Databases, Data Warehousing and Information Retrieval

Grigorios Loukides, PhD
Email: grigorios.loukides@kcl.ac.uk

Contact Details

- **Lecturer:** Dr Grigorios Loukides
- Room: (N)5.13, Bush House
- Contact Hours: Thursday 12.30-14.30
- Email: grigorios.loukides@kcl.ac.uk

- **TAs**: Diego Sempreboni, Akkapon Wongkoblap, Parvin Sadigova
- Emails: (firstname.lastname@kcl.ac.uk)

How to Get Help

General Questions about the Module/Clarifications

1. Ask questions in the lectures, labs and tutorials
2. Ask questions in the Question Forum on KEATS

Personal Questions

1. See me during my office hours
2. Send me an email

Module Overview

Three parts

- Database analysis and design (Entity Relationship Model, Normalization)
- Database implementation using SQL (Structured Query Language)
- Advanced topics: Data warehouses, Information Retrieval and no-SQL databases

Module Assessment

- Written Examination (January) 80%
- 2 Individual Courseworks:
  - Database Implementation (10%)
    Hand Out: 12 October
    Hand In: 16 November
  - Database Design and Optimization (10%)
    Hand Out: 12 November
    Hand In: 10 December

Expectations

- Attending a lecture is more than watching it online if you do not attend, you miss out!
  - Lecture recordings are a study and revision aid.
  - Watching lectures online is NOT a replacement for attending lectures.
  - Statistically, there is a clear and direct link between attendance and attainment: Students who do not attend lectures do less well in exams.

Expectations of inclusive behaviour

- The Department of Informatics is committed to providing an inclusive learning and working environment.
- Staff and students are expected to behave respectfully to one another - during lectures, outside of lectures and when communicating online or through email.
- We won't tolerate inappropriate or demeaning comments related to gender, gender identity and expression, sexual orientation, disability, physical appearance, race, religion, age, or any other personal characteristic.
- If you witness or experience any behaviour you are concerned about, please speak to someone about it. This could be one of your lecturers, your personal tutor, a programme administrator, the Informatics equality & diversity lead (Elizabeth Black), or any other member of staff you feel comfortable talking to.
- The College also has a range of different support and reporting procedures that you might find helpful: `kcl.ac.uk/harassment`

Principal Objectives

This module will allow you store data systematically employing modern database technologies, and will equip you with fundamental understanding and skills to independently study advanced data warehousing and information retrieval solutions.

Teaching and Learning Methods

Weekly teaching arrangements:

- Lecture + Tutorial (3 hours per week)
- Practical (2 hours per week)

Suggested Books

- Introduction to Database Systems (8th edition), by C.J. Date, Pearson Publishing, 2003.
- Fundamentals of Database Systems (7th edition), by Ramez Elmasri and Shamkant B. Navathe, Pearson Publishing, 2015.
- Data Warehousing in the Age of Big Data (1st edition), by Krish Krishnan, O'Reilly Media, 2013.
- Introduction to Information Retrieval (1st edition), by Christopher D. Manning, Prabhakar Raghavan and Hinrich Schtze, Cambridge University Press, 2008.

Session Objectives

In this session, you will learn:

- The difference between data and information
- What a database is, the various types of databases, and why they are valuable assets for data science
- How modern databases evolved from file systems
- The main components of the database system
- The main functions of a database management system (DBMS)

Why Databases?

- Databases solve many of the problems encountered in data management
- Used in almost all modern settings involving data management:
  - Business
  - Research
  - Administration
- Important to understand how databases work and interact with other applications

Exercise

- Find examples of situations in which you interact with databases on a daily basis

