LEARNING UNIT I: DATA WAREHOUSING

Introduction to Data Warehouse:

A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision-making process.

Subject-Oriented: A data warehouse can be used to analyze a particular subject area. For example, "sales" can be a particular subject.

Integrated: A data warehouse integrates data from multiple data sources. For example, source A and source B may have different ways of identifying a product, but in a data warehouse, there will be only a single way of identifying a product.

Time-Variant: Historical data is kept in a data warehouse. For example, one can retrieve data from 3 months, 6 months, 12 months, or even older data from a data warehouse. This contrasts with a transactions system, where often only the most recent data is kept. For example, a transaction system may hold the most recent address of a customer, where a data warehouse can hold all addresses associated with a customer.

Non-volatile: Once data is in the data warehouse, it will not change. So, historical data in a data warehouse should never be altered.

Data Warehouse Design Process:

A data warehouse can be built using a top-down approach, a bottom-up approach, or a combination of both.

- The top-down approach starts with the overall design and planning. It is useful in cases where the technology is mature and well known, and where the business problems that must be solved are clear and well understood.
- The bottom-up approach starts with experiments and prototypes. This is useful in the early stage of business modeling and technology development. It allows an organization to move forward at considerably less expense and to evaluate the benefits of the technology before making significant commitments.
- In the combined approach, an organization can exploit the planned and strategic nature of the top-down approach while retaining the rapid implementation and opportunistic application of the bottom-up approach.

The warehouse design process consists of the following steps:

- Choose a business process to model, for example, orders, invoices, shipments, inventory,

account administration, sales, or the general ledger. If the business process is organizational and involves multiple complex object collections, a data warehouse model should be followed. However, if the process is departmental and focuses on the analysis of one kind of business process, a data mart model should be chosen.

- Choose the grain of the business process. The grain is the fundamental, atomic level of data to be represented in the fact table for this process, for example, individual transactions, individual daily snapshots, and so on.

- Choose the dimensions that will apply to each fact table record. Typical dimensions are time, item, customer, supplier, warehouse, transaction type, and status.
- Choose the measures that will populate each fact table record. Typical measures are numeric additive quantities like dollars sold and units sold.

DATA WAREHOUSE ARCHITECTURE:

Data warehouse is a subject oriented, integrated, time-variant, and non-volatile collection of data. This data helps analysts to take informed decisions in an organization.

Steps for the Design and Construction of Data Warehouse:

This subsection presents a business analysis framework for data warehouse design. The basic steps involved in the design process are also described.

The Design of a Data Warehouse:

A Business Analysis Framework Four different views regarding the design of a data warehouse must be considered:

- the top-down view,
- the data source view,
- the data warehouse view,
- the business query view.

The top-down view allows the selection of relevant information necessary for the data warehouse. The data source view exposes the information being captured, stored and managed by operational systems. The data warehouse view includes fact tables and dimension tables

Finally, the business query view is the Perspective of data in the data warehouse from the viewpoint of the end user.

THREE-TIER DATA WAREHOUSE ARCHITECTURE:

The bottom tier is ware-house database server which is almost always a relational database system.

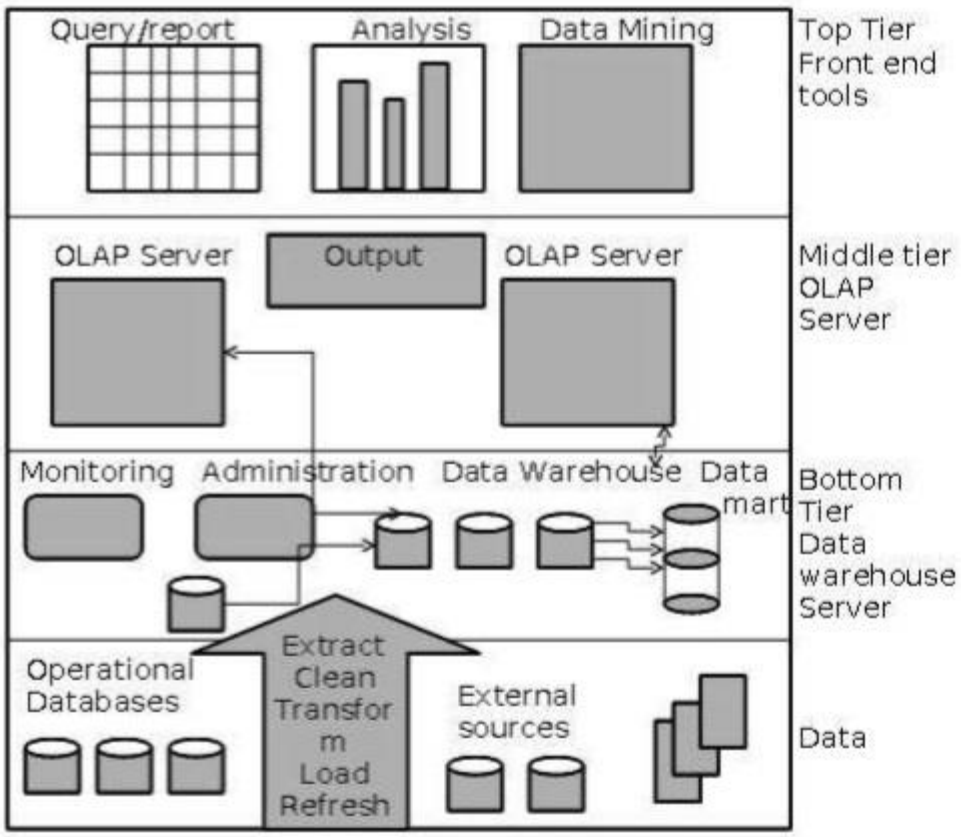
The middle tier is an OLAP server which is typically implemented using either

a Relational OLAP (ROLAP) model.

a Multidimensional OLAP (MOLAP) model.

The top tier is a client, which contains query and reporting tools, analysis tools, and/or data mining tools (e.g., trend analysis, prediction, and so on).

Three-tier Data warehouse architecture



From the architecture point of view, there are three data warehouse models:

- the enterprise warehouse.
- the data mart
- the virtual warehouse

Enterprise warehouse:

An enterprise warehouse collects all of the information about subjects spanning the entire organization.

It provides corporate-wide data integration, usually from one or more operational systems or external information providers, and is cross-functional in scope.

It typically contains detailed data as well as summarized data, 13 and can range in size from a few gigabytes to hundreds of gigabytes, terabytes, or beyond.

Data mart:

A data mart contains a subset of corporate-wide data that is of value to a specific group of users.

The scope is connected to specific, selected subjects.

For example: a marketing data mart may connect its subjects to customer, item, and sales.

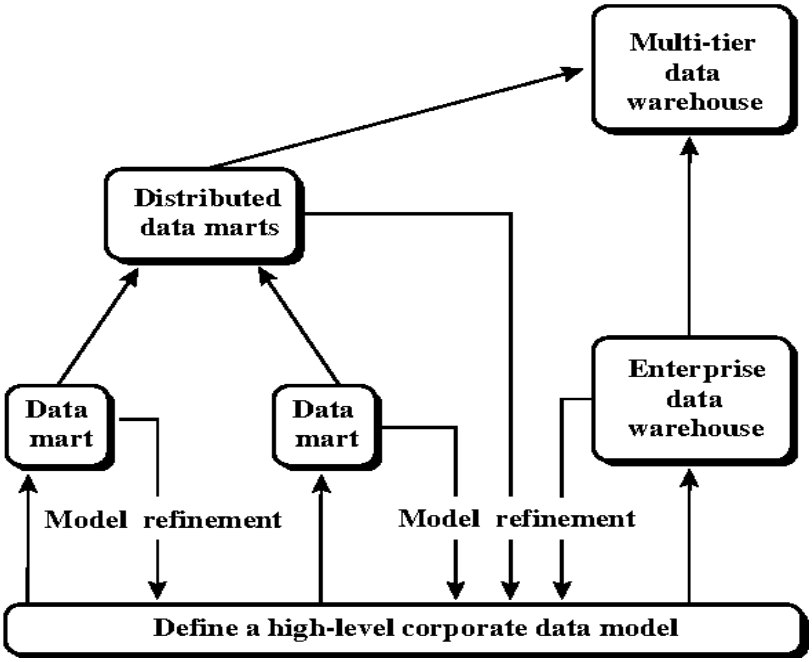
The data contained in data marts tend to be summarized. Depending on the source of data, data marts can be categorized into the following two classes:

(i) **Independent data marts** are sourced from data captured from one or more operational systems or external information providers, or from data generated locally within a particular department or geographic area.

(ii) **Dependent data marts** are sourced directly from enterprise data warehouses.

Virtual warehouse:

A virtual warehouse is a set of views over operational databases. For efficient query processing, only some of the possible summary views may be materialized.
A virtual warehouse is easy to build but requires excess capacity on operational database servers.
Figure: A recommended approach for data warehouse development. Data warehouse Back-End Tools and Utilities the ETL (Extract Transformation Load) process



OVERVIEW OF MULTIDIMENSIONAL DATA MODEL:

Multidimensional data model stores data in the form of data cube. Mostly, data warehousing supports two or three-dimensional cubes.
A data cube allows data to be viewed in multiple dimensions.
The dimensions are entities with respect to which an organization wants to keep records.
-**For example,** in store sales record, dimensions allow the store to keep track of things like monthly sales of items and the branches and locations.
A multidimensional database helps to provide data-related answers to complex business queries quickly and accurately.

Data warehouses and Online Analytical Processing (OLAP) tools are based on a multidimensional data model.

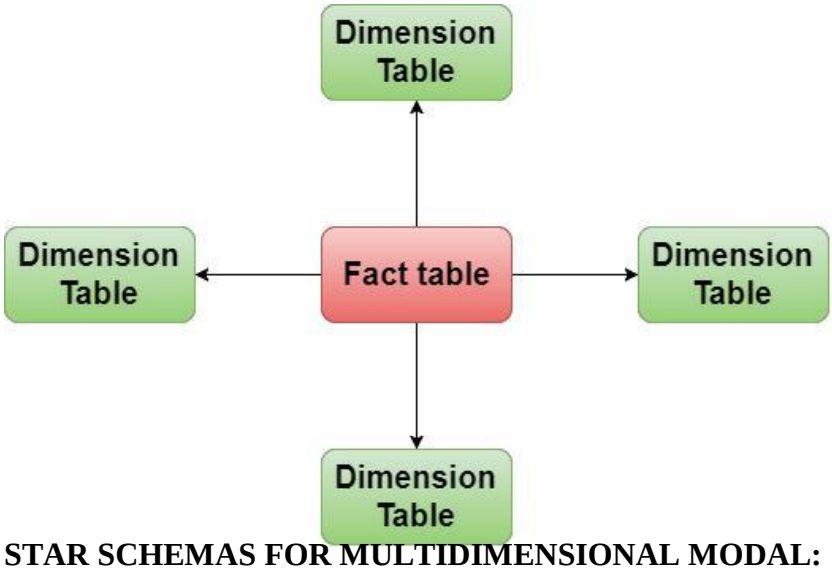
OLAP in data warehousing enables users to view data from different angles and dimensions.

