ECON485 Introduction to Database Systems

Lecture 01 – Data Management, Databases, Data Warehouses and Popular DBMS

Challenges in Data

- The amount of data increases exponentially over time
- Data are scattered throughout organizations
- Data are generated from multiple sources (internal, personal, external)
- New sources of data (e.g., blogs, podcasts, videocasts, and RFID tags and other wireless sensors)
- Data Degradation (e.g., customers move to new addresses, change their names, etc.)
- Data Rot
- Data security, quality, and integrity are critical
- Legal requirements change frequently and differ among countries and industries

New Approaches in Data

Streaming Data

Inter-organizational Data Sharing

Open Data

Outsourcing of Data Processing

Data Anonymization Services

Data Governance and Master Data Management

Data Governance is an approach to managing information across an entire organization involving a formal set of unambiguous rules for creating, collecting, handling, and protecting its information.

We have two concepts central to data governance.

Master Data Management is **a strategy for data governance** involving a process that spans all organizational business processes and applications providing companies with the ability to store, maintain, exchange, and synchronize a consistent, accurate, and timely for the company's master data.

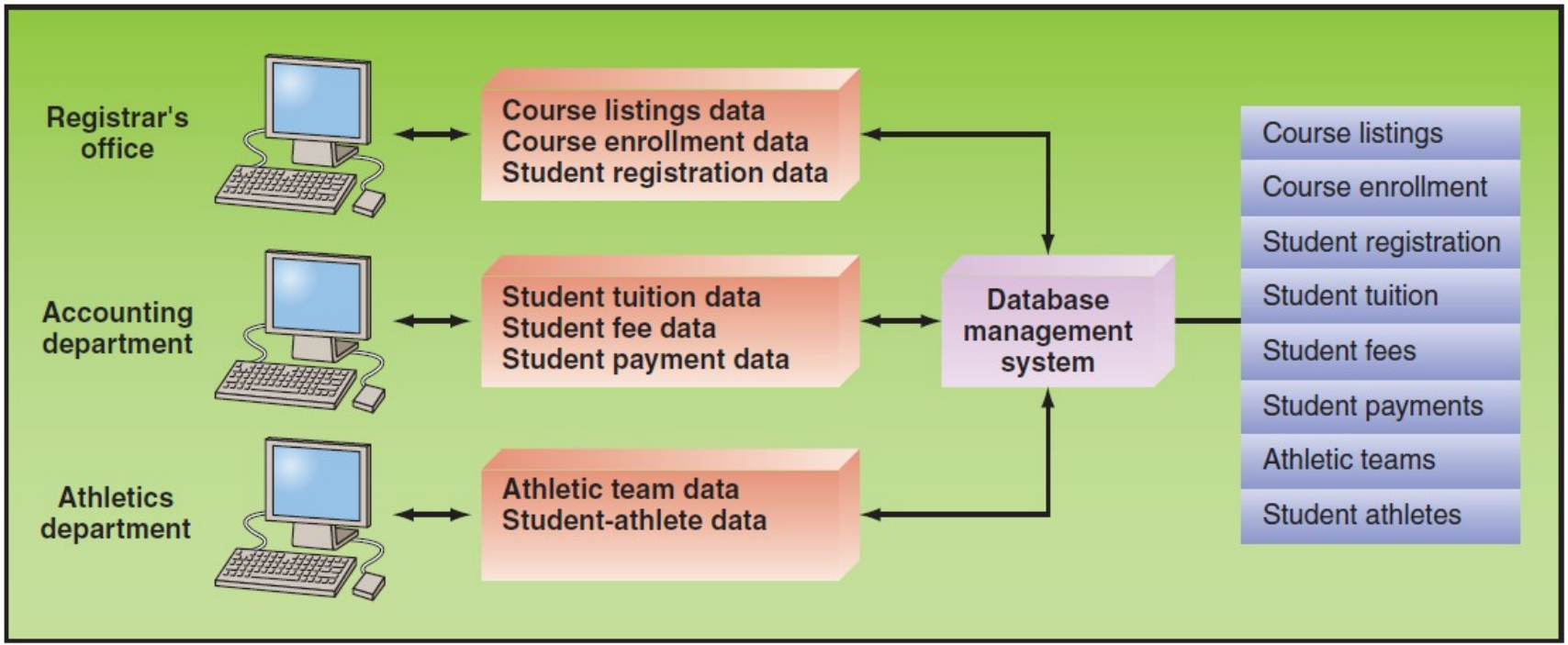
Master Data is a **set of core data** (e.g., customer, product, employee, vendor, geographic location, etc.) that span the enterprise information systems.