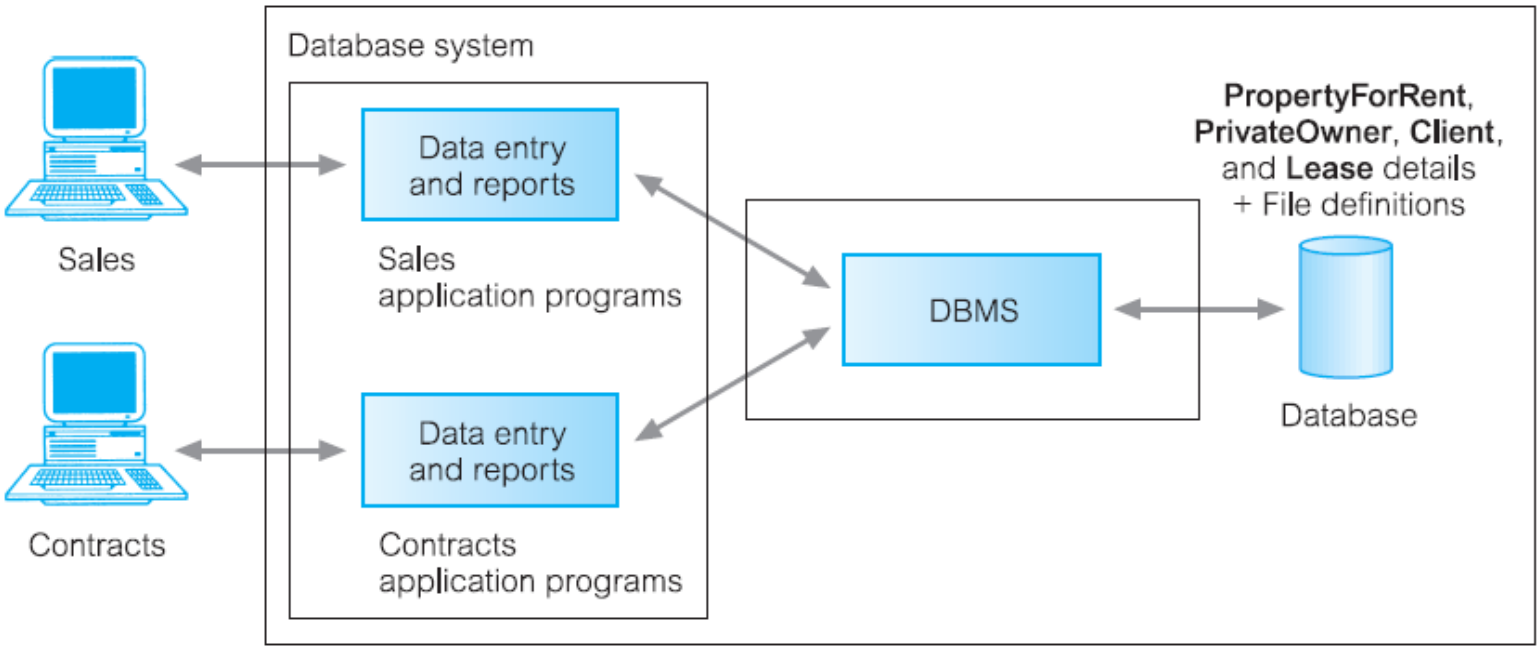
Data vs. Information

- Data are raw facts
- Information is the result of processing raw data to reveal meaning
  - Data: building blocks of information
- Information requires context to reveal meaning
- Data are the foundation of information, which is the bedrock of knowledge

Introducing the Database

- Database: shared, integrated computer structure that stores a collection of:
  - End-user data: raw facts of interest to end user
  - Metadata: data about data
    - Provides description of data characteristics and relationships in data
    - Complements and expands value of data
- Database management system (DBMS): collection of programs
  - Manages structure and controls access to data

Database Environment



File Systems

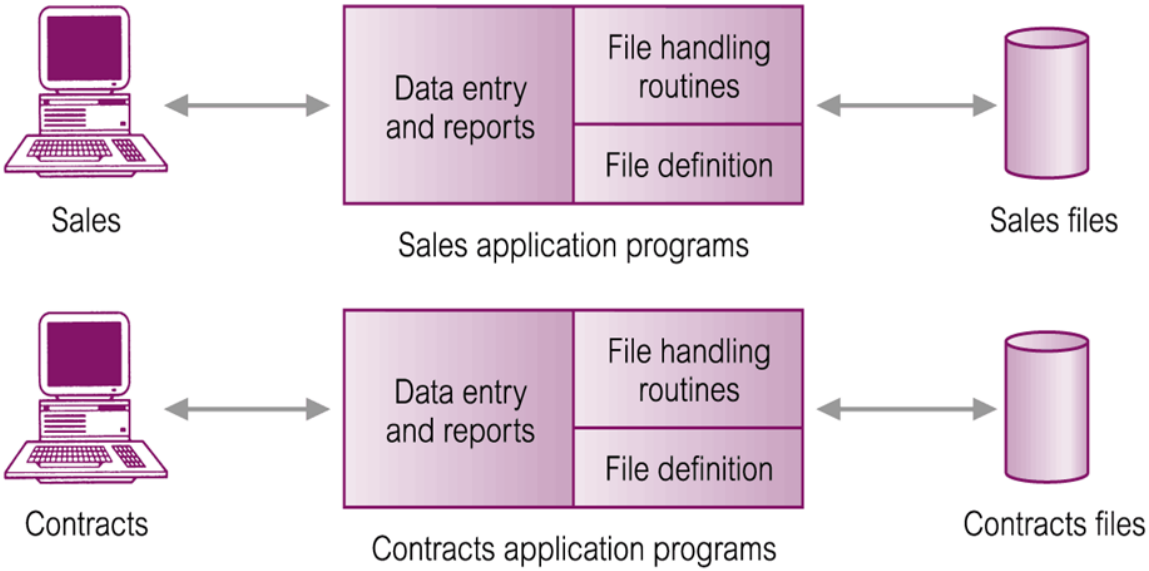


Figure 1.5 File-based processing.

Sales Files

PropertyForRent (propertyNo, street, city, postcode, type, rooms, rent, ownerNo)
PrivateOwner (ownerNo, fName, lName, address, telNo)
Client (clientNo, fName, lName, address, telNo, prefType, maxRent)

Contracts Files

Lease (leaseNo, propertyNo, clientNo, rent, paymentMethod, deposit, paid, rentStart, rentFinish, duration)
PropertyForRent (propertyNo, street, city, postcode, rent)
Client (clientNo, fName, lName, address, telNo)

Each program maintains its own set of data

Limitations of File Systems

- Separation and isolation of data
  - Users of one program may be unaware of potentially useful data held by other programs
- Duplication of data
- Data dependence (File structure is defined in the program code)
- Incompatible file formats
- Fixed Queries/Proliferation of application programs
- ...

Exercise

PROJ_NUM	PROJ_NAME	EMP_NUM	EMP_NAME	JOB_CODE	JOB_CHG_HOUR	PROJ_HOURS	EMP_PHONE
1	Hurricane	101	John D. Newson	EE	$85.00	13.3	653-234-3245
1	Hurricane	105	David F. Schwann	CT	$60.00	16.2	653-234-1123
1	Hurricane	110	Anne R. Ramoras	CT	$60.00	14.3	615-233-5568
2	Coast	101	John D. Newson	EE	$85.00	19.8	653-234-3254
2	Coast	108	June H. Sattlemeir	EE	$85.00	17.5	905-554-7812
3	Satellite	110	Anne R. Ramoras	CT	$62.00	11.6	615-233-5568
3	Satellite	105	David F. Schwann	CT	$26.00	23.4	653-234-1123
3	Satelite	123	Mary D. Chen	EE	$85.00	19.1	615-233-5432
3	Satellite	112	Allecia R. Smith	BE	$85.00	20.7	615-678-6879

What data redundancies do you detect? How could those redundancies lead to anomalies?

