Module 1

Introduction

Data information and knowledge

Data: It is the unorganized and unprocessed fact and figures. It can be a number, a word, a picture or a recorded sound. It has no meaning itself.

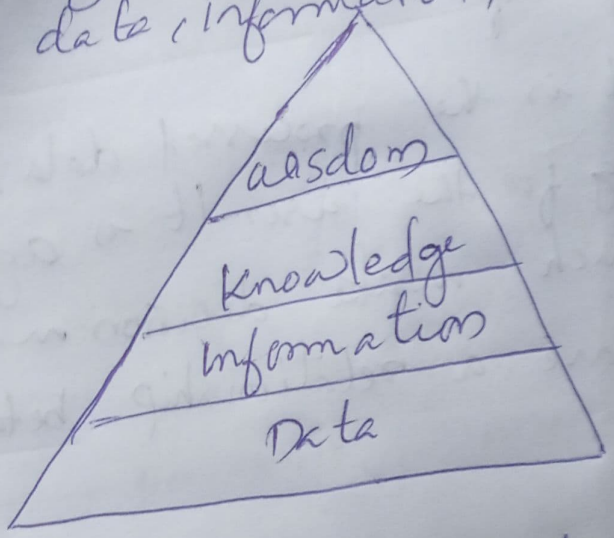
Information: It is the processed data, it has a meaning for the user. It is aggregation of data which make decision making easier. It also have a relationship between pieces of data.

Knowledge: It is defined as expertise and skills acquired by a person through experience or education or it is the theoretical or practical understanding of a subject. Knowledge is the ability to take an action.

Wisdom:
It is the ability of the human being to judge between right and wrong / good and bad. It is the understanding and realising of people, things, events or situation

Resulting in most appropriate action.

These relationship is also known as information hierarchy, knowledge hierarchy or knowledge pyramid. It refers to the relationship b/w data, information, knowledge and wisdom. It is also known as DIKW hierarchy. The 4 components in the hierarchy are data, information, knowledge and wisdom.