Schemas for Multidimensional Data Model:-

- Star Schema
- Snowflakes Schema
- Fact Constellations Schema

Star Schemas for Multidimensional Model:

The simplest data warehouse schema is star schema because its structure resembles a star.
Star schema consists of data in the form of facts and dimensions.
The fact table present in the center of star and points of the star are the dimension tables.
In star schema fact table contain a large amount of data, with no redundancy.
Each dimension table is joined with the fact table using a primary or foreign key.



STAR SCHEMAS FOR MULTIDIMENSIONAL MODAL:

Fact Tables:

A fact table has two types of columns: one column of foreign keys (pointing to the dimension tables) and other of numeric values.

Fact Table	
PK	id Dimension Table
PK	id Dimension Table
PK	id Dimension Table

Dimension Tables:

Dimension table is generally small in size as compared to a fact table. The primary key of a dimension table is a foreign key in a fact table.

Example of Dimension Tables:-

Time dimension table
Product dimension table
Employee dimension table
Geography dimension table
The main characteristics of star schema are that it is easy to understand and small number of tables can join.

SNOWFLAKE SCHEMAS FOR MULTIDIMENSIONAL MODEL:

The snowflake schema is a more complex than star schema because dimension tables of the snowflake are normalized.
The snowflake schema is represented by centralized fact table which is connected to multiple dimension table and this dimension table can be normalized into additional dimension tables.

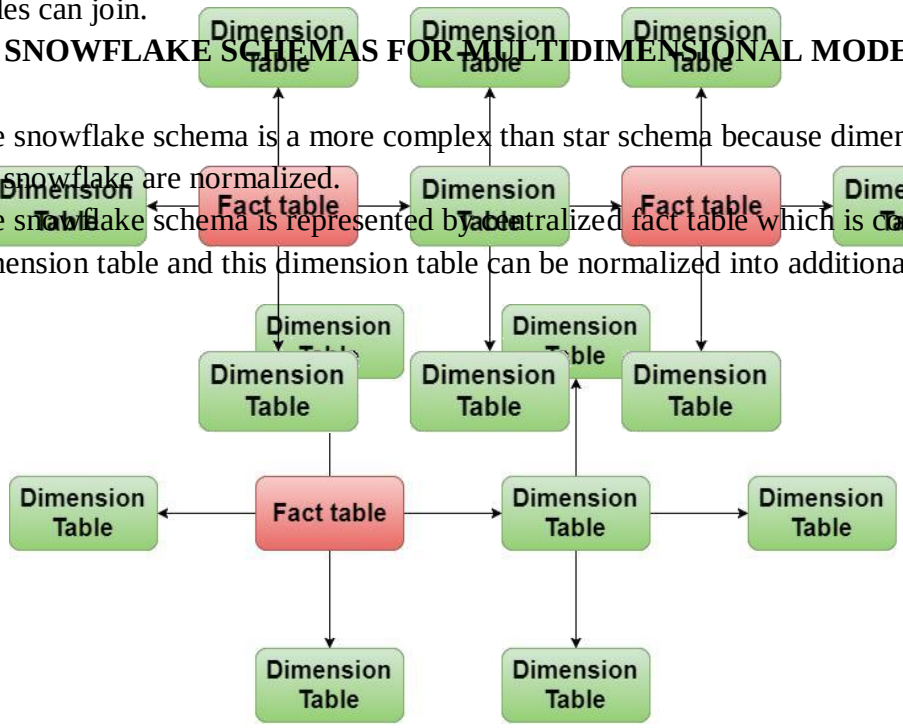


Fig: Snowflake Schemas for Multidimensional Model

The major difference between the snowflake and star schema models is that the dimension tables of the snowflake model are normalized to reduce redundancies.

FACT CONSTELLATION SCHEMAS FOR MULTIDIMENSIONAL MODEL:

A fact constellation can have multiple fact tables that share many dimension tables.
This type of schema can be viewed as a collection of stars, Snowflak and hence is called a galaxy schema or a fact constellation.

Fact constellation Schemas for Multidimensional Modal:

The main disadvantage of fact constellation schemas is its more complicated design.

Meta Data Repository:

Metadata are data about data. When used in a data warehouse, metadata are the data that define warehouse objects. Metadata are created for the data names and definitions of the given warehouse. Additional metadata are created and captured for time stamping any extracted data, the source of the extracted data, and missing fields that have been added by data cleaning or integration processes.
A metadata repository should contain the following:

- A description of the structure of the data warehouse, which includes the warehouse schema, view, dimensions, hierarchies, and derived data definitions, as well as data mart locations and contents.
- Operational metadata, which include data lineage (history of migrated data and the sequence of transformations applied to it), currency of data (active, archived, or purged), and monitoring information (warehouse usage statistics, error reports, and audit trails).
- The algorithms used for summarization, which include measure and dimension definition algorithms, data on granularity, partitions, subject areas, aggregation, summarization, and predefined queries and reports.
- The mapping from the operational environment to the data warehouse, which includes source databases and their contents, gateway descriptions, data partitions, data extraction, cleaning, transformation rules and defaults, data refresh and purging rules, and security (user authorization and access control).
- Data related to system performance, which include indices and profiles that improve data access and retrieval performance, in addition to rules for the timing and scheduling of refresh, update, and replication cycles.
- Business metadata, which include business terms and definitions, data ownership information, and charging policies.

Schema Design:

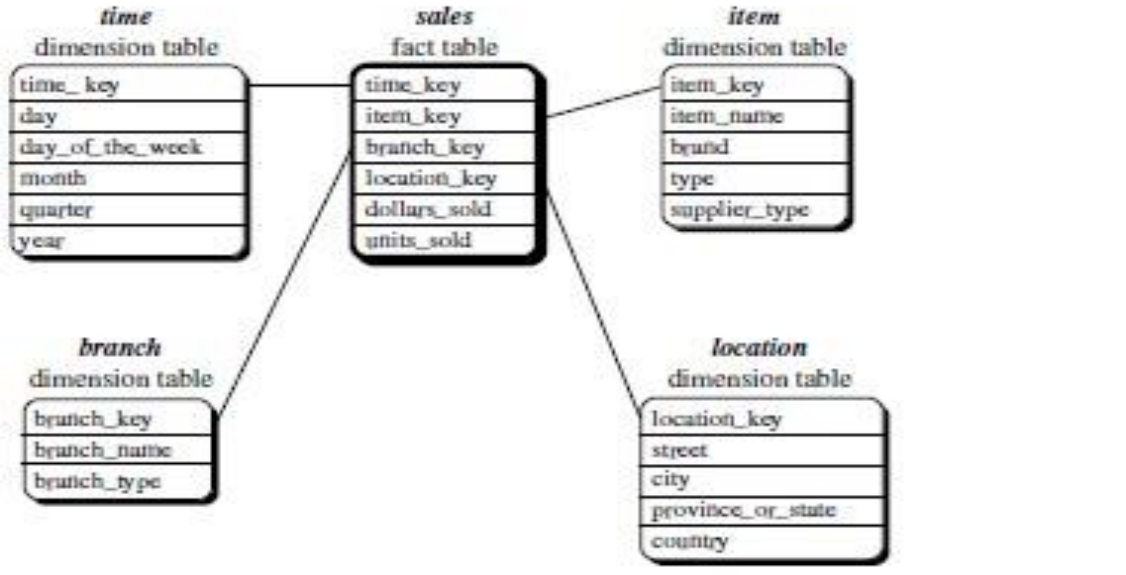
Stars, Snowflakes, and Fact Constellations: Schemas for Multidimensional Databases The entity relationship data model is commonly used in the design of relational databases, where a database schema consists of a set of entities and the relationships between them. Such a data model is appropriate for online transaction processing. A data warehouse, however, requires a concise, subject-oriented schema that facilitates on-line data analysis. The most popular data model for a data warehouse is a multidimensional model. Such a model can exist in the form of

a star schema, a snowflake schema, or a fact constellation schema. Let's look at each of these schema types. Star schema: The most common modeling paradigm is the star schema, in which the data warehouse contains (1) a large central table (fact table) containing the bulk of the data, with no redundancy, and (2) a set of smaller attendant tables (dimension tables), one for each dimension. The schema graph resembles a starburst, with the dimension tables displayed in a radial pattern around the central fact table.

Star schema:

A star schema for All Electronics sales is shown in Figure. Sales are considered along four dimensions, namely, time, item, branch, and location. The schema contains a central fact table for sales that contains keys to each of the four dimensions, along with two measures: dollars sold and units sold. To minimize the size of the fact table, dimension identifiers (such as time key and item key) are system-generated identifiers. Notice that in the star schema, each dimension is represented by only one table, and each table contains a set of attributes. For example, the location dimension table contains the attribute set {location key, street, city, province or state, country}. This constraint may introduce some redundancy.

For example, "Vancouver" and "Victoria" are both cities in the Canadian province of British Columbia. Entries for such cities in the location dimension table will create redundancy among the attributes province or state and country, that is, (..., Vancouver, British Columbia, Canada) and (..., Victoria, British Columbia, Canada). Moreover, the attributes within a dimension table may form either a hierarchy (total order) or a lattice (partial order).



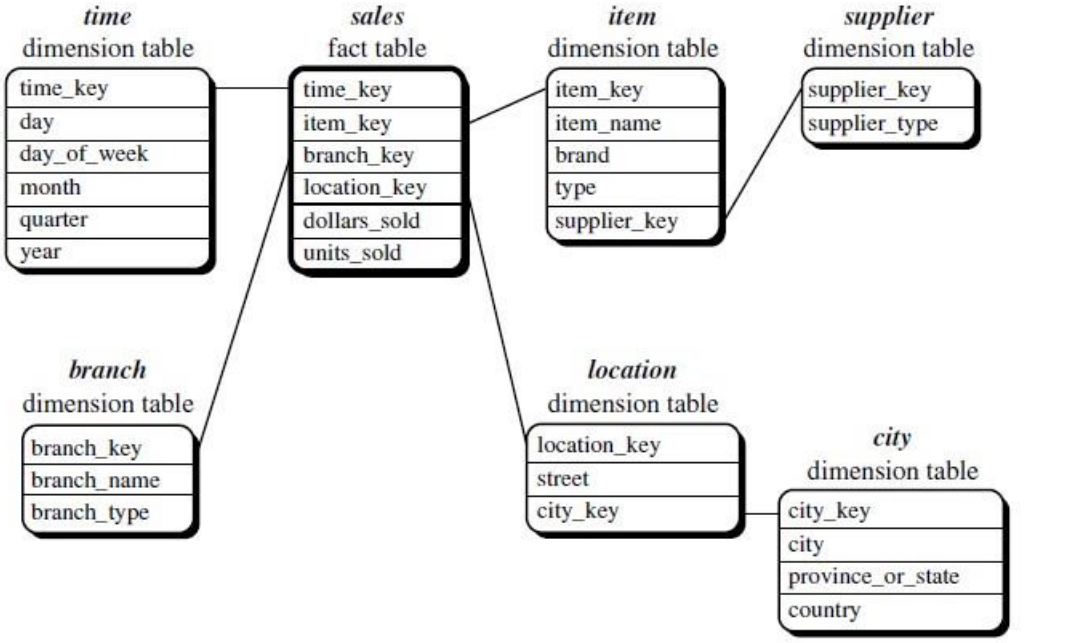
Star schema of a data warehouse for sales.