Database Systems

Database system consists of logically related data stored in a single logical data repository

- May be physically distributed among multiple storage facilities
- Eliminates most of file systems problems

Role of the DBMS

DBMS is the intermediary between the applications and the database. It enables:

- Defining (describing the structure)
- Constructing (populating by data)
- Manipulating (querying, updating)
- Preserving consistency
- Protecting from misuse (security, authentication)
- Recovering from failure
- Concurrent usage of a database

Types of Databases

- Number of users
- Database location(s)
- Expected type and extent of use

Number of Users

- Single-user database supports only one user at a time
  - Desktop database: single-user; runs on PC
- Multiuser database supports multiple users at the same time
  - Workgroup and enterprise databases

Location

- Centralized database: data located at a single site
- Distributed database: data distributed across several different sites

Usage

- Operational database: supports a companys day-to-day operations
  - Transactional or production database
- Data warehouse: stores data used for tactical or strategic decisions

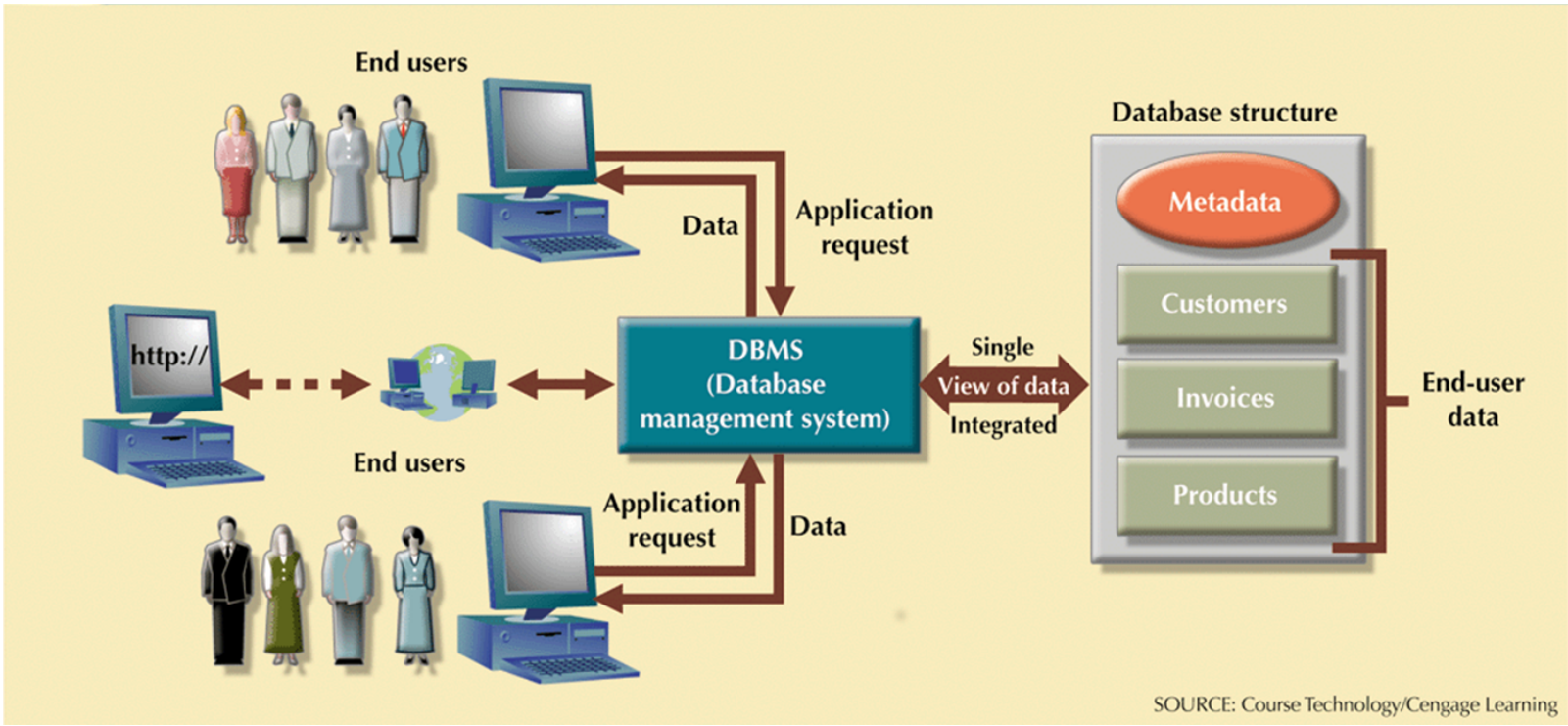
Types of Databases

PRODUCT	NUMBER OF USERS			DATA LOCATION		DATA USAGE		XML
	SINGLE USER	MULTIUSER						
		WORKGROUP	ENTERPRISE	CENTRALIZED	DISTRIBUTED	OPERATIONAL	ANALYTICAL	
MS Access	X	X		X		X		
MS SQL Server	X[3]	X	X	X	X	X	X	X
IBM DB2	X[3]	X	X	X	X	X	X	X
MySQL	X	X	X	X	X	X	X	X
Oracle RDBMS	X[3]	X	X	X	X	X	X	X

Database Languages

- **Data Definition Language (DDL)** used to specify the database structure
- **Data Manipulation Language (DML)** used to both read and update the database:
  - The part of a DML that involves data retrieval is called a query language

