Assignment 1 Set A

Course Code :- CAP446

Date of Submission :- 27-9-21

Name :- Sirjanpreet Kaur

Roll no. :- B56

Section :- D2112

Registration no. :- 12107974

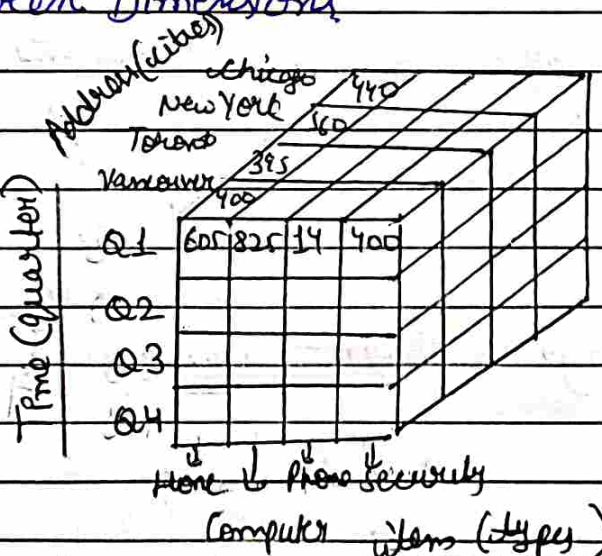
Assignment 1 Set-A

Ques:- What is a data cube? Explain its various operations by taking a real-life example or a case study.

Ans:- A data cube is a multidimensional data model that stores the optimized, summarized or aggregated data which is defined by dimensions and facts. Data cube stores the precomputed data and eases online analytical processing. Data stored in data cube is represented in dimensions and facts.

Dimensions are entities with respect to which organization wants to keep records.

Facts are numerical measures. It is the quantities by which we want to analyze relationships between Dimensions.

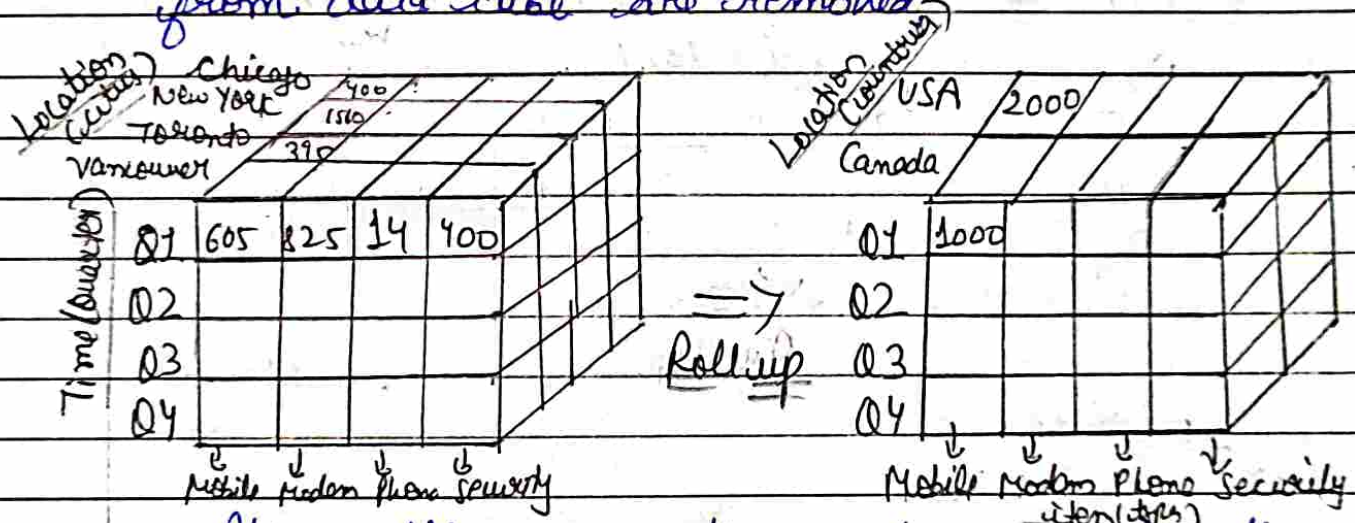


In the above data cube, we can store the sales data in many dimensions like sales at all location, sales of all time.