Snowflake schema.:

A snowflake schema for All Electronics sales is given in Figure Here, the sales fact table is identical to that of the star schema in Figure. The main difference between the two schemas is in the definition of dimension tables.

The single dimension table for item in the star schema is normalized in the snowflake schema, resulting in new item and supplier tables. For example, the item dimension table now

contains the attributes item key, item name, brand, type, and supplier key, where supplier key is linked to the supplier dimension table, containing supplier key and supplier type information. Similarly, the single dimension table for location in the star schema can be normalized into two new tables: location and city. The city key in the new location table links to the city dimension.
Notice that further normalization can be performed on province or state and country in the snowflake schema

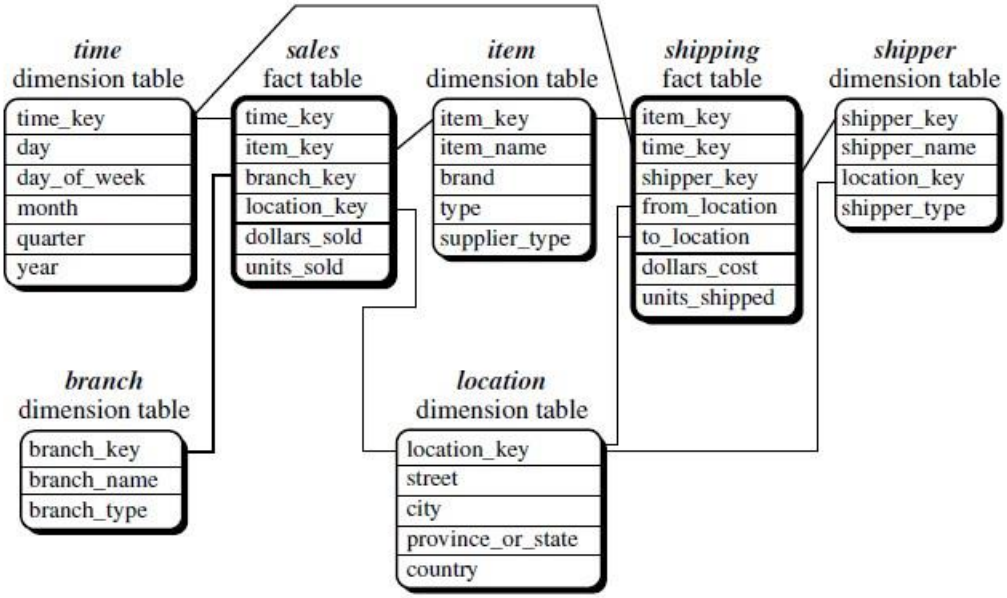


Snowflake schema of a data warehouse for sales.

Fact constellation.

A fact constellation schema is shown in Figure. This 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, and two measures: dollars cost and units shipped.
A fact constellation schema allows dimension tables to be shared between fact tables. For example, the dimensions tables for time, item, and location are shared between both the sales and shipping fact tables.
In data warehousing, there is a distinction between a data warehouse and a data mart.

A data warehouse collects information about subjects that span the entire organization, such as customers, items, sales, assets, and personnel, and thus its scope is enterprise-wide. For data warehouses, the fact constellation schema is commonly used, since it can model multiple, interrelated subjects. A data mart, on the other hand, is a department subset of the data warehouse that focuses on selected subjects, and thus its scope is department wide. For data marts, the star or snowflake schema are commonly used, since both are geared toward modeling single subjects, although the star schema is more popular and efficient.



5 Fact constellation schema of a data warehouse for sales and shipping.

Meta Data Repository:

Metadata are data about data. When used in a data warehouse, metadata are the data that define warehouse objects. Metadata are created for the data names and definitions of the given warehouse. Additional metadata are created and captured for time stamping any extracted data, the source of the extracted data, and missing fields that have been added by data cleaning or integration processes.

A metadata repository should contain the following:

- A description of the structure of the data warehouse, which includes the warehouse schema, view, dimensions, hierarchies, and derived data definitions, as well as data mart locations and contents.
- Operational metadata, which include data lineage (history of migrated data and the sequence of transformations applied to it), currency of data (active, archived, or purged), and monitoring information (warehouse usage statistics, error reports, and audit trails).
- The algorithms used for summarization, which include measure and dimension definition algorithms, data on granularity, partitions, subject areas, aggregation, summarization, and predefined queries and reports.
- The mapping from the operational environment to the data warehouse, which includes source databases and their contents, gateway descriptions, data partitions, data extraction, cleaning, transformation rules and defaults, data refresh and purging rules, and security (user authorization and access control).
- Data related to system performance, which include indices and profiles that improve data access and retrieval performance, in addition to rules for the timing and scheduling of refresh, update, and replication cycles.

- Business metadata, which include business terms and definitions, data ownership information, and charging policies.

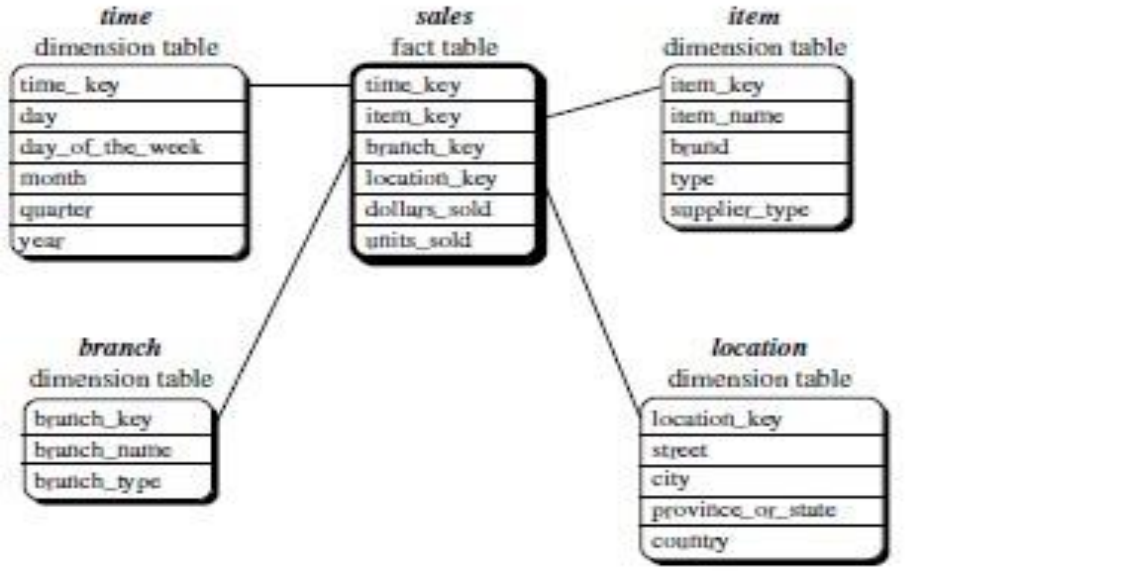
Schema Design:

Stars, Snowflakes, and Fact Constellations: Schemas for Multidimensional Databases The entity relationship data model is commonly used in the design of relational databases, where a database schema consists of a set of entities and the relationships between them. Such a data model is appropriate for online transaction processing. A data warehouse, however, requires a concise, subject-oriented schema that facilitates on-line data analysis. The most popular data model for a data warehouse is a multidimensional model. Such a model can exist in the form of a star schema, a snowflake schema, or a fact constellation schema. Let's look at each of these schema types. Star schema: The most common modeling paradigm is the star schema, in which the data warehouse contains (1) a large central table (fact table) containing the bulk of the data, with no redundancy, and (2) a set of smaller attendant tables (dimension tables), one for each dimension. The schema graph resembles a starburst, with the dimension tables displayed in a radial pattern around the central fact table.

Star schema:

A star schema for All Electronics sales is shown in Figure. Sales are considered along four dimensions, namely, time, item, branch, and location. The schema contains a central fact table for sales that contains keys to each of the four dimensions, along with two measures: dollars sold and units sold. To minimize the size of the fact table, dimension identifiers (such as time key and item key) are system-generated identifiers. Notice that in the star schema, each dimension is represented by only one table, and each table contains a set of attributes. For example, the location dimension table contains the attribute set {location key, street, city, province or state, country}. This constraint may introduce some redundancy.

For example, "Vancouver" and "Victoria" are both cities in the Canadian province of British Columbia. Entries for such cities in the location dimension table will create redundancy among the attributes province or state and country, that is, (..., Vancouver, British Columbia, Canada) and (..., Victoria, British Columbia, Canada). Moreover, the attributes within a dimension table may form either a hierarchy (total order) or a lattice (partial order).



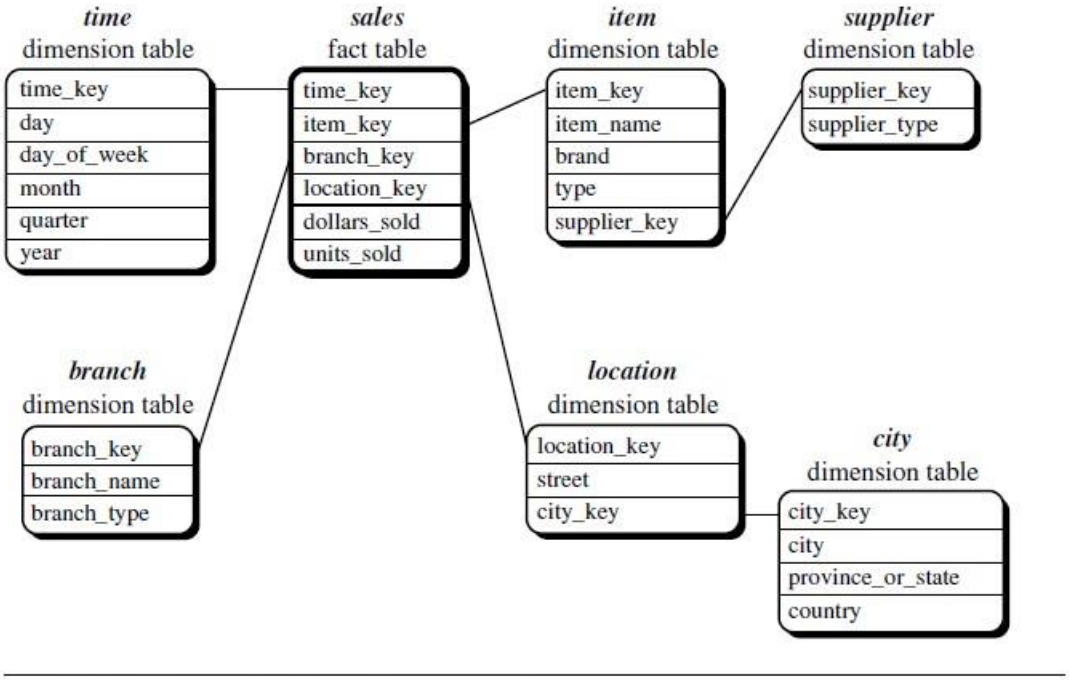
Star schema of a data warehouse for sales.