Master Data is considered a core asset of organizations because without this data set, it works the organization cannot function.

Discussion: Ransomware, how it works, when it works.

The Database Approach

A data file is a collection of logically related records.



The Database Approach

Database Management Systems (DBMS) Minimize:

- Data Redundancy
- Data Isolation
- Data Inconsistency

Maximize:

- Data Security
- Data Integrity
- Data Independence

The Database Approach

In all databases there is a data hierarchy:

Bit (binary digit): represents the smallest unit of data a computer can process and it consists only of a 0 or a 1.

Byte: A group of eight bits represents a single character (letter, number, or symbol).

Field: A column of data containing a logical grouping of characters into a word, a small group of words (e.g., last name, social security number, etc.).

Record: A logical grouping of related fi elds in a row (e.g., student's name, the courses taken, the date, and the grade).

Data File: logical grouping of related records is called a data file or a table similar in appearance to a speadsheet in Excel consisting of multiple columns and multiple rows.

Database: logical grouping of related data files (aka database tables).

The Database Approach

Different database systems have different designs on how the data files and databases interact.

Some database management systems are designed so that there are multiple databases managed.

Some systems are designed so that there is always a single database, but there is also an intermediate layer between the database and data file (table) layers.

The Database Approach

Database Management System (DBMS) is a set of programs that provide users with tools to create and manage a database.

Note that managing a database is much harder than creating it.

The famous Relational Database Model is based on the concept of two-dimensional tables and is usually designed with a number of related tables with each of these tables contains records (listed in rows) and attributes (listed in columns).

The relationships are actually between individual records, however as the tables represent the records' structure, we usually discuss relationships as between tables.

The Database Approach

The data model on the other hand is a diagram that represents entities in the database and their relationships.

Entity: a person, place, thing, or event (e.g., customer, an employee, or a product).

Record: generally describes an entity and an **instance of an entity** refers to each row in a relational table. Entity models are conceptual, but individual records represent instances of real world entities after which the database entity models were created.

Attribute: each characteristic or quality of a particular entity. Some attributes are calculated from other attributes. These are called derived attributes.

Most data models are person models or product models.

In these models, places, time, and durations ae usually attributes.

However place (location) based models and time based (temporal) models are also becoming significant.

The Database Approach

Relationships are important.

In order to keep track of relationships the concept of “keys” has been created.

Primary Key: a field in a database that uniquely identify each record so that it can be retrieved, updated, and sorted.

Secondary Key: a field that has some identifying information, but typically does not identify the record with complete accuracy and therefore cannot serve at the Primary Key.

Foreign Key: a field (or group of fields) in one table that uniquely identifies a row of another table. It is used to establish and enforce a link between two tables.

A Sidenote on Big Data

Most organizations claim they need Big Data or they have Big Data.

Big Data is defined as diverse, high volume, high-velocity information assets that require new forms of processing to enable enhanced decision making, insight discovery, and process optimization. (Gartner definition).

Big data can also be defined as **vast** datasets that (1) exhibit **variety**, (2) include **structured, unstructured, and semi-structured data**, (3) are **generated at high velocity** with an **uncertain pattern**, (4) **do not fit neatly** into "traditional" structured, relational databases, and (5) can be captured, processed, transformed, and analyzed in a reasonable amount of time only by **sophisticated information systems**.

