

Introduction to Data Access and Storage

```
$ echo "Data Sciences Institute"
```

Introduction:

→ **Welcome**

What is SQL?

Data Modelling

Welcome / About Us

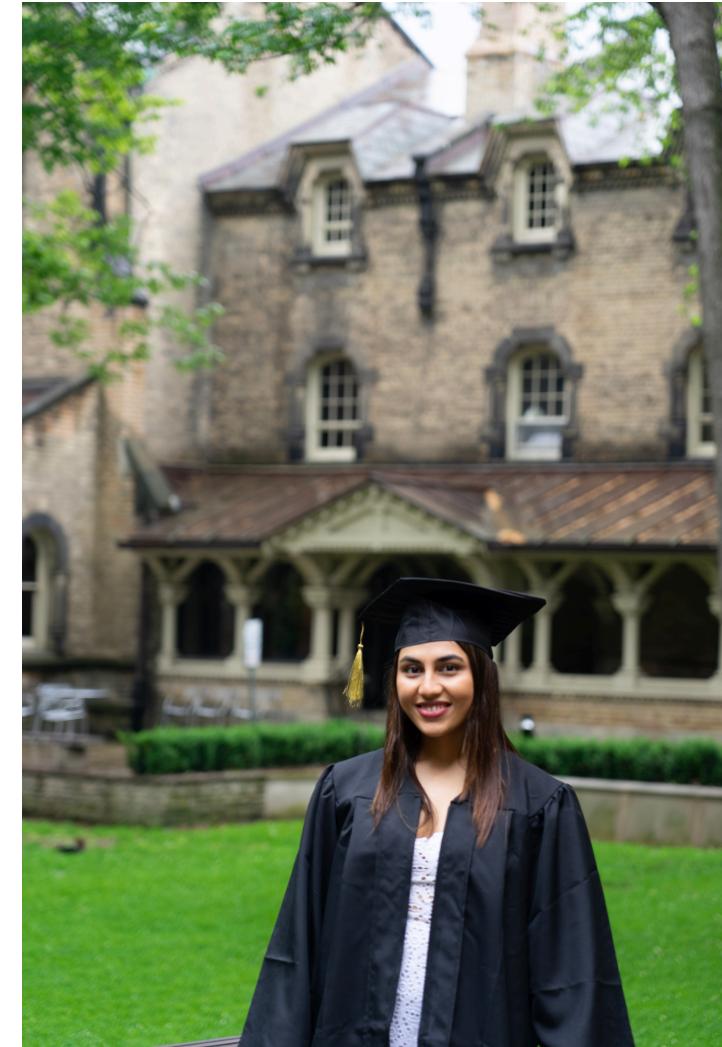
About Us (Thomas)

- Started with SQL in first job after graduating
- Data Analyst in US healthcare until moving to Canada in 2018
- Transitioned to Data Engineer at Plan Canada
- Pursued Master of Information at UofT for R and Python focus
- Currently working on AI Governance and Ethics at Dataiku
- *Have a 2 year old Samoyed named Alto who takes up a lot of my free time* ➡



About Us (Ananya)

- UofT Computer Science and Statistics graduate, with minor in Mathematics
- Co-op at Citi (Software Development) and BMO (Business Analytics)
- RBC Capital Markets BSA using SQL for Salesforce CRM data, transitioning to Supervisory and Regulatory role soon
- Former Women in Computer Science president at UofT
- Involved with Toronto Wxmen in Data Science and RWomen at RBC



About Us (Amanda)

- UofT Statistics undergraduate (entering 4th year)
- Self-taught SQL through online resources
- Currently in summer research program under UofT Statistics department
- Considering MSc in Statistics
- Hobbies: piano, puzzle, Netflix



About Us (Sidra)

- Bachelor's in Biomedical Engineering from NED-UET
- Postgraduate diploma in data analytics from Toronto School Of Management.
- Machine Learning Software Foundations Certificate from DSI-University Of Toronto.
- 4+ years of experience in business intelligence and data analysis.
- Engaged in Continuously Learning and Upskilling my skills



