Data Management and Visualization

Dr. Ashish Kumar Jha

Session 1

Data Management and databases

Agenda

What is Database

Database Management Systems

Relational Data Model

Relationship between the information systems lifecycle and the database system development lifecycle.

Key Terms

Data

- Data is raw (unprocessed) facts that have some relevancy to an individual or organization.

Information

- Information is data that has been processed or given some structure that brings meaning to an individual or organization.

Database

- Shared collection of logically related data (and a description of this data), designed to meet the information needs of an organization.

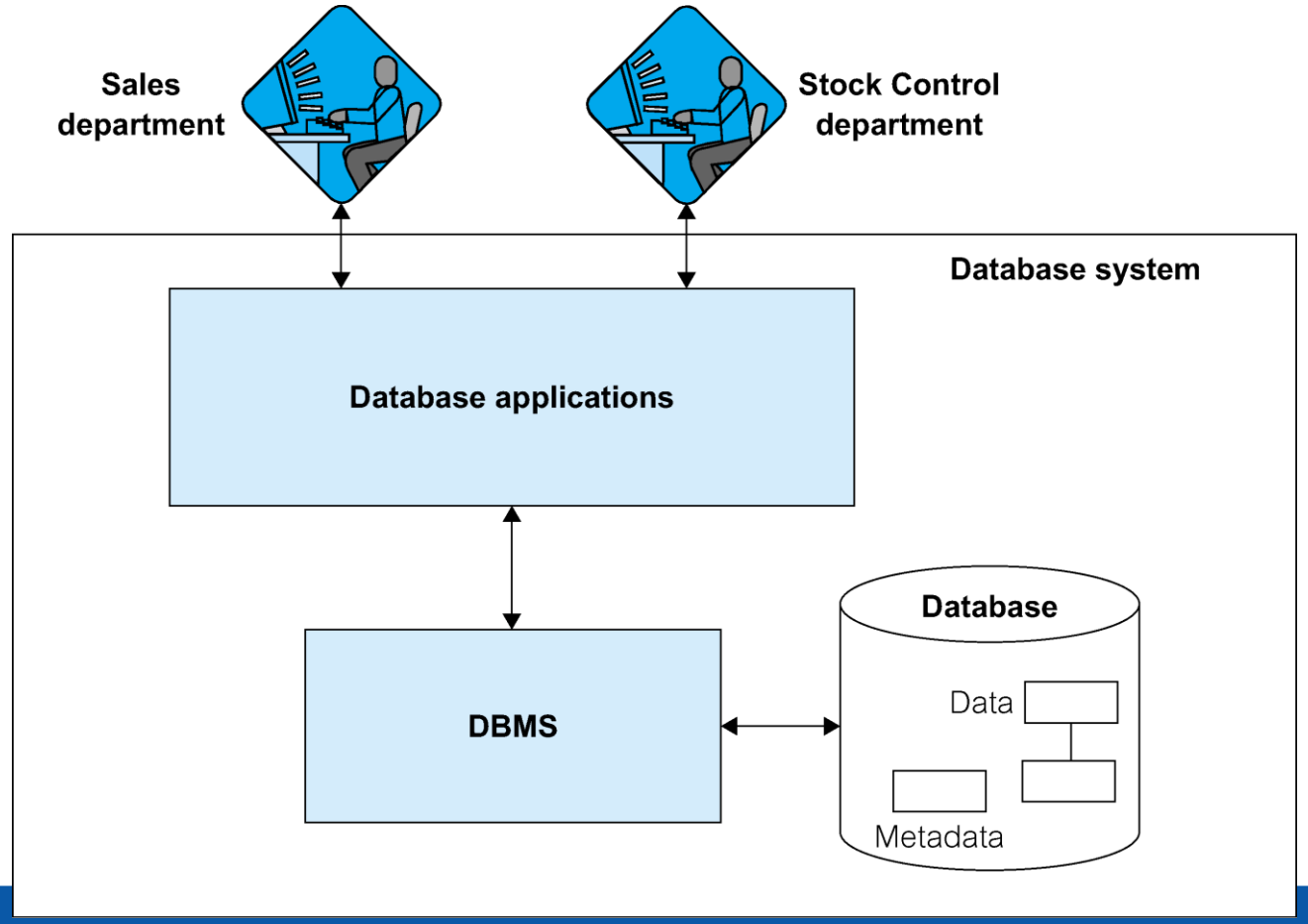
DBMS

- A software system that enables users to define, create, and maintain the database and that provides controlled access to this database.

Examples of Database Systems

- Purchases from the supermarket
- Purchases using your credit card
- Booking a holiday at the travel agents
- Using the local library
- Renting a DVD
- Using the Internet

Sales and Stock Control Departments



Aspects of DBMS

Database Design

- Schema, Views, Access
- Normalization
- Entity Relationships

Database Queries

- Codes to access, store or modify data from database

Database Applications

- Transactions
- Warehouses
- Data Mining

Database Administration and Security

Views

Allows each user to have his or her own view of the database.

A view is essentially some subset of the database.

Benefits include:

- Provides a level of security.
- Provides a mechanism to customize the appearance of the database.
- Presents a consistent, unchanging picture of the structure of the database, even if the underlying database is changed.

Components of DBMS Environment

Hardware

– Can range from a PC to a network of computers.

Software

– DBMS, operating system, network software (if necessary) and also the application programs.

Data

– Used by the organization and a description of this data called the schema.