Most data is created by designed business processes in organizations. Therefore those processes create more or less well organized data. Therefore most data resides in "traditional" databases and are easily processed using the relationships.

Most data is not necessarily big data. However big data is **valuable**.

A Sidenote on Big Data (Cnt’d)

Big Data

Volume: incredible volume of data.

Velocity: The rate at which data flow into an organization is rapidly increasing and it is critical because it increases the speed of the feedback loop between a company and its customers.

Variety: Big Data formats change rapidly and can include include satellite imagery, broadcast audio streams, digital music files, Web page content.

Big Data can come from untrusted sources.

Big Data is dirty: Dirty data refers to inaccurate, incomplete, incorrect, duplicate, or erroneous data.

Big Data changes, especially in data streams: Organizations must be aware that data quality in an analysis can change, or the data itself can change, because the conditions under which the data are captured can change.

A Sidenote on Big Data (Cnt'd)

Many organizations are moving towards NoSQL databases to process Big Data.

NoSQL databases can (1) manipulate structured as well as unstructured data and (2) inconsistent or missing data providing an alternative for organizations that have (3) more and (4) different kinds of data in addition to the traditional, structured data that fit neatly into the rows and columns of relational databases.

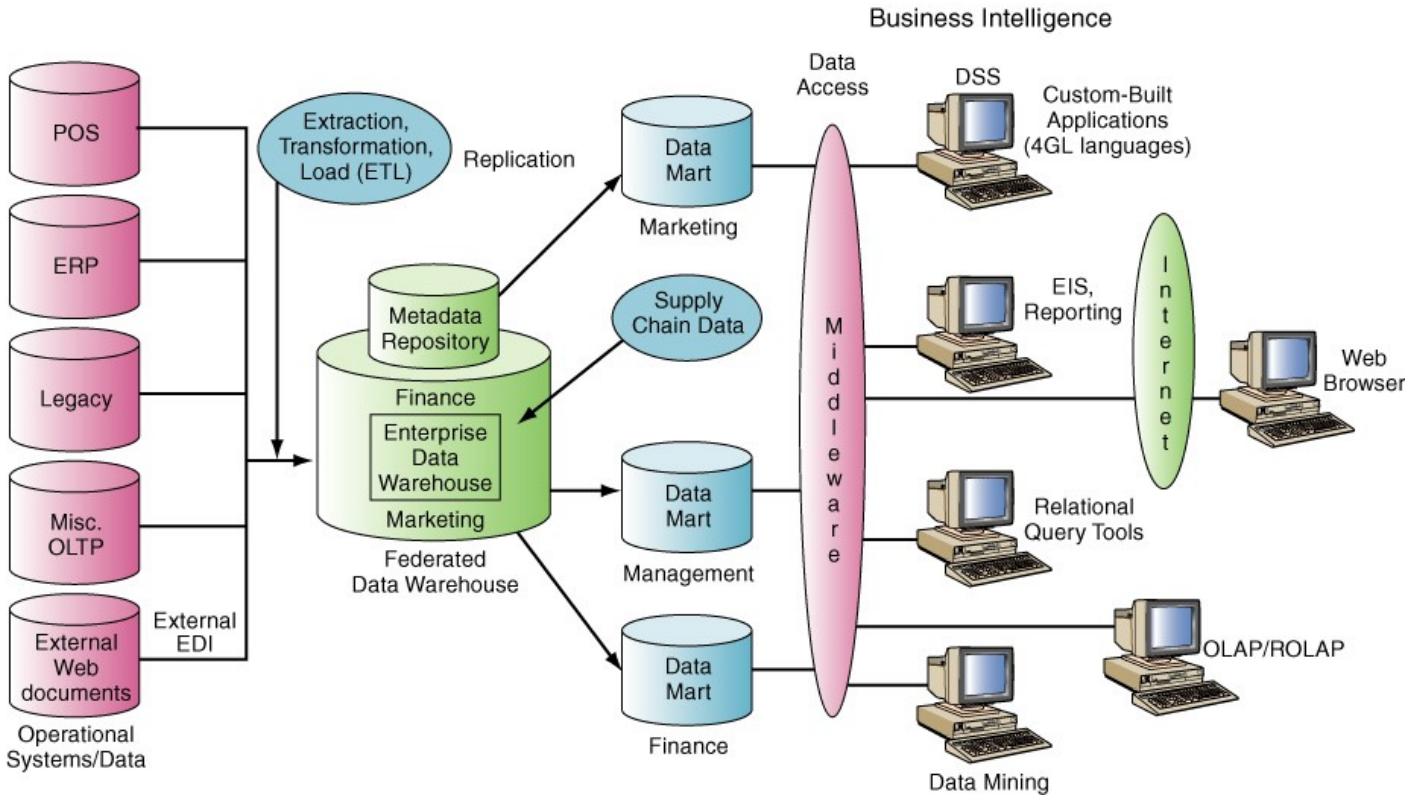
Example: M&A activity requires two organizations with similar but different data hierarchies to merge.