Snowflake schema.:

A snowflake schema for All Electronics sales is given in Figure Here, the sales fact table is identical to that of the star schema in Figure. The main difference between the two schemas is in the definition of dimension tables.

The single dimension table for item in the star schema is normalized in the snowflake schema, resulting in new item and supplier tables. For example, the item dimension table now contains the attributes item key, item name, brand, type, and supplier key, where supplier key is linked to the supplier dimension table, containing supplier key and supplier type information. Similarly, the single dimension table for location in the star schema can be normalized into two new tables: location and city. The city key in the new location table links to the city dimension.

Notice that further normalization can be performed on province or state and country in the snowflake schema

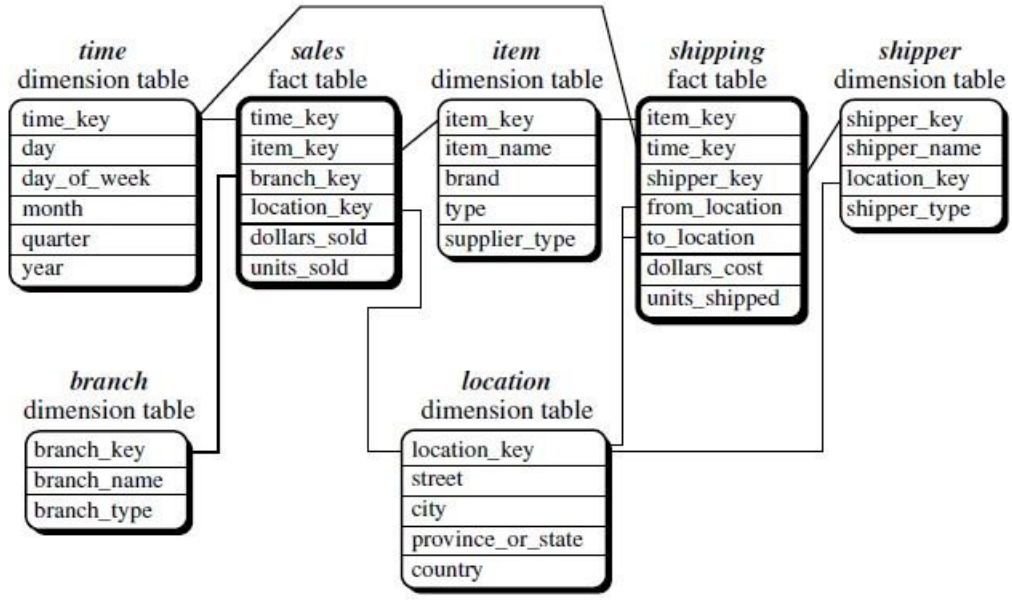


Snowflake schema of a data warehouse for sales.

Fact constellation.

A fact constellation schema is shown in Figure. This 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, and two measures: dollars cost and units shipped.
A fact constellation schema allows dimension tables to be shared between fact tables. For example, the dimensions tables for time, item, and location are shared between both the sales and shipping fact tables.
In data warehousing, there is a distinction between a data warehouse and a data mart.

A data warehouse collects information about subjects that span the entire organization, such as customers, items, sales, assets, and personnel, and thus its scope is enterprise-wide. For data warehouses, the fact constellation schema is commonly used, since it can model multiple, interrelated subjects. A data mart, on the other hand, is a department subset of the data warehouse that focuses on selected subjects, and thus its scope is department wide. For data marts, the star or snowflake schema are commonly used, since both are geared toward modeling single subjects, although the star schema is more popular and efficient.



5 Fact constellation schema of a data warehouse for sales and shipping.

Measures: Their Categorization and Computation:

"How are measures computed?" To answer this question, we first study how measures can be categorized.1 Note that a multidimensional point in the data cube space can be defined by a set of dimension-value pairs, for example, htime = "Q1", location = "Vancouver",item = "computer"i. A data cube measure is a numerical function that can be evaluated at each point in the data cube space. A measure value is computed for a given point by aggregating the data corresponding to the respective dimension-value pairs defining the given point. Measures can be organized into three categories (i.e., distributive, algebraic, holistic), based on the kind of aggregate functions used.

Distributive: An aggregate function is distributive if it can be computed in a distributed manner as follows. Suppose the data are partitioned into n sets. We apply the function to each partition, resulting in n aggregate values. If the result derived by applying the function to the n aggregate values is the same as that derived by applying the function to the entire data set (without partitioning), the function can be computed in a distributed manner. For example, count() can be computed for a data cube by first partitioning the cube into a set of sub-cubes, computing count() for each sub-cube, and then summing up the counts obtained for each Sub-cube. Hence, count() is a distributive aggregate function. For the same reason, sum(), min(), and max() are distributive aggregate functions. A measure is distributive if it is obtained by applying a distributive aggregate function. Distributive measures can be computed efficiently because they can be computed in a distributive manner.

OLAP(Online analytical Processing):

- OLAP is an approach to answering multi-dimensional analytical (MDA) queries swiftly.
- OLAP is part of the broader category of business intelligence, which also encompasses relational database, report writing and data mining.

- OLAP tools enable users to analyze multidimensional data interactively from multiple perspectives.

OLAP consists of three basic analytical operations:

- ➢ Consolidation (Roll-Up)
- ➢ Drill-Down
- ➢ Slicing and Dicing

- Consolidation involves the aggregation of data that can be accumulated and computed in one or more dimensions. For example, all sales offices are rolled up to the sales department or sales division to anticipate sales trends.
- The drill-down is a technique that allows users to navigate through the details. For instance, users can view the sales by individual products that make up a region's sales.

- Slicing and dicing is a feature whereby users can take out (slicing) a specific set of data of the OLAP cube and view (dicing) the slices from different viewpoints.

Types of OLAP:

1. Relational OLAP (ROLAP):

- ROLAP works directly with relational databases. The base data and the dimension tables are stored as relational tables and new tables are created to hold the aggregated information. It depends on a specialized schema design.
- This methodology relies on manipulating the data stored in the relational database to give the appearance of traditional OLAP's slicing and dicing functionality. In essence, each action of slicing and dicing is equivalent to adding a "WHERE" clause in the SQL statement.
- ROLAP tools do not use pre-calculated data cubes but instead pose the query to the standard relational database and its tables in order to bring back the data required to answer
- the question.
- ROLAP tools feature the ability to ask any question because the methodology does not limit to the contents of a cube. ROLAP also has the ability to drill down to the lowest level of detail in the database.

2. Multidimensional OLAP (MOLAP):

- MOLAP is the 'classic' form of OLAP and is sometimes referred to as just OLAP.
- MOLAP stores this data in an optimized multi-dimensional array storage, rather than in a relational database. Therefore it requires the pre-computation and storage of information in the cube - the operation known as processing.

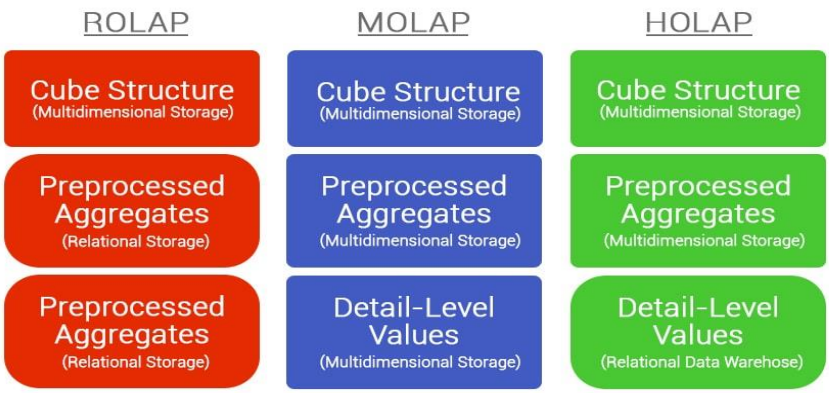
- MOLAP tools generally utilize a pre-calculated data set referred to as a data cube. The data cube contains all the possible answers to a given range of questions.
- MOLAP tools have a very fast response time and the ability to quickly write back data into the data set.

3. Hybrid OLAP (HOLAP):

- There is no clear agreement across the industry as to what constitutes Hybrid OLAP, except that a database will divide data between relational and specialized storage.
- For example, for some vendors, a HOLAP database will use relational tables to hold the larger quantities of detailed data, and use specialized storage for at least some aspects of the smaller quantities of more-aggregate or less-detailed data.
- HOLAP addresses the shortcomings of MOLAP and ROLAP by combining the capabilities of both approaches.
- HOLAP tools can utilize both pre-calculated cubes and relational data sources.

TYPES OF OLAP SERVERS:

ROLAP versus MOLAP versus HOLAP :



DATA WAREHOUSE IMPLEMENTATION:

Efficient Computation of Data Cubes Data cube can be viewed as a lattice of cuboids

The bottom-most cuboid is the **base cuboid**

The top-most cuboid (**apex**) contains only one cell

How many cuboids in an n-dimensional cube with L levels? Materialization of data cube

Materialize every (cuboid) (full materialization), none (no materialization), or some (partial materialization)
Selection of which cuboids to materialize
Based on size, sharing, access frequency, etc. Cube Operation
Cube definition and computation in DMQL define cube sales [item, city, year]:
sum(sales_in_dollars) compute cube sales
Transform it into a SQL-like language (with a new operator cube by, introduced by Gray et al. '96)

SELECT item, city, year, SUM (amount) FROM SALES CUBE BY item, city, year

Need compute the following Group-Bys (date, product, customer), (date, product),(date, customer), (product, customer), (date), (product), (customer) 23
Cube Computation:

ROLAP-Based Method

Efficient cube computation methods
ROLAP-based cubing algorithms (Agarwal et al'96)
Array-based cubing algorithm (Zhao et al'97)
Bottom-up computation method (Bayer & Ramarkrishnan'99)

ROLAP-based cubing algorithms :

Sorting, hashing, and grouping operations are applied to the dimension attributes in order to reorder and cluster related tuples

Grouping is performed on some sub aggregates as a ―partial grouping step‖

Aggregates may be computed from previously computed aggregates, rather than from the base fact table Multi-way Array Aggregation for Cube
Computation
Partition arrays into chunks (a small sub cube which fits in memory).

Compressed sparse array addressing: (chunk_id, offset) Compute aggregates in multiway by visiting cube cells in the order which minimizes the number of times to visit each cell, and reduces memory access and storage cost.

Indexing OLAP data:

The bitmap indexing method is popular in OLAP products because it allows quick searching in data cubes.

The bitmap index:
It is an alternative representation of the record ID (RID) list.

In the bitmap index for a given attribute, there is a distinct bit vector, By, for each value v in the domain of the attribute.

If the domain of a given attribute consists of n values, then n bits are needed for each entry in the bitmap index.

Join index:

The join indexing method gained popularity from its use in relational database query processing.

Traditional indexing maps the value in a given column to a list of rows having that value.