Welcome / Course Content

Course Content

- Data Modelling, Data Structures, Schemas, Basic Data Management, Normal Forms
- Basic SQL Syntax
- Essential SQL Syntax
- Advanced Techniques
- Importing and Exporting Data to and from SQL
- SQL's relationship to the Machine Learning Pipeline
- Data Stewardship, Ethics, and SQL in the Wild

GitHub Repo

<https://github.com/UofT-DSI/sql>

- Schedule
- These slides (HTML & PDF)
- Our database for live coding
- All in-class code
- Homework (and answers, posted afterwards)
- Assignment details and rubrics
- Due Dates



It is crucial you visit the repo throughout the course, as I may make minor changes and push new content.

Course Content

- This course is an *Introduction* to SQL
- At the end of the course, I hope you will:
 - Feel comfortable with SQL
 - Know how to search for the right thing on Stack Overflow
 - Read documentation
- We won't cover advanced topics like:
 - Stored Procedures, Triggers, Jobs
 - DBA work (monitoring, server setup, etc)
 - Complex ETL or tooling



The screenshot shows the DB Browser for SQLite application. On the left, there is a code editor window containing SQL queries. On the right, there is a results pane displaying the output of the queries.

SQL Queries:

```
1 select * from patients
2 where address_postcode like '%High St%'
3 and date > 2023-01-01
4 and type = 'House'
5
6 select * from patients
7 where address_postcode like '%Northmembers Dr%'
8
9 select * from patients
10 where address_postcode like '%Westgate Lane'
11
12
13 select * from patients
14 where address_postcode like '%Westgate Lane'
15
16 select * from patients
17 where patient_id = 1000000000
18
19 select * from patients
20 where patient_id = 1000000001
21
22 select * from patients
23 where patient_id = 1000000002
24
25 select * from patients
26 where patient_id = 1000000003
27
28 select * from patients
29 where patient_id = 1000000004
30
31 select * from patients
32 where patient_id = 1000000005
33
34 select * from patients
35 where patient_id = 1000000006
36
37 select * from patients
38 where patient_id = 1000000007
39
40 select * from patients
41 where patient_id = 1000000008
42
43 select * from patients
44 where patient_id = 1000000009
```

Results:

patient_id	name	address	postcode	date	type
1000000000	Michael Jackson	123 Main Street	High St 123	2023-01-01	House
1000000001	John Smith	456 Elm Street	High St 456	2023-01-02	House
1000000002	Jane Doe	789 Oak Street	Northmembers Dr 789	2023-01-03	House
1000000003	David Miller	101 Pine Street	Westgate Lane 101	2023-01-04	House
1000000004	Sarah Johnson	202 Cedar Street	Westgate Lane 202	2023-01-05	House
1000000005	Robert Wilson	303 Elm Street	High St 303	2023-01-06	House
1000000006	Emily Davis	404 Cedar Street	Westgate Lane 404	2023-01-07	House
1000000007	James Brown	505 Elm Street	High St 505	2023-01-08	House
1000000008	Sarah Johnson	606 Cedar Street	Westgate Lane 606	2023-01-09	House
1000000009	Robert Wilson	707 Elm Street	High St 707	2023-01-10	House

Software:
DB Browser for SQLite

Homework (required)

- Six homeworks total
- At the end of each SQL module, I will provide you with 3-10 queries to write on your own
 - Covers the topics of the modules
- Review answers in Office Hours with myself or course support
 - I may go over one or two queries at the start of class if desired
- Designed to be relatively easy
 - Reaffirms what we wrote together
 - Doing work on your own helps reinforce the learning
- ChatGPT probably won't help you much

Grading

- Pass/Fail. Do the work, pass the course :)
- One Assignment: Data Model Design, 40%, due May 25
- Homework: 48% (8% each), due on Thursdays and Saturdays
- Class Attendance: 12% (2% each)
 - Let myself or course support know if you are unable to attend a lesson
 - Code along!! Best way to learn.

What questions do you have about the course?

Welcome / Quick Technical Check

Quick Technical Check

Let's make sure everyone has DB Browser for SQLite installed.