Components of DBMS Environment

Procedures

- Instructions and rules that should be applied to the design and use of the database and DBMS.

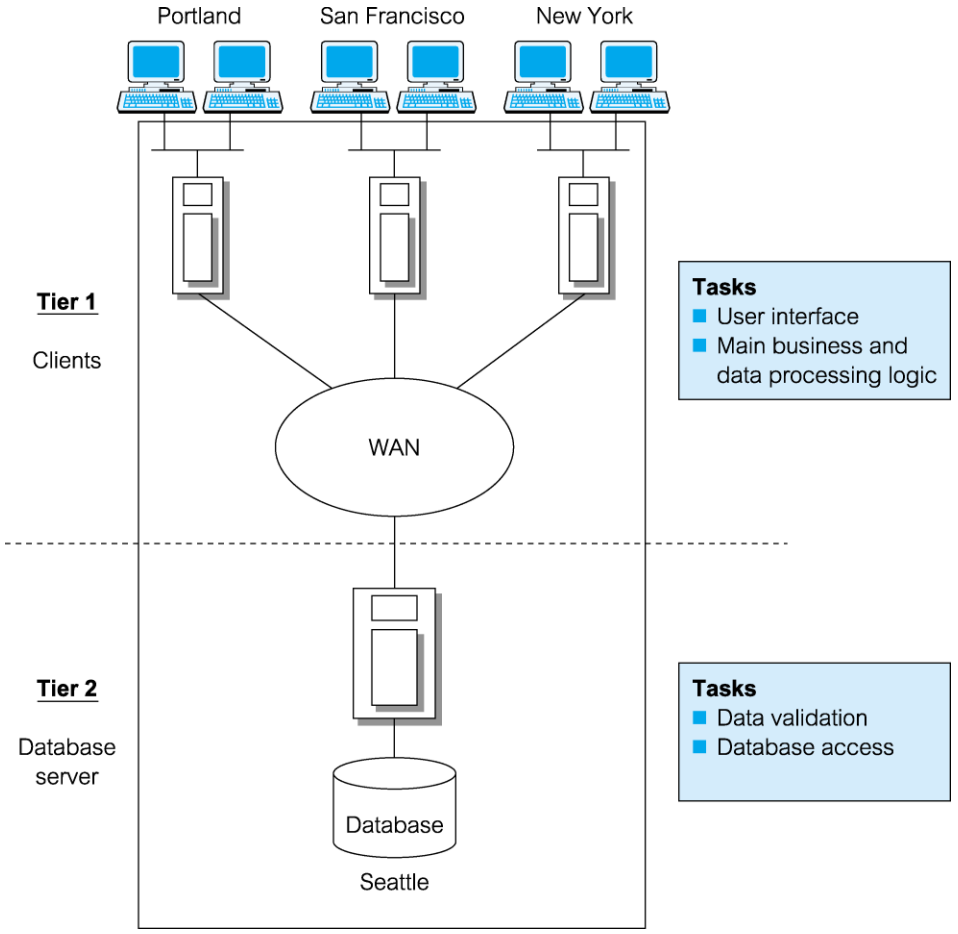
People

- Includes database designers, DBAs, application programmers, and end-users.

Two-Tier Client-Server

Client manages main business and data processing logic and user interface.

Server manages and controls access to database.



Database Design

The structure of the database is determined during the database design.

A system must be data-driven to satisfy an organization's information needs.

A well designed database produces a system that provides the correct information for the decision-making process to succeed.

Historical Perspective

File-Based Systems are a collection of application programs that perform services for the end users (e.g. reports).

Each program defines and manages its own data.

Limitations of File-Based Systems

- **Separation and isolation of data**
  - Each program maintains its own set of data.
  - Users of one program may be unaware of potentially useful data held by other programs.

- **Duplication of data**
  - Same data is held by different programs.
  - Wasted space and potentially different values and/or different formats for the same item.

Limitations of File-Based Approach

- **Data dependence**
  - File structure is defined in the program code.
- **Incompatible file formats**
  - Programs are written in different languages, and so cannot easily access each other's files.
- **Fixed Queries/Proliferation of application programs**
  - Programs are written to satisfy particular functions.
  - Any new requirement needs a new program.

DBMS Development

Arose because:

- Definition of data was embedded in application programs, rather than being stored separately and independently.
- No control over access and manipulation of data beyond that imposed by application programs.

Result:

- The database and Database Management System (DBMS).

DBMS Development

First-generation

- Hierarchical and Network

Second generation

- Relational

Third generation

- Object-Relational
- Object-Oriented

Functions of a DBMS

- Data storage, retrieval, and update.
- A user-accessible catalog.
- Transaction support.
- Concurrency control services.
- Recovery services.
- Authorization services.
- Support for data communication.
- Integrity services.
- Services to promote data independence.
- Utility services.

Pros and Cons of DBMSs

Advantages

- Control of data redundancy
- Data consistency
- Sharing of data
- Improved data integrity
- Improved maintenance through data independence.

Disadvantages

- Complexity
- Cost of DBMS
- Cost of conversion
- Performance
- Higher impact of a failure

What is a Data Model

Integrated collection of concepts for describing data, relationships between data, and constraints on the data used by an organization.

Has three components:

- a structural part;
- a manipulative part;
- a set of integrity rules.

Relational Model Terminology

Relation: table with columns and rows.