However, organizations still carry out their day to day operations on traditional relational databases. So we cannot simply forget about them.

In this course we will focus mainly on relational databases and then discuss NoSQL in detail.

To better understand the need for NoSQL and how to use NoSQL, one has to have some level of understanding on how the structured paradigm has problems.

Data Warehousing



Data Warehousing

A Data Warehouse is a repository of historical data that are organized by subject to support decision makers in the organization.

Data warehouses provide **a single version of truth**.

A Data Mart is a low-cost, scaled-down version of a data warehouse that is designed for the end-user needs in a strategic business unit (SBU) or an individual department.

An (Open) Data Portal is a data mart that organizations provide to third parties.

Data Warehousing

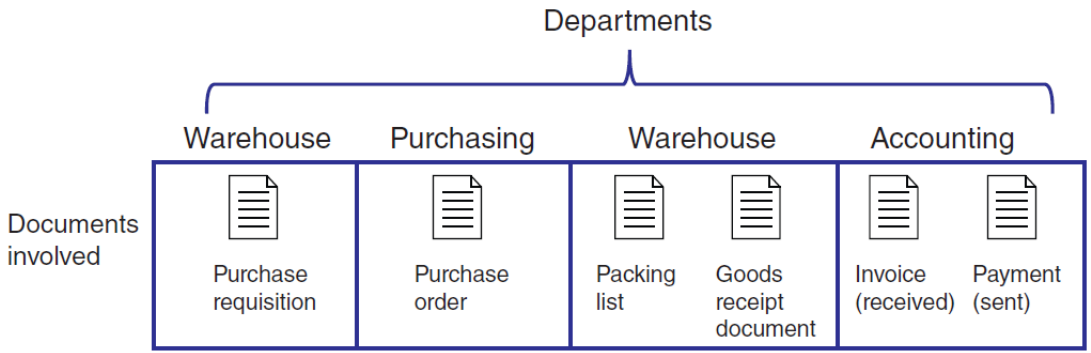
Properties of data warehouses:

Organized by **business dimension or subject**: Data are organized by subject. For example, by customer, vendor, product, price level, and region.

This arrangement differs from transactional systems, where data are organized by business process, such as order entry, inventory control, and accounts receivable.

Example. Buying airline tickets (transactional). Qualifying for frequent flier benefits (business dimension).

Example: Purchasing a computer (transactional). Analyzing who will probably need a new computer (business dimension).



Data Warehousing

Properties of data warehouses:

Organized by **business dimension or subject**: Data are organized by subject. For example, by customer, vendor, product, price level, and region.

This arrangement differs from transactional systems, where data are organized by business process, such as order entry, inventory control, and accounts receivable.