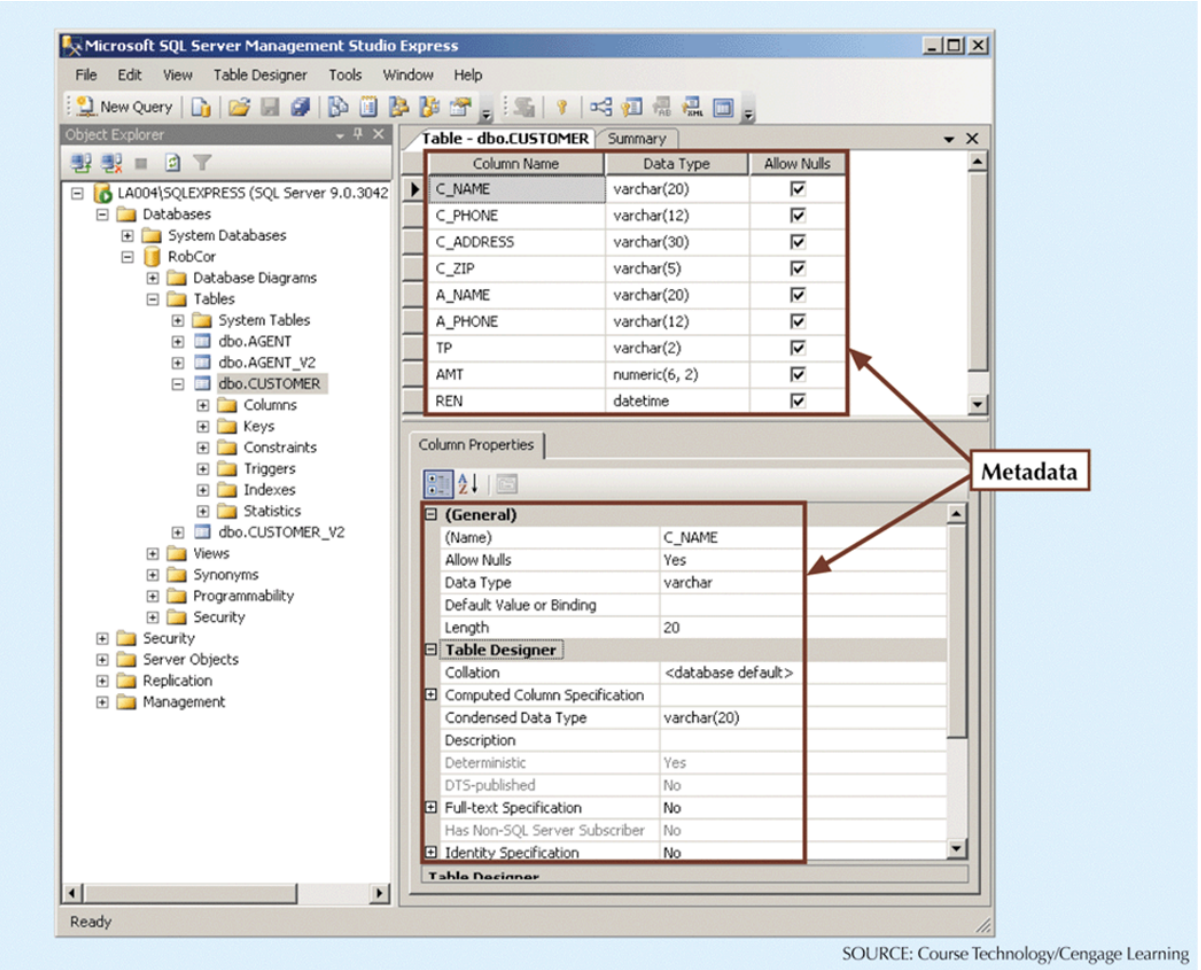
Database management system (DBMS)



DBMS Functions I

- Most functions are transparent to end users
  - Can only be achieved through the DBMS
- Data dictionary management
  - DBMS stores definitions of data elements and relationships (metadata) in a data dictionary
  - DBMS looks up required data component structures and relationships
  - Changes automatically recorded in the dictionary
  - DBMS provides data abstraction and removes structural and data dependency