- If not, please download it here: <https://sqlitebrowser.org/dl/>

For live coding:

- Please download/fork the FarmersMarket.db from our GH repo:
 - [FarmersMarket.db](#)
- Open it in SQLite with the "Open Database" button and navigate to wherever you have saved it

Good to go?

If not, please message your course support

Introduction:

Welcome

→ [What is SQL?](#)

Data Modelling

What is SQL? / SQL

SQL Fundamentals

- SQL: Structured Query Language
 - Pronounced as either S.Q.L. (ess-cue-ell) or “sequel”
- SQL is a *query* language rather than a programming language
 - Querying is closer to telling a computer *what you want*, rather than *what is has to do*
 - SQL code is often less reproducible than other programming languages because it's domain specific
 - Some SQL code, especially more advanced procedural code, is reproducible within the same flavour
 - SQL's domain is databases and is based on set theory

SQL Fundamentals

- Designed to manage data within Relational Database Management Systems (RDBMs),
e.g.
 - MSSQL
 - Oracle DB
 - MySQL/MariaDB
 - PostgreSQL

SQL Formatting

- Like other programming/query languages, SQL has reserved keywords/commands to perform instructional operations
 - Generally, these keywords are written in all caps: `SELECT`
 - Most modern interpreters no longer require this, but it is the expected standard
- All statements/queries should end with a semicolon
 - A few SQL constructs (like common table expressions, we'll get to these later) require them, otherwise they are optional
 - I'll almost certainly forget to use them
 - There's some debate over whether or not it's best practice

SQL Formatting

- In SQL, white space and/or line breaks do not matter
 - Readability is important
 - Try to keep SQL statements to a reasonable screen width
 - Use sensible line breaks
 - Offset subqueries with indents

SQL Formatting

- Code is commented in/out with `--` rather than `#`
- Code blocks can be commented out with `/* */`

```
/*
somecode spanning
multiple lines

*/
```

What is SQL? / Flavours

Flavours

- RDBMs differ from one to the next:
 - different keywords
 - e.g. return only 10 rows:
`SELECT TOP 10...` vs `SELECT ... LIMIT 10`
 - different syntax
 - e.g. not equal:
`!=` or `<>` (or both)
 - other, more nuanced/complex differences
 - e.g. architecture, data types, etc

Flavours

- We are using ***SQLite***:
 - Super easy to get setup
 - Requires almost no overhead
 - Open source, *free*
 - Similar enough in syntax to learn on
 - Used all over the world and in many applications
 - e.g. Firefox uses a SQLite backend to write a user's history locally

Flavours

- Broad observations about Open Source systems:
 - Excellent at what they are designed for
 - Varying data types (SQLite has some unique ones!)
 - Not every command exists, but workarounds are usually possible
 - Some utilize RDBMs that feel extremely outdated

Flavours

- Broad observations about enterprise systems:
 - Powerful and designed to handle edge cases
 - Feel a bit more refined
 - Can be version dependent
 - Tend to "lock in" businesses/organizations
 - Migration is costly, sometimes outrageously so
 - Newer players (Snowflake, Databricks, etc) and cloud providers (Azure, AWS, GCP, etc) offer a lot more functionality than just database querying
 - Sometimes use different terminology to describe SQL tasks

What is SQL? / Environments for SQL

Databases

- Relational databases are a collection of tables, views, procedural code, and other SQL-assisting artefacts
 - Generally the data stored in a database will be related to a real-world concept
 - Backends to data-collecting systems are often databases
 - e.g. CRMs, EMR software, ERPs, web-based applications
 - Usually not connected to other databases unless deemed necessary
 - Often transactional, meaning data is actively being written to by frontend systems
 - Tables are normalized

Databases

- There are also non-relational databases, often referred to as NoSQL
 - We won't cover these
 - Common tools include: Amazon DynamoDB, Azure CosmosDB, MongoDB, Google Cloud Datastore