In contrast, join indexing registers the joinable rows of two relations from a relational database.

For example: if two relations R(RID;A) and S(B; SID) join on the attributes A and B, then the join index record contains the pair (RID; SID), where RID and SID are record identifiers from the R and S relations, respectively. Efficient processing of OLAP queries

Determine which operations should be performed on the available cuboids. This involves transforming any selection, projection, roll-up (group-by) and drill-down operations specified in the query into corresponding SQL and/or OLAP operations. For example, slicing and dicing of a data cube may correspond to selection and/or projection operations on a materialized cuboid.

Determine to which materialized cuboid(s) the relevant operations should be applied. This involves identifying all of the materialized cuboids that may potentially be used to answer the query.

OLAP:

Online Analytical Processing is based on the multidimensional data model that allow user to extract and view data from different points of view.

OLAP data stored in multidimensional data.

OLAP OPERATIONS

Since OLAP servers are based on multidimensional view of data, we will discuss OLAP operations in multidimensional data.

Here is the list of OLAP operations:

- Roll-up
- Drill-down,
- Slice and dice,
- Pivot (rotate)

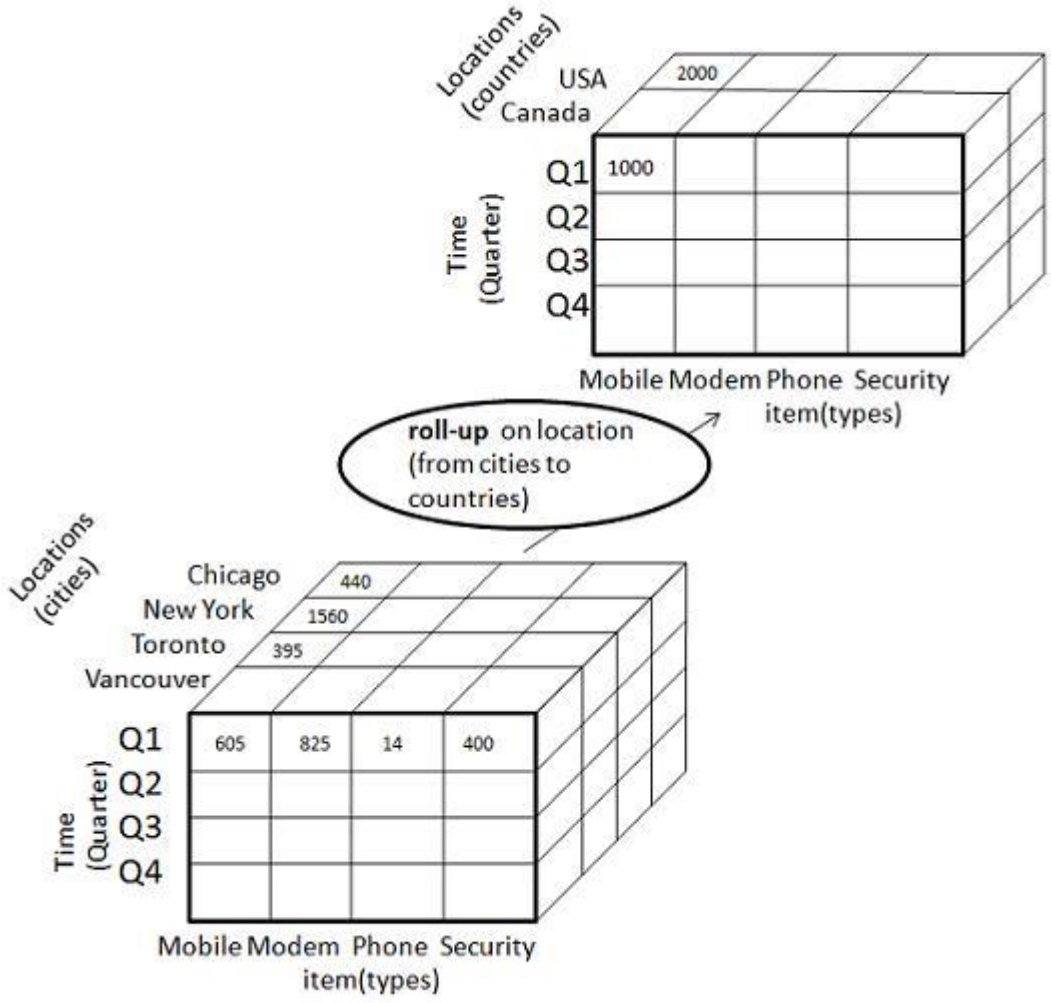
Roll-up

Roll-up performs aggregation on a data cube in any of the following ways −

By climbing up a concept hierarchy for a dimension

By dimension reduction

The following diagram illustrates how roll-up works.



Roll-up is performed by climbing up a concept hierarchy for the dimension location.

Initially the concept hierarchy was "street < city < province < country".

On rolling up, the data is aggregated by ascending the location hierarchy from the level of city to the level of country.

The data is grouped into cities rather than countries.

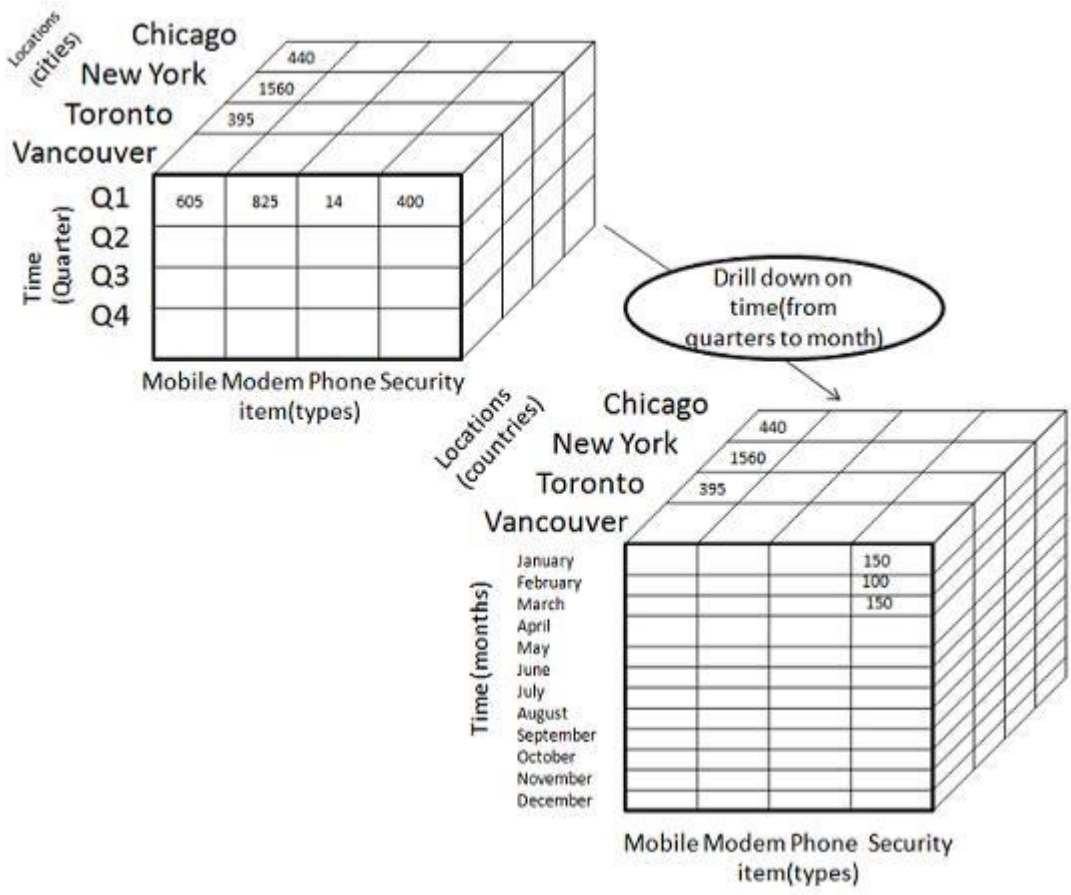
When roll-up is performed, one or more dimensions from the data cube are removed.

Drill-down

Drill-down is the reverse operation of roll-up. It is performed by either of the following ways:

- By stepping down a concept hierarchy for a dimension
- By introducing a new dimension.

The following diagram illustrates how drill-down works



Drill-down is performed by stepping down a concept hierarchy for the dimension time.

Initially the concept hierarchy was "day < month < quarter < year."

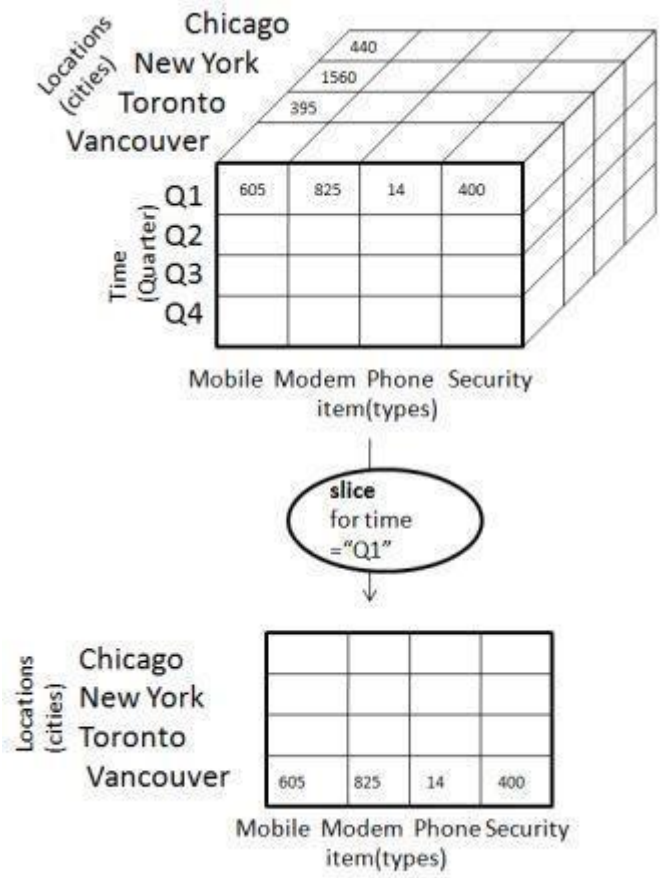
On drilling down, the time dimension is descended from the level of quarter to the level of month.

When drill-down is performed, one or more dimensions from the data cube are added.

It navigates the data from less detailed data to highly detailed data.

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.