Operations on Data Cube

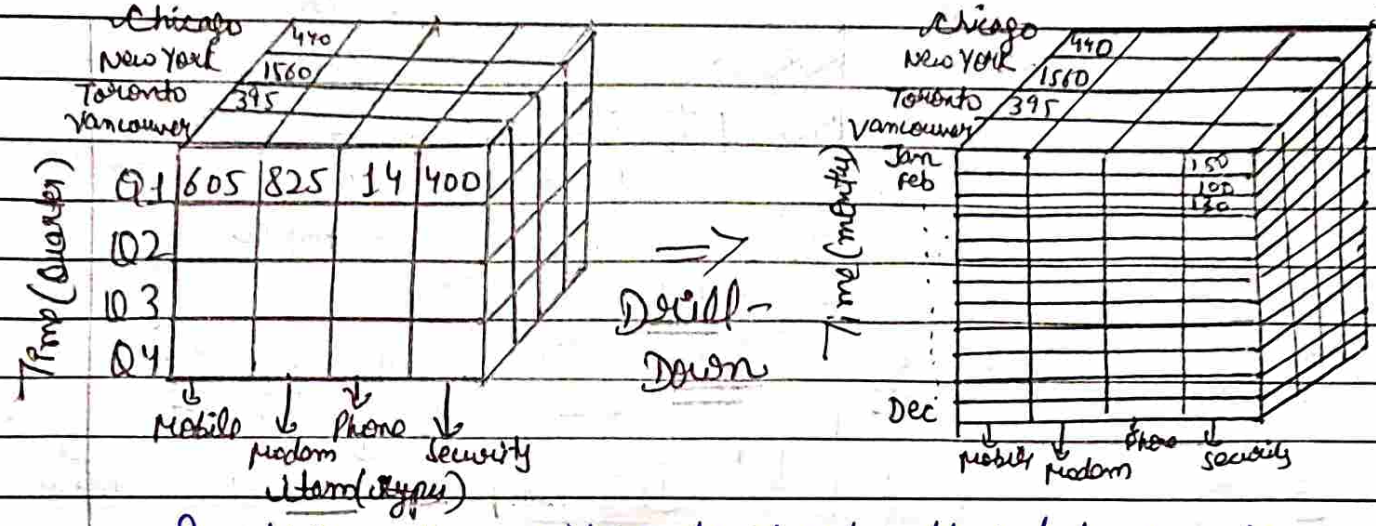
The operations can be conducted in order to view data from different angles. There are four operations that can be implemented are as follows

1. **Roll up:-** Roll-up operation summarizes or aggregates the dimensions. Roll up is performed by climbing up a concept hierarchy for dimension location. When roll-up operation is performed then one or more dimensions from data cube are removed