Data Warehouses and Data Marts

- Data Warehouses are highly structured collections of (usually tabular) data
 - Data has been processed for a specific purpose, e.g. analytics
 - Data has been centralized
 - Often with the assistance of ETL (Extract, Transform, Load) tools
 - Have a lot of overhead, require governance, and strict rigidity
 - Tables are denormalized
 - Very common for enterprises, but losing traction in many industries

Data Warehouses and Data Marts

- Data Marts are created from Data Warehouses to focus on a single subject
 - Designed to make Data Warehouses easier to use
 - Data is structured, but flexibility is driven by the purpose of the Data Mart
 - Some denormalized tables may be normalized or undergo even greater normalization
 - *The subject/purpose of the Data Mart might drive these types of decisions*
 - Common for enterprises that have Data Warehouses

Data Lakes and Data Swamps

- Data Lakes allow on-demand access to raw, semi-structured, structured, and unstructured data
 - Not defined by a specific purpose
 - Highly scalable
 - Can be transactional, if systems are designed to produce outputs into Data Lakes
 - Often data sources for machine learning pipelines live in Data Lakes
 - Inexpensive compared to Data Warehouses
 - Increasingly common for enterprises to shift towards Data Lakes, especially with support from newer tools like Snowflake, Databricks, etc, which can maximize the analytical value of a Data Lake

Data Lakes and Data Swamps

- Data Swamps...are poorly governed Data Lakes
 - Lack of documentation, lack of governance, poorly designed Data Lakes become Swamps
 - Avoid building these

What questions do you have?

Introduction:

Welcome

What is SQL?

→ Data Modelling

Data Modelling Topics (its a lot!)

- Relational Database Management Systems
- Data Models
- Structure of Data
- Constraints
- Entity Relationship Diagrams
- Attributes of an ERD: Entities & Relationships
- Relationship Examples
- Conceptual, Logical, Physical Models
- Assignment 1: Design a Logical Model

Relational Database Management Systems

- Relational Database Management Systems (RDBMs) are software designed to:
- Store large amounts of data
- Utilize a query language to allow easy retrieval of the data
- Allow multiple users to access the data simultaneously
- Manage permissions for data access
- Mitigate data corruption and unauthorized access

Relational Database Management Systems

- Generally, data is stored in a *database*
 - A database is a collection of information
 - Within a database, a collection of objects (e.g. tabular data "tables") is stored

Relational Database Management Systems

- RDBMs allow users to define interactions between these objects, such as:
 - Establish the relationship between objects
 - Define procedural scripts to query specific data or trigger an action
 - Schedule routine work (e.g. procedures to run, maintenance, etc)

Data Models

- A data model is a notation for describing data or information
- Data models consist of:
 - Structure of the data
 - Operations
 - Constraints on the data
 - Relationships

Structure of Data

- SQL is comprised of tables

Breed	Affectionate w/ Family	Good w/ Other Dogs	Shedding	Coat Type	Coat Length	Playfulness	Energ
Pugs	5	4	4	Smooth	Short	5	3
Akitas	3	1	3	Double	Medium	3	4
Samoyeds	5	3	3	Double	Long	5	4

Structure of Data

- Tables have Attributes and Observations
 - In SQL we call Attributes "Columns"
 - e.g. Breed, Coat Type, Coat Length
 - and Observations "Rows"
 - e.g. Samoyed, Double, Long
- SQL databases require tables to be named
 - e.g. We can call this table "breed_traits"

Structure of Data

- Columns are defined (and restricted, i.e. constrained) by data types
- Common ones include:
 - `INT` (integers: 1,2,3,-1,-2,-3)
 - most systems conserve storage space specifying their range
 - `FLOAT` , `DECIMAL` , `REAL` (decimal: 5.5, 3.33333)
 - `VARCHAR` , `NVARCHAR` , `TEXT` (text strings, with a maximum length associated: 'abc')
 - `DATE` , `DATETIME` , `TIME` (dates and times: '2023-01-09', '11:11:11.000')
- These may vary slightly by flavour (in SQLite they are simpler and less restricted)