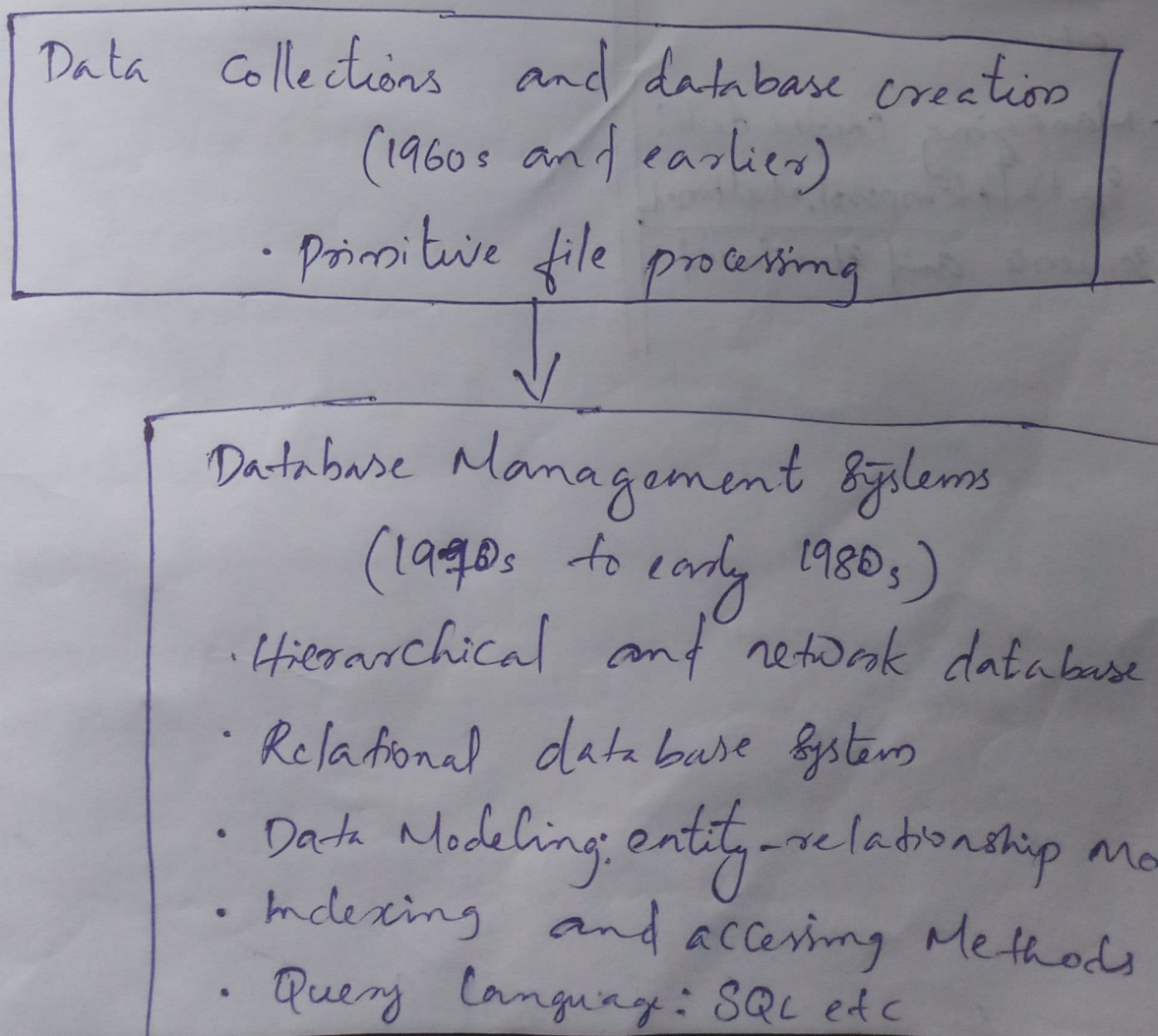
knowledge management

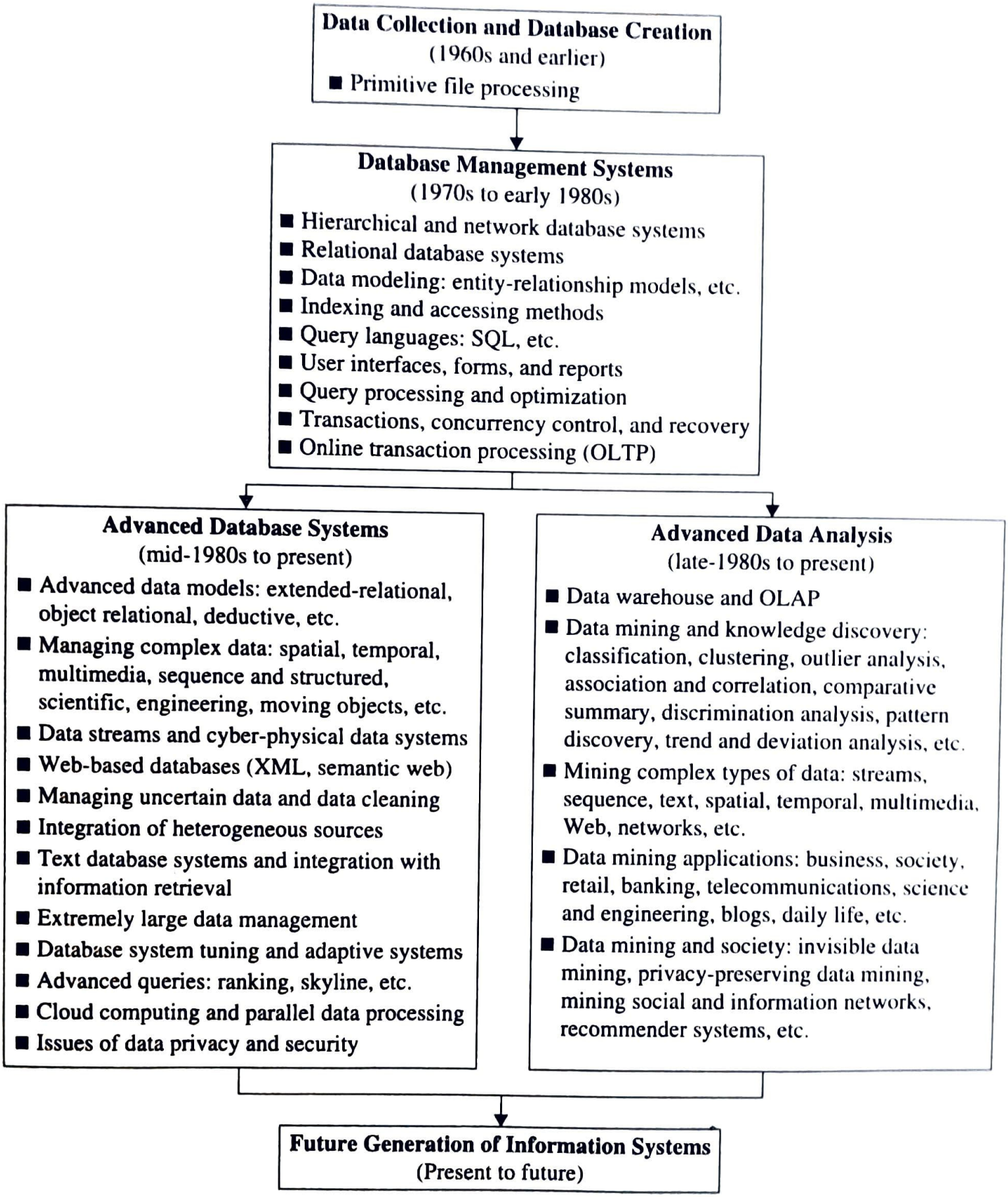
It is the process that governs the creation, dessimination and utilisation of knowledge aimed at the success of an organization. It is also known as the process of creating, storing, sharing and reusing organizational knowledge.

Datamining

It is also known as knowledge Discovery from Data (KDD). Data mining is the process of Collecting useful data and patterns from enormous data for various applications.

Data mining provides powerful tools for automatically uncover valuable information from the tremendous amount of data and to transform such data into organized knowledge. Data mining turns a large collection into knowledge




Data Collection and Database Creation
(1960s and earlier)
■ Primitive file processing
Database Management Systems
(1970s to early 1980s)
■ Hierarchical and network database systems
■ Relational database systems
■ Data modeling: entity-relationship models, etc.
■ Indexing and accessing methods
■ Query languages: SQL, etc.
■ User interfaces, forms, and reports
■ Query processing and optimization
■ Transactions, concurrency control, and recovery
■ Online transaction processing (OLTP)
Advanced Database Systems
(mid-1980s to present)
■ Advanced data models: extended-relational, object relational, deductive, etc.
■ Managing complex data: spatial, temporal, multimedia, sequence and structured, scientific, engineering, moving objects, etc.
■ Data streams and cyber-physical data systems
■ Web-based databases (XML, semantic web)
■ Managing uncertain data and data cleaning
■ Integration of heterogeneous sources
■ Text database systems and integration with information retrieval
■ Extremely large data management
■ Database system tuning and adaptive systems
■ Advanced queries: ranking, skyline, etc.
■ Cloud computing and parallel data processing
■ Issues of data privacy and security
Advanced Data Analysis
(late-1980s to present)
■ Data warehouse and OLAP
■ Data mining and knowledge discovery: classification, clustering, outlier analysis, association and correlation, comparative summary, discrimination analysis, pattern discovery, trend and deviation analysis, etc.
■ Mining complex types of data: streams, sequence, text, spatial, temporal, multimedia, Web, networks, etc.
■ Data mining applications: business, society, retail, banking, telecommunications, science and engineering, blogs, daily life, etc.
■ Data mining and society: invisible data mining, privacy-preserving data mining, mining social and information networks, recommender systems, etc.
Future Generation of Information Systems
(Present to future)