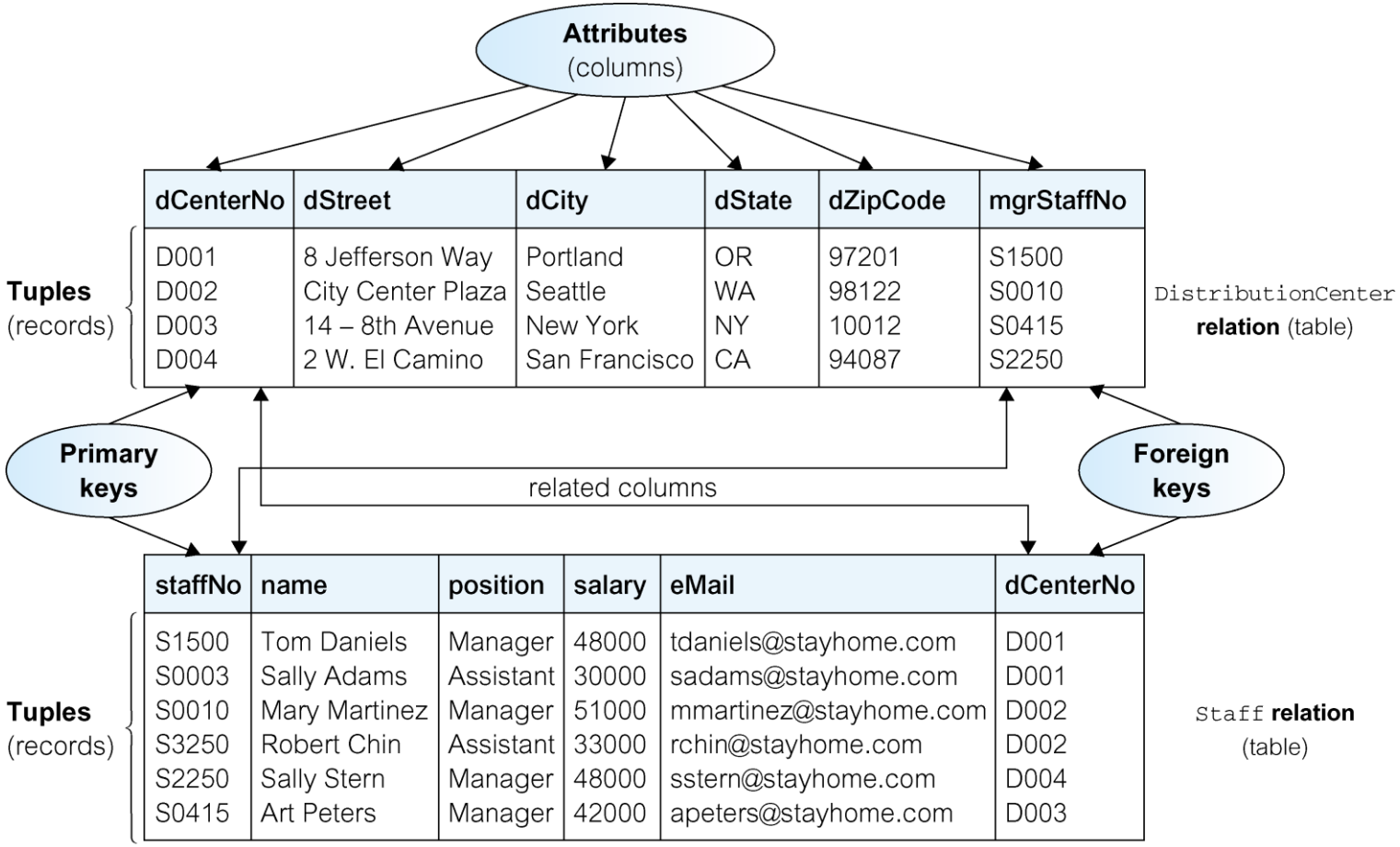
Attribute: named column of a relation.

Domain: set of allowable values for one or more attributes.

Tuple: a record of a relation.

Relational Database - collection of normalized relations with distinct relation names.

Instances of Distribution Center and Staff relations



Domains for some attributes of distribution Center and staff relations

Attribute	Domain name	Meaning	Domain definition
dCenterNo	DCenter_Numbers	Set of all possible distribution center numbers.	Character: size 4, range D001 – D999
dStreet	Street_Names	Set of all possible street names.	Character: size 60
staffNo	Staff_Numbers	Set of all possible staff numbers.	Character: size 5, range S0001 – S9999
position	Staff_Positions	Set of all possible staff positions.	Manager or Assistant
salary	Staff_Salaries	Possible values of staff salaries.	Monetary: 8 digits, range $10 000.00 – $100 000.00

Properties of Relational Tables

Table name is distinct from all other table names in the database.

Each cell of table contains exactly one atomic (single) value.

Each column has a distinct name.

Values of a column are all from the same domain.

Each record is distinct; there are no duplicate records.

Order of columns has no significance.

Order of records has no significance, theoretically.

Relational Keys

Superkey

- A column, or a set of columns, that uniquely identifies a record within a table.

Candidate Key

- A superkey that contains only the minimum number of columns necessary for unique identification.
- In each record, values of the candidate key uniquely identify that record (*uniqueness*).
- No proper subset of the candidate key has the uniqueness property (*irreducibility*).

Relational Keys

Composite Key

- A key consists of more than one column

Primary Key

- The candidate key that is selected to identify records uniquely within table.