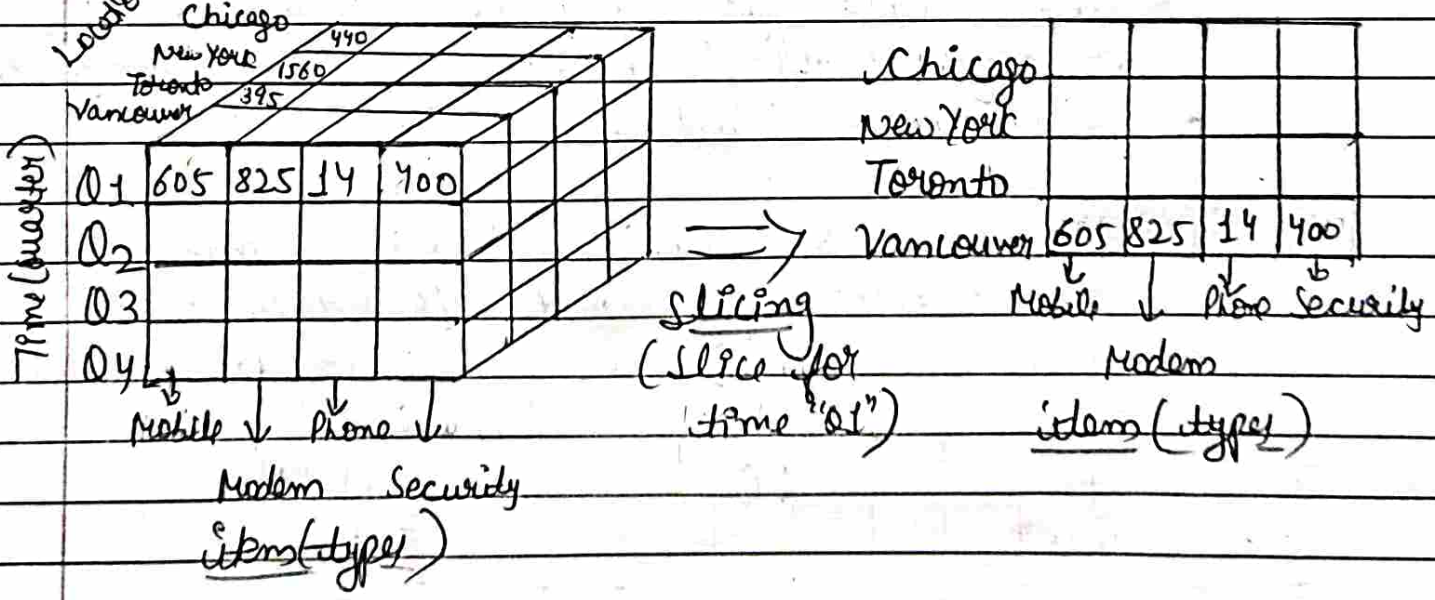
After rolling up, the quarters remains the same but the cities summed up in countries. The cities comes in USA are summed up in USA and the cities comes in Canada are summed up in Canada. So it is aggregated into countries.

2. **Drill-down:-** Drill-down is the reverse operation of roll-up. In this we divide data attributes into sub-data attributes. It navigate the data from less detailed data to highly detailed data. One or more dimensions are added into data cube



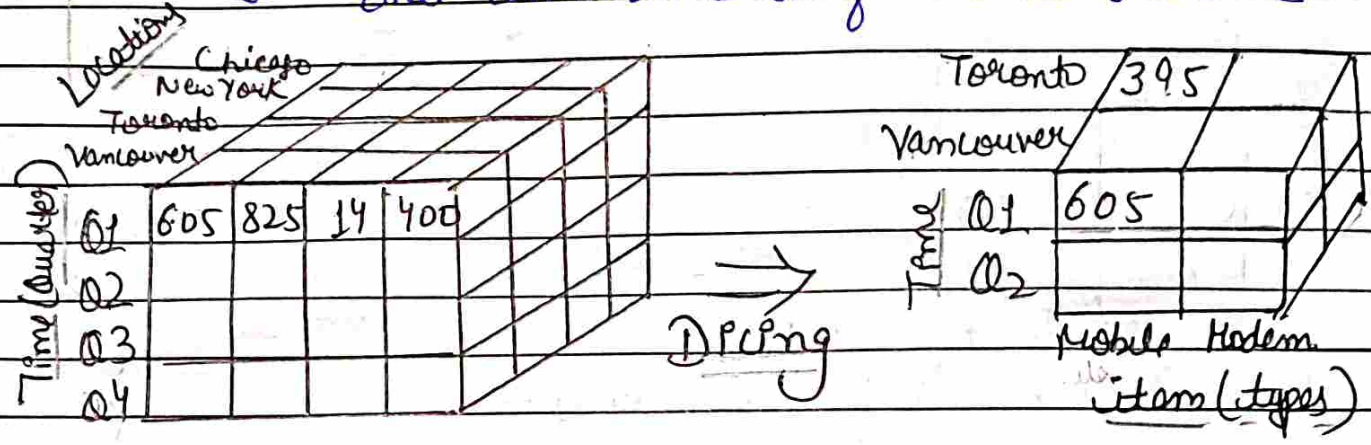
In this, the cities don't get affected by quarters are affected. The data is divided with respect to months. This gives clearity about what is happening in the particular month in the particular city.