DBMS Functions I

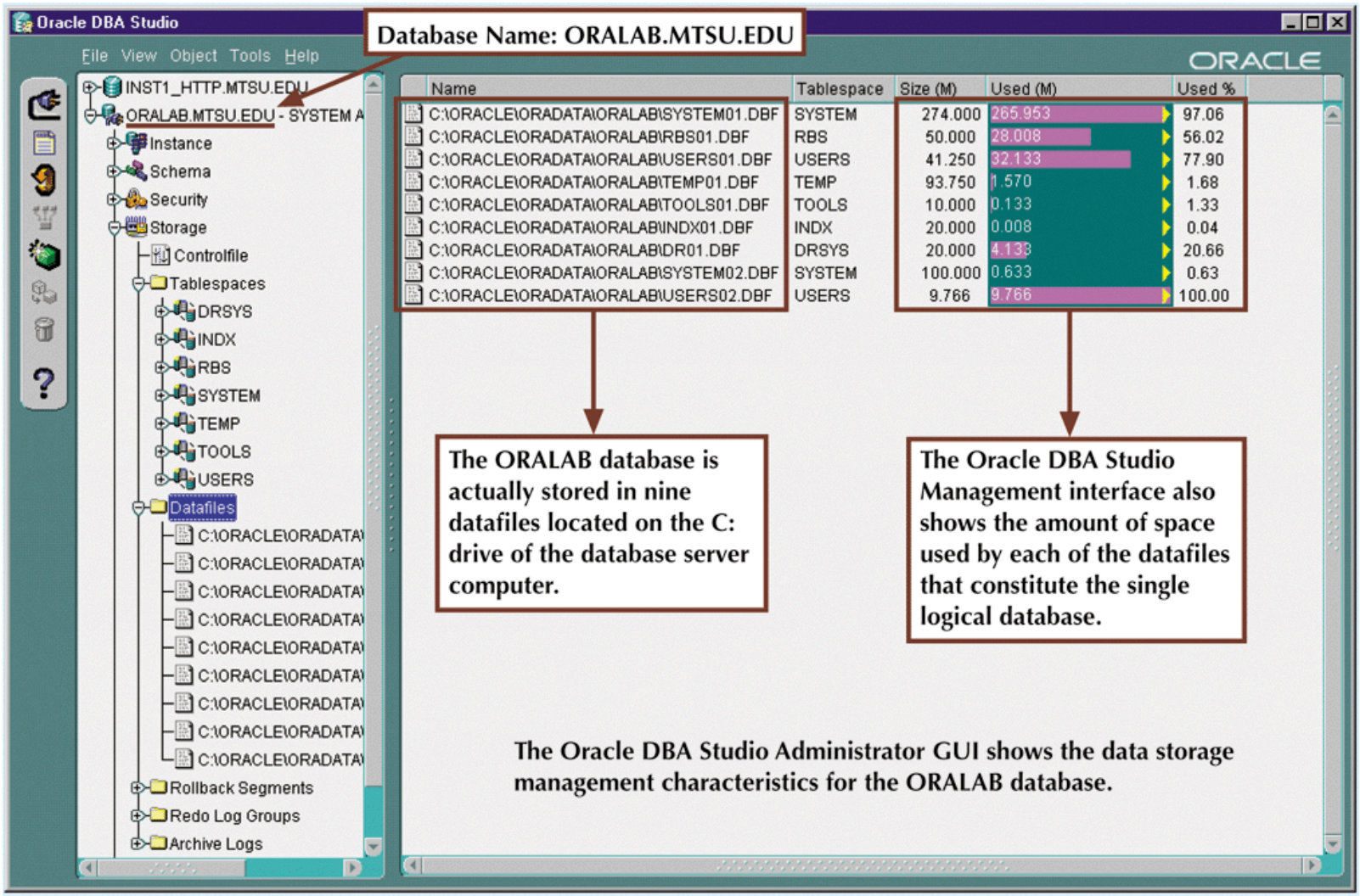


SOURCE: Course Technology/Cengage Learning

DBMS Functions II

- Data storage management
  - DBMS creates and manages complex structures required for data storage
  - Also stores related data entry forms, screen definitions, report definitions, etc.
  - Performance tuning: activities that make the database perform more efficiently
  - DBMS stores the database in multiple physical data files

DBMS Functions II



DBMS Functions III

- Data transformation and presentation
  - DBMS transforms data entered to conform to required data structures
  - DBMS transforms physically retrieved data to conform to users logical expectations
- Security management
  - DBMS creates a security system that enforces user security and data privacy
  - Security rules determine which users can access the database, which items can be accessed, etc.

DBMS Functions IV

- Multiuser access control
  - DBMS uses sophisticated algorithms to ensure concurrent access does not affect integrity
- Backup and recovery management
  - DBMS provides backup and data recovery to ensure data safety and integrity
  - Recovery management deals with recovery of database after a failure
    - Critical to preserving databases integrity

DBMS Functions V

- Data integrity management
  - DBMS promotes and enforces integrity rules
    - Minimizes redundancy
    - Maximizes consistency
  - Data relationships stored in data dictionary used to enforce data integrity
  - Integrity is especially important in transaction-oriented database systems

DBMS Functions VI

- Database access languages and application programming interfaces
  - DBMS provides access through a query language
  - Query language is a nonprocedural language
  - Structured Query Language (SQL) is the de facto query language
    - Standard supported by majority of DBMS vendors

DBMS Functions VII

- Database communication interfaces
  - Current DBMSs accept end-user requests via multiple different network environments
  - Communications accomplished in several ways:
    - End users generate answers to queries by filling in screen forms through Web browser
    - DBMS automatically publishes predefined reports on a Web site
    - DBMS connects to third-party systems to distribute information via e-mail

Exercise

How do you convince a group of friends who run a small business using a file-based approach data management that they should manage their business data using database technology?

Advantages of database systems

- Improved data sharing
- Improved data security
- Better data integration
- Minimized data inconsistency
- Improved data access
- Improved decision making
- Increased end-user productivity

Disadvantages of database systems

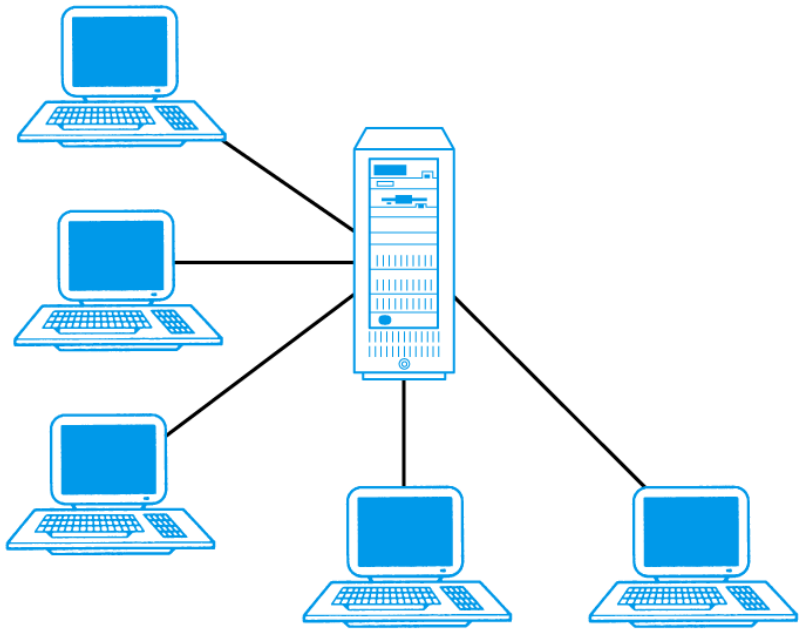
- Increased costs
- Complexity
- Vendor dependence
- Frequent upgrade/replacement cycles

Multi-User DBMS Architectures

- Tele-processing
- File Server
- Two-Tier Client-Server
- Three-Tier Client-Server