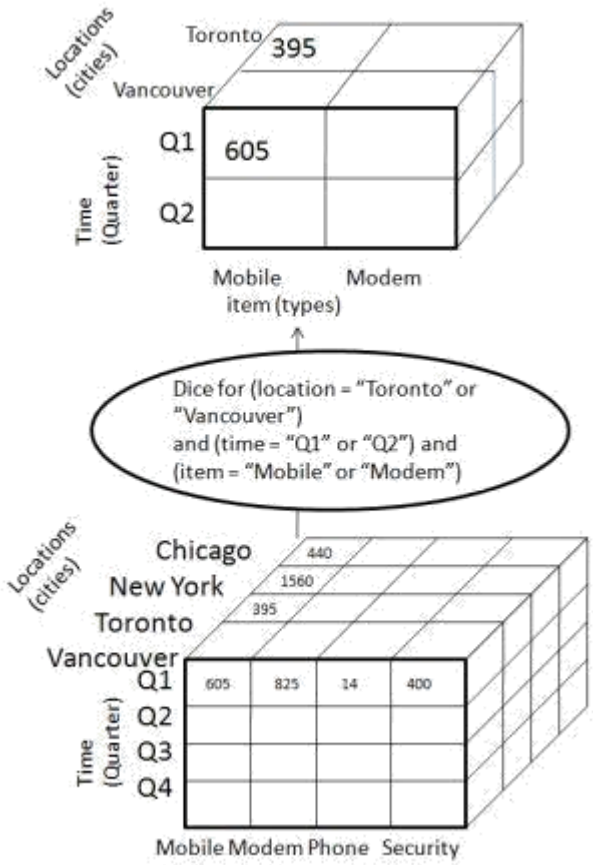


Here Slice is performed for the dimension "time" using the criterion time "Q1".

It will form a new sub-cube by selecting one or more dimensions.

Dice

Dice 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 based on the following selection criteria involves three dimensions.

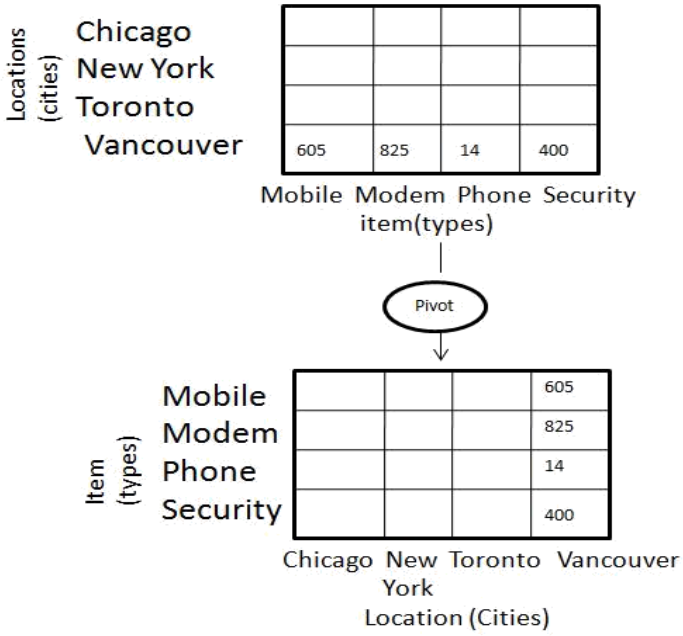
(location = "Toronto" or "Vancouver")

(time = "Q1" or "Q2")

(item =" Mobile" or "Modem")

Pivot

The pivot operation is also known as rotation. It rotates the data axes in view in order to provide an alternative presentation of data. Consider the following diagram that shows the pivot operation.



FROM DATA WAREHOUSING TO DATA MINING:

"How do data warehousing and OLAP relate to data *mining*.We also introduce on-line analytical mining (OLAM), a powerful paradigm that integrates OLAP with data mining technology.

FROM ON-LINE ANALYTICAL PROCESSING TO ON LINE ANALYTICAL MINING (OLAM):

OLAM stands for Online analytical mining. It is also known as OLAP Mining. It integrates online analytical processing with data mining and mining knowledge in multi-dimensional databases. There are several paradigms and structures of data mining systems.

Various data mining tools must work on integrated, consistent, and cleaned data. This requires costly pre-processing for data cleaning, data transformation, and data integration. Thus, a data warehouse constructed by such pre-processing is a valuable source of high-quality information for both OLAP and data mining. Data mining can serve as a valuable tool for data cleaning and data integration.

OLAM is particularly important for the following reasons which are as follows −

High quality of data in data warehouses − Most data mining tools are required to work on integrated, consistent, and cleaned information, which needs costly data cleaning, data integration, and data transformation as a pre-processing phase. A data warehouse constructed by such pre-processing serves as a valuable source of high-quality data for OLAP and data mining. Data mining can also serve as a valuable tool for data cleaning and data integration.

Available information processing infrastructure surrounding data warehouses − Comprehensive data processing and data analysis infrastructures have been or will be orderly constructed surrounding data warehouses, which contains accessing, integration, consolidation, and transformation of various heterogeneous databases, ODBC/OLE DB connections, Web-accessing and service facilities, and documenting and OLAP analysis tools. It is careful to create the best use of the available infrastructures instead of constructing everything from scratch.

OLAP-based exploratory data analysis − Effective data mining required exploratory data analysis. A user will be required to traverse through a database, select areas of relevant information, analyze them at multiple granularities, and display knowledge/results in multiple forms.

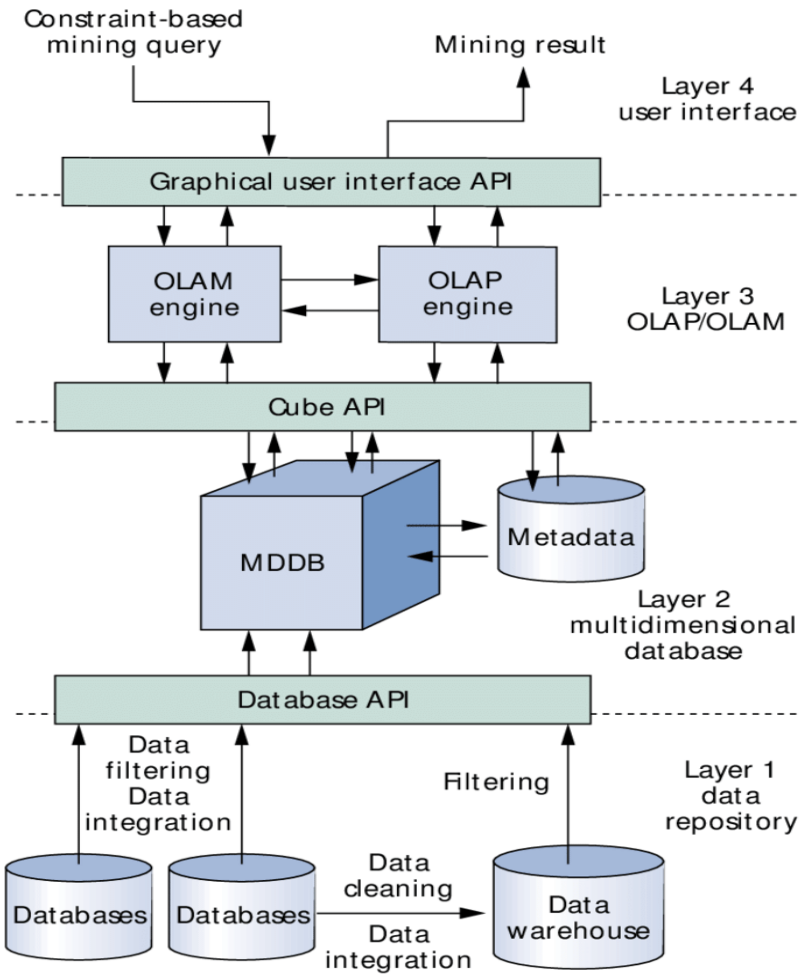
Online analytical mining supports facilities for data mining on multiple subsets of data and at several levels of abstraction, by drilling, pivoting, filtering, dicing, and slicing on a data cube and some intermediate data mining outcomes.

On-line selection of data mining functions − It supports a user who cannot understand what type of knowledge they would like to mine. By integrating OLAP with various data mining functions, online analytical mining provides users with the flexibility to choose desired data mining functions and swap data mining tasks dynamically.

ARCHITECTURE FOR ON-LINE ANALYTICAL MINING(OLAM):

- An OLAM engine performs analytical mining in data cubes in a similar manner as an OLAP engine performs on-line analytical processing.
- An integrated OLAM and OLAP architecture is shown in Figure, where the OLAM and OLAP engines both accept users' on-line queries via a User GUI API and work with the data cube in the data analysis via a Cube API.
- A metadata directory is used to guide the access of the data cube. The data cube can be constructed by accessing and/or integrating multiple databases and/or by filtering a data warehouse via a Database API which may support OLEDB or ODBC connections.

Since an OLAM engine may perform multiple data mining tasks, such as concept description, association, classification, prediction, clustering, time-series analysis, etc, it usually consists of multiple, integrated data mining modules and is more sophisticated than an OLAP engine.


Constraint-based mining query
Mining result
Layer 4 user interface
Graphical user interface API
OLAM engine
OLAP engine
Layer 3 OLAP/OLAM
Cube API
MDDB
Metadata
Layer 2 multidimensional database
Database API
Data filtering Data integration
Filtering
Layer 1 data repository
Databases
Databases
Data cleaning
Data integration
Data warehouse

LEARNING UNIT II: DATA PREPROCESSING

Definition - What does Data Preprocessing mean?

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing. Data preprocessing is used database-driven applications such as customer relationship management and rule-based applications (like neural networks). Data goes through a series of steps during preprocessing:

Data Cleaning: Data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data.

Data Integration: Data with different representations are put together and conflicts within the data are resolved.

Data Transformation: Data is normalized, aggregated and generalized.

Data Reduction: This step aims to present a reduced representation of the data in a data warehouse.

Data Discretization: Involves the reduction of a number of values of a continuous attribute by dividing the range of attribute intervals.

Why Data Pre-processing? Data preprocessing prepares raw data for further processing. The traditional data preprocessing method is reacting as it starts with data that is assumed ready for analysis and there is no feedback and impart for the way of data collection. The data inconsistency between data sets is the main difficulty for the data preprocessing.

1 . Data Cleaning.

Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

(i). Missing values

Ignore the tuple: This is usually done when the class label is missing (assuming the mining task involves classification or description). This method is not very effective, unless the tuple

contains several attributes with missing values. It is especially poor when the percentage of missing values per attribute varies considerably.

Fill in the missing value manually: In general, this approach is time-consuming and may not be feasible given a large data set with many missing values.

Use a global constant to fill in the missing value: Replace all missing attribute values by the same constant, such as a label like ―Unknown". If missing values are replaced by, say, ―Unknown", then the mining program may mistakenly think that they form an interesting concept, since they all have a value in common - that of ―Unknown". Hence, although this method is simple, it is not recommended.

Use the attribute mean to fill in the missing value: For example, suppose that the average income of All Electronics customers is $28,000. Use this value to replace the missing value for income.

Use the attribute mean for all samples belonging to the same class as the given tuple: For example, if classifying customers according to credit risk, replace the missing value with the average income value for customers in the same credit risk category as that of the given tuple.

Use the most probable value to fill in the missing value: This may be determined with inference-based tools using a Bayesian formalism or decision tree induction. For example, using the other customer attributes in your data set, you may construct a decision tree to predict the missing values for income.

(ii). Noisy data

Noise is a random error or variance in a measured variable.

1. Binning methods:

Binning methods smooth a sorted data value by consulting the neighbourhood", or values around it. The sorted values are distributed into a number of 'buckets', or bins. Because binning methods consult the neighbourhood of values, they perform local smoothing.