Example. Buying airline tickets (transactional). Qualifying for frequent flier benefits (business dimension).

Use **online analytical processing (OLAP)**: Data are collected from multiple systems and then integrated around subjects.

Time variant: Data warehouses and data marts maintain historical data (i.e., data that include time as a variable).

Relationships may also be defined in terms of time.

Example: The account balance of a supplier. Stock tickers.

Data Warehousing

Properties of data warehouses (cnt'd):

Nonvolatile: Data warehouses and data marts are nonvolatile—that is, users cannot change or update the data.

- This is to ensure data integrity, but also has some advantages regarding performance.
- Unless otherwise known, one can safely assume that for any record, there is one writer and many readers.

Multidimensional: Typically the data warehouse or mart uses a multidimensional data structure.

- Recall that relational databases store data in two-dimensional tables.
- Multidimensionality appears when multiple data tables are use together or when two versions of the same table are used.
- Multidimensionality also appears when the same table is used to present multiple business views at the same time.

Data Warehousing

In a typical data warehouse we have:

Source Systems: Systems that provide a source of organizational data.

Sources can be active, ie. writing to databases, creating records or passive, ie. being read by programs that are writing to databases.

Data Integration: reflects the growing number of ways that source system data can be handled.

Typically organizations need to Extract, Transform, and Load (ETL) data from source system into a data warehouse or data mart.

ETL may be a hard task to perform, consuming much resources.

Data Warehousing

In a typical data warehouse we have (cnt'd):

Storing the Data: A variety of architectures can be used to store decision-support data and the most common architecture is one central enterprise data warehouse, without data marts.

- In addition to the high level architecture of data storage, there are qualities to consider for a selected architecture.

Metadata: data maintained about the data within the data warehouse. (e.g., database, table, and column names; refresh schedules; and data-usage measures.)

- Metadata is becoming increasingly important.

Data Warehousing

In a typical data warehouse we have (cnt'd):

Data Quality: quality of the data in the warehouse must meet users' needs. If it does not, users will not trust the data and ultimately will not use it.

- Some of the data can be improved with data-cleansing soft ware, but the better, long-term solution is to improve the quality at the source system level.

Governance: To ensure that BI is meeting their needs, organizations must implement governance to plan and control their BI activities.

- Governance requires that people, committees, and processes be in place.

Users: There are many potential BI users, including IT developers; frontline workers; analysts; information workers; managers and executives; and suppliers, customers, and regulators.

Data Warehousing

OLAP creates a data cube.

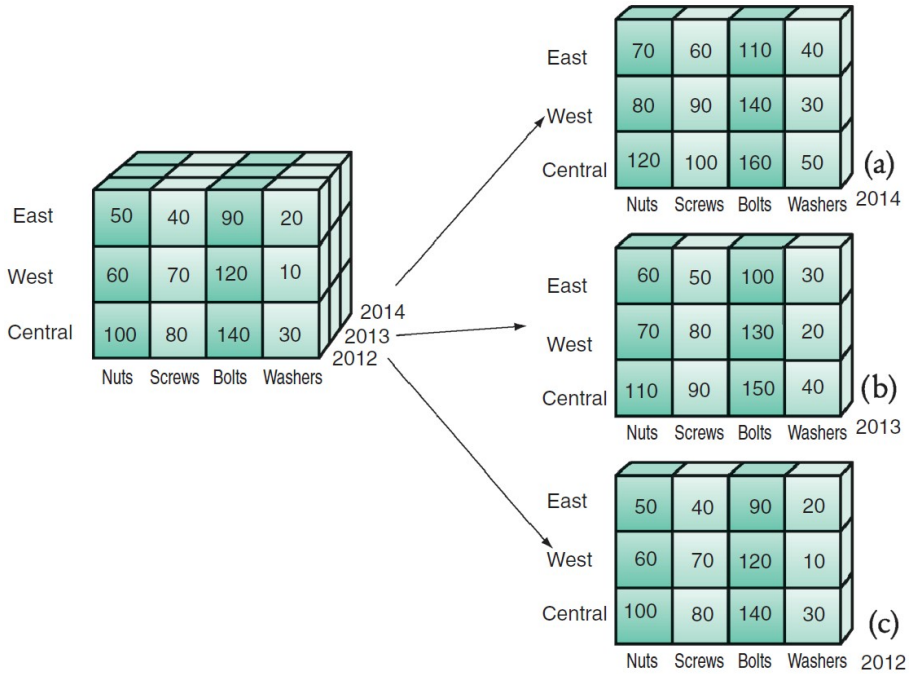
When you have a cube, you run many queries, including variations of the same query on the cube. This takes **a lot** of time.

