1.1 Database basics

Data

Data is numeric, textual, visual, or audio information that describes real-world systems. Data is collected and processed to aid in a variety of tasks, such as forecasting weather, analyzing financial investments, and tracking the global spread of pandemics.

Data can vary in several important ways:

- **Scope.** The amount of data produced and collected can vary. Ex: A small business might track an inventory of a few thousand items, but a large commerce website might track billions of items.
- **Format.** Data may be produced as numbers, text, image, audio, or video. Ex: A phone's proximity sensor generates raw numbers, and a satellite captures images.
- **Access.** Some data sources are private while others are made publicly available. Ex: A retail company may use private customer data to discover purchasing behavior patterns, but a government may be required by law to share certain data sets.

Historically, data was mostly ***analog***, encoded as continuous variations on various physical media. Ex: Audio was recorded as vibrations impressed on vinyl disks. Images were recorded as chemicals on celluloid tapes. Today, data is mostly ***digital***, encoded as zeros and ones on electronic and magnetic media.

The shift from analog to digital data facilitated the rise of large computer databases.

PARTICIPATION ACTIVITY | 1.1.1: Examples of public data sets.

Animation content:

undefined

Animation captions:

1. Data.gov provides thousands of U.S. government data sets. Precipitation data can be used to visualize rainfall intensity.
2. Kaggle.com allows users to find and publish data sets. The Financial Tweet data set can show who tweets on similar topics.
3. data.Nasa.gov provides data sets in aerospace and other related sciences. A data set of light measurements describes astronomical phenomena.

Sources: Rainfall map, Social network diagram, Hubble mosaic of the Crab Nebula from Wikipedia.org

PARTICIPATION ACTIVITY 1.1.2: Public data sets.

These websites offer public data sets:

- data.gov
- cancer.gov/research
- kaggle.com
- data.nasa.gov
- opendata.cityofnewyork.us

Click on each link, review the website, and drag the website name below to the matching description.

If unable to drag and drop, refresh the page.

kaggle.com **data.gov** **cancer.gov/research** **opendata.cityofnewyork.us** **data.nasa.gov**

	Provides more than 250,000 U.S. government data sets to support research and application development.
	Collects and reports data and information relative to all forms of cancer.
	Owned by Google, supports an online community that allows users to find and publish data sets.
	Provides numerous data sets in categories such as aerospace, earth science, and space science.
	Data collected by the New York City government to support continuous monitoring and improvements to NYC and residents' health.

Reset

Databases

A ***database*** is a collection of data in a structured format. In principle, databases can be stored on paper or even clay tablets. In practice, however, modern databases are invariably stored on computers. The database structure ensures that similar data is stored in a standardized manner.

Many modern databases contain trillions of bytes of data and support thousands of simultaneous users. Consequently, databases must be managed with sophisticated software tools:

- A ***database system***, also known as a ***database management system*** or ***DBMS***, is software that reads and writes data in a database. Database systems ensure data is secure, internally consistent, and available at all times. These functions are challenging for large databases with many users, so database systems are complex.
- A ***query*** is a request to retrieve or change data in a database. A ***query language*** is a specialized programming language, designed specifically for database systems. Query languages read and write data efficiently, and differ significantly from general-purpose languages such as Python, Java, and C++.
- A ***database application*** is software that helps business users interact with database systems. Many databases are complex, and most users are not familiar with query languages. Consequently, direct database access is usually not feasible. Instead, programmers write applications to simplify the user experience and ensure data access is efficient and secure.

Database software is organized in layers. Applications interact with a query language on one layer, and a query language interacts with a database system on another layer. Other software layers, such as the operating system, are beyond the scope of this material.

Terminology

*The term **database** sometimes refers to a database system rather than the data managed by the system. The meaning is usually clear from context. In this material, **database** is used both ways.*

PARTICIPATION ACTIVITY | 1.1.3: A bank database.

Animation content:

There is a box named Banking database system. In this box is another box named Bank database. In that box there is another box named Customer bank accounts. In that box there is an unnamed table with two columns named Customer and Balance. Slide 2. There is a figure labeled Bank customer. Below is a box labeled Banking application. There is a line from the box Banking application to the box Banking database system. A money bag appears next to the box Banking application and move along the line to the box Banking database system. Then the value 3889 dollars and 10 cents in the unnamed table turns red. The value 3889 dollars and 10 cents is duplicated and moved next to the box Banking application.

Animation captions:

1. Banking data is stored in a database and is managed by a database system.
2. A bank customer uses an application to perform bank transactions.
3. Banking transactions cause the database system to modify the bank's database and update the user's account.

PARTICIPATION ACTIVITY | 1.1.4: Database in a database system.

Refer to the animation above.

1) Where would a bank customer's account data be directly stored?

- In the Banking database system
- In the customer banking application
- In the Bank database

2) Which of the following manages the Bank database?

- The customer banking application
- The banking database system
- The computer the Banking database system resides on

3) Which of the following would prevent unauthorized access to the Bank database?

- The banking database system
- The Bank database
- The customer banking application

4) How would a bank customer access their bank data?

- Through the banking database system
- Through the customer banking application
- Through the Bank Database

Database roles

People interact with databases in a variety of roles:

- A ***database administrator*** is responsible for securing the database system against unauthorized users. A database administrator enforces procedures for user access and database system availability.
- A ***database designer*** determines the format of each data element and the overall database structure. Database designers must balance several priorities, including storage, response time, and support for rules that govern the data. Since these priorities often conflict, database design is technically challenging.
- A ***database programmer*** develops computer programs that utilize a database. Database programmers write applications that combine database query languages and general-purpose programming languages. Query languages and general-purpose languages have significant differences, so database programming is a specialized challenge.
- A ***database user*** is a consumer of data in a database. Database users request, update, or use stored data to generate reports or information. Database users usually access the database via applications but can also submit queries directly to the database system.

For simple databases with a limited amount of data and few users, one person may assume several roles. Ex: The database administrator might also be a database designer or database programmer. For large, complex databases, each person usually takes on just one role.

PARTICIPATION ACTIVITY 1.1.5: Database roles.

Animation content:

Slide 1. There is a box named database system and a figure named Database designer. The is an arrowing pointing from the Database designer to the Banking database system box and is labeled Designs. The words Create bank database move from the Database designer to the banking database system and disappears. Next the words create customer and account move from the Database designer to the Banking and database system. Slide 2. A new figure named Database administrator appears. A line goes from Database administrator to the Banking database system and is labeled Manages. The words add user move from the Database designer to the banking database system and disappears. The words backup data move from the Database designer to the banking database system and disappears. Slide 3. A new figure named End user appears. A line goes from End user to the Banking database system and is labeled Application request. The words change address move from the End user to the banking database system and disappears. The words transfer money move from the End user to the banking database system and disappears.

Animation captions:

1. The database designer establishes the structure of the database and determines the data to be collected and stored.
2. The database administrator ensures the database is available and secure.
3. A database programmer uses query languages and programming languages to develop applications for database users.
4. Database users are the primary consumers of database data through applications or with query languages.

PARTICIPATION ACTIVITY 1.1.6: Roles.

1) Which role is responsible for providing access to the database?

- Database administrator
- Database designer
- Database programmer
- Database users

2) Which role is responsible for defining the detailed database design?

- Database administrator
- Database designer
- Database programmer
- Database users

3) Which role uses an application to query a database and generate a report?

