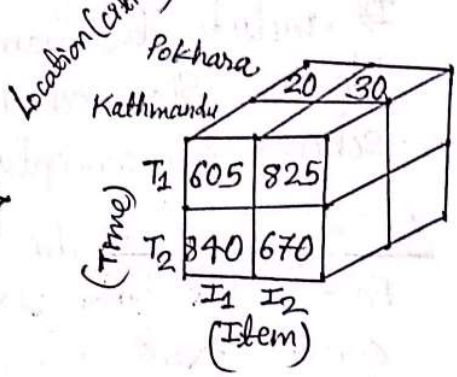
Note: For figure just reverse the figure that we draw in roll-up. i.e, first draw Cube 2 then generate (draw) Cube 1.

3) **Slice:** The slice operation selects one particular dimension from a given cube and provides a new sub-cube. Consider the following diagram that shows how slice works. Here, Slice is performed for the dimension "time" using the criterion time = "T_1". It will form a new sub-cube by selecting any one.

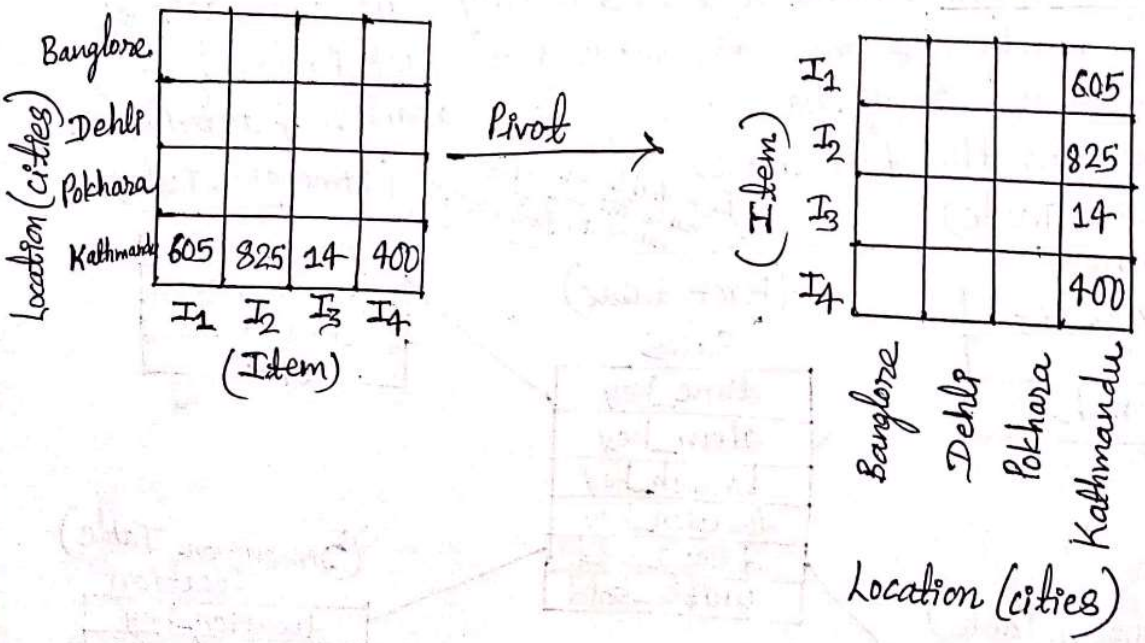


4) Dice: Dice operation selects two or more dimensions from a given cube and provides a new sub-cube. Consider the following diagram that shows the dice operation. The dice operation on the cube is based on the selection criteria involving three dimensions: (location = "Kathmandu" or "Pokhara"), (time = "T_1" or "T_2"), and (item = "I_1" or "I_2").

Same First figure that we used in slice (take refrence of that fig.) → Dice operation →



5) Pivot: The pivot operation is also known as rotation. It rotates the data axes in order to provide an alternative presentation of data. Consider the following diagram that shows the pivot operation. Normally pivoting is performed after slicing.



[Imp]

#Conceptual Modeling of Data Warehouse:

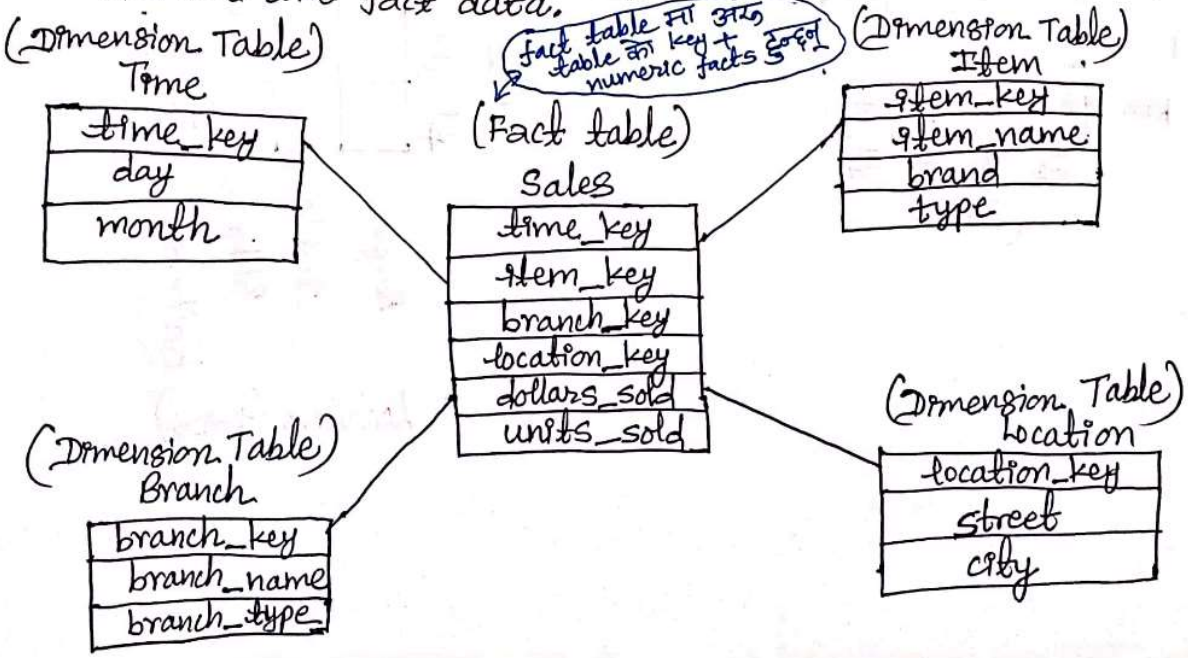
→ Normally यो unit को short question सोध्दा long सोधे maximum chance यो topic को सोध्दा / अरु topic short को लागि

A conceptual data model recognizes the highest-level relationships between the different entities. The goal of conceptual data warehouse modeling is to develop a schema for logical representation of data stored in data warehouse. Schema is a logical description of the entire data warehouse. It includes the name and description of records and aggregates. We use Star schema, Snowflake schema, and Fact-Constellation schema for conceptual modeling of data warehouse.

1) Star Schema: This schema contains two types of tables: Fact Table and Dimension Tables. Fact Table lies at the center point and dimension tables are connected with fact table such that star shape is formed.

Fact Tables: A fact table typically has two types of columns: foreign keys to dimension tables and measures that contain numeric facts. Those facts contain aggregates of data at specified level.