In this example, the data for price are first sorted and partitioned into equal-depth bins (of depth 3). In smoothing by bin means, each value in a bin is replaced by the mean value of the bin. For example, the mean of the values 4, 8, and 15 in Bin 1 is 9. Therefore, each original value in this bin is replaced by the value 9. Similarly, smoothing by bin medians can be employed, in which each bin value is replaced by the bin median. In smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value.

(i).Sorted data for price (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34
(ii).Partition into (equi-width) bins: Bin 1: 4, 8, 15

Bin 2: 21, 21, 24

Bin 3: 25, 28, 34

(iii).**Smoothing by bin means:**

Bin 1: 9, 9, 9,

Bin 2: 22, 22, 22

Bin 3: 29, 29, 29

(iv).Smoothing by bin boundaries:

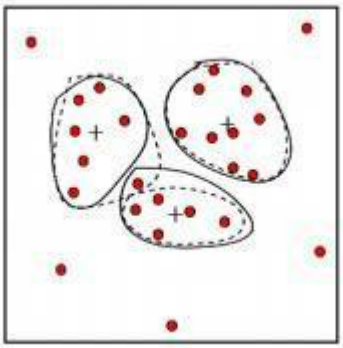
Bin 1: 4, 4, 15

Bin 2: 21, 21, 24

Bin 3: 25, 25, 34

2. Clustering:
Outliers may be detected by clustering, where similar values are organized into groups or clusters.

Figure: Outliers may be detected by clustering analysis.



Combined computer and human inspection: Outliers may be identified through a combination of computer and human inspection. In one application, for example, an information-theoretic measure was used to help identify outlier patterns in a handwritten character database for classification.

Regression: Data can be smoothed by fitting the data to a function, such as with regression.

Linear regression involves finding the ―best" line to fit two variables, so that one variable can be used to predict the other.

Multiple linear regression is an extension of linear regression, where more than two variables are involved and the data are fit to a multidimensional surface.

(iii). Inconsistent data:

There may be inconsistencies in the data recorded for some transactions. Some data inconsistencies may be corrected manually using external references. For example, errors made at data entry may be corrected by performing a paper trace. This may be coupled with routines designed to help correct the inconsistent use of codes.

2. DATA TRANSFORMATION:

In data transformation, the data are transformed or consolidated into forms appropriate for mining. Data transformation can involve the following:

Normalization, where the attribute data are scaled so as to fall within a small specified range, such as -1.0 to 1.0, or 0 to 1.0.

There are three main methods for data normalization :

min-max normalization

z-score normalization

Normalization by decimal scaling.

(i).Min-max normalization performs a linear transformation on the original data. Suppose that minA and maxA are the minimum and maximum values of an attribute A. Min-max normalization maps a value v of A to v0 in the range [new minA; new maxA] by computing

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A.$$

(ii).z-score normalization (or zero-mean normalization), the values for an attribute A are normalized based on the mean and standard deviation of A. A value v of A is normalized to v0 by computing where mean A and stand dev A are the mean and standard deviation, respectively, of attribute A.

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

(iii). **Normalization by decimal scaling** normalizes by moving the decimal point of values of attribute A. The number of decimal points moved depends on the maximum absolute value of A. A value v of A is normalized to v0by computing where j is the smallest integer such that

$$Max(|v'|) < 1.$$

Smoothing, which works to remove the noise from data? Such techniques include binning, clustering, and regression.

(i). Binning methods:

Binning methods smooth a sorted data value by consulting the neighbourhood, or values around it. The sorted values are distributed into a number of 'buckets', or bins. Because binning methods consult the neighbourhood of values, they perform local smoothing. Figure illustrates some binning techniques.

In smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value.

(i).Sorted data for price (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

(ii).Partition into (equi-width) bins:

Bin 1: 4, 8, 15

Bin 2: 21, 21, 24

Bin 3: 25, 28, 34

(iii).Smoothing by bin means:

Bin 1: 9, 9, 9,

Bin 2: 22, 22, 22

Bin 3: 29, 29, 29

(iv).Smoothing by bin boundaries:

Bin 1: 4, 4, 15

Bin 2: 21, 21, 24

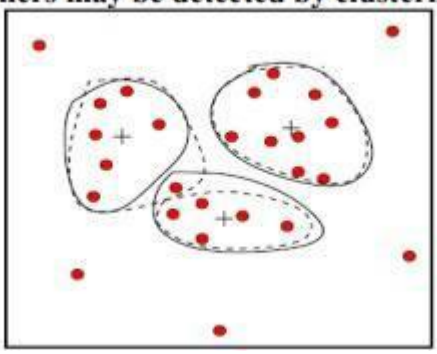
Bin 3: 25, 25, 34

(ii). Clustering:

Outliers may be detected by clustering, where similar values are organized into groups or clusters. Intuitively, values which fall outside of the set of clusters may be considered outliers.

Figure: Outliers may be detected by clustering analysis.

Figure: Outliers may be detected by clustering analysis.



Aggregation: where summary or aggregation operations are applied to the data. For example, the daily sales data may be aggregated so as to compute monthly and annual total amounts.

Generalization of the data: where low level or 'primitive' (raw) data are replaced by higher level concepts through the use of concept hierarchies. For example, categorical attributes, like street, can be generalized to higher level concepts, like city or county.

3. DATA REDUCTION:

Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data.

Strategies for data reduction include the following:

Data cube aggregation:where aggregation operations are applied to the data in the construction of a data cube.

Dimension reduction: where irrelevant, weakly relevant or redundant attributes or dimensions may be detected and removed.

Data compression, where encoding mechanisms are used to reduce the data set size.

Numerosity reduction: where the data are replaced or estimated by alternative, smaller data representations such as parametric models (which need store only the model parameters instead of the actual data), or nonparametric methods such as clustering, sampling, and the use of histograms.

Discretization and concept hierarchy generation: where raw data values for attributes are replaced by ranges or higher conceptual levels. Concept hierarchies allow the mining of data at multiple levels of abstraction, and are a powerful tool for data mining.

Data Cube Aggregation

The lowest level of a data cube
the aggregated data for an individual entity of interest

e.g., a customer in a phone calling data warehouse.
Multiple levels of aggregation in data cubes
Further reduce the size of data to deal with
Reference appropriate levels
Use the smallest representation which is enough to solve the task

Queries regarding aggregated information should be answered using data cube, when possible

LEARNING UNIT III: DATA MINING

DEFINITION OF DATA MINING

Data Mining is defined as extracting information from huge sets of data. In other words, we can say that data mining is the procedure of mining knowledge from data. The information or knowledge extracted so can be used for any of the following applications

- Market Analysis
- Fraud Detection
- Customer Retention
- Production Control
- Science Exploration

Major Sources of data:

Business –Web, E-commerce, Transactions, Stocks - Science – Remote Sensing, Bio informatics, Scientific Simulation - Society and Everyone – News, Digital Cameras, YouTube. Need for turning data into knowledge – Drowning in data, but starving for knowledge.

Definition of Data Mining?

Extracting and 'Mining' knowledge from large amounts of data. "Gold Mining from rock or sand" is same as "Knowledge mining from data"

Other terms for Data Mining:

Knowledge Mining

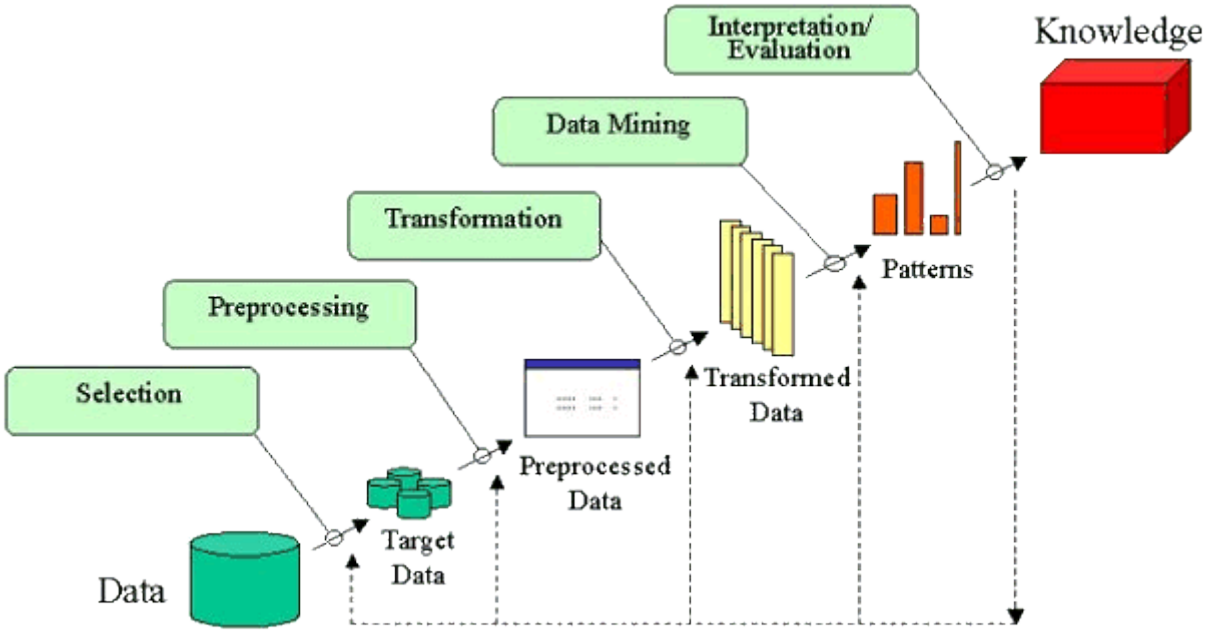
Knowledge Extraction o Pattern Analysis

KNOWLEDGE DISCOVERY (KDD) PROCESS:

Several Key Steps:
Data processing

- **Data cleaning (**remove noise and inconsistent data)
- **Data integration** (multiple data sources maybe combined)
- **Data selection** (data relevant to the analysis task are retrieved from database)
- **Data transformation** (data transformed or consolidated into forms)
- **Data mining** (an essential process where intelligent methods are applied to extract data patterns)

- **Pattern evaluation** (identify the truly interesting patterns)
- **Knowledge presentation** (mined knowledge is presented to the user with visualization or representation techniques)



DATA MINING ON WHAT KIND OF DATA (TYPES OF DATA):

RELATIONAL DATABASES:

A database system, also called a database management system (DBMS), consists of a collection of interrelated data, known as a database, and a set of software programs to manage and access the data.

A relational database: is a collection of tables, each of which is assigned a unique name.

Each table consists of a set of attributes (*columns* or *fields*) and usually stores a large set of tuples (*records* or *rows*).

Each tuple in a relational table represents an object identified by a unique *key* and described by a set of attribute values.

A semantic data model, such as an entity-relationship (ER) data model, is often constructed for relational databases.

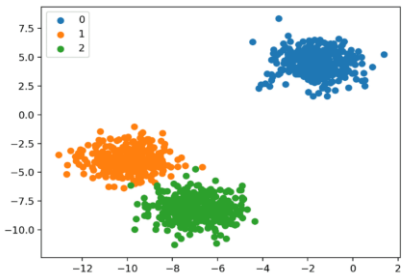
An **ER data model** represents the database as a set of entities and their relationships.