Structure of Data

- Data types are important:
 - They affect operation speed, storage size, data validity
 - Speed: it's computationally less expensive to compute smaller values
 - Storage: small is usually better, but the wrong size will affect systems; e.g. Psy's Gangnam Style exceeded $2,147,483,647$ (32-bit signed, $2^{32}/2-1$) views, causing YouTube to expand the view counter to 64-bit (~9.2b if signed)
 - Validity: ensures columns contain the right type of data for operations, e.g. avoiding $5 + 'ten' = ??$

Constraints

- Data Models also specify constraints
- Constraints are rules that must be followed:
 - Referential-Integrity constraints
 - Ensure that values in one table have corresponding values in another table
 - Attribute Constraints
 - Ensure that certain types of values are always consistent within columns
 - May also ensure whether values are unique, not missing, etc

Constraints

- **NULL and NOT NULL**
 - If a value can be missing or not
- **UNIQUE**
 - All values are different
- **PRIMARY KEY (PK)**
 - Ensures each value in a column is unique within the table (e.g. an ID field)
 - One PK per table
 - Cannot be NULL
 - Ensures database integrity by restricting record deletion

Constraints

- **FOREIGN KEY (FK)**
 - Creates a linkage between a column in one table and a column in another table
 - Generally, foreign keys are linked to primary keys
 - Sometimes share the same name as the linked column, but this isn't required
 - Linkage requires data types to be the same
 - As many FKS as needed per table
 - May be NULL
 - Record can be deleted

Entity Relationship Diagrams

- Entity Relationship Diagrams (ERDs) are diagrams depicting the structure of tables within a database
 - This both *identifies the tables and describes their relationships*
- ERDs are useful for:
 - Database design
 - Debugging
 - Writing logical, consistent, and efficient queries

Entity Relationship Diagrams

- There are three levels of detail for ERD depictions:
 - Conceptual model
 - Logical model
 - Physical model

Attributes of an ERD Entity

- For a given table:
 - Name
 - Relationship to another table
 - Column Names
 - Column Types
 - Primary Keys (if present)
 - Foreign Keys (if present)

Attributes of an ERD Relationship

- Defines which columns are related
- Defines what type of relationship exists:
 - One-to-One
 - One-to-Many
 - Many-to-Many

Relationship Examples

One-to-One: where a given row within a table is associated with only a single row in another table

Table 1: Country — Table 2: Capital City

Table 1:Country		Table 2:Capital
Canada	1:1	Ottawa
USA	1:1	Washington DC
Mexico	1:1	Mexico City

Relationship Examples

One-to-Many: where a given row within a table can be referenced by multiple rows in another table

Relationship Examples

Table 1: Country — Table 2: States

Table 1:Country		Table 2:States
Canada	1: ∞	Alberta
Canada	1: ∞	British Columbia
Canada	1: ∞	...(11 more rows)
USA	1: ∞	Alabama
USA	1: ∞	Alaska
USA	1: ∞	...(48 more rows)
...

Relationship Examples

Many-to-Many: where multiple rows within a table can be referenced by multiple rows in another table

For this example, consider different ways to define "European" membership, such as whether or not a country: 1) is a member of the EU, 2) uses the Euro, or 3) has abolished border controls (Schengen Agreement)

Table 1: Country — Table 2: Membership

Country	Country ID
Slovenia	001
Sweden	002
Switzerland	003
...(more countries)	...

Membership	Member ID
EU	10
Eurozone	11
Schengen	12

Relationship Examples

Table: Different countries align differently within these definition

Country ID		Member ID
001	$\infty:\infty$	10
001	$\infty:\infty$	11
001	$\infty:\infty$	12
002	$\infty:\infty$	10
002	$\sim:\sim$	12

We can create additional Many-to-Many relationships if we created a table including NATO/UN membership, because many non-European countries are NATO/UN members.