Query results are **cached**.

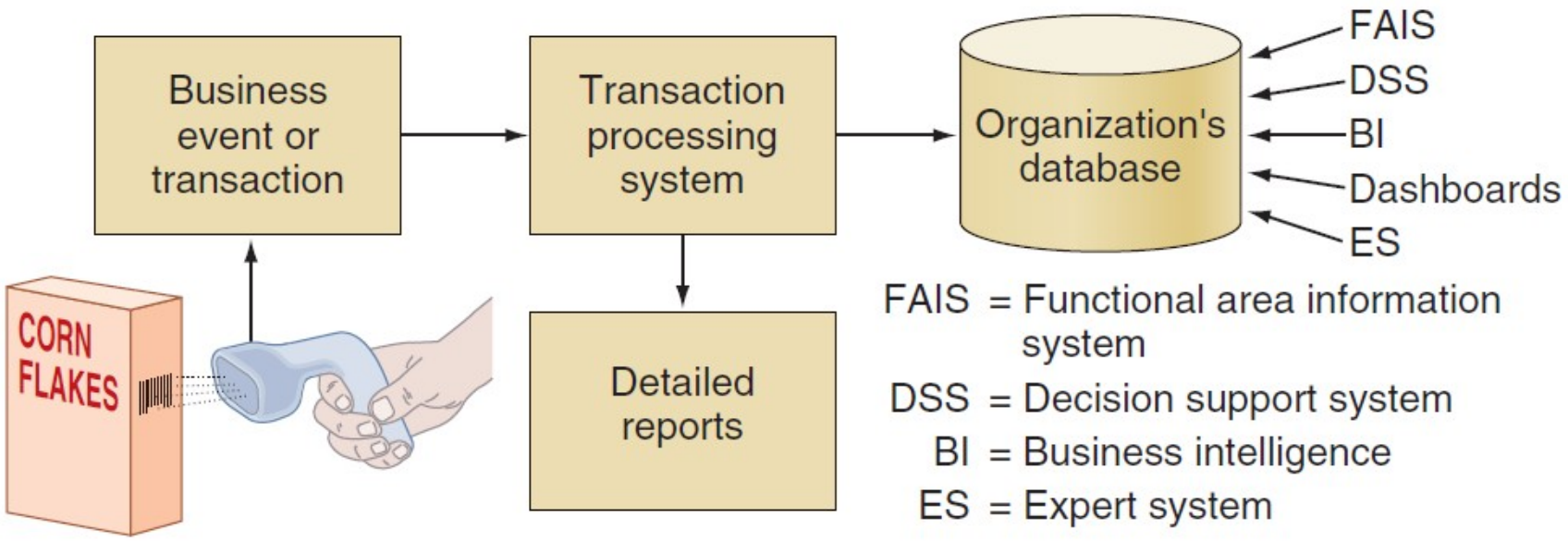
Then when users run a particular query, the query is compared to the existing queries.

If there is a match, the result is already cached.

Cached results usually have a lifetime that depends on the business process that adds to or updates the database.



Databases Are Part of the Whole Information System



Databases Are Part of the Whole Information System

ACCOUNTING	FINANCE	HUMAN RESOURCES	PRODUCTION/ OPERATIONS	MARKETING	
Profitability Planning	Financial Planning	Employment Planning, Outsourcing	Product Life Cycle Management	Sales Forecasting, Advertising Planning	STRATEGIC
Auditing, Budgeting	Investment Management	Benefits Administration, Performance Evaluation	Quality Control, Inventory Management	Customer Relations, Sales Force Automation	TACTICAL
Payroll, Accounts Payable, Accounts Receivable	Manage Cash, Manage Financial Transactions	Maintain Employee Records	Order Fulfillment, Order Processing	Set Pricing, Profile Customers	OPERATIONAL

Popular DBMS's

According to Statista, Top 10 of DBMS's in the relational world

- Big Two – Oracle and MySQL
- Third - Microsoft SQL Server
- Fourth - PostgreSQL
- Then comes IBM DB2, SQLite, Microsoft Access, MariaDB, Apache Hive and Microsoft Azure SQL

See – https://bit.ly/2YH7Dnb

We will compare the top 3

Oracle vs MySQL (and MariaDB)

Oracle

Commercial license, cost based on size of server resources dedicated to DBMS.

Oracle runs on Linux predominantly but also supports other major OS's.

Supports a variety of indexes, so that efficient processing of a variety of data types including multimedia files is possible.

Supports distributed architecture, so is extremely reliable.

Supports data partitioning.

Data location is abstracted. Oracle allows interacting with the database without knowing the physical storage medium or location of the data.

Has many built-in tools to assist database administrators.

Top database for government and finance.

MySQL (and MariaDB)

Open Source.

MySQL runs on almost any platform.

Supports indexing for text ad numbers only. This is not a problem for databases storing typical commercial data.

Does not support distributed architecture, but still very reliable.

Does not support data partitioning.

Extremely high performance, in particular when inserting new records.

Top database for e-commerce applications.

MySQL's parent company was acquired by MYSQL. A spin-off project was "forked" within 24 hours and became MariaDB.

Oracle vs Microsoft SQL Server

Oracle

Commercial license, cost based on size of server resources dedicated to DBMS.

Oracle runs on Linux predominantly but also supports other major OS's.

Every new database connection is automatically treated like a transaction.

Almost all commands are executed in parallel.

Can undo changes within a transaction, in almost any setting and environment.

Very detailed error handling.

Everything is organized by a schema, so that all is well managed. The subset collection of these database schemas are shared between all the schemas and users.

Oracle offers automation support through its database upgrade assistant.

Microsoft SQL Server

Commercial license.

Runs on Windows and Linux (very recent).

Transactions are explicitly defined.

Most commands are executed serially.

Cannot undo changes to data in a transaction once "commit" command is sent.

Simpler error handling.

Organizes everything by database names.

SQL Server also offers automation support through the SQL upgrade advisor. It does not, however, support parallel query execution and every database has its own unshared file disk on the server.

Oracle vs Microsoft SQL Server

MySQL (and MariaDB)

- Open Source.
- Runs on almost anything.
- Has multiple alternative database engines that can be replaced easily.
- Can be installed and operated on very small spaces.
- Allows query cancellation.
- Uses non-blocking I/O to access database.

Microsoft SQL Server

- Commercial license.
- Runs on Windows and Linux.
- Has a single database engine.
- Expects a large amount of operational storage space.
- It doesn't allow query cancellation mid-way in the process.
- Uses blocking I/O to access database.

Installing MySQL

Linux (Ubuntu)

Single command on the command line.

```
$ sudo apt update && sudo apt install mysql-server
$ sudo systemctl status mysql
```

A “good” result includes “active (running)” in the response text.

Also to harden the DBMS against typical security risks.

```
$ sudo mysql_secure_installation
```

Installing MySQL

Windows

Download installer at (~450MB) – https://dev.mysql.com/downloads/installer/

Run the installer.

Using MySQL

Next week we will be designing and creating tables for an small business application.