3. **Slicing** :- The slice Operation selects one particular dimension from a cube and provides a new sub-cube



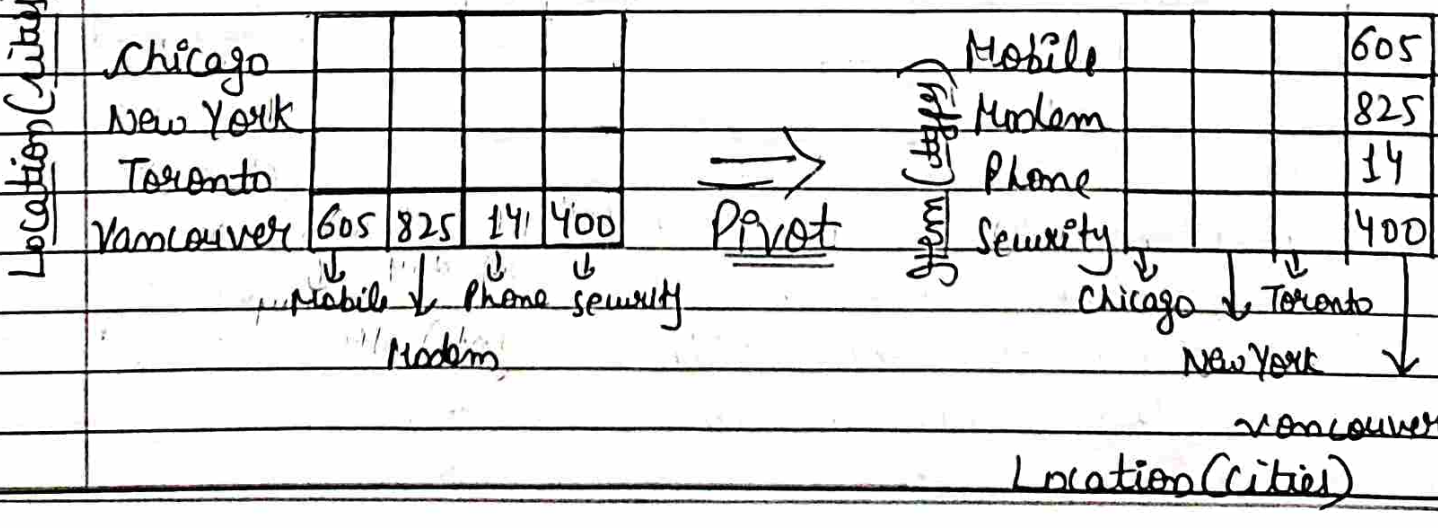
In this, we want to slice data only for "Q1". The data shown in above figure is for Q1 for all the four cities separately. It basically form new subcube by selecting one or more dimension

4. **Dicing:-** The dice operation select more than one dimension to form a subcube



In this, we divide data into three dimensions. In this, we dice only for location Toronto and Vancouver and dice only for "Q1" and "Q2" and we want item only mobile or modem. Now, we can only see data for Toronto, vancouver for "Q1" and "Q2", and the product shown is Mobile or Modem. It selects criteria based on three dimensions.

5. **Pivot:-** Pivot operation is also known as Rotation. It rotates the data cube in order to view data cube from different dimensions



Ques 2:- What can be the major issues in the process of data mining? Support your answer with relevant examples.

Ans: Data Mining is the process of extracting knowledgeable data from Database. However, Data mining is dynamic and fast-expanding field, but it faces some major issues in research of Data Mining. The major issues are given below:-

1. Mining Methodology and User Interaction Issues:-

(i) Mining different kinds of knowledge that means different clients want different kind of information so it is difficult to extract different kinds of data that can meet the requirements of client.

(ii) Interactive mining of knowledge at multiple levels of abstraction because interactive mining allows users to focus the search patterns from different angles that is difficult to know.

(iii) Background knowledge is required to guide discovery process.

(iv) Handling Incomplete data in large Databases is difficult. Because most of data is incorrect that can be due to human error or failure of any instrument.