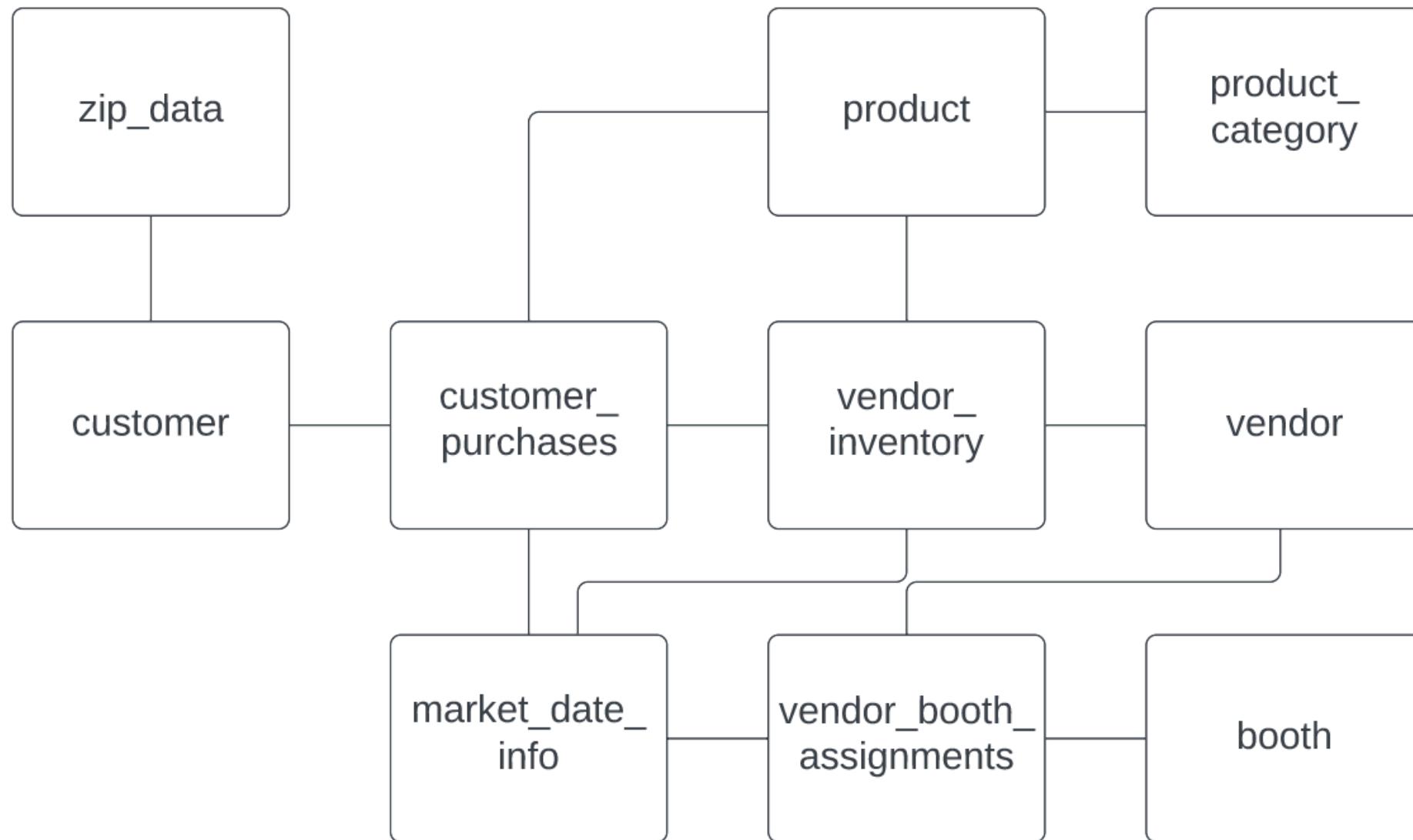
Relationship Examples

We'd create additional Many-to-Many relationships if we created a course table, because A and Ananya are involved in multiple courses!

