Introduction to SQL

Structured Query Language

Why do we care about SQL?

- Lots of data is available in SQL databases
- You may be tasked with using this data and you have to be able to get it out of the database.
- Efficient storage and retrieval of records
- Good for storing MASSIVE amounts of data.
- You may want to make your own SQL database at some point

The Database Management System (DBMS)

- Place where we store and manage data needed to support an app or website.
- There are many DBMS for different purposes. Two common DBMS are referred to as SQL or NoSQL.
  - SQL databases are often relational and are commonly used to store business and website data. Data is in tables with columns and rows.
  - NoSQL databases are “object” databases where each column can contain object data often in JSON format.
- SQL is the Structured Query Language we use to manage the data in the system.
- Oracle, MySQL, and SQL Server are the top 3 Relational DBMS.
  - https://db-engines.com/en/ranking
- We will be using SQLite for its simplicity and Google BigQuery because it contains open source data.

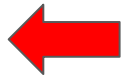
The Data

- In a relational database, data is stored in tables also known as entities.
  - Each table is designed for a specific purpose
  - Tables look like Excel worksheets or Pandas DataFrames with rows and columns.
    - Columns are also called fields or attributes.
    - Rows are also called records and have a unique identifier similar to Pandas indexes.

Column

Sample table:

EmployeeId	FirstName	LastName	Sex	BirthDate
111	Prisha	Agarwal	Female	1975-Aug-01
222	Miguel	Garcia	Male	1987-May-15

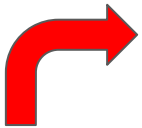


Row

Data Normalization

- Most SQL databases are relational meaning one table relates to another table through a unique identifier.
- Relationships are designed to reduce duplicate data - data normalization.

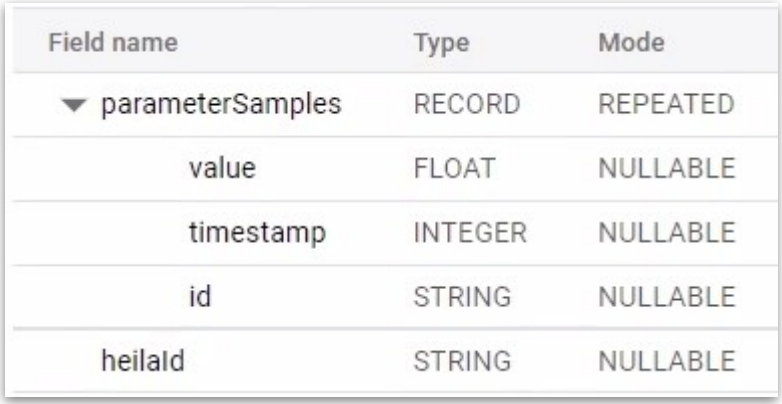
EmployeeId	FirstName	LastName	Sex	BirthDate
111	Prisha	Agarwal	Female	1975-Aug-01
222	Miguel	Garcia	Male	1987-May-15



AddressId	EmployeeId	Address	City	State
444	222	123 My Street	Los Alamos	NM
555	222	75 Avenue East	San Dimas	CA

Database Normalization and Data Science

- Goals are different
  - Speedy access across massive datasets
  - Easy querying
  - Data visualization
- Data may be denormalized containing duplicate information to avoid joining tables.
- NoSQL or hierarchical data may be more common
  - Example is storing data from IoT devices in a single column.



Field name	Type	Mode
▼ parameterSamples	RECORD	REPEATED
value	FLOAT	NULLABLE
timestamp	INTEGER	NULLABLE
id	STRING	NULLABLE
heilaId	STRING	NULLABLE

Data Types

A data type describes the data that we will be storing in the database.

Database designers usually choose the data type that consumes the **least** amount of space in the system but still meets the requirements for their applications.

Common data types fall into the following categories:

- Character data - names, addresses, descriptions
- Numeric data - row ids, classification codes/encoded data, continuous values like temperature, cost, weight
- Dates and times - birth dates, application dates, sample times
- Binary data - pdf files, images, device readings

Why do I care about data types?

- How you query the data will vary depending on the data type.
- You can avoid truncating query results if you know how long a character field/column is.
- If combining data from multiple tables, it is helpful to know the data length for similar fields in the different tables.
- Models often require data in encoded numeric format.
- Certain data types may be necessary for preparing graphical visualizations.
- Numeric data types must be known for proper calculations to avoid unexpected rounding errors.
- And the list goes on….

The Language - Structured Query Language (SQL)

- SQL (pronounced Sequel or S-Q-L) is the language used to manage SQL databases and their data.
  - SQL varies a little bit for each DBMS.
  - SQL sounds like regular spoken english. Designed to be intuitive and easy to learn.
- SQL has different sub-languages
  - DDL - Data Definition Language - is used to create and define entities in a database.
  - DCL - Data Control Language - is used to control access to objects in a database; permissions.
  - DML - Data Manipulation Language - is used for Creating Reading Updating Deleting data in the database
    - Also called CRUD operations.
- We will be focusing a little on DDL, but mostly on DML for reading data out of SQL databases.

Data Definition Language - DDL

Common DDL commands are:

CREATE - create a database object; table, index, view, procedure

ALTER - change a database object

DROP - remove/delete a database object

This is mostly outside the scope of this class.

Data Control Language - DCL

Common DCL commands are:

Object permissions

GRANT - give users access to specified database objects

REVOKE - remove user access to specified database objects

Transaction level handling

COMMIT - after a series of data transactions are complete, finalize them

ROLLBACK - while a series of data transactions are in process, you can undo them

This is beyond the scope of this class.

Data Manipulation Language - DML

Common DML commands are:

INSERT - Create new records in database tables

SELECT - Read records from database tables

UPDATE - Update existing records in database tables

DELETE - Delete existing records in database tables

In this class you will be be primarily using the SELECT command.

Let’s get started!