Alternate Keys

- Candidate keys that are not selected to be primary key.

Foreign Key

- Column, or set of columns, within one table that matches the candidate key of some (possibly the same) table.

Relational Integrity

Nulls

- Represents value for a column that is currently unknown or not applicable for record.
- Deals with incomplete or exceptional data.
- Represents the absence of a value and is not the same as zero or spaces, which are values.

Relational Integrity

Entity Integrity

- In a base table, no column of a primary key can be null.

Referential Integrity

- If a foreign key exists in a table, either foreign key value must match a candidate key value of some record in its home table or foreign key value must be wholly null.

Integrity Constraints

- Rules that define or constrain some aspect of the data used by the organization.

Relational Languages

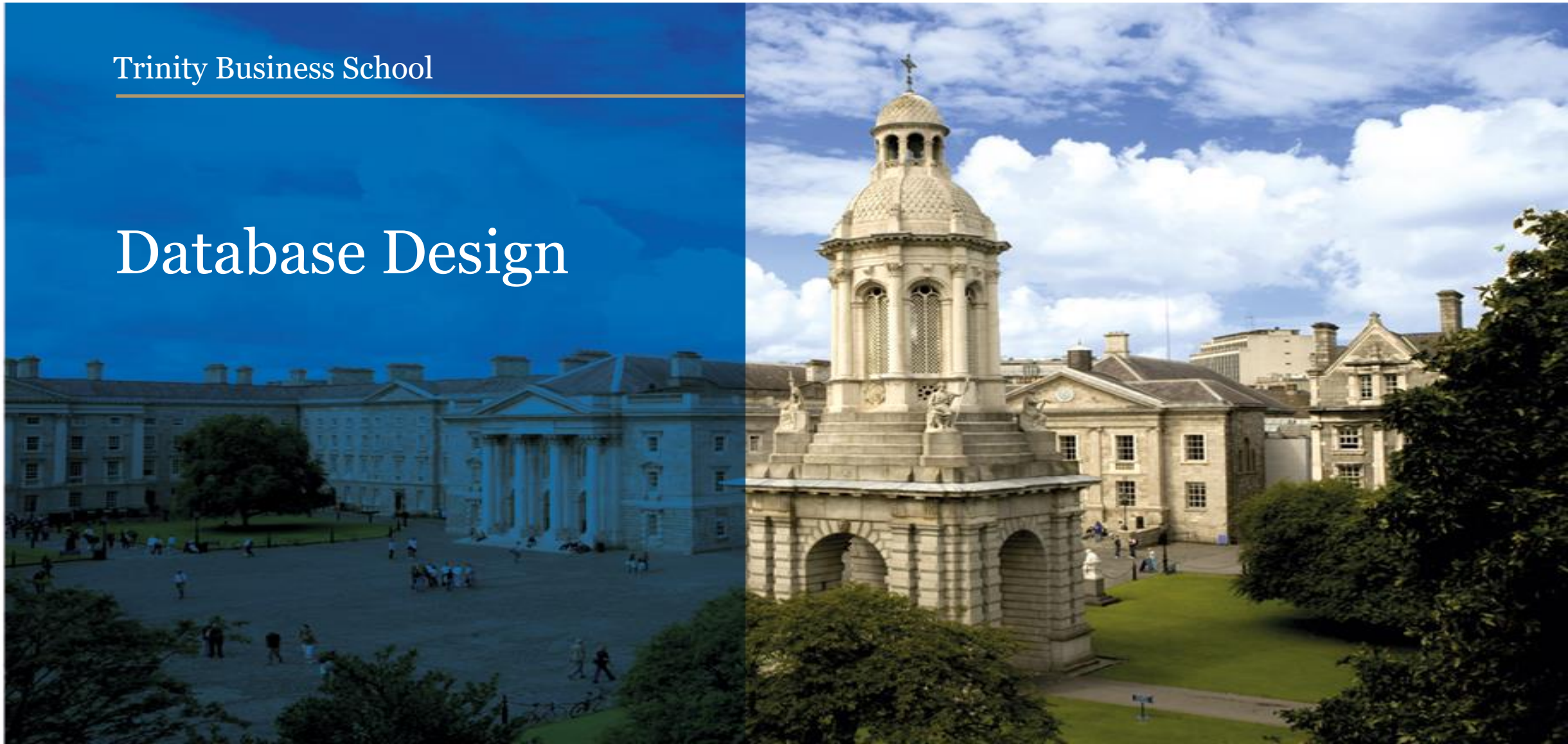
Two main languages are

- SQL (Structured Query Language)
- QBE (Query By Example)

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Trinity Business School

Database Design



Information system

Resources that enable collection, management, control, and dissemination of data/information throughout an organization.

Database is fundamental component of and Information Systems (IS). Development/usage should be viewed from perspective of the wider requirements of the organization.