- Database designer
- Database administrator
- Database programmer
- Database users

Exploring further:

- Data and Reality, by William Kent
- Data.gov
- Kaggle
- NASA's Data Portal

1.2 Database systems

File systems and database systems

Small databases that are shared by one or two users can be managed in a text file or spreadsheet. Text files and spreadsheets are inadequate, however, as databases grow in size, complexity, and use. Large, complex databases that are shared by many users have special requirements:

- **Performance**. When many users and applications simultaneously access large databases, query response time degrades rapidly. Database systems maintain fast response times by structuring data properly on storage media and processing queries efficiently.
- **Authorization**. Many database users should have limited access to specific tables, columns, or rows of a database. Database systems authorize individual users to access specific data.

- **Security**. Database systems ensure authorized users only access permissible data. Database systems also protect against hackers by encrypting data and restricting access.
- **Rules**. Database systems ensure data is consistent with structural and business rules. Ex: When multiple copies of data are stored in different locations, copies must be synchronized as data is updated. Ex: When a course number appears in a student registration record, the course must exist in the course catalog.
- **Recovery**. Computers, database systems, and individual transactions occasionally fail. Database systems must recover from failures and restore the database to a consistent state without loss of data.

File systems are not designed for these demanding requirements. The limitations of file systems became clear as business adopted computers in the 1960s. Since then, database systems have replaced file systems for large databases with many users.

PARTICIPATION ACTIVITY | 1.2.1: Limitations of file systems.

Animation content:

Animation captions:

1. A list of bank transactions is stored in the text file bookkeeping.txt. Each persons' transactions include a date, type, and the amount of money paid or received.
2. Two programs access the text file. One adds new transactions, and the other calculates account balances.
3. When program A writes transactions quickly to the file, Program B misses the $4000 deposit and calculates Raul's balance incorrectly.
4. Program A may write an erroneous transaction.
5. A lack of adequate security could allow unauthorized users to access the file.

PARTICIPATION ACTIVITY | 1.2.2: Limitations of file systems.

Refer to the animation above.

1) If program A writes to bookkeeping.txt the same time that program B reads from bookkeeping.txt, what can potentially go wrong?

- Nothing can go wrong.
- Program A may be writing only partial data.
- Program B may be reading only partial data.

2) Why was an unauthorized user permitted to perform the transaction that added $999999 to bookkeeping.txt?

- The user had read access to the text file.
- The user had write access to the text file.
- The user had no access to the text file.

3) What is wrong with the transaction posted on 13/31 for the amount +6z@yy?

- The data in the transaction contains invalid pieces of data.
- The transaction data is valid and would correctly update the bookkeeping.txt file.

Transactions

Transaction management is a particularly challenging requirement for database systems.

A ***transaction*** is a group of queries that must be either completed or rejected as a whole. Execution of some, but not all, queries results in inconsistent or incorrect data. Ex: A debit-credit transaction transfers funds from one bank account to another. The first query removes $100 from one account and the second query deposits $100 in another account. If the first query succeeds but the second fails, $100 is mysteriously lost. The transaction must process either both queries or neither query.

When processing transactions, database systems must:

- **Ensure transactions are processed completely or not at all**. A computer or application might fail while processing a transaction. When failing to process a transaction, the database system must reverse partial results and restore the database to the values prior to the transaction.
- **Prevent conflicts between concurrent transactions**. When multiple transactions access the same data at the same time, a conflict may occur. Ex: Sam selects a seat on a flight. Maria purchases the same seat in a separate transaction before Sam completes his transaction. When Sam clicks the 'purchase' button, his seat is suddenly unavailable.
- **Ensure transaction results are never lost**. Once a transaction completes, transaction results must always be saved on storage media, regardless of application or computer failures.

The above requirements are supported in sophisticated transaction management subsystems of most database systems.

PARTICIPATION ACTIVITY 1.2.3: Transactions.

Animation content:

Animation captions:

1. Two programs access a bank database. The database tracks customer deposits, credits, and account balances.
2. Program A requests the database transfer $50 from Raul to Mai.
3. Transaction 1 deducts $50 from Raul's account and adds $50 to Mai's account.
4. Program B requests that Raul transfer $200 to Kawika.
5. $200 is deducted from Raul's account.
6. Kawika closes his account before Transaction 2 completes. The database reverses the $200 deduction.

PARTICIPATION ACTIVITY 1.2.4: Transactions.

Match the transaction behavior to the described situation.

If unable to drag and drop, refresh the page.

Transaction processed completely or not at all | **Ensure transaction results are never lost**

Prevent conflicts between concurrent transactions

	A program is adding a penalty fee to an account that is below $1000 while another program is adding $2000 to the same account.
	A streaming video service's database is disabled due to a lightning strike. The IT department verifies that all database transactions performed before the lightning strike were saved on storage media.
	Maria purchases an airline ticket, but a server failure causes the ticket to become unavailable before the transaction completes. The database must reverse any partial changes.

Reset

Architecture

The ***architecture*** of a database system describes the internal components and the relationships between components. At a high level, the components of most database systems are similar:

- The ***query processor*** interprets queries, creates a plan to modify the database or retrieve data, and returns query results to the application. The query processor performs ***query optimization*** to ensure the most efficient instructions are executed on the data.
- The ***storage manager*** translates the query processor instructions into low-level file-system commands that modify or retrieve data. Database sizes range from megabytes to many terabytes, so the storage manager uses ***indexes*** to quickly locate data.
- The ***transaction manager*** ensures transactions are properly executed. The transaction manager prevents conflicts between concurrent transactions. The transaction manager also restores the database to a consistent state in the event of a transaction or system failure.
- The ***log*** is a file containing a complete record of all inserts, updates, and deletes processed by the database. The transaction manager writes log records before applying changes to the database. In the event of a failure, the transaction manager uses log records to restore the database.

- The ***catalog***, also known as a ***data dictionary***, is a directory of tables, columns, indexes, and other database objects. Other components use catalog information to process and execute queries.

Database systems have different capabilities, and component details vary significantly. Ex: Some database systems do not support transactions and therefore have no transaction manager. Ex: The storage manager implementation depends on the physical structure of data on storage media.

PARTICIPATION ACTIVITY | 1.2.5: Database system architecture.

Animation content:

Slide 1. There is a box named database system. Inside this box there are three more boxes named transaction manager query processor and storage manager. Slide 2. A new box appears to the left of box database system and is connected to box database system by a line. Text appears next to box application that states Retrieve overdue accounts. The text moves from box application to below box query processor. Slide 3. New text appears that states Optimized retrieval instructions and moves to below box storage manager. Slide 4. A new box named disk appears and has three terms data data dictionary and indexes inside. The box disk is to the right of box database system and they are connected by a line. Text appears below the box disk and states Lopez Cordeiro. Slide 5. The term query appears to the right of box transaction manager and has an arrow pointing from query to text log that appears in box disk. An arrow appears that goes from text Lopez Cordeiro to box application. The text is duplicated and moves along the line to box application.

Animation captions:

1. A database system is composed of a query processor, storage manager, transaction manager, log, and catalog.
2. An application sends queries to the query processor.
3. The query processor uses information from the catalog to perform query optimization.
4. The storage manager translates the query processor instructions into file-system commands and uses an index to quickly locate the requested data.
5. The transaction manager logs insert, update, and delete queries, and the result is sent back to the application.

PARTICIPATION ACTIVITY | 1.2.6: Database system architecture.

1) The query processor has direct access to the database data on storage media.

- True
- False

2) Without query optimization, the storage manager cannot retrieve the database data.