Conceptual Models

- Define the tables (objects/entities) and their relationships
- Our Farmers Market database:
 - 10 tables
 - Relationships between these tables
 - e.g. product and product_category: *what type of thing a product is*
 - product and customer_purchases: *what products a customer has bought*
 - product and vendor_inventory: *what products each vendor has available*
 - Not all tables share a relationship to one another, but all tables have at least one relationship

Farmers Market Conceptual Model



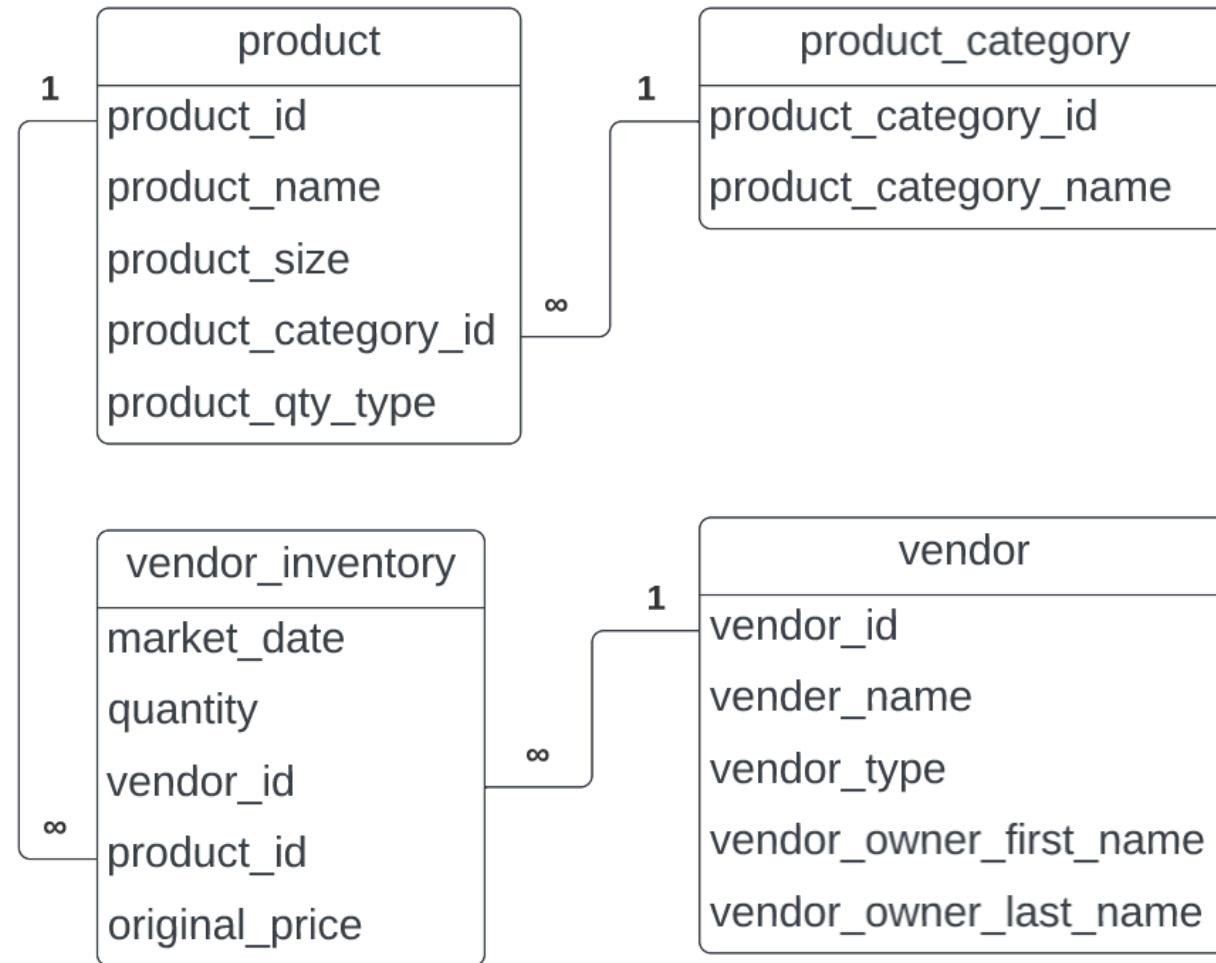
Logical Models

- Add additional detail to the conceptual model by adding column names for each table
- Often indicate the type of relationship
 - One-to-One
 - One-to-Many
 - Many-to-Many