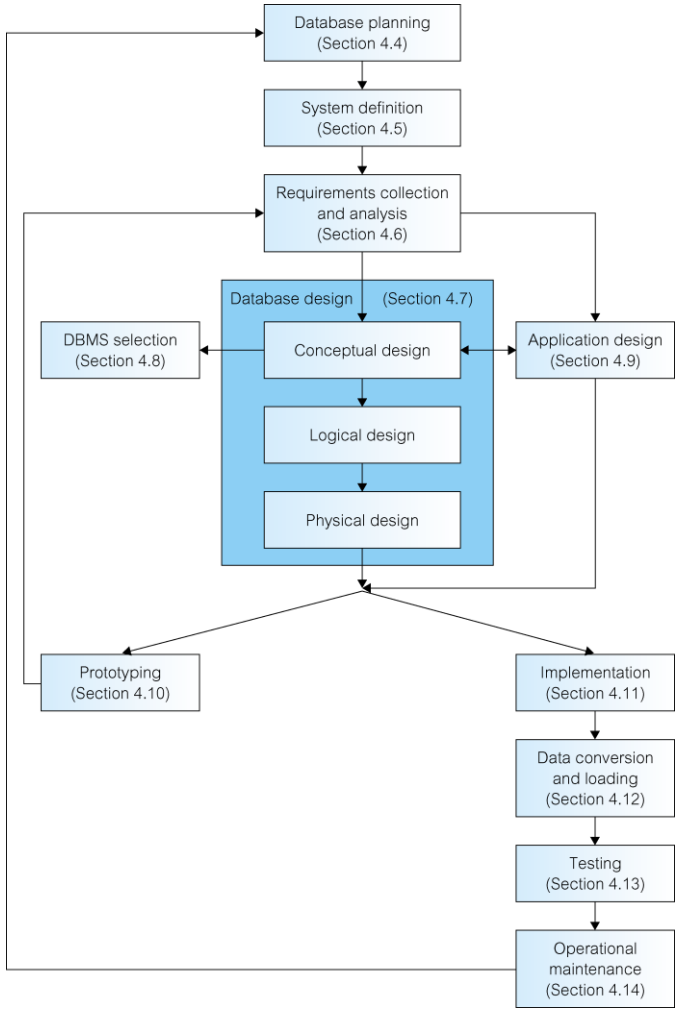
Structured approach to development of the database component of an IS is required.

Information system lifecycle

A structured approach to software development called information systems (IS) lifecycle or software development lifecycle (SDLC).

- Waterfall
- Spiral
- Agile

Database system development lifecycle - stages



Stages of database system development lifecycle

Database planning

System definition

Requirements collection and analysis

Database design

DBMS selection (optional)

Stages of database system development lifecycle

Application design

Prototyping (optional)

Implementation

Data conversion and loading

Testing

Operational maintenance.

Database planning

Management activities that allow stages of database system development lifecycle to be realized as efficiently and effectively as possible.

Should be integrated with overall IS strategy of the organization.

Includes creation of the mission statement and mission objectives for the database system.

Mission statement

Those driving database project normally define the mission statement.

Defines major aims of database system.

Helps clarify purpose of the database system and provides clearer path towards the efficient and effective creation of required database system.

Mission objectives

Once mission statement is defined, mission objectives are defined.

Each objective should identify a particular task that the database system must support.

Should also include additional information that specifies the work to be done, the resources with which to do it, and the money to pay for it all.

Database planning

Database planning may also include development of standards that govern:

- how data will be collected,
- how the format should be specified,
- what necessary documentation will be needed,
- how design and implementation should proceed.

System definition

Describes scope and boundaries of database system, including its major user views.

Describes how database system will interface with other parts of the organization's information system.

System definition

User view defines what is required of a database system from the perspective of:

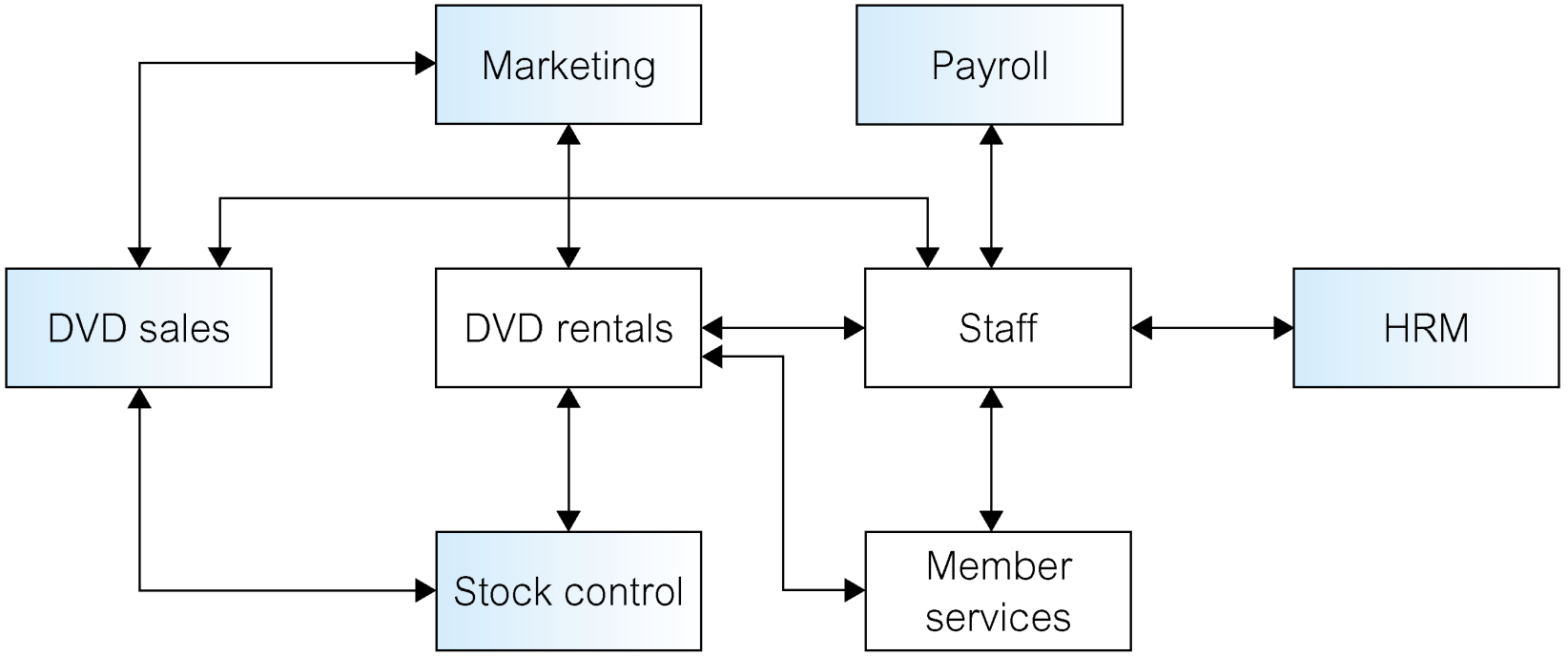
- a particular job (such as Manager or Supervisor) or
- business application area (such as marketing, personnel, or stock control).
- Database system may have one or more user views.