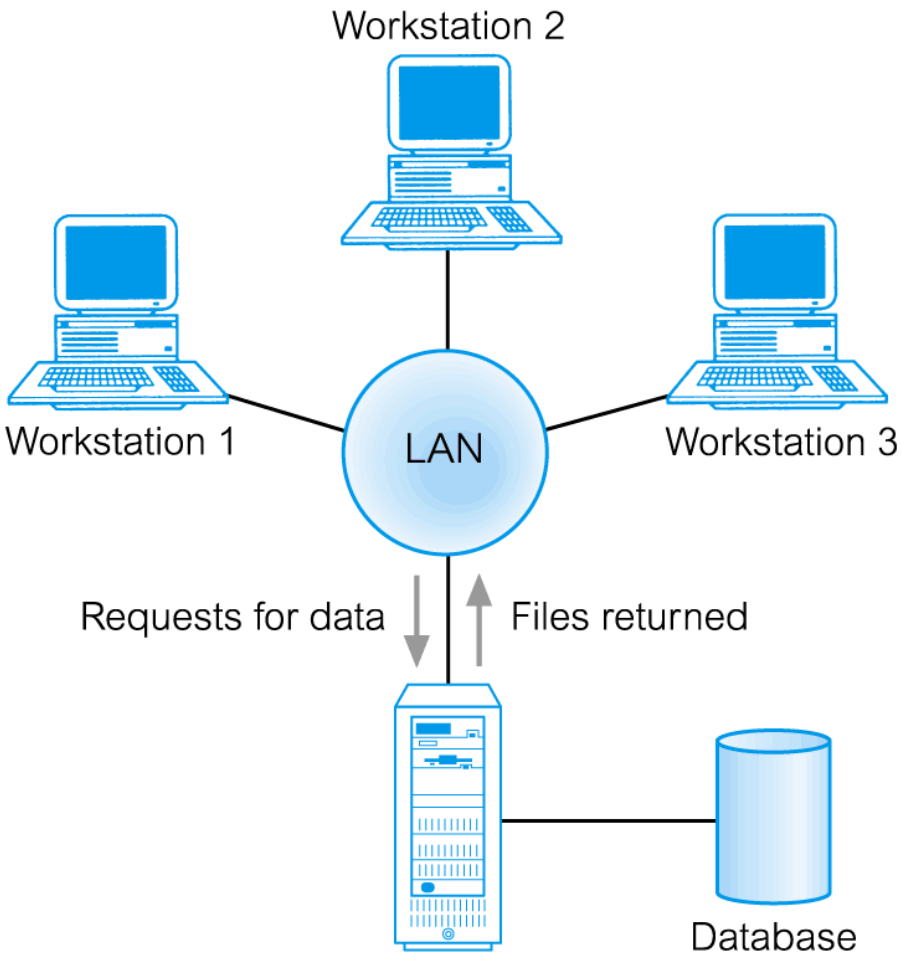
Tele-processing

Single mainframe with a number of terminals attached



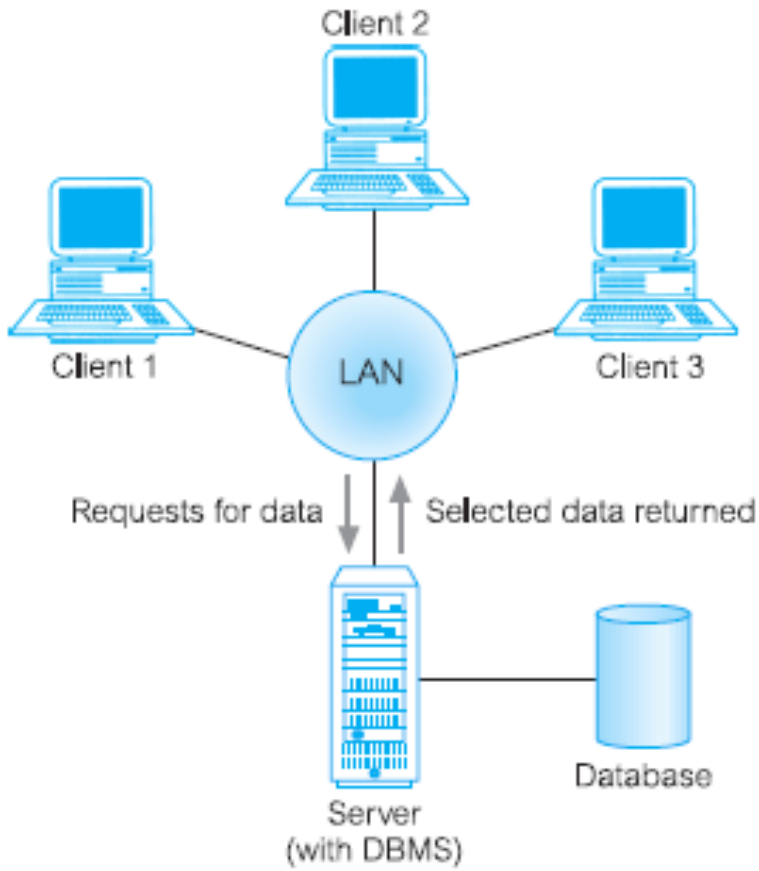
File Server

- Database resides on file-server.
- DBMS and applications run on each workstation