2. Performance Issues:-

(i) Efficiency and scalability:- To extract information from huge amount

of data, data mining algorithms must be efficient and scalable

(ii) Parallel, distributed, and incremental mining algorithms:- These algorithms partition the data into pieces and each piece is processed in parallel. The patterns from each partition are merged.

3. Diversity of DataBase Types:- The wide diversity of database types brings about challenges to data mining. These include:

(i) Handling complex types of Data:- There are many kinds of data stored in databases and data warehouses. It is not possible for one system to mine all these kinds of data. So different data mining techniques should be constructed for different kinds of data.

(ii) Mining dynamic, networked, and global data repositories:- Many sources of data are connected by various kinds of network and forming distributed, heterogeneous global information systems and networks. The discovery of knowledge from different sources of structured, unstructured data poses great challenges to data mining

4. Data Mining and Society:-

(i) Social impacts of data mining:- It is important to study the impact of data mining on society. The use of data and data protection are need to be

concerned.

(ii) Privacy-preserving data mining:- Data mining help in scientific discovery, security protection, economy recovery. But it poses the risk of disclosing an individual's personal information. So it needs to be observe data sensitivity and preserve people's privacy while performing data mining.

(iii) Invisible data mining:- We cannot expect everyone in society to have knowledge about data mining techniques. Many systems have in-built data mining functions, one can simply use data mining results by clicking on mouse etc. This is done without any knowledge to user. For example, while purchasing items online, user may be unaware of the buying patterns of the customer that is collecting the data by store, that may be used to recommend other items for purchase in the future.

Ques 3:- How are interesting patterns detected from a dataset? Illustrate an example to support your answer.

Ans:- Various kinds of knowledge can be mined from dataset. Not all patterns are interesting. Only a small fraction of patterns would be interesting to any user. A pattern is interesting only if it is:

1. easily understood by humans
2. valid on new data with certainty
3. potentially useful
4. novel

An Interesting pattern is the one that represents knowledge. The first step to detect interesting pattern in data mining is.

1. By Organize and evaluate the data :- By organizing and evaluating the data, you will be able to find the relevant data. By prioritizing the data, we will be able to separate the noisy data from the data set and relevant information is provided to us. The ability to evaluate and analyze our large data sets is challenging but it will provide advantage to our business

2. By Discovering Frequent Itemsets :- Discovering frequent patterns designed to be applied on a transactional database in transaction mode by customers in stores. A transaction is defined a set of distinct items. For example, in transaction database, there are four transactions :- the items are "bread, butter", "bread, milk", "bread", "milk" and "butter".

T1 : bread, butter, spinach
T2 : butter, salmon
T3 : bread, milk, butter
T4 : bread, bread, milk

As we see, milk and butter is common. So the store put milk, butter alongside so that people can buy milk along with bread. This is market strategy.

3. Discovering Sequential Rules:- A sequence database contains some sequences. For example

ID	Sequences
seq 1	$\langle \{a,b\}, \{c\}, \{f\}, \{g\}, \{e\} \rangle$
seq 2	$\langle \{a,d\}, \{c\}, \{b\}, \{a,b,e,f\} \rangle$
seq 3	$\langle \{a\}, \{b\}, \{f\}, \{e\} \rangle$
seq 4	$\langle \{b\}, \{f,g,e\} \rangle$

We assume that each sequence represent what customer has bought in supermarket. In seq 2, indicates that second customer bought items "a" and "d" together, then bought item item "c", then bought "b", and then bought "a", "b", "e" and "f" together. So the shopkeeper put items together so that customer get another item that is nearby the product that he/she is buying.