System definition

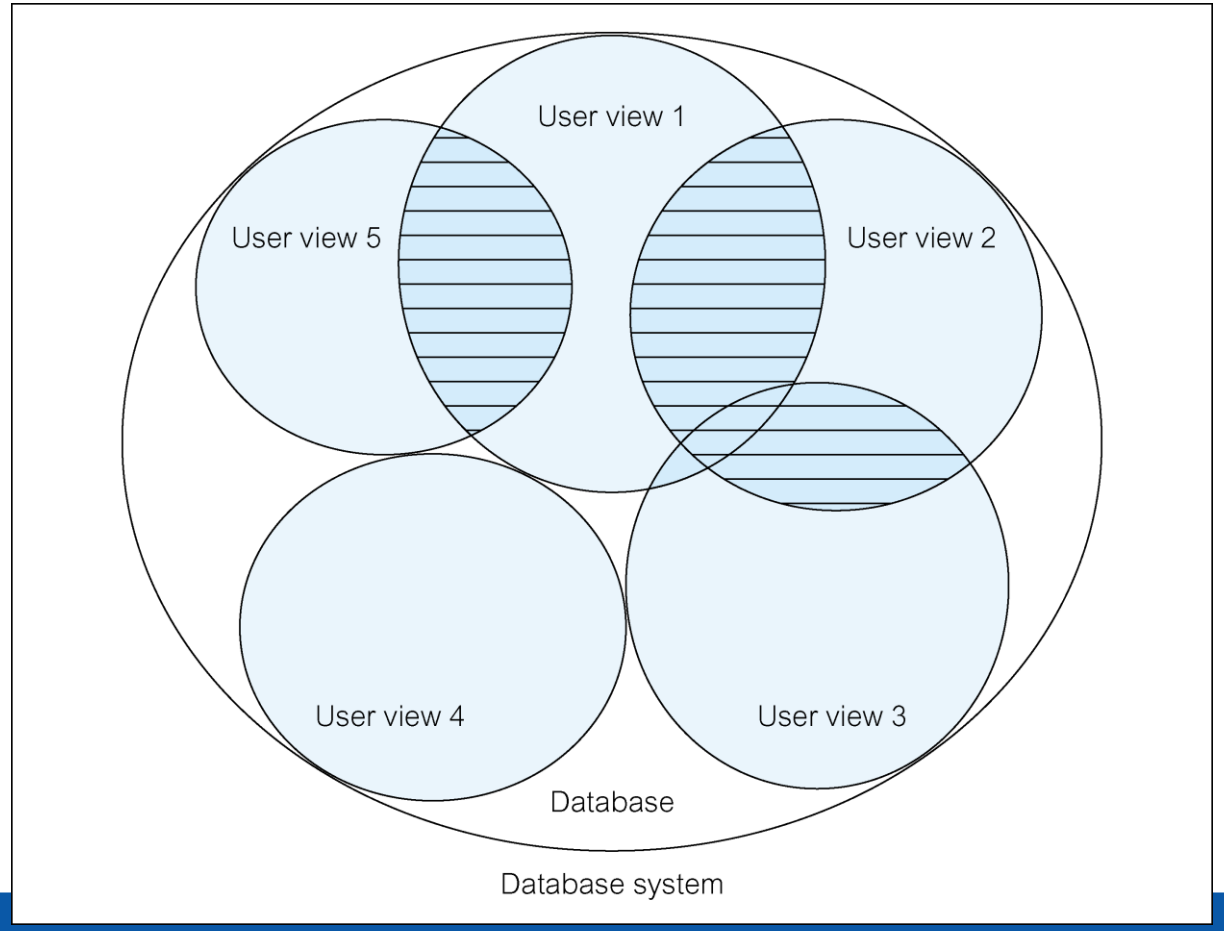
Identifying user views helps ensure that no major users of the database are forgotten when developing requirements for new application.

User views also help in development of complex database system allowing requirements to be broken down into manageable pieces.

Extended version of the *StayHome Online case study*



Database system with multiple user views



Requirements collection and analysis

Process of collecting and analyzing information about the organization to be supported by the database system, and using this information to identify the requirements for the new system.