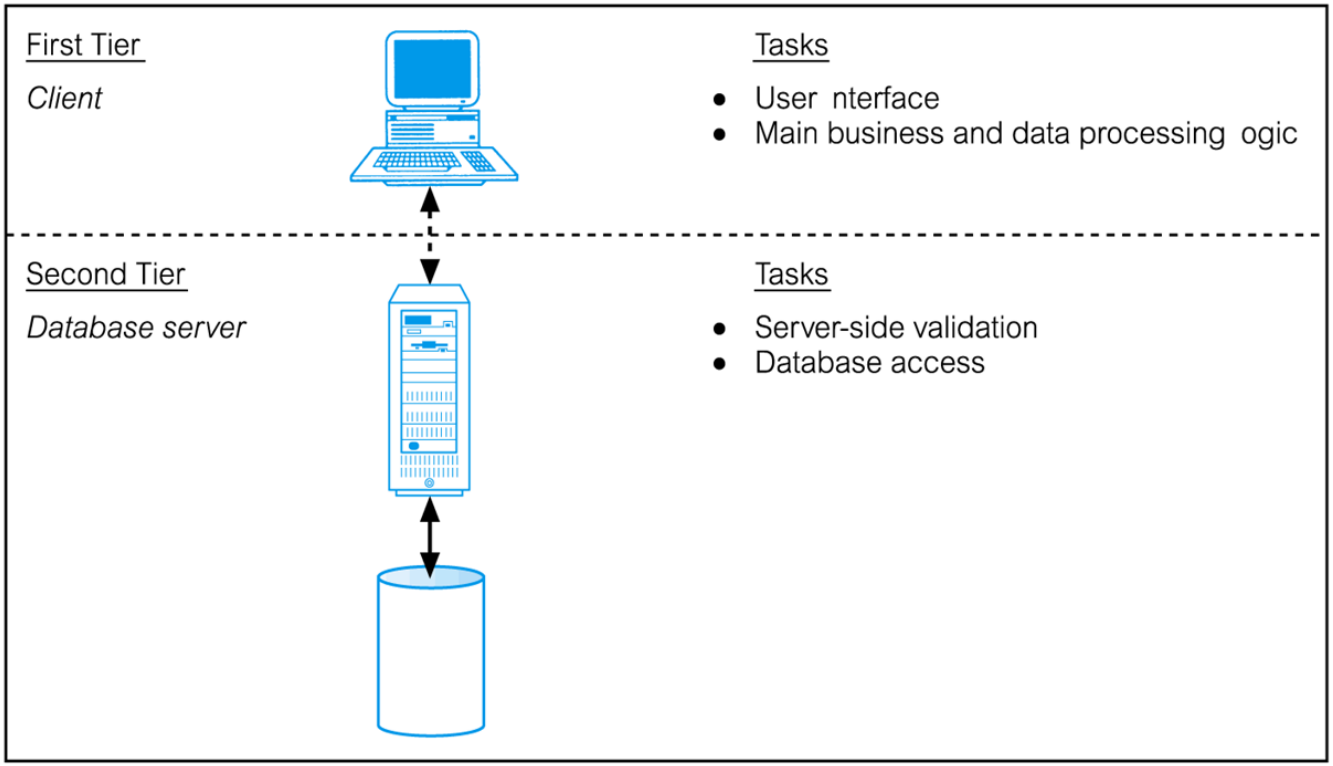
Client-Server Architecture

- Client-server refers to the way in which software components interact to form a system.
- A client process requires some resource, and a server provides the resource



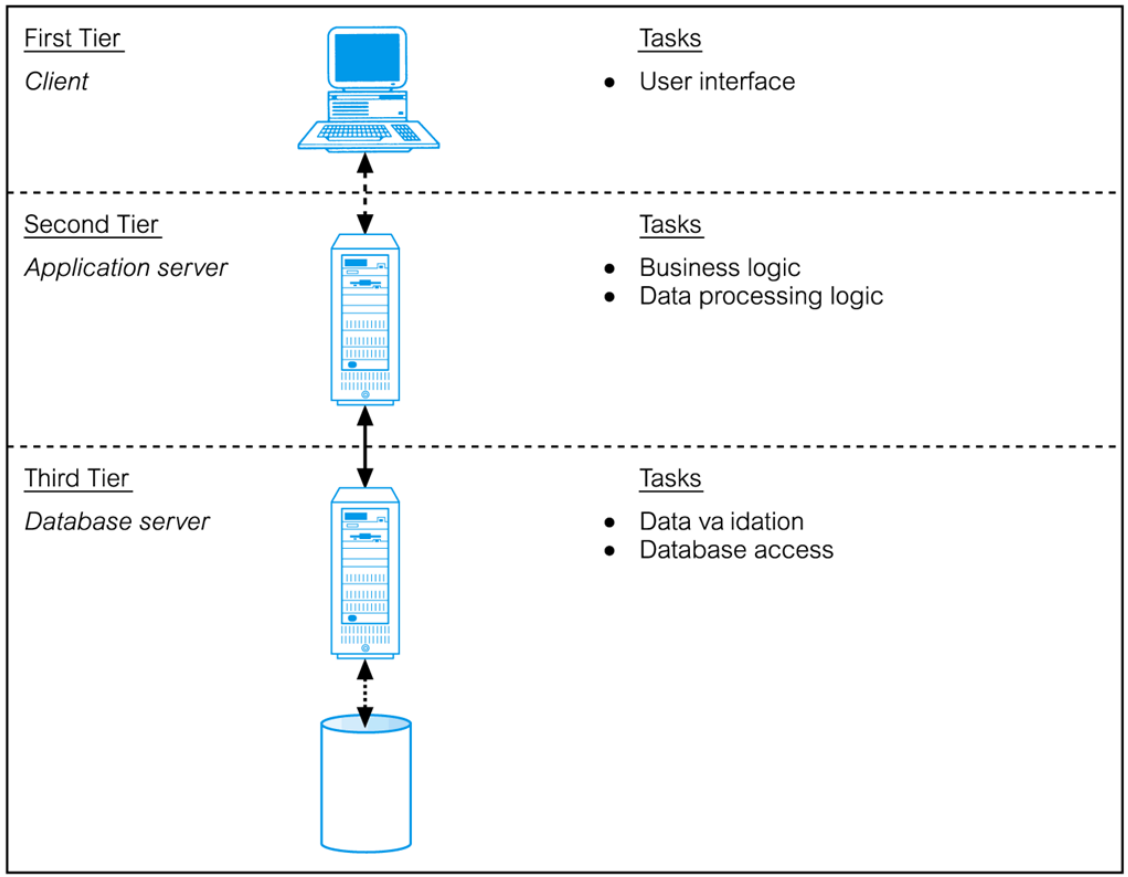
Two-Tier Client-Server

- Client (tier 1) manages user interface and runs applications
- Server (tier 2) holds database and DBMS