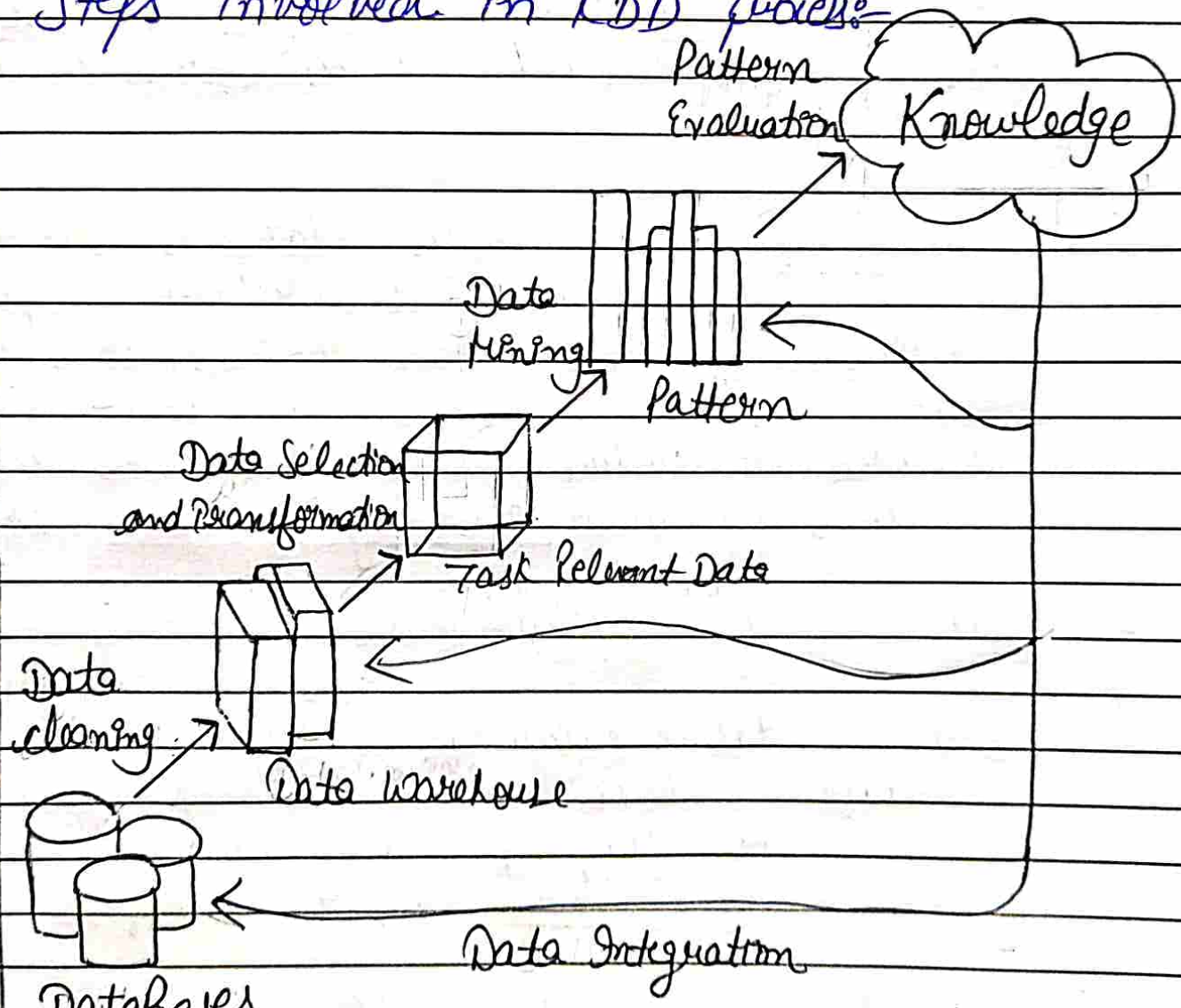
4. Implementations of pattern mining algorithms:- By applying various mining algorithms, we can detect various interesting patterns from dataset. It has wide range of applications.

So, these are the ways that how can we detect interesting patterns from a dataset

Ques 4: What is KDD process? State each step with one example.

Ans:- KDD is Knowledge Discovery from Data. It refers to the procedure of discovering knowledge in Data and implement Data Mining techniques so that particular information can be extracted from large data. KDD is knowledge extraction from data

Steps Involved in KDD process:-



b Data Cleaning:- In this, redundant data can be eliminated. Data cleaning is removing of noisy and irrelevant data from Data Warehouse. We remove unwanted data.

Data that is missing is eliminated. Data can be cleaned by Data discrepancy detection and Data transformation tools.