Warehousing

a repository of multiple heterogeneous data sources organized under a unified schema at a single site to facilitate management decision making. This technology includes data cleaning, data integration, online analytical processing (OLAP). It is an analysis technique with functionalities such as summarization, consolidation, and aggregation. It also support views information from different angles.

Data mining

It is also known as knowledge mining from data, data or pattern analysis, data archeology, knowledge extraction. It also known as knowledge Discovery from Data (KDD). The knowledge Discovery process contains an iterative of the following steps:-

1) Data cleaning - Remove noise and inconsistent data

2) Data Integration – Multiple data sources can be combined.

3) Data selection – Here relevant datas are retrieved from the database using analysis process.

4) Data transformation – Datas are transformed and consolidated into forms appropriate for mining by performing Summary and aggregation operations.

5) Data mining – It is the process used for extracting data patterns.

6) Pattern evaluation – used for identifying interesting patterns representing knowledge based on interesting measures.

7) Knowledge presentation – Here visualization and knowledge representation techniques are used to present mined knowledge to users.

Step 1-4 represent different forms of data preprocessing where data are prepared for mining. Data mining step may interact with users or a knowledge base. The interesting patterns

stored as new knowledge in the knowledge base.

Data mining is the process of discovering interesting Patterns and knowledge on large amount of data. The data sources include databases, datawarehouses, web, Information Repositories etc.

Which kind of data can be mined?

The basic forms of data for mining applications are

i) database data
ii) datawarehouse data
iii) transactional data

Data mining can also be applied to other forms of data like data streams, sequence data, graph/networked data, Spatial data, text data, multimedia data and www.

i) Database data

A database system also called database management system (DBMS) consist of a collection of interrelated data known as database and a set of software programs to manage and access the data. The software program provide mechanisms for defining database structure and data storage for specifying and managing concurrent, shared or distributed data access for ensuring consistency and security of the information stored.

A relational database is a collection of tables each of which have a unique name. Each table consist of a set of attributes (columns or field) and a large set of tuples (records/rows). Each tuple in a relational table represent and object identified by a unique key and a set of attribute values. An entity relationship model (ER model) is used for relational database. It represent database as a set of

tables and their relations.

Relational data can be accessed by database queries written in a relational query language.

ii) Data warehouses

It is a repository of information collected from multiple sources stored under a unified schema usually reside at a single time. It can be constructed using a process of data cleaning, data integration, data transformation, data loading and periodic data refreshing.

A data warehouse can be modelled by a multidimensional data structure known as data cube, in which each dimension corresponds to an attribute or a set of attribute in the schema and each cell stores the value of some aggregate measures like count or sum. A data cube provides a multidimensional view of data allows the pre-computation and fastest access of summarized data

iii) Transactional data

It captures a data transaction in each record like customer's purchase, flight booking etc. It always includes a unique transaction integrity no. and a list of items making of the transactions. A transactional data base have additional table which contain other information related to the transaction.

iv) Other kind of data

There are many other kind of data that have versatile forms and structures. They are i) time related / sequenced data
eg: historical data, stock exchange etc

ii) data streams
eg: video and sensor data

iii) Spatial data
eg: map

iv) Engineering and design data
eg: design of buildings, integrated circuit

v) Hypertext and multimedia data
include image, video and audio.

vi) Graph and networked data.

What kind of patterns can be mined?

There are a number of data mining functionalities. This include 1) characterization and discrimination.

2) mining of different pattern associations and correlation

3) Classification and regression

4) clustering and analysis

5) Outlier analysis