Information is gathered for each major user view including:

- a description of data used or generated;
- details of how data is to be used/generated;
- any additional requirements for new database system.

Requirements collection and analysis

Information is analyzed to identify requirements for new database system.

Another important activity is deciding how to manage database system with multiple user views.

Three main approaches:

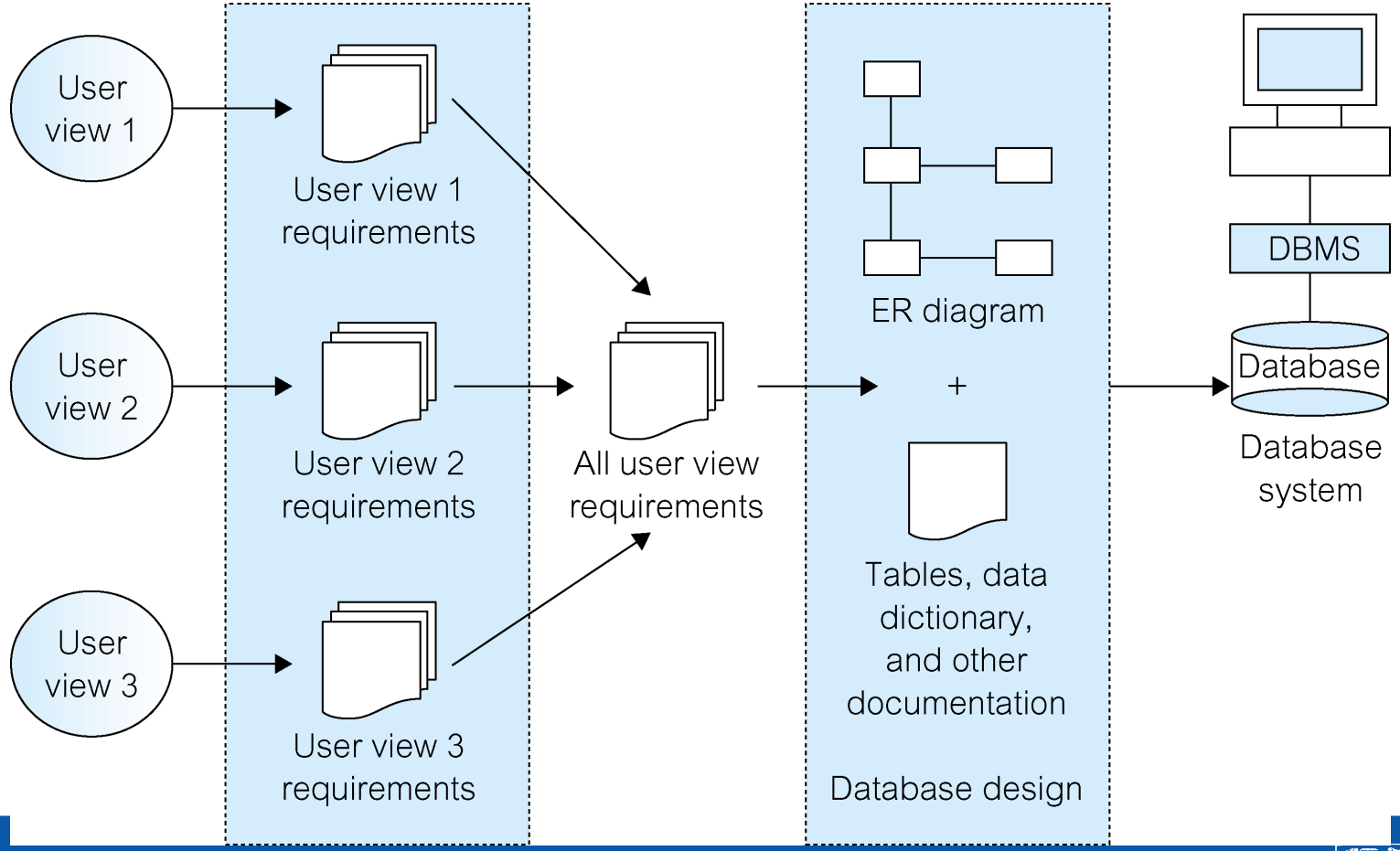
- centralized approach;
- view integration approach;
- combination of both approaches.

Centralized approach

Requirements for each user view are merged into a single set of requirements for the new database system.

A data model representing all user views is created during the database design stage.