2. Data Integration:- In this, data from heterogenous sources are collected in a combined source. Issues may arise in Data Integration while combining data in Data warehouse. To solve this issue, correlation analysis is used. Once the issue is resolved, data is stored in Data warehouse.

3. Data Selection:- In this, data is selected from Data warehouse to produce knowledgeable information. In this, it is decided that what data is selected and what is not selected to produce relevant data. Data selection is done by using Clustering, Regression, etc.

4. Data Transformation:- Once the data is selected, then data is to be transform. Data in datawarehouse is in different forms, so data is to be transformed into appropriate form. Data Mapping and code generation are techniques are used to transform data.

5. Data Mining:- In this, intelligent methods are applied to extract data patterns. In this interesting patterns are defined, relations are defined and we come out with Interesting patterns. Classification or characterization is used for Data Mining.

6. Pattern Evaluation :- Patterns are evaluated to produce the meaningful data. It is defined as identifying strictly increasing patterns to represent knowledge based on given measures. Summarization and Visualization is used to make data understandable by user.

7. Knowledge Representation :- After Pattern Evaluation, it is need to represent knowledge. Knowledge Representation is representing of data after various steps such as from Data cleaning to Pattern Evaluation. The knowledge that is extracted is used to generate reports, tables, charts, graphs etc.

In this whole process, we firstly extract data from the Data Source. Then the extracted data is cleaned by using various tools. After that, the data is integrated from various sources and combined in Data Warehouse. Then Data Selection is performed to select the relevant data from the Data Warehouse. The selected data is then transformed into appropriate form required by mining procedure. Data Mining transforms relevant data into patterns. To understand the patterns, pattern evaluation is defined that is used to represent data by using various tools. Hence the knowledge is extracted that is used to generate charts, graphs, reports. So, it is Knowledge Discovery from the Data.

Ques:- How is an OLAP different and complex in working than OLTP? Justify your answer with an example

Ans:- OLAP is different and complex in working than OLTP in the following ways

	OLAP	OLTP
1.	OLAP as online analytical processing consist of a type of software that is used for data analysis for business decisions	OLTP is online transaction processing that provides transaction-oriented applications in three-tier architecture
2.	OLAP is a category of software technology that enables analysts, managers to analyze the complex data derived from data warehouse	OLTP perform online transactions and query processing. It covers day-to-day operations of an organization
3.	OLAP is used for analyzing data. The people who use OLAP must have knowledge about data warehouse	OLTP is accessed by users, clerks, client and IT professionals. OLTP is customer-oriented
4.	OLAP have historic data and can't be modified	OLTP have current data and can be changed

	OLAP	OLTP
5.	OLAP's main operation is to extract multi-dimensional data for analysis	OLTP's main operation is to insert, delete and update data in databases
6.	OLAP's has long and less frequent transactions	OLTP has short but frequent transactions.
7.	The processing time for OLAP is more than OLTP.	The processing time for OLTP is less than OLAP.
8.	The queries of OLAP are more complex than OLTP.	The queries of OLTP are less complex than OLAP
9.	In OLAP, the transaction is less frequent and it does not bother much about data integrity.	In OLTP, the transactions are frequent, if it fails in middle, it may harm data integrity.
10.	In OLAP, only read only queries are there, they operate on huge volumes of data and the queries are complex	In OLTP, there is short and atomic transactions It requires recovery mechanism and concurrence control.

	OLAP	OLTP
11.	OLAP allows hundreds of users	Databases allows thousands of users
12.	Query throughput is the performance metric	Transaction throughput is the performance metric
13.	OLAP creates a single platform for all type of business analytical needs which includes planning, budgeting, and analysis	It administers daily transactions of an organizations
14.	Example: (i) Recommendation system like cookies. When we search the data on any topic, then the previous page that we visited is come up first. (ii) Videos recommendation on youtube. YouTubes gives recommendation about videos that we often watch.	Example (i) Banking – ATM. ATM is example of OLTP. It is based on ACID properties. (ii) OLTP is used for online banking, sending text message, online shopping, add book to cart etc.