Data Mining Techniques

There are a wide array of data mining techniques used in data science and data analytics. Your choice of technique depends on the nature of your problem, the available data, and the desired outcomes. Predictive modelling is a fundamental component of mining data a nd is widely used to make predictions or forecasts based on historical data patterns. You may also employ a combination of techniques to gain comprehensive insights from the data. Top -10 data mining techniques:

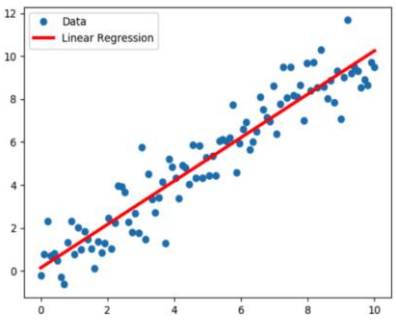
1. **Classification**

Classification is a technique used to categorize data into predefined classes or categories based on the features or attributes of the data instances. It involves training a model on labeled data and using it to predict the class labels of new, unseen data instances.



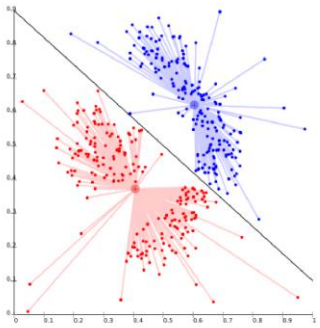
2. **Regression**

Regression is employed to predict numeric or continuous values based on the relationship between input variables and a target variable. It aims to find a mathematical function or model that best fits the data to make accurate predictions.



3. **Clustering**

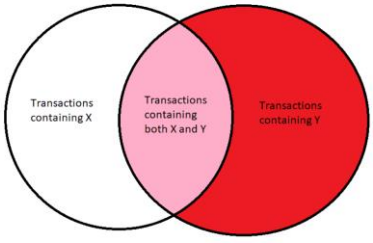
Clustering is a technique used to group similar data instances together based on their intrinsic characteristics or similarities. It aims to discover natural patterns or structures in the data without any predefined classes or labels.



Source: Wikipedia

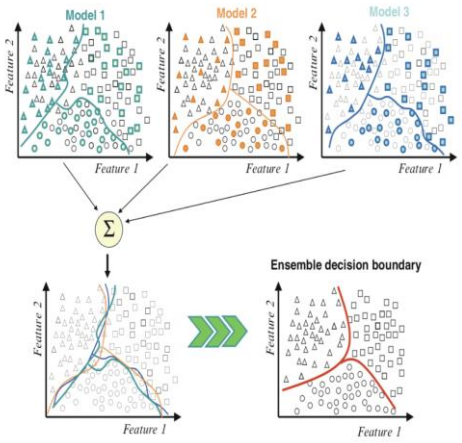
4. **Association Rule**

Association rule mining focuses on discovering interesting relationships or patterns among a set of items in transactional or market basket data. It helps identify frequently co -occurring items and generates rules such as "if X, then Y" to reveal associations between items. This simple Venn diagram shows the associations between itemsets X and Y of a dataset.



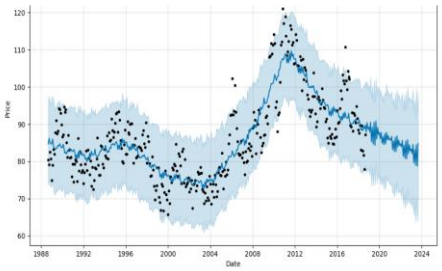
Source: Wikipedia

5. **Anomaly Detection**
Anomaly detection, sometimes called outlier analysis, aims to identify rare or unusual data instances that deviate significantly from the expected patterns. It is useful in detecting fraudulent transactions, network intrusions, manufacturing defects, or any other abnormal behavior.



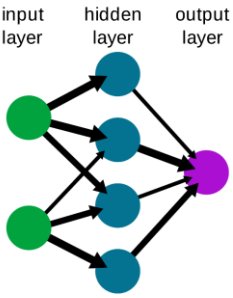
Source: Ensemble Machine Learning

6. **Time Series Analysis**
Time series analysis focuses on analyzing and predicting data points collected over time. It involves techniques such as forecasting, trend analysis, seasonality detection, and anomaly detection in time-dependent datasets.

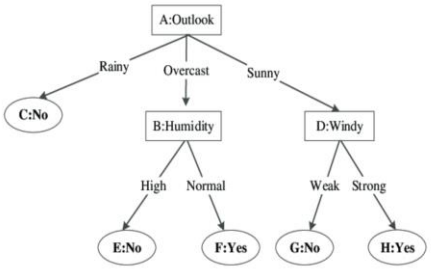


Source: Data Science Stack Exchange

7. **Neural Networks**
Neural networks are a type of machine learning or AI model inspired by the human brain's structure and function. They are composed of interconnected nodes (neurons) and layers that can learn from data to recognize patterns, perform classification, regression, or other tasks.



8. **Decision Trees**
Decision trees are graphical models that use a tree-like structure to represent decisions and their possible consequences. They recursively split the data based on different attribute values to form a hierarchical decision-making process.



Source: ResearchGate

9. **Ensemble Methods**
Ensemble methods combine multiple models to improve prediction accuracy and generalization. Techniques like Random Forests and Gradient Boosting utilize a combination of weak learners to create a stronger, more accurate model.

10. **Text Mining**
Text mining techniques are applied to extract valuable insights and knowledge from unstructured text data. Text mining includes tasks such as text categorization, sentiment analysis, topic modeling, and information extraction, enabling your organization to derive meaningful insights from large volumes of textual data, such as customer reviews, social media posts, emails, and articles.

Difference between Data mining and Text mining

Data mining can be understood as a process of data extraction from a huge data set. The data is extracted to acquire knowledge about certain data sets to be further used for learning and processing purposes.

Data mining involves the following steps:

1. **Business Understanding:** Business understanding refers to a process of comprehending each feature of a topic and work.
2. **Data Selection:** It is used to pick the best data set for performing data extraction.
3. **Data Preparation:** It prepares the extracted data to undergo further improvement.
4. **Modeling:** It remodels the input data based on user requirements.
5. **Evaluation:** It thoroughly reviews the complete process to check for possible faults or data leakage within the process. It plays an important role in data
6. **Deployment:** Once everything is evaluated, the data is ready for deployment and can be further utilized.

Applications of Data Mining

- **Market Analysis**

Market analysis is one such application of data science that helps analyze the current status of the market. As a result, it enables an individual in decision-making in terms of investments and business strategies for generating profit.

- **Fraud Detection**

Frauds can be easily detected with the help of fraud detection by extracting more and more information related to any particular instance and then formulating a decision whether it is legal or illegal.

- **Customer Retention**

It extracts customer's information based on their interests and offers them exciting deals to buy any particular product. These strategies not only help in providing a high level of customer satisfaction but also maintain a healthy relationship with them.

- **Science Exploration**

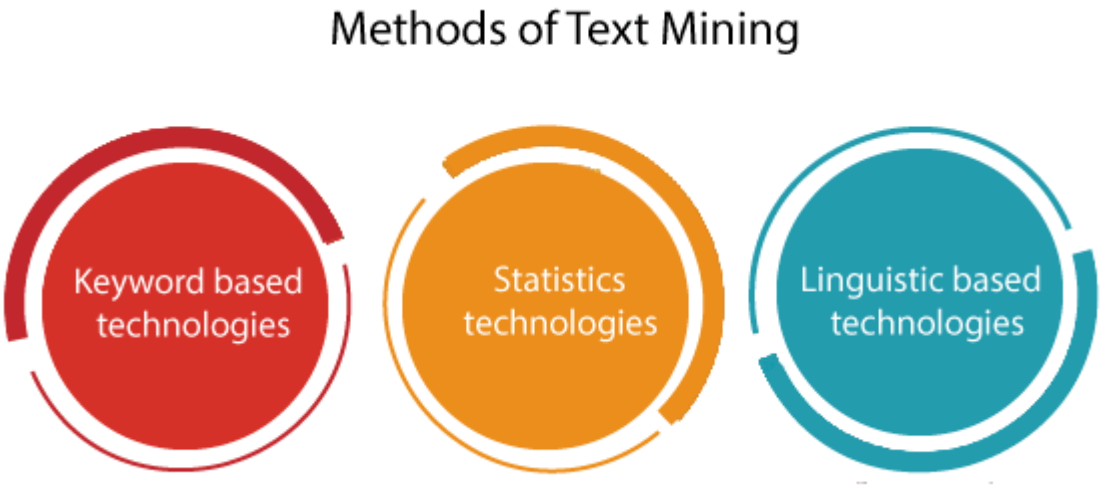
With the help of data mining, we can extract previous experiments or test case's knowledge and further utilize it to work proficiently. In this way, the errors can be minimized by learning from preceding mistakes and utilized for producing better results.

Text Mining

Text Mining is also known as text data mining. It refers to the process of extracting high-quality data from the text. High-quality data is usually extracted through the discovering of patterns and trends such as statistical pattern learning.

Text analysis includes pattern recognition, information extraction, information retrieval, data mining techniques involve association analysis, visualization, and predictive analytics.

Text Mining comprises a wide range of methods; the primary three methods are given below.



1. Keyword-based technologies
2. Statistics technologies
3. Linguistic based technologies

Keyword-based technologies

In keyword-based technologies, the input is based on the keywords selected in the text extracted as a series of character strings.

Statistics technologies

Statistics technologies refer to the system which is based on machine learning. It has a training set of documents used as a model to categorize and manage text.

Linguistic based technologies

Linguistic-based technologies are a method based on a language processing system. The output f the text analysis gives an understanding of the structure of the text, logic, and grammar employed.

Application of Text Mining

Risk Management

Risk management is the process of identifying risk, quantifying that risk, and then employing different types of strategies to manage that risk. Preliminary risk analysis is usually a primary cause of failure of any industry. Primarily in the financial industry, where adoption of risk management software based on text mining can enhance the capability to reduce risk.

Customer care services

Customer service is the act of taking care of the customer's needs by providing and delivering professional, helpful, high-quality service and assistance before, during, and after the customer's requirements are met. Nowadays, text analytics software is adopted to enhance customer experience using various sources of information such as trouble tickets, surveys, and reviews to improve the management, quality, and speed in resolving problems.

Difference between Data Mining and Text Mining

Data Mining	Text Mining
Data mining is a process to extract useful information from huge datasets.	Text Mining is a part of data mining that includes the processing of text from huge documents.
In data mining, we get the stored data in a structured format.	In text mining, we get the stored data in an unstructured format.
It allows the mining of mixed data.	It allows mining of text only.
Data processing is done directly.	Data processing is done linguistically.
It is a homogeneous process.	It is a heterogeneous process.
Pre-defined databases and sheets are used to collect the information.	The text is used to gather high-quality data.
The statistical method is used for data evaluation.	Computational linguistic principles are used to evaluate the text.