Centralized approach



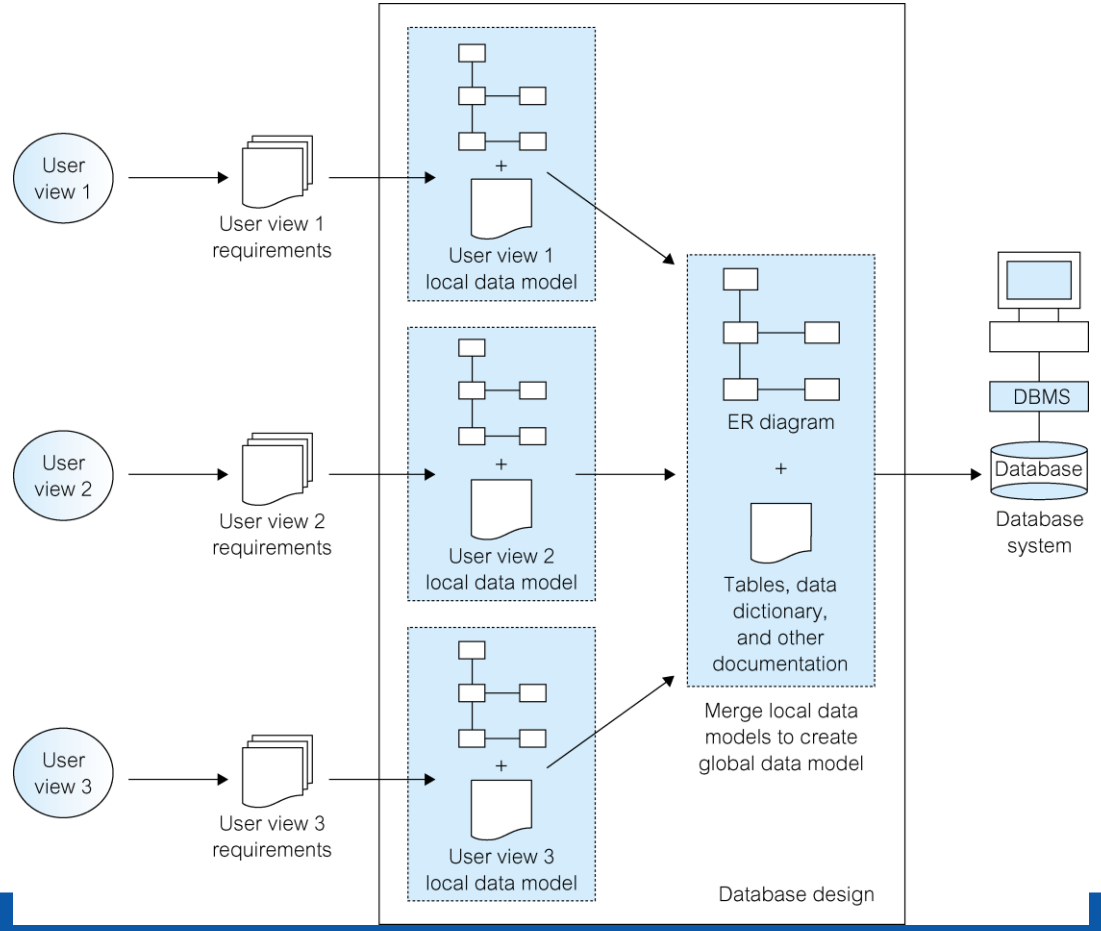
View integration approach

Requirements for each user view remain as separate lists. Data models representing each user view are created and then merged during the database design stage.

Data model representing one or more but not all user views is called a local data model.

Local data models are then merged to produce a global data model to represent all user views.

View integration approach



Database design

Process of creating a design that will support the organization's mission statement and objectives for the required database system.

Three main phases of database design:

- conceptual database design,
- logical database design,
- physical database design.

DBMS selection

Selection of an appropriate DBMS to support the database system.

Undertaken at any time prior to logical design provided sufficient information is available regarding system requirements.

Application design

Design of user interface and application programs that use and process the database.

Database and application design are parallel activities.

Transaction is an action, or series of actions, carried out by a single user or application program that accesses or changes content of the database.

Should define and document the high-level characteristics of the transactions required.

Application design

Important characteristics of transactions:

- data to be used by the transaction;
- functional characteristics of the transaction;
- output of the transaction;
- importance to the users;
- Expected rate of usage.

Three main types of transactions:

- retrieval transactions
- update transactions
- mixed transactions

Guidelines for form/report design

- Meaningful title
- Comprehensible instructions
- Logical grouping and sequencing of fields
- Visually appealing layout of the form/report
- Familiar field labels
- Consistent terminology and abbreviations
- Consistent use of color
- Visible space and boundaries for data-entry fields
- Convenient cursor movement
- Error correction for individual characters and entire fields
- Error messages for unacceptable values
- Optional fields marked clearly
- Explanatory messages for fields
- Completion signal

Prototyping

Building working model of a database system.

Purpose is to:

- to identify features of a system that work well, or are inadequate;
- to suggest improvements or even new features;
- to clarify the users' requirements;
- to evaluate feasibility of a particular system design.

Prototyping

There are two prototyping strategies:

- Requirements prototyping determines the requirements of a proposed database system and then the prototype is discarded.
- Evolutionary prototyping is used for the same purposes, but the prototype is not discarded and with further development becomes the working database system.

Implementation

Physical realization of the database and application designs.

Use DDL to create database schemas and empty database files.

Use DDL to create user views.

Use 3GL or 4GL to create the application programs, which includes database transactions.

Use DDL to implement security and integrity controls. However, some may be defined using DBMS utilities or operating system.

Data conversion and loading

Transferring any existing data into new database and converting any existing applications to run on new database.

only required when a new database system is replacing an old system.

common for a DBMS to have a utility that loads existing files into the new database.

May be possible to convert and use application programs from the old system for use by the new system.

Testing

Process of running the database system with the intent of finding errors.

Use carefully planned test strategies and realistic data.

Testing cannot show absence of faults; it can show only that software faults are present.

Demonstrates that database and application programs appear to be working according to requirements.

Operational maintenance

Process of monitoring and maintaining the database system following installation and involves:

- monitoring performance of system. If performance falls, may require tuning or reorganization of the database.
- maintaining and upgrading database system (when required).
- incorporating new requirements into database system.