Three-Tier Client-Server

- Thin Client (tier 1) manages user interface
- Application Server (tier 2) runs applications
- Database Server (tier 3) holds database and DBMS



Data Models

- Data model is an abstraction
- Data models:
  - Relatively simple representations of complex real-world data structures
  - Often graphical

Evolution of Data Models

GENERATION	TIME	DATA MODEL	EXAMPLES	COMMENTS
First	1960s–1970s	File system	VMS/VSAM	Used mainly on IBM mainframe systems Managed records, not relationships
Second	1970s	Hierarchical and network	IMS, ADABAS, IDS-II	Early database systems Navigational access
Third	Mid-1970s	Relational	DB2 Oracle MS SQL Server MySQL	Conceptual simplicity Entity relationship (ER) modeling and support for relational data modeling
Fourth	Mid-1980s	Object-oriented Object/ relational (O/R)	Versant Objectivity/DB DB2 UDB Oracle 11g	Object/relational supports object data types Star Schema support for data warehousing Web databases become common
Fifth	Mid-1990s	XML Hybrid DBMS	dbXML Tamino DB2 UDB Oracle 11g MS SQL Server	Unstructured data support O/R model supports XML documents Hybrid DBMS adds object front end to relational databases Support large databases (terabyte size)
Emerging Models: NoSQL	Late 2000s to present	Key-value store Column store	SimpleDB (Amazon) BigTable (Google) Cassandra (Apache)	Distributed, highly scalable High performance, fault tolerant Very large storage (petabytes) Suited for sparse data Proprietary API

Conclusion

In this session we have covered:

- File Systems
- Database
  - Definition
  - Types
  - Languages
  - DBMS

MySQL and MariaDB

MySQL

- Is a DBMS
- Is relational (more about this next week)
- Was acquired by Oracle

MariaDB

- Is a drop-in replacement for MySQL
- Is a community-developed fork of MySQL

Names MySQL and MariaDB are used interchangeably

Linux

Linux is a Unix-like computer operating system assembled under the model of free and open-source software development and distribution

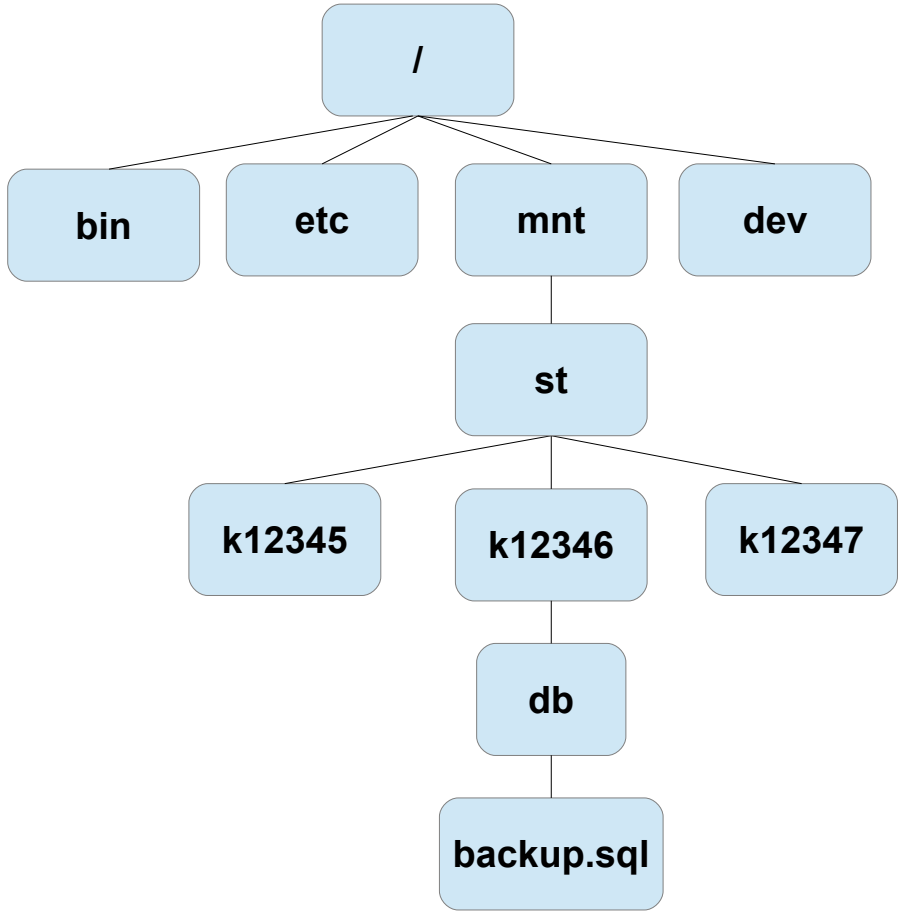
- Linux has been used for many computing platforms
  - PC, PDA, Supercomputer,
- Not only character user interface but graphical user interface is available
- Commercial vendors moved in Linux itself to provide freely distributed code.

CentOS Linux

CentOS (from Community Enterprise Operating System) is a Linux distribution that attempts to provide a free, enterprise-class, community-supported computing platform

Directory Tree

When you log on the the Linux OS using your username you are automatically located in your home directory. Your current directory is your home directory



Suggested Readings

- Chapters 1 and 2 of Fundamentals of Database Systems. Elmasri & Navathe.
- Chapters 1 and 2 of Database systems: a practical approach to design, implementation, and management. Connolly, Thomas M; Begg, Carolyn