- True
- False

3) The catalog allows the storage manager to quickly locate the requested data.

- True
- False

4) Every database query must be logged by the transaction manager to recover the database in the event of a system failure.

- True
- False

Products

Most leading database systems are relational. A ***relational database*** stores data in tables, columns, and rows, similar to a spreadsheet. All data in a column has the same format. All data in a row represents a single object, such as a person, place, product, or activity.

All relational database systems support the SQL query language. ***SQL*** stands for Structured Query Language and includes statements that read and write data, create and delete tables, and administer the database system.

Relational systems are ideal for databases that require an accurate record of every transaction, such as banking, airline reservation systems, and student records. The growth of the internet in the 1990s generated massive volumes of online data, called ***big data***, often with poorly structured or missing information. Relational systems were not initially designed for big data and, as a result, many non-relational systems have appeared since 2000. The newer non-relational systems are called ***NoSQL***, for 'not only SQL', and are optimized for big data.

Prior to 2000, most database systems were commercial products, developed by for-profit companies and licensed for a fee. Since 2000, an alternative licensing model, called open source, has become popular. ***Open source*** software is software that anyone can inspect, copy, and modify with no licensing fee.

NoSQL and open source systems have proliferated, and hundreds of database systems are now available. The website db-engines.com ranks systems by tracking product references on social media, internet searches, job websites, and technical websites. Internet references are an imperfect measure of product utilization, but do provide a general indication of interest and activity.

Figure 1.2.1: Leading database products.

Product	Sponsor	Type	License	DB-Engines rank (May 2020)
Oracle Database	Oracle	Relational	Commercial	1
MySQL	Oracle	Relational	Open source	2
SQL Server	Microsoft	Relational	Commercial	3
PostgreSQL	PostgreSQL Global Development Group	Relational	Open source	4
MongoDB	MongoDB	NoSQL	Open source	5

PARTICIPATION ACTIVITY 1.2.7: Database system categories.

Match the category to the description.

If unable to drag and drop, refresh the page.

Commercial **Open source** **Relational** **NoSQL**

MySQL but not MongoDB

	Optimized for big data generated on the internet
	SQL Server but not MySQL
	Allows programmers to inspect and modify source code

Reset

Exploring further:

- DB-Engines database systems ranking

1.3 Query languages

Common queries

A database system responds to queries written in a query language. A ***query*** is a command for a database that typically inserts new data, retrieves data, updates data, or deletes data from a database. A ***query language*** is a computer programming language for writing database queries.

PARTICIPATION ACTIVITY | 1.3.1: Insert, select, update, and delete database queries.

Animation content:

Slide 1. There is a box named Bank database. In this box is another box named Account. Account has the names and balances for two people. Slide 2. Text appears and states Insert into account Ethan Carr 5000. The text Ethan Carr 5000 appears in the box Account. Slide 3. New text appears and states Select name from Account where Balance is greater than 3000. The names Raul Lopez and Ethan Carr are highlighted with balances 3300 and 5000 respectively. The names appear below the text. Slide 4. Text appears and states Update Account Set Raul Lopez's balance equals 4500 dollars. The number for Raul Lopez in the account box is changed from 300 to 4500 Slide 5. Text appears and states Delete Mai Shiraishi from Account. Mia Shirashi 2500 is crossed out in box Account.

Animation captions:

1. A bank database stores the names and balances for two accounts: Raul and Mai.
2. An insert query inserts new data into the database. Ethan's new account is inserted into the database.
3. A select query retrieves information from the database. The query retrieves the names of individuals that have a balance more than $3000.
4. An update query changes existing data in the database. Raul's balance is changed from 3300 to 4500.
5. A delete query removes data from the database. Mai's account is removed.

Terminology

The four common queries are sometimes referred to as ***CRUD*** *operations, an acronym for Create, Read, Update, and Delete data.*

PARTICIPATION ACTIVITY 1.3.2: Queries.

Refer to the animation above.

1) Only one of the queries does not change the database contents.
   - True
   - False

2) Given the data in the Bank database, a select query for accounts with negative balances would return nothing.
   - True
   - False

3) The insert, select, update, and delete queries are the only types of commands necessary to interact with a database system.
   - True
   - False

4) An update query cannot update data that isn't in the database.

- True
- False

Writing queries with SQL

Structured Query Language, or ***SQL***, is the standard query language of relational database systems. The SQL standard is sponsored by the American National Standards Institute (ANSI) and the International Organization for Standardization (ISO). SQL is pronounced either 'S-Q-L' or 'seekwəl', but the preferred pronunciation is 'S-Q-L'.

SQL was first developed at IBM in the 1970s as an experimental query language for a prototype relational database. At the time, IBM was the dominant computer company, so SQL became the dominant relational query language. Today, all relational database systems support SQL.

Terminology

*The term **NoSQL** refers to a new generation of non-relational databases. NoSQL originally meant 'does not support SQL'. However, many NoSQL databases have added support for SQL, and 'NoSQL' has come to mean 'not only SQL'.*

An SQL ***statement*** is a database command, such as a query that inserts, selects, updates, or deletes data:

- ***INSERT*** inserts rows into a table.
- ***SELECT*** retrieves data from a table.
- ***UPDATE*** modifies data in a table.
- ***DELETE*** deletes rows from a table.

The SQL language contains many other statements for creating and deleting databases, creating and deleting tables, assigning user permissions, and so on.

PARTICIPATION ACTIVITY 1.3.3: SQL statements: INSERT, SELECT, UPDATE, and DELETE.

Animation content:

Slide 1. A new table named Account appears and has three columns named ID Name and Balance. Slide 2. Two rows are added to the table. Row one's values are 831 Raul Lopez and 3300. Row two's values are 572 Mai Shiraishi 2500. Slide 3. Two lines of code appear. The first line of code states INSERT INTO Account. The second line of code states VALUES left parenthesis 290 comma apostrophe Ethan Carr apostrophe comma 5000 right parenthesis semicolon. A new row is added to the table Account with the values 290 Ethan Carr and 5000. Slide 4. Three new lines of code appear. The first line of code states SELECT Name. the second line of code states FROM Account. The third line of code states WHERE Balance is greater than 3000 semicolon. The values Raul Lopez and Ethan Carr are highlighted in the table Account. The values are copied and moved below the code. Slide 5. Three lines of code appear. The first line of code states UPDATE Account. The second line of code states SET Balance - 4500. The third line of code states WHERE ID = 831. The value in row on in column Balance int table Account changes from 3300 to 4500. Slide 6. Two lines of code appear. The first line of code states DELETE FROM Account. The second line of code states WHERE ID = 572. The second row in table Account gets a line crossed through it.

Animation captions:

1. A bank database has an Account table with three columns: ID, Name, and Balance.
2. The Account table has two rows that store Raul and Mai's account data.
3. The INSERT statement adds a new row with Ethan's account data.
4. The SELECT statement retrieves the names of accounts with a balance larger than $3000.
5. The UPDATE statement change's Raul's balance from 3300 to 4500.
6. The DELETE statement deletes Mai from the database.

PARTICIPATION ACTIVITY 1.3.4: SQL statements.

Refer to the Account table below.