Logical Models

- Our (partial) Farmers Market database:
 - product (5 columns) shares a One-to-Many relationship with vendor_inventory (5 columns) on product_id
 - product_category (2 columns) shares a One-to-Many relationship with product (5 columns) on product_category_id
 - _____ shares a _____ relationship with _____ on vendor_id

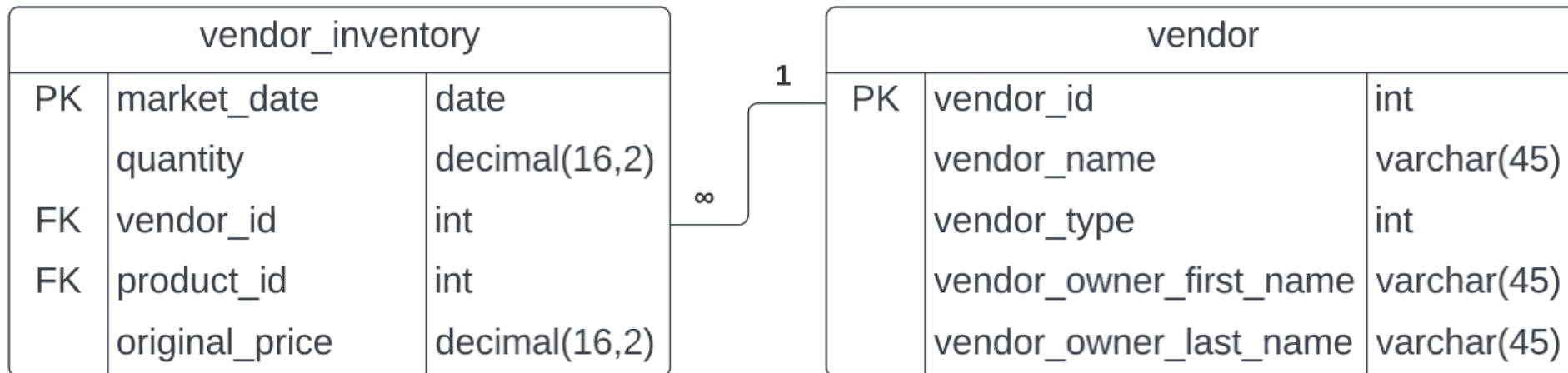
Farmers Market Logical Model (partial)



Physical Models

- Add additional detail to the logical model by adding key type and column data type
- Our (partial) Farmers Market database:
 - vendor_id (int) is the PK for vendor, which shares a One-to-Many relationship with vendor_inventory on vendor_id (FK)
 - product_id (int) is a FK for vendor_inventory (so *elsewhere in this diagram, we'd connect this to a PK of another table*)
 - market_date (date) is the PK for vendor_inventory
 - why?  Think, Pair, Share

Farmers Market Physical Model (partial)



What questions do you have about Data Modelling?

Assignment 1

Assignment 1: Design a Logical Model

Q1) Create a logical model for a small bookstore. 

At the minimum it should have employee, order, sales, customer, and book entities (tables). Determine sensible column and table design based on what you know about these concepts. Keep it simple, but work out sensible relationships to keep tables reasonably sized. Include a date table. There are several tools online you can use, I'd recommend [Lucidchart](#) or [Draw.io](#).

Assignment 1: Design a Logical Model

Q2) We want to create employee shifts, splitting up the day into morning and evening. Add this to the ERD.

Assignment 1: Design a Logical Model

Q3) The store wants to keep customer addresses. Propose two architectures for the CUSTOMER_ADDRESS table, one that will retain changes, and another that will overwrite. Which is type 1, which is type 2? *Hint, search type 1 vs type 2 slowly changing dimensions.*

Bonus: Are there privacy implications to this, why or why not?

Assignment 1: Design a Logical Model

Q4) Review the AdventureWorks ERD here: [link](#)

Highlight at least two differences between it and your ERD. Would you change anything in yours?

What questions do you have about anything from today?