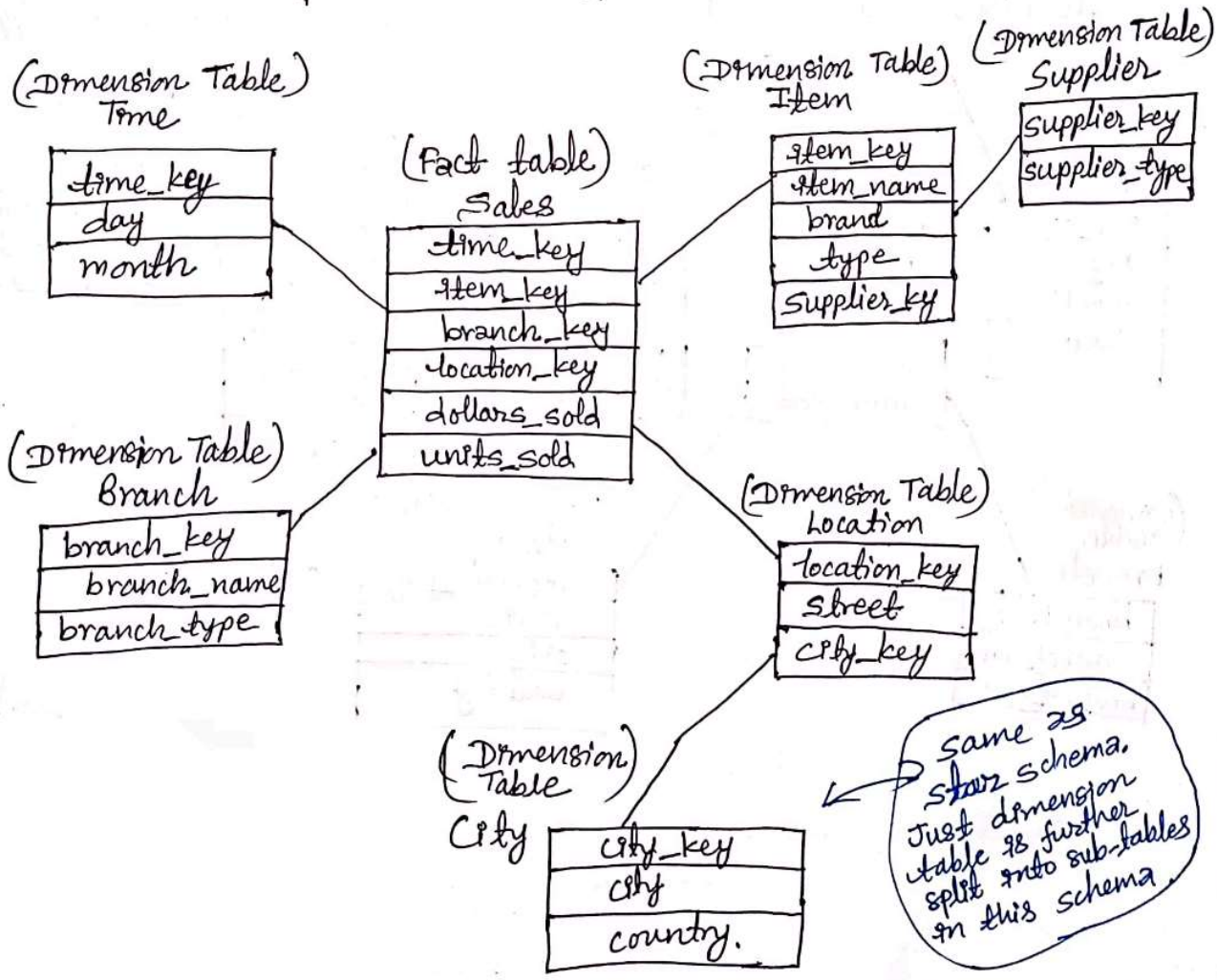
Dimension Tables: Dimension tables usually have a relatively small number of records compared to fact tables, but each record may have a very large number of attributes to describe the fact data.



Advantage: Since star schema contains de-normalized dimension tables, it leads to simpler queries due to lesser number of join operations and it also leads to better system performance.

Disadvantage: It is difficult to maintain integrity of data and data redundancy is also high in star schema due to de-normalized tables. ~~It is~~

2) Snowflake Schema: The snowflake schema is a variant of the star schema model, where some dimension tables are normalized, thereby further splitting the data into additional tables. The resulting schema graph forms a shape similar to a snowflake. For example, the item dimension table in star schema is normalized and split into two dimension tables, namely item and supplier table.



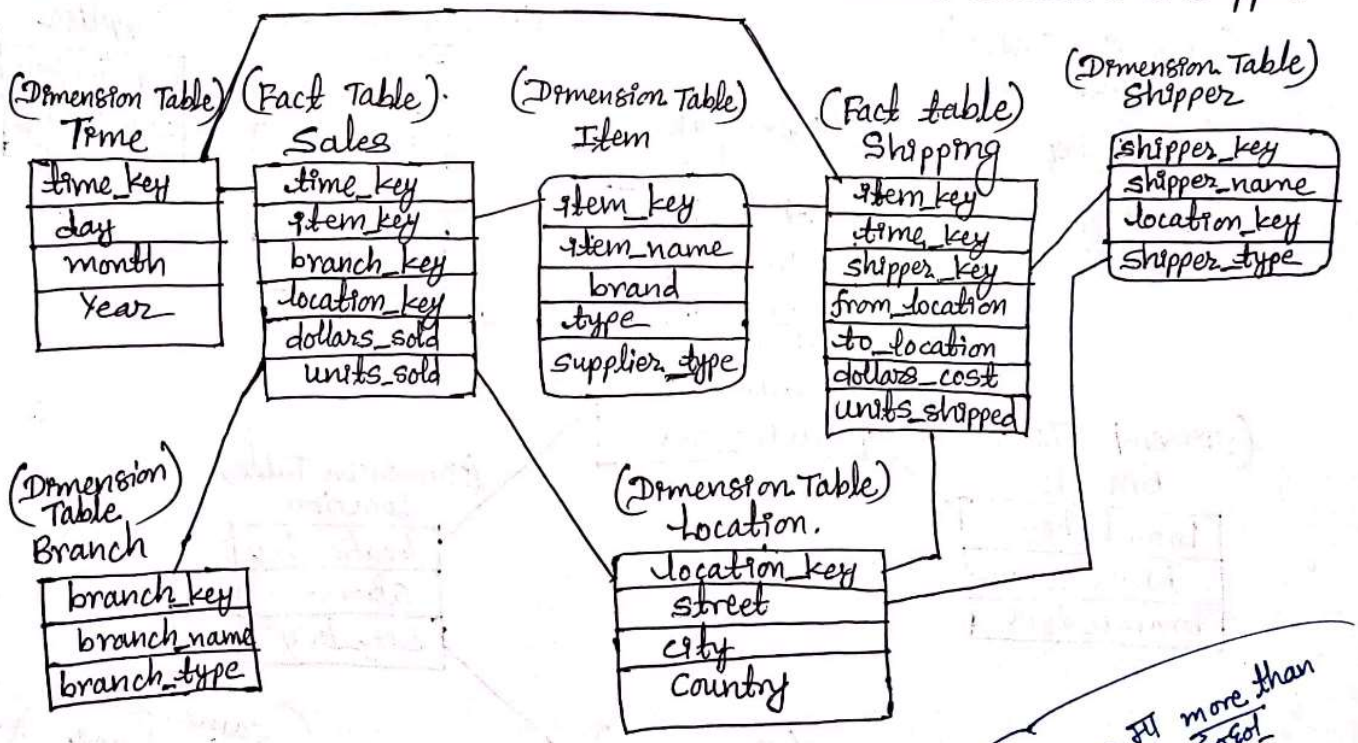
Advantages and Disadvantages here are just opposite to star schema.

Advantage: Due to normalization table is easy to maintain integrity and saves storage space.

Disadvantage: More joins will be needed to execute query due to further split of dimension tables and system performance may be adversely impacted.

3) Fact-Constellation Schema: This kind of schema can be viewed as a collection of stars, and hence is also called a galaxy schema. This schema allows dimension tables to be shared between fact tables.

For example, following schema specifies two fact tables, sales and shipping. The sales table definition is identical to that of the star schema. The shipping table has five dimensions or keys: item key, time key, shipper key, from location, and to location. It also contains two measures: dollars cost and units shipped.

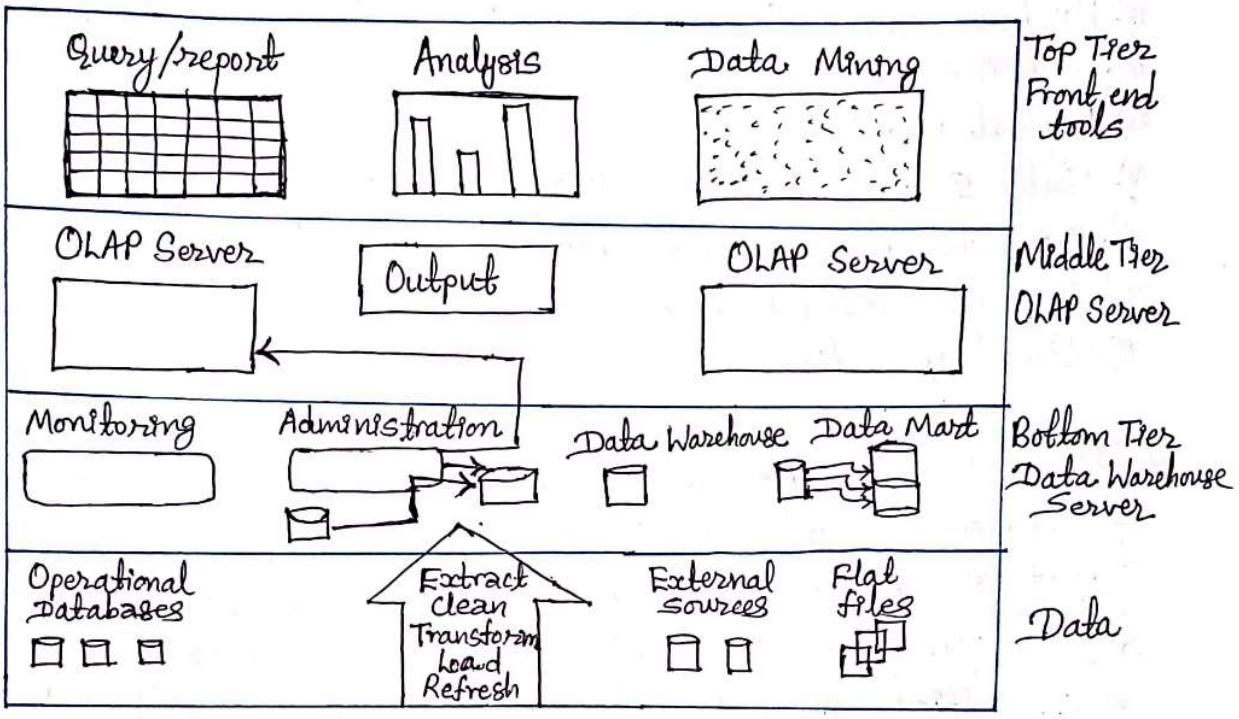


Note: It is collection of star schema. So, advantages and disadvantages of this schema are same as of star schema.

दो schema में more than 1 fact table होता। I dimension table एक से ज्यादा fact से जुड़ा होता। fact और connection होता है। for e.g. time and shipping, Sales, Location के shipping etc.

#Architecture of data warehouse: [Imp],

Generally a data warehouses adopt three-tier architecture: Bottom, Middle, and Top Tier.



Data comes to Bottom Tier from operational databases, External sources, and Flat files. Data is extracted, cleaned, transformed, loaded, and refreshed ~~before sending to~~ in the Bottom Tier.

Bottom Tier: Bottom tier of architecture is data warehouse server. Back end tools and utilities are used to feed data into the bottom tier from operational databases and other sources.

Middle Tier: We have OLAP Server in the middle tier, which can be implemented by either ROLAP or MOLAP. ROLAP is Relational OLAP and, MOLAP is Multidimensional OLAP.

Top-Tier: This tier is the front-end client layer. This layer holds query tools, reporting tools, analysis tools, and data mining tools. These tools are helpful in generating trend analysis, prediction and so on.

#Data Warehouse Implementation:

The series of activities that are essential to create a fully functioning Data Warehouse are as follows:

i) Requirements analysis and capacity planning.
ii) Hardware integration
iii) Modeling
iv) Physical modeling
v) Data Sources identifying and connecting
vi) Data will be extracted, transformed, and loaded. (i.e, ETL operations).
vii) Testing data warehouses.
viii) User application.

#Data Marts: [Imp]

A data mart is a subset of a data warehouse focused on a particular line of business, department, or subject area. Data marts make specific data available to a defined group of users, which allows those users to quickly access critical insights without wasting time searching through an entire data warehouse. For example, many companies may have a data mart that aligns with a specific department in the business, such as finance, sales or marketing. The primary purpose of a data mart is to partition a smaller set of data from a whole to provide easier data access for the end consumers.

#Components of Data Warehouse: [Imp]

A typical data warehouse has four main components: a central database, ETL tools, metadata, and access tools.

1) Central Database: A central database serves as the foundation of data warehouse. This database is traditionally implemented on the RDBMS technology. Because of Big Data, real-time performance. etc, in-memory databases are rapidly gaining popularity.

iii) Data Integration/ETL Tools: Data is pulled from source systems for rapid analytical consumption using a variety of data integration approaches such as ETL (extract, transform, load).

iii) Metadata: Metadata is data about data that describes data warehouse. It is used for building, maintaining, managing, and using data warehouse. Technical metadata describes how to access data, where it resides, and how it is structured. Business metadata adds context to our data.

iv) Access tools: Access tools allow users to interact with data in data warehouse. Examples of access tools include: query and reporting tools, application development tools, data mining tools, and OLAP tools.

#Need for data warehousing:

Data warehouse is needed due to following reasons:

→ To integrate data from multiple sources in one repository.
→ To enable business users to view summarized data from different angle.
→ To store historical data from past.
→ To help managers to make better decisions.
→ To reduce time needed for analysis and reporting.

#Trends in Data Warehousing:

Parallel Processing: Analysts need to analyze large volume of data stored in data warehouse and need to produce results fast. Uniprocessor systems may not be sufficient in many cases therefore data warehouse systems need to support parallel processing. It can be achieved either by using parallel processor or by using query processing technique.

Query Tools: Data warehouse systems need to provide query tools to users so that users can specify task, provide feedback, and seek more explanation from the system. Such tools must be user friendly.

Data Fusion: It is a technology dealing with merging of data from different sources. It has wider scope and includes real-time merging of data from instruments and monitoring systems.

Software Agents: Software agent is a program that is executed in certain environment autonomously and is capable of making decisions based on data obtained from environment and from other agents. Such agents needs to be integrated into data warehouse systems to provide alerts about predefined business conditions to users.