Account

ID	Name	Balance
831	Raul Lopez	3300
572	Mai Shiraishi	2500
290	Ethan Carr	5000

1) What is Braden Smith's balance in the following INSERT statement?

```
INSERT INTO Account
VALUES (800, 'Braden Smith',
200);
```

- 800
- 200
- Unknown

2) Which name is retrieved by the following SELECT statement?

```
SELECT Name
FROM Account
WHERE Balance < 3000;
```

- Raul Lopez
- Mai Shiraishi
- Ethan Carr

3) Whose balance does the following UPDATE statement change?

```
UPDATE Account
SET Balance = 850
WHERE ID = 290;
```

- Raul Lopez
- Mai Shiraishi
- Ethan Carr

4) Who is deleted by the following DELETE statement?

```
DELETE FROM Account
WHERE ID = 999;
```

- Mai Shiraishi
- No one
- Everyone

CHALLENGE ACTIVITY | 1.3.1: Query languages.

543058.505174.qx3zqy7

Start

Country

Name	Capital	IndependenceYear	Population
Fiji	Suva	1970	880000
Belgium	Brussels	1831	11430000


```
INSERT INTO Country
VALUES ('Chile', 'Santiago', 1810, 18730000);
```

What is the new row's Population?

1 2 3 4

Check Next

Creating tables with SQL

The SQL ***CREATE TABLE*** statement creates a new table by specifying the table and column names. Each column is assigned a ***data type*** that indicates the format of column values. Data types can be numeric, textual, or complex. Ex:

- INT stores integer values.
- DECIMAL stores fractional numeric values.
- VARCHAR stores textual values.
- DATE stores year, month, and day.

Some data types are followed by one or two numbers in parentheses, indicating the size of the data type. Ex: VARCHAR(10) indicates ten characters. DECIMAL(10, 3) indicates ten significant digits, including three after the decimal point.

PARTICIPATION ACTIVITY 1.3.5: Creating an Employee table.

Animation content:

Slide 1. There is code that states CREATE TABLE Employee left parenthesis with right parenthesis semicolon being five lines down. Text appears that states table name and points to the text Employee in the first line of code. An empty table named Employee appears to the right of the code. Slide 2. Four lines of code appear between the two lines of existing code. The first line of code states ID SMALL INT comma. The second line of code states Name VARCHAR left parenthesis 60 right parenthesis comma. The third line of code states Birth Date DATE comma. The fourth line of code states Salary DECIMAL left parenthesis 7 comma 2 right parenthesis. The words ID Name Birth Date and Salary are labeled column names and the words SMALL INT VAR CHAR left parenthesis 60 right parenthesis DATE and DECIMAL left parenthesis 7 comma 2 right parenthesis are labeled data types. The table Employee gets four columns named ID Name Birth Date and Salary.

Animation captions:

1. The CREATE TABLE statement names the new table "Employee".
2. The column names and data types are separated by commas.

PARTICIPATION ACTIVITY 1.3.6: Creating tables with SQL.

Refer to the animation above.

1) The Employee table is created with 4 different data types.

- True
- False

2) Only the ID column stores numbers.

- True
- False

3) The BirthDate column stores only a date and no time.

- True
- False

4) The Employee table is an empty table once created.

- True
- False

PARTICIPATION ACTIVITY | 1.3.7: Query the Movie table.

The SQL statements below create a Movie table and insert some movies. The SELECT statement selects all movies.

Press the Run button to produce a result table. Verify the result table displays five movies.

Modify the SELECT statement to only select movies released after October 31, 2015:

```
SELECT *
FROM Movie
WHERE ReleaseDate > '2015-10-31';
```

Then run the SQL again and verify the new query returns only three movies, all with release dates after October 31, 2015.

Hint: Move the semicolon that follows `FROM Movie` to the end of the statement.

```
CREATE TABLE Movie (
  ID INT,
  Title VARCHAR(100),
  Rating VARCHAR(5),
  ReleaseDate DATE
);

INSERT INTO Movie VALUES
  (1, 'Rogue One: A Star Wars Story', 'PG-13', '2016-12-10'),
  (2, 'Hidden Figures', 'PG', '2017-01-06'),
  (3, 'Toy Story', 'G', '1995-11-22'),
  (4, 'Avengers: Endgame', 'PG-13', '2019-04-26'),
  (5, 'The Godfather', 'R', '1972-03-14');

-- Modify the SELECT statement:
SELECT *
FROM Movie;

```


Run Reset code

▶ View solution

1.4 Database design and programming

Analysis

A ***database design*** is a specification of database objects such as tables, columns, data types, and indexes. Database design also refers to the process used to develop the specification.

For small, simple databases, the database design process can be informal and unstructured. For large, complex databases, the process has three phases:

1. Analysis
2. Logical design
3. Physical design

The ***analysis*** phase specifies database requirements without regard to a specific database system. Requirements are represented as entities, relationships, and attributes. An entity is a person, place, activity, or thing. A relationship is a link between entities, and an attribute is a descriptive property of an entity.

Terminology

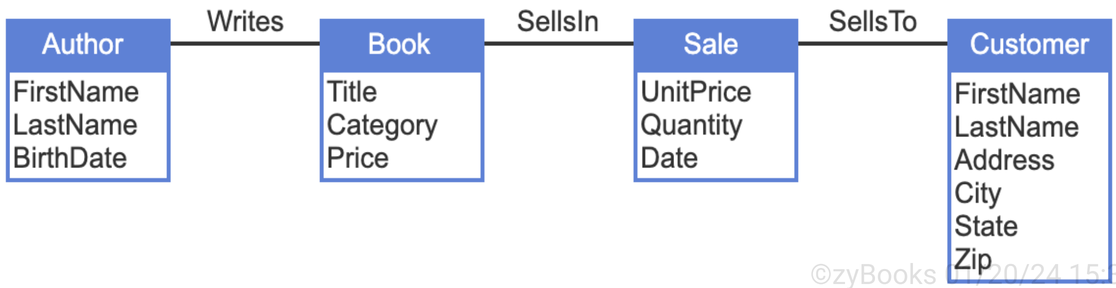
Analysis *has many alternative names, such as conceptual design, entity-relationship modeling, and requirements definition.*

Entities, relationships, and attributes are depicted in ***ER diagrams***:

- Rectangles represent entities. Entity names appear at the top of rectangles.
- Lines between rectangles represent relationships.
- Text inside rectangles and below entity names represent attributes.

ER diagrams are usually supplemented by textual descriptions of entities, relationships, and attributes.

Figure 1.4.1: ER diagram.



PARTICIPATION ACTIVITY 1.4.1: Analysis.

Refer to the ER diagram above.

1) What is 'Writes'?
- Entity
- Relationship
- Attribute

2) What is 'Category'?
- Entity
- Relationship
- Attribute

3) What is 'Sale'?
- Entity
- Relationship
- Attribute

4) Is there a direct relationship between Customer and Book?
- Yes
- No

Logical design

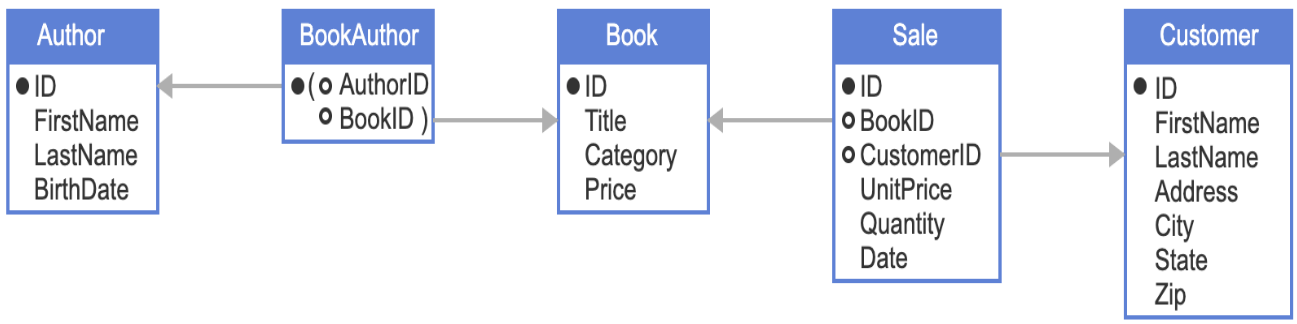
The ***logical design*** phase implements database requirements in a specific database system. For relational database systems, logical design converts entities, relationships, and attributes into tables, keys, and columns. A ***key*** is a column used to identify individual rows of a table. Tables, keys, and columns are specified in SQL with CREATE TABLE statements.

The logical design is depicted in a ***table diagram***. Table diagrams are similar to ER diagrams but more detailed:

- Rectangles represent tables. Table names appear at the top of rectangles.
- Text within rectangles and below table names represents columns.
- Solid bullets (●) indicate key columns.
- Empty bullets (○) and arrows indicate columns that refer to keys.

The logical design, as specified in SQL and depicted in a table diagram, is called a database ***schema***.

Figure 1.4.2: Table diagram.



PARTICIPATION ACTIVITY 1.4.2: Logical design.

Refer to the ER and table diagrams above.

1) CustomerID is:

- ○ A table
- ○ A key
- ○ A column that refers to a key

2) The arrow from Sale to Book corresponds to which column of Sale?

- ○ ID
- ○ BookID
- ○ CustomerID

3) What element of the ER diagram does the BookAuthor table implement?

- Writes
- Book
- Author

4) Logical design is:

- A process only
- A specification only
- Either a process or a specification

5) A schema is depicted in:

- An ER diagram
- A table diagram

Physical design

The ***physical design*** phase adds indexes and specifies how tables are organized on storage media. Ex: Rows of a table may be sorted on the values of a column and stored in sort order. Physical design is specified with SQL statements such as CREATE INDEX and, like logical design, is specific to a database system.

Physical design can be depicted in diagrams. However, logical design is more important for database users and programmers, so physical design diagrams are not commonly used.

In relational databases, logical and physical design affect queries differently. Logical design affects the query result. Physical design affects query processing speed but never affects the query result. The principle that physical design never affects query results is called ***data independence***.

Data independence allows database designers to tune query performance without changes to application programs. When database designers modify indexes or row order, applications run faster or slower but always generate the same results.

Prior to relational databases, most database systems did not support data independence. Performance tuning often forced time-consuming changes to applications. Data independence is a major advantage of relational databases and contributed to the rapid adoption of relational technology in the 1980s.

PARTICIPATION ACTIVITY | 1.4.3: Data independence.

Animation content:

undefined

Animation captions:

1. The initial physical design sorts Book rows by ID.
2. The SQL query selects book titles that cost more than $20.00.
3. A new physical design sorts Book rows by Title.
4. The new physical design has an index on Price.
5. The query scans the index rather than the table, and thus is faster. However the result is the same.

PARTICIPATION ACTIVITY | 1.4.4: Database design process.

Match the term to the description.

If unable to drag and drop, refresh the page.

Analysis **Physical design** **Database design** **Logical design**

	Implementation of database requirements as tables, keys, and columns in a specific database system
	The overall process of determining and implementing database requirements
	Specification of database requirements without regard to implementation
	Affects query performance but not query results

Reset

Programming

Because of data independence, relational database applications can be programmed before the physical design is in place. Applications may run slowly but will generate correct results.

SQL is the standard relational query language but lacks important programming features. Ex: Most SQL implementations are not object-oriented. To write a database program, SQL is usually combined with a general-purpose programming language such as C++, Java, or Python.

To simplify the use of SQL with a general-purpose language, database programs typically use an application programming interface. An ***application programming interface***, or ***API***, is a library of procedures or classes that links a host programming language to a database. The host language calls library procedures, which handle details such as connecting to the database, executing queries, and returning results. Ex: JDBC is a library of Java classes that access relational databases.

Dozens of database APIs are available. Each programming language supports a different API. Major programming languages like C++ and Java support several APIs.

PARTICIPATION ACTIVITY | 1.4.5: Database programming with Python.

Animation content:

undefined

Animation captions:

1. The Book table contains book ID, title, category, and price.
2. The Python code fragment uses the Connector/Python API to access the MySQL database system.
3. A cursor object helps extract query results. The bookCursor object connects to the book database.
4. bookQuery contains a SELECT query that selects the title and category for books that cost more than $20.00
5. The execute() method executes the SELECT query.
6. Each pass through the for loop fetches one query result row into the resultRow variable.
7. resultRow has an element for each result column. The print statements print both elements.

PARTICIPATION ACTIVITY | 1.4.6: Database programming.

1) Each host language, such as Java or C++, has a different API.

- True
- False

2) In the animation above, the cursor helped extract individual rows of the query result for processing in a loop.

- True
- False

3) SQL is a general-purpose programming language.

- True
- False

1.5 Structured Query Language

Structured Query Language

Structured Query Language (***SQL***) is a high-level computer language for storing, manipulating, and retrieving data. SQL is the standard language for relational databases, and is commonly supported in non-relational databases. SQL is pronounced either 'S-Q-L' or 'seekwəl', but the preferred pronunciation is 'S-Q-L'.

The SQL standard is published jointly by the American National Standards Institute (ANSI) and the International Organization for Standardization (ISO). The standard was first published in 1986 and has since evolved through many versions. For details on the standard, see the link in Exploring Further, below.

Relational databases generally support most important elements of the SQL standard. However, most relational databases do not support the entire standard and many support custom extensions. This material describes standard SQL in most cases. Where MySQL differs from the standard, this material describes MySQL syntax and calls out the differences.

PARTICIPATION ACTIVITY 1.5.1: SQL interaction with a database system.

Animation content:

Slide 1. A small box named Database designer appears on the left and a large box named Database system appears on the right. An arrow named SQL commands appears going from the Database designer box to the Database system box. The words CREATE DATABASE appear next to the Database designer box and move along the SQL commands arrow to the Database system box. A new box named World database appears inside the Database system box. The words CREATE

TABLE appear next to the Database designer box and move along the SQL commands arrow to the Database system box. A new table named City appears inside the World database box. The words CREATE TABLE appear next to the Database designer box and move along the SQL commands arrow to the Database system box. A new table named Country appears inside the World database box. The words CREATE TABLE appear next to the Database designer box and move along the SQL commands arrow to the Database system box. A new table named CountryLanguage appears inside the World database box. Slide 2. A small box named Database user appears to the left of the Database system box. An arrow named SQL commands appears from the Database user box to the Database system box. The words INSERT Kabul appear next to the Database user box and move along the SQL commands arrow to the Database system box. Kabul appears in row one of the City table. The words SELECT Afghanistan appear next to the Database user box and move along the SQL commands arrow to the Database system box. Afghanistan then appears in row one of the Country table and then moves underneath the Database user box.

Animation captions:

1. A database designer uses SQL to create a database and the database tables.
2. A database user uses SQL to insert, retrieve, update, and delete data from the tables.

PARTICIPATION ACTIVITY 1.5.2: SQL.

1) SQL commands can create databases and tables.

- True
- False

2) A database designer and database user both use SQL.

- True
- False

3) All database systems use identical SQL statements.

- True
- False

SQL syntax

An SQL ***statement*** is a complete command composed of one or more clauses. A ***clause*** groups SQL keywords like SELECT, FROM, and WHERE with table names like City, column names like Name, and conditions like Population > 100000. *An SQL statement may be written on a single line, but good practice is to write each clause on a separate line.*

PARTICIPATION ACTIVITY | 1.5.3: Three clauses in a SELECT statement.

Animation content:

Slide 1. The code SELECT Name appears. Text appears to the left of the code and states SELECT clause. An arrow appears from this text to the code. Slide 2. The code FROM City appears below the previous code. Text appears below the previous text and states FROM clause. An arrow appears from this text to the second line of code. Slide 3. The code WHERE Population is greater than 100000 appears below the previous code. Text appears below the previous text and states WHERE clause. An arrow appears from this text to the third line of code. Slide 4. A semicolon is added to the end of the third line of code and a right bracket encompassing all three lines of code appears with the text Statement appearing to the right of the bracket.

Animation captions:

1. The SELECT clause starts the statement. Name is a column name.
2. The FROM clause must follow the SELECT clause. City is a table name.
3. The WHERE clause is optional. When included, the WHERE clause must follow the FROM clause. Population > 100000 is a condition.
4. The three clauses ending in a semicolon is a statement. The statement retrieves the names of all cities that have a population greater than 100,000 people.

In MySQL, all SQL statements end with a semicolon. SQL keywords like SELECT, FROM, WHERE, etc. are not case sensitive. Ex: SELECT and select are equivalent. However, identifiers like column names and table names are case sensitive in many database systems. This material uses capital letters for SQL keywords so the keywords stand out from other syntactic parts. The table below summarizes various syntactic features of SQL.

Table 1.5.1: Summary of SQL syntax features.

Type	Description	Examples

Literals	Explicit values that are string, numeric, or binary.
	Strings must be surrounded by single quotes or double quotes.
		`'String'`
		`"String"`
		`123`
		`x'0fa2'`

	Binary values are represented with `x'0'` where the 0 is any hex value.	
Keywords	Words with special meaning.	SELECT, FROM, WHERE
Identifiers	Objects from the database like tables, columns, etc.	City, Name, Population
Comments	Statement intended only for humans and ignored by the database when parsing an SQL statement.	`-- single line comment` `/* multi-line Comment */`

All database systems recognize single quotes for string literals, but some also recognize double quotes. This material uses single quotes to ensure the SQL statements are compatible with all database systems.

PARTICIPATION ACTIVITY 1.5.4: SQL syntax.

1) The INSERT statement adds a student to the Student table. How many clauses are in the INSERT statement?

```
INSERT INTO Student
VALUES (888, 'Smith', 'Jim',
3.0);
```

- 1
- 2
- 3

2) The SQL statement below is used to select students with the last name "Smith". What is wrong with the statement?

```
SELECT FirstName
FROM Student
WHERE LastName = Smith;
```


- The literal "Smith" must be surrounded by single quotes or double quotes.
- The WHERE clause should be removed.
- The last name "Smith" may not exist in the database.

3) What is wrong with the SQL statement below?

```
SELECT FirstName
from Student
```

- The keyword from must be written FROM.
- The WHERE clause is missing.
- A terminating semicolon is missing.

4) What is wrong with the SQL statement below?

```
SELECT Gpa
--FROM Student;
```

- The FROM clause is a comment.
- The WHERE clause is missing.
- Gpa should be a table name.

SQL sublanguages

The SQL language is divided into five sublanguages:

- ***Data Definition Language*** (DDL) defines the structure of the database.
- ***Data Query Language*** (DQL) retrieves data from the database.

- ***Data Manipulation Language*** (DML) manipulates data stored in a database.
- ***Data Control Language*** (DCL) controls database user access.
- ***Data Transaction Language*** (DTL) manages database transactions.

PARTICIPATION ACTIVITY | 1.5.5: SQL sublanguages.

Animation content:

Slide 1. A cylinder named Database appears containing the words City and Country. The caption DDL appears. A line strikes through Country and a new word Personnel appears below Country. Slide 2. The caption DQL appears. The Personnel table appears next to the caption with one row. Slide 3. The caption DML appears. The values of the row in Personnel change and a new row is inserted. Slide 4. The caption DCL appears next to a table with columns Username and Role. The table has three rows. A line strikes through the second row. Slide 5. The caption DTL appears. An arrow appears from the Personnel table to the Database cylinder. The arrow is labeled saved.

Animation captions:

1. DDL creates, alters, and drops tables.
2. DQL selects data from a table.
3. DML inserts, updates, and deletes data in a table.
4. DCL grants and revokes permissions to and from users.
5. DTL commits data to a database, rolls back data from a database, and creates savepoints.

PARTICIPATION ACTIVITY | 1.5.6: SQL sublanguages.

Match the SQL sublanguage to the statement behavior.

If unable to drag and drop, refresh the page.

DQL DDL DTL DCL DML

	Insert a data row into table product.
	Rollback database changes.
	Select all rows from table Product.

	Grant all permissions to user 'tester'.
	Create table Product.

Reset

Exploring further:

- SQL standard (Wikipedia)

1.6 MySQL

MySQL

This material uses MySQL as a reference relational database system. Although the material is relevant to all relational databases, SQL syntax and many activities are based on MySQL.

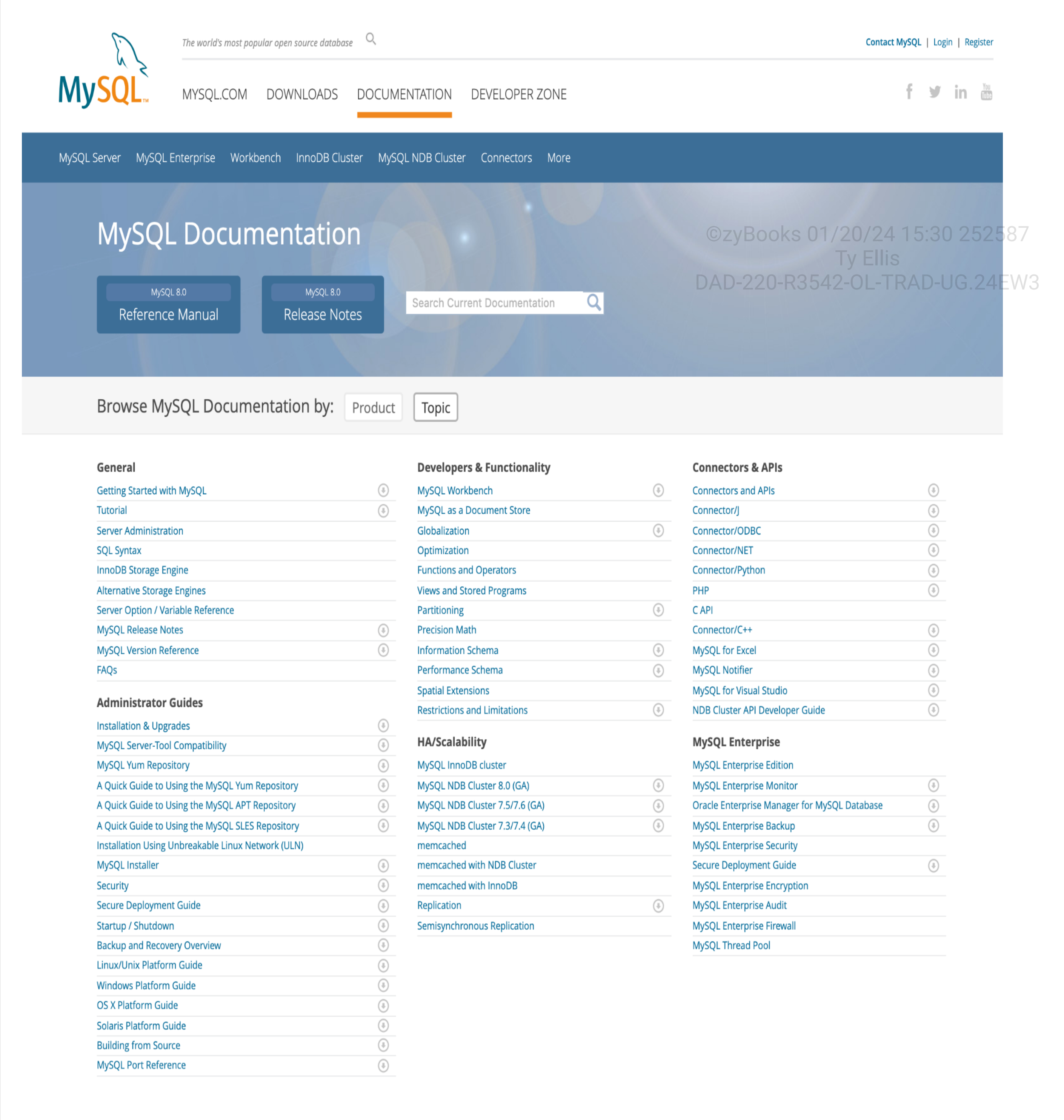
MySQL is a leading relational database system sponsored by Oracle. MySQL is relatively easy to install and use, yet has many advanced capabilities. MySQL runs on all major operating systems, including Linux, Unix, Mac OS, and Windows. For these reasons, MySQL is one of the most popular database systems.

MySQL is available in two editions:

- ***MySQL Community***, commonly called ***MySQL Server***, is a free edition. MySQL Server includes a complete set of database services and tools, and is suitable for non-commercial applications such as education.
- ***MySQL Enterprise*** is a paid edition for managing commercial databases. MySQL Enterprise includes MySQL Server and additional administrative applications.

This book is based on MySQL Server release 8.0. Forthcoming versions of this book will be upgraded to MySQL Server 8.1, released in July 2023. Complete documentation for MySQL Server 8.0 is available online.

Figure 1.6.1: MySQL documentation (dev.mysql.com/doc).



Instructions for downloading and installing MySQL Server 8.0 on Windows or Mac OS are available from the 'Exploring further' links below.

©zyBooks 01/20/24 15:30 252587
Ty Ellis
DAD-220-R3542-OL-TRAD-UG.24EW3

When installing MySQL Server, the user must enter a password for the ***root account***, the administrative account that has full control of MySQL. Other database user accounts may optionally be created. After installation, MySQL Server runs as a service in the background. MySQL Server automatically starts and stops when the operating system starts and stops.

PARTICIPATION ACTIVITY 1.6.1: MySQL.

1) Refer to the website db-engines.com. What is the overall MySQL ranking, compared to all database systems?

- 1
- 2
- 5

2) What account can create other user accounts?

- MySQL account
- Root account
- All accounts can create user accounts

3) Content in this material only applies to the MySQL database system.

- True
- False

MySQL Command-Line Client

The ***MySQL Command-Line Client*** is a text interface included in the MySQL Server download. The Command-Line Client allows developers to connect to the database server, perform administrative functions, and execute SQL statements.

To run the Command-Line Client, a user must first open a Command Prompt on Windows or a Terminal on a Mac:

- Windows: Click the Start button in the Taskbar, type "cmd", then click Command Prompt.
- Mac: Click on the Terminal application, usually found in the Applications > Utilities folder.

When MySQL Command-Line Client is started with the root account, the user is prompted to enter the root account password. Then Command-Line Client attempts to connect to the database server running on the local machine.

PARTICIPATION ACTIVITY | 1.6.2: Using the MySQL Command-Line Client.

Animation content:

Slide 1. A command line appears and the first line states Enter password. Slide 2. More lines of appear and state. Welcome to the MySQL monitor. Commands end with semicolon or backslash g.

Your MySQL connection id is 3. Dot. Dot. Dot. Type apostrophe help semicolon apostrophe or apostrophe backslash h apostrophe for help. Type '\c' to clear the current input statement. My s q l greater than. Slide 3. Text is added to the last line. The text that is added states USE world semicolon. A new line appears and states my s q l greater than SELECT asterisk FROM City semicolon.

Animation captions:

1. From a command-line prompt, the user starts the MySQL Command-Line Client. The -u option names the account, and -p indicates a password must be entered.
2. The user enters the root password established during installation.
3. After a successful login, the user is presented a startup screen and a mysql prompt.
4. Commands are entered one by one on the command line. The SELECT statement returns the first 10 rows from the city table.

The animation above shows the user typing SQL commands that use the 'world' database, a database that is usually installed with MySQL. The world database contains three tables: city, country, and countrylanguage. Users can practice entering SQL statements that work with and manipulate the world database. Some installations do not include the world database, so users must download and install the world database from MySQL.com separately.

MySQL Server returns an ***error code*** and description when an SQL statement is syntactically incorrect or the database cannot execute the statement.

Figure 1.6.2: MySQL error codes.

```
mysql> SELECT FROM city;
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that
corresponds to
your MySQL Server version for the right syntax to use near 'FROM city' at line 1

mysql> INSERT INTO city VALUES (123, 'Amsterdam', 'NLD', 'Noord-Holland', 731200);
ERROR 1062 (23000): Duplicate entry '123' for key 'PRIMARY'
```

PARTICIPATION ACTIVITY 1.6.3: Introduction to MySQL.

1) The root account password is set when installing MySQL.

- True
- False

2) The database server must be manually started each time the user runs the MySQL Command-Line Client.

- True
- False

3) The MySQL Command-Line Client provides a graphical interface for interacting with the database server.

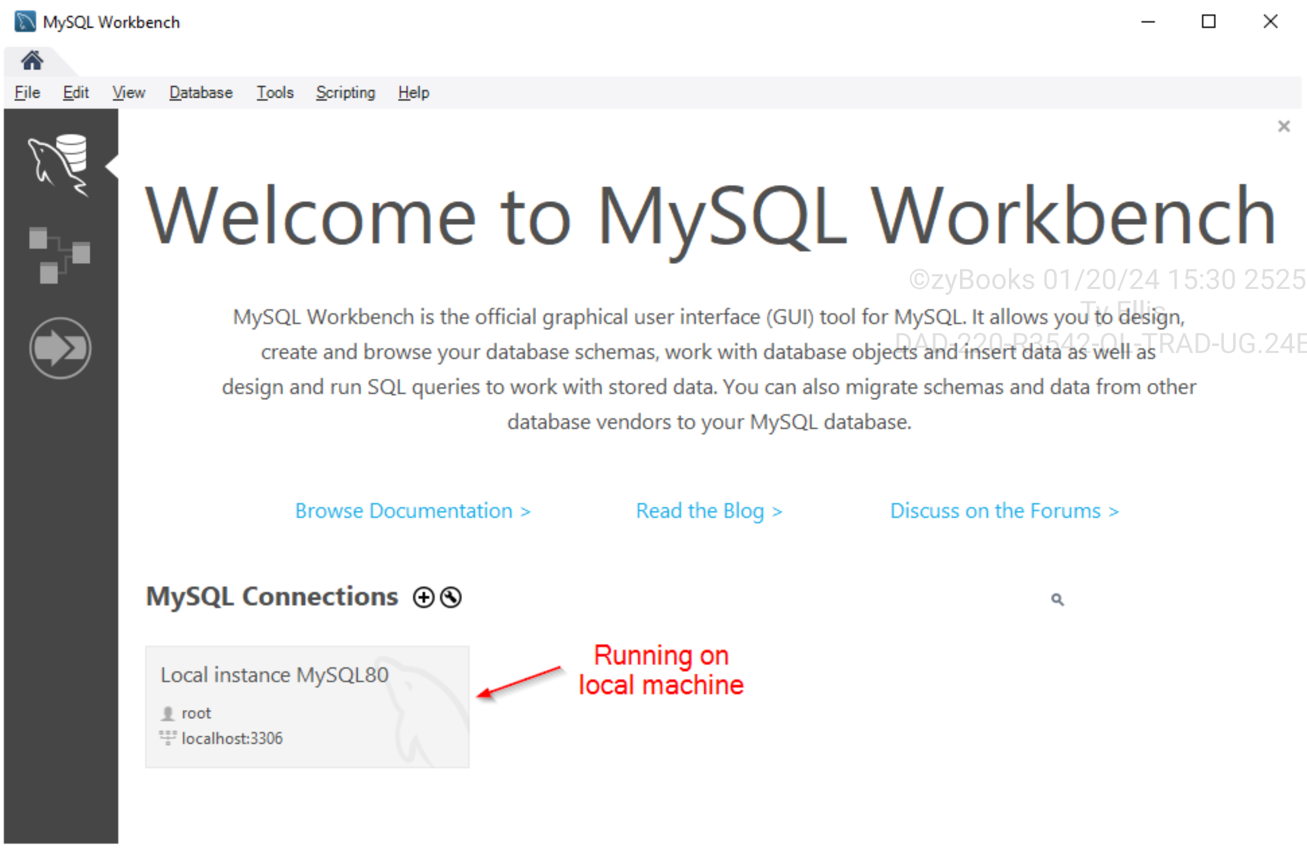
- True
- False

MySQL Workbench

Some developers prefer to interact with MySQL Server via a graphical user interface. ***MySQL Workbench*** is installed with MySQL Server and allows developers to execute SQL commands using an editor. When MySQL Workbench is started, the user can connect to MySQL Server running on the local machine or on the network.

The figure below shows the MySQL Workbench home screen on Windows. The Mac version has some minor differences. Clicking on the box labeled **Local Instance MySQL80** connects to MySQL Server running on the same computer as MySQL Workbench.

Figure 1.6.3: MySQL Workbench home screen.

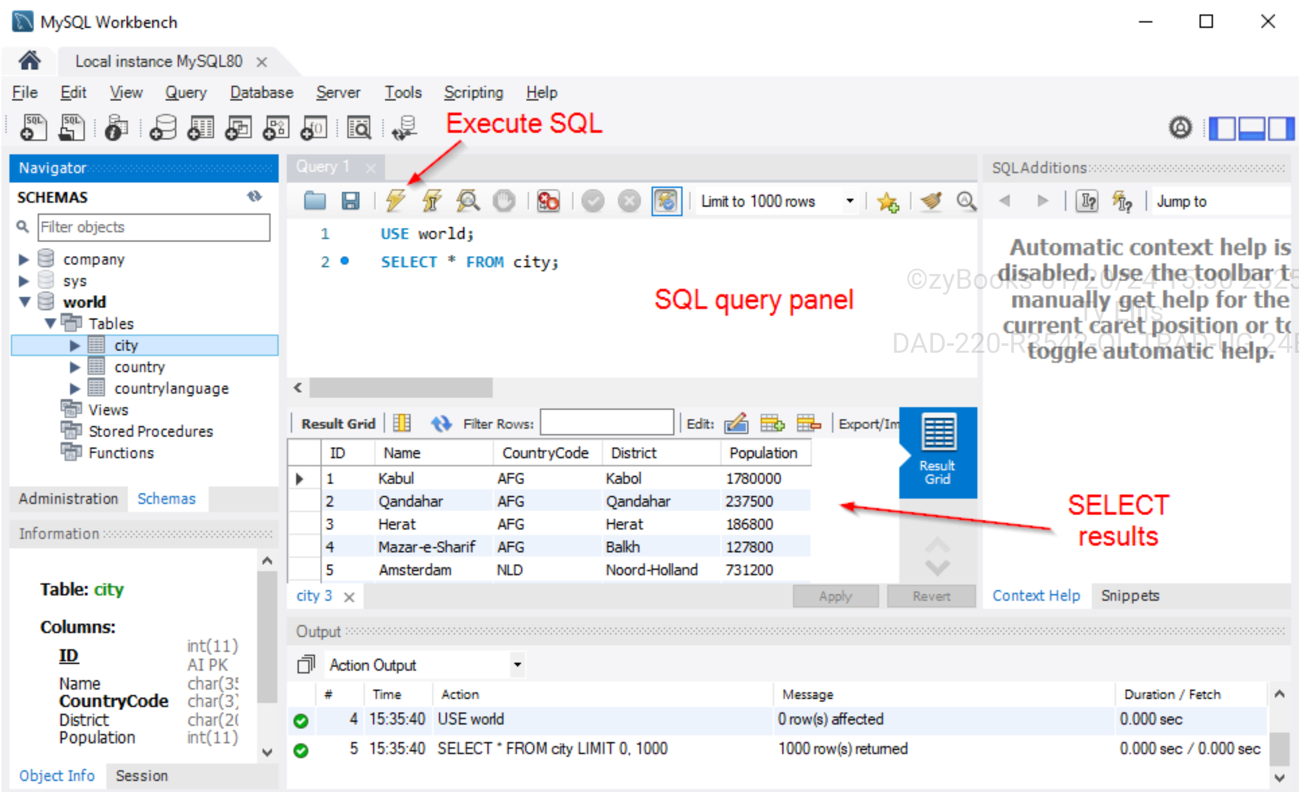


After connecting to MySQL server, Workbench shows the **Navigator** sidebar on the left with two tabs:

1. The **Administration** tab shows various administrative options, like checking the server's status, importing/exporting data, and starting/stopping the MySQL server.
2. The **Schemas** tab shows a list of available databases. A database can be expanded to show the database's tables.

The figure below shows the world database's three tables: city, country, and countrylanguage. The query panel is where the user enters SQL statements. Pressing the lightning bolt icon executes the SQL statements and shows the results below the query panel. A summary of the executed statements is shown at the bottom of the window.

Figure 1.6.4: MySQL Workbench executing a SELECT statement.



PARTICIPATION ACTIVITY 1.6.4: MySQL Workbench.

1) MySQL Workbench and MySQL Command-Line Client both allow the user to type SQL statements.

- True
- False

2) The SQL statements in the SQL query panel are not executed until the lightning bolt is clicked.

- True
- False

3) The MySQL Workbench screenshot above shows the columns that make up the City table.

- True
- False

Exploring further:

- MySQL download
- MySQL installation video (Windows)
- MySQL installation video (Mac)
- MySQL documentation - home page
- MySQL documentation - installation
- MySQL documentation - Workbench
- MySQL documentation - problems and common errors
- MySQL documentation - Sakila database installation