Lecture 01: Introduction to Data Engineering

DATA 503: Fundamentals of Data Engineering

Lucas P. Cordova, Ph.D.

2026-01-12

This lecture covers the introduction to data engineering.

Table of contents

1 Setting the stage

1.1 What We'll Discuss

- What data engineering is and why it exists
- The data engineering pipeline
- Relational databases as a workhorse

- SQL as your first "superpower"

1.2 The setting

- In this lecture, our data world is Dunder Mifflin.
- We will use familiar Office characters, events, and business problems as examples.
- The goal is not TV trivia.
- The goal is to make abstract engineering ideas feel concrete.

1.3 Quick question

If you had to define data engineering in one sentence, what would you say?

- Write your sentence in 15 seconds.
- Share with the person next to you.

1.4 What counts as "data" at Dunder Mifflin?

Examples include:

- Sales calls, quotes, invoices
- Customer records and contacts
- Warehouse inventory and shipping logs
- HR data (hiring, training, performance)
- Emails and calendar invites
- "Prank events" if Dwight is logging them

1.5 Mini-quiz

Which role is primarily responsible for making raw data reliable and accessible for others?

A. Data Analyst
B. Data Scientist
C. Data Engineer
D. Product Manager

Answer: C

2 The point of data engineering

2.1 The data engineer's job

- Build systems that move and shape data so it can be used reliably.
- Make data easy to find, trustworthy, and fast to access.
- Reduce chaos so others can do analysis, reporting, and ML.

2.2 In The Office terms

Michael wants a dashboard in 10 minutes.

- "How many sales did we make this week?"
- "Which customers are at risk of churning?"
- "What does the warehouse backlog look like?"

Your job is to make those questions answerable without manual spreadsheet heroics.

2.3 Where things break

Common failure modes:

- Data is missing or duplicated.
- Definitions are inconsistent.
- The report takes 40 minutes to run.
- Nobody knows which table to trust.
- The pipeline fails silently on a Tuesday.

2.4 Think-pair-share: "The spreadsheet problem"

Prompt:

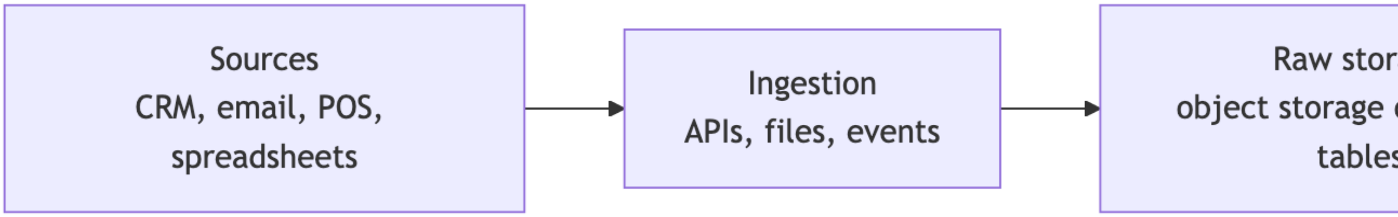
- Think of a time a spreadsheet became the system of record.
- What went wrong?
- What would you build instead?

Directions:

- Think (1 minute)
- Pair (3 minutes)
- Share (3 to 4 pairs with the room)

2.5 A pipeline mental model

A pipeline is a repeatable path from sources to usable outputs.



2.6 The "why" in one slide

Data engineering exists because:

- Data is produced by many systems.
- Data changes over time.
- Data is messy.
- Organizations still need answers on demand.

3 Data quality and trust

3.1 The Five V's (Scranton edition)

- Volume
  - How much data: every order, every call, every invoice
- Velocity
  - How fast it arrives: live sales calls vs nightly shipments
- Variety
  - Tables, PDFs, emails, phone call logs, images
- Veracity
  - Can we trust it: typos, duplicates, missing values
- Value
  - Does it help decisions: pricing, staffing, inventory planning

3.2 Quick question

Which “V” is usually hardest in your experience?

- Raise a hand for:
  - Volume
  - Velocity
  - Variety
  - Veracity
  - Value

3.3 Veracity is usually the silent killer

A simple example:

- Sales reps enter customer names manually.
- “Prince Family Paper” becomes:
  - Prince Family Paper
  - Prince Family Papers
  - Prince Family Papeer

Now “top customers” depends on spelling.

3.4 Data quality is not just correctness

Also think about:

- Consistency across systems
- Timeliness
- Completeness
- Lineage (where it came from)
- Observability (how you know it is working)

4 ETL, ELT, and the lifecycle

4.1 ETL vs ELT

ETL:

- Extract

- Transform
- Load

ELT:

- Extract
- Load
- Transform (inside the warehouse)

4.2 Why the difference matters

ETL is often:

- Great for strict control and smaller volumes
- Easier to reason about transformations

ELT is often:

- Faster to iterate for analytics teams
- More flexible once data is centralized

4.3 Batch vs streaming

Batch:

- “Run the daily sales rollup at 2am”
- Often cheaper and simpler

Streaming:

- “Update the live sales leaderboard every minute”
- More complex but lower latency

4.4 Think-pair-share: choose a mode

Scenario:

- Corporate asks for a daily report of sales by rep.
- Michael asks for a live leaderboard on a TV in the office.

Questions:

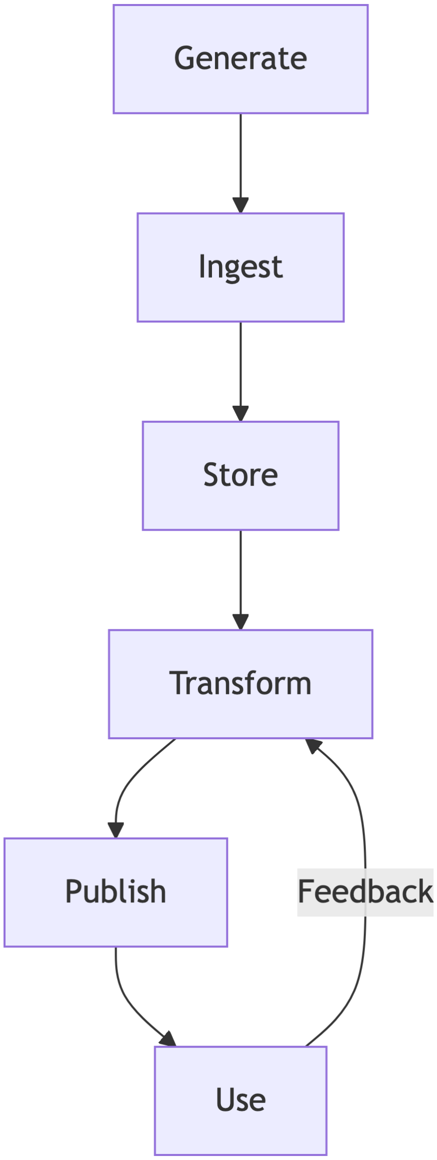
- Which use case is batch?
- Which use case is streaming?
- What do you lose when you choose batch?

4.5 What “production” means

A pipeline is production when:

- It runs on a schedule or event.
- It is monitored.
- Failures alert the right humans.
- Data contracts are stable enough that changes are managed.

4.6 A small lifecycle picture



4.7 Break (10 minutes)

During break:

- Pick one question you want answered about data engineering.
- Write it down.

- We will collect a few when we return.

5 Relational databases

5.1 Why relational databases still matter

Relational databases remain a core tool because:

- Tables match how many business questions are asked.
- SQL is powerful and widely supported.
- Constraints and relationships reduce duplication and ambiguity.
- They are a reliable foundation for analytics and applications.

5.2 When a relational database is a good fit

- You have structured entities (customers, orders, employees).
- You care about relationships and integrity.
- You need precise querying and joins.
- You want constraints (unique, foreign keys).

5.3 Our tiny Dunder Mifflin dataset

We will pretend we have these tables:

- employees
- customers
- orders
- order_items
- products
- episodes (optional, for fun)

5.4 Example: employees

Columns:

- employee_id (PK)
- full_name
- role
- branch
- hire_date

ℹ Note

Primary Key (PK) is a unique identifier for each row in the table. More on this later.

5.5 Sample employees data

employee_id	full_name	role	branch	hire_date
1	Michael Scott	Regional Manager	Scranton	1992-03-15
2	Dwight Schrute	Assistant Regional Manager	Scranton	1995-04-01
3	Jim Halpert	Sales Representative	Scranton	1999-08-01
4	Pam Beesly	Receptionist	Scranton	2000-01-03
5	Stanley Hudson	Sales Representative	Scranton	1990-09-10
6	Phyllis Vance	Sales Representative	Scranton	2000-02-14
7	Kevin Malone	Accountant	Scranton	1998-06-15
8	Oscar Martinez	Accountant	Scranton	1996-11-20
9	Angela Martin	Head of Accounting	Scranton	1994-05-05
10	Creed Bratton	Quality Assurance	Scranton	1993-12-01

5.6 Example: customers

Columns:

- customer_id (PK)
- customer_name

- industry
- address_line_1
- address_line_2
- city
- state
- zip_code
- country

5.7 Example: orders and order_items

orders:

