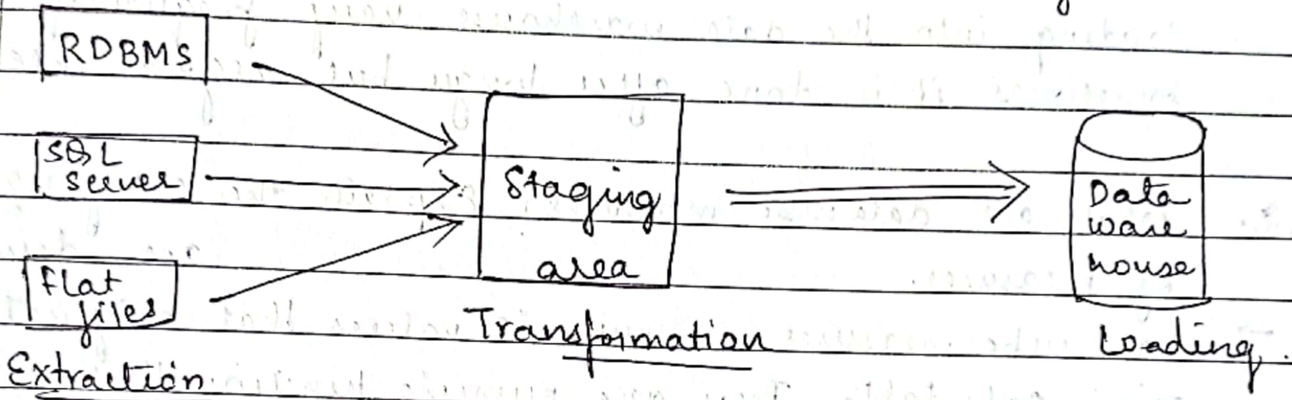
15/5/22 DMDW - Assignment - 01

Q1. What is ETL? Explain the steps in ETL.

→ ETL is a process in Data warehousing and it stands for Extract, Transform and load. It is a process in which an ETL tool extracts the data from various data source systems, transforms it in the staging area, and then finally, loads it into Data warehouse system.



⇒ Steps involved in ETL:

1. Extraction: The first step in the ETL process is extraction. In this step, data from various source systems is extracted which can be in various formats like relational db, No SQL, XML and flat files into the staging area. It is important to extract the data from various source systems and store it into the staging area first and not directly into the data warehouse because the extracted data is in various formats and can be corrupted also.

2. Transformation:- The second step of ETL process is Transformation. In this step, a set of rules or functions are applied on the extracted data to convert it into a single standard format. It may involve following processes. ① filtering - loading only certain attributes into the data warehouse ② cleaning - filling up the NULL values with some default values etc; ③ joining :- joining multiple attributes into one ④ Splitting: splitting a single

attribute into multiple attributes.

⑤ Sorting: sorting tuples on the basis of some attribute.

③. Loading:: The third and final step of ETL process is loading. In this step, the transformed data is finally loaded into the data warehouse. Sometimes the data is updated by loading into the data warehouse very frequently and sometimes it is done after longer but regular intervals.

Q2. What are datacube measures? Explain the categorization of measures.

→ Data cube measures are numeric values that are derived from originated table. They are numeric function that can be evaluated at each point in the data cube space.

→ Data cubes are mainly categorized into 2 categories:

1)→ Multi-dimensional data cube:- Most OLAP products are developed based on a structure where the cube is patterned as a multidimensional array.

2) ROLAP :: Relational OLAP make use of relational database mode rather than multidimensional db. It gets data related answers on the request of user submitted queries and it doesn't require storage and pre-computation of data.

Q3. Explain data warehouse models

→ ① Enterprise Data Warehouse (EDW):- It is a centralized warehouse. It provides a ~~unified~~ decision support service across the enterprise. It offers a unified approach for organizing & representing data. It also provide the ability to classify data according to the subject and give acces according to those divisions.

② Operational data store: ~~Ø~~ It is a datastore required when neither datawarehouse nor OLTP systems support organisations reporting needs. In OPS, data warehouse is refreshed in real time. Hence it is widely preferred for routine activities like storing records of the Employees.

③ Data Mart: It is a subset of datawarehouse. It is specially designed for a particular line of business, such as sales, / finance. In an independent date mart, data can collect directly from sources.

Q4. Explain efficient processing of OLAP queries

→ Efficient processing of OLAP queries:

→ The purpose of materializing cuboids and constructing OLAP index structures is to speed up query processing in data cubes. 2 steps are involved:

1) Determine which operations should be performed on the available cuboids: This involves transforming any selection, projection, roll-up and drill-down operations specified in the query into corresponding SQL and/or OLAP operations
Eg:- Slicing Ee Dicing of a data.

2) Determine to which materialized cuboids the relevant operations should be applied: This involves identifying all the materialized cuboids that may potentially be used to answer the query, estimating the costs of using the remaining materialized cuboids and selecting the cuboid with the least cost.

Q5. What is meta data in data warehousing ? What does it contain ?

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

→ metadata is the roadmap to the datawarehouse and it acts as a directory

→ Metadata in the datawarehouse defines the warehouse objects.

→ meta data consists of following:

1) Business metadata:- It contains the data ownership information, business defn and changing policies.

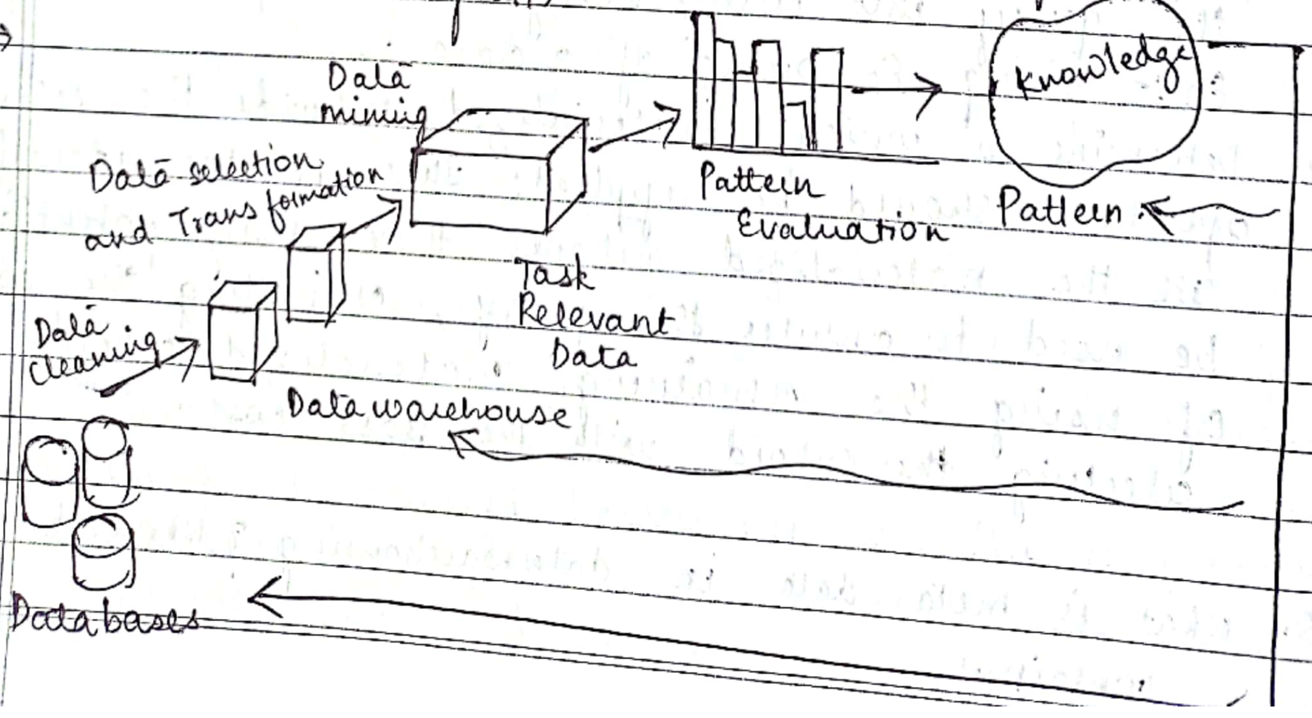
2) Operational metadata:- It includes currency of data and data lineage.

3) Data for mapping from operational environment to datawarehouse : It includes the source databases and their contents, data extraction etc;

4) Algorithms for summarization:- It includes dimension algorithms, data on granularity, aggregation etc;

Q6. Explain the knowledge data discovery (KDD) with a neat diagram

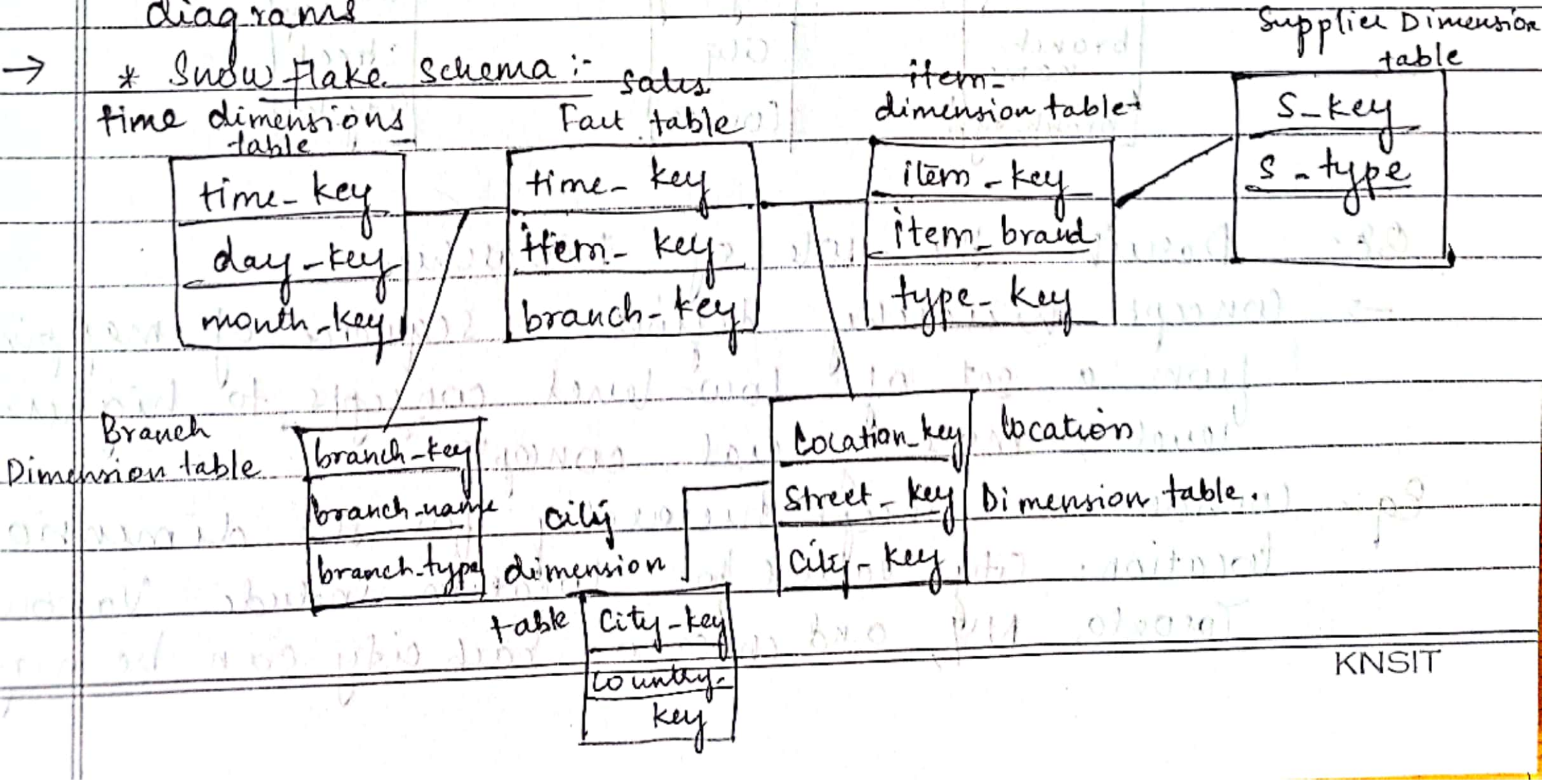
→



1. Data cleaning :- is defined as removal of noisy and Irrelevant data from collection
2. Data Integration :- Data Integration is defined as heterogeneous data from multiple sources combined in a common source.
3. Data selection : It is a process where data relevant to the analysis ~~asho~~ is decided and retrieved from the data collection
4. Data Transformation :- is defined as the process of transforming data into appropriate form required by mining procedure.
5. Data mining : it is defined as clever techniques that are applied to extract patterns potentially useful.
6. Pattern Evaluation :- It is defined as identifying strictly increasing patterns representing knowledge based.
7. Knowledge representation :- is defined as a technique which utilizes visualization tools to represent data mining results

Q7: Explain the snowflake and star schemas with relevant diagrams

→ * Snow flake Schema :-



→ Some dimension tables in the snowflake schema are normalized

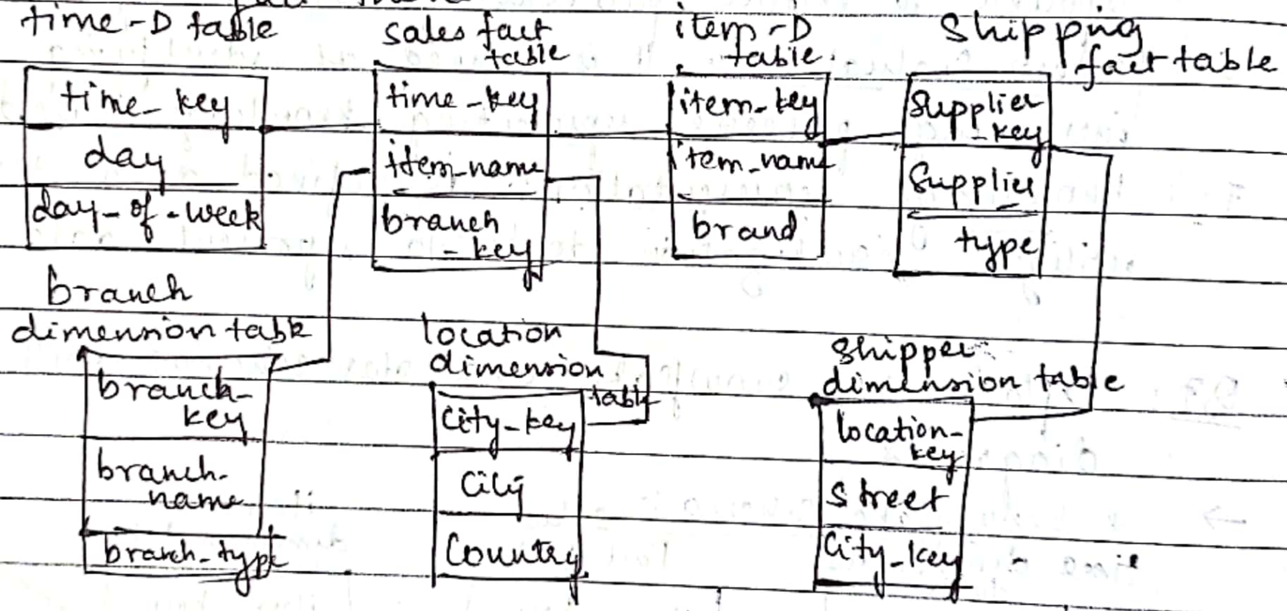
→ The normalization splits up the data in additional tables

→ Unlike star schema, the dimensions table in a snowflake schema are normalized.

⇒ Fact constellation schema

→ A fact conetellation has multiple fact tables. It is also known as galaxy schema

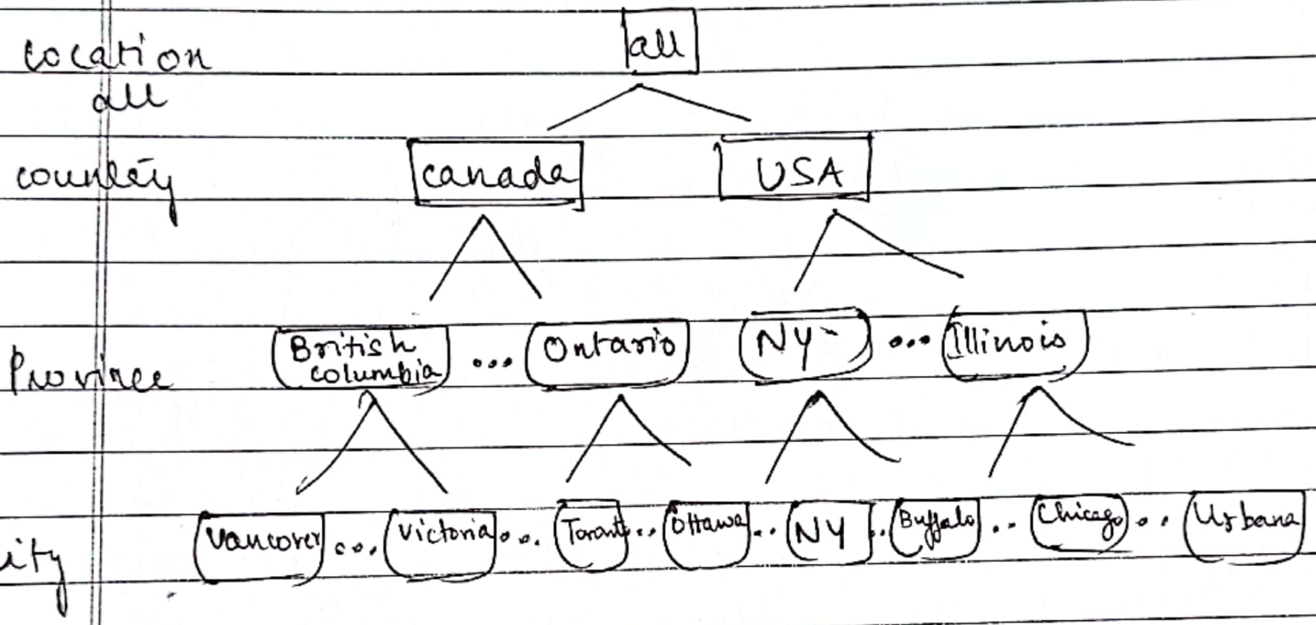
→ It is also possible to share dimension tables Between fact tables.



Q8. Describe the role of hierarchies

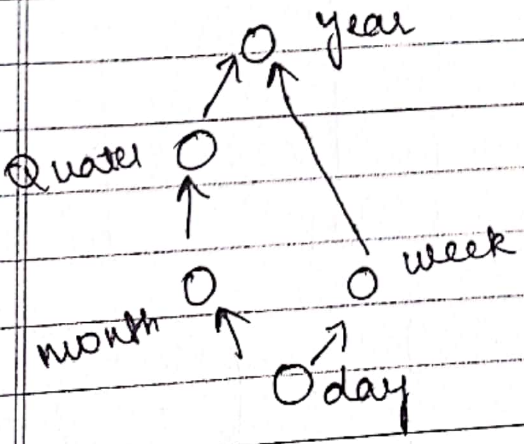
→ Concept hierarchy defines a sequence of mappings from a set of low-level concepts to higher-level, more general concepts.

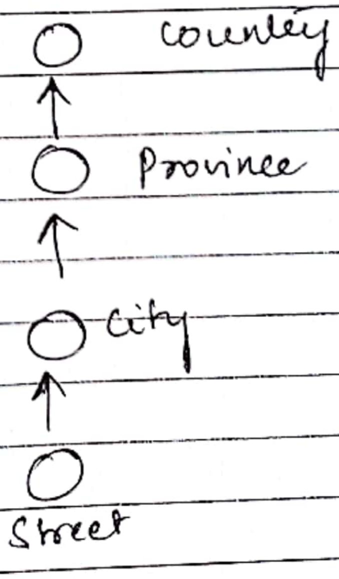
Eg:- Consider a concept hierarchy for the dimension location. City values for location include Vancouver, Toronto, NY, and chicago. Each city can be mapped to the province/state to which it belongs. The provinces and states can in turn be mapped to the country. These mappings form a concept of hierarchy for the dimension location, mapping a set of low-level concepts to higher-level, more general concepts. ~~to~~, This concept is illustrated as follows.



→ A concept hierarchy i,e ~~total~~/ partial order among attributes in a db schema is called schema hierarchy.

→ Lattice structure ~~A concept~~





→ concept hierarchies may also be defined by discretizing / grouping values for a given dimension/ attribute, resulting in a set-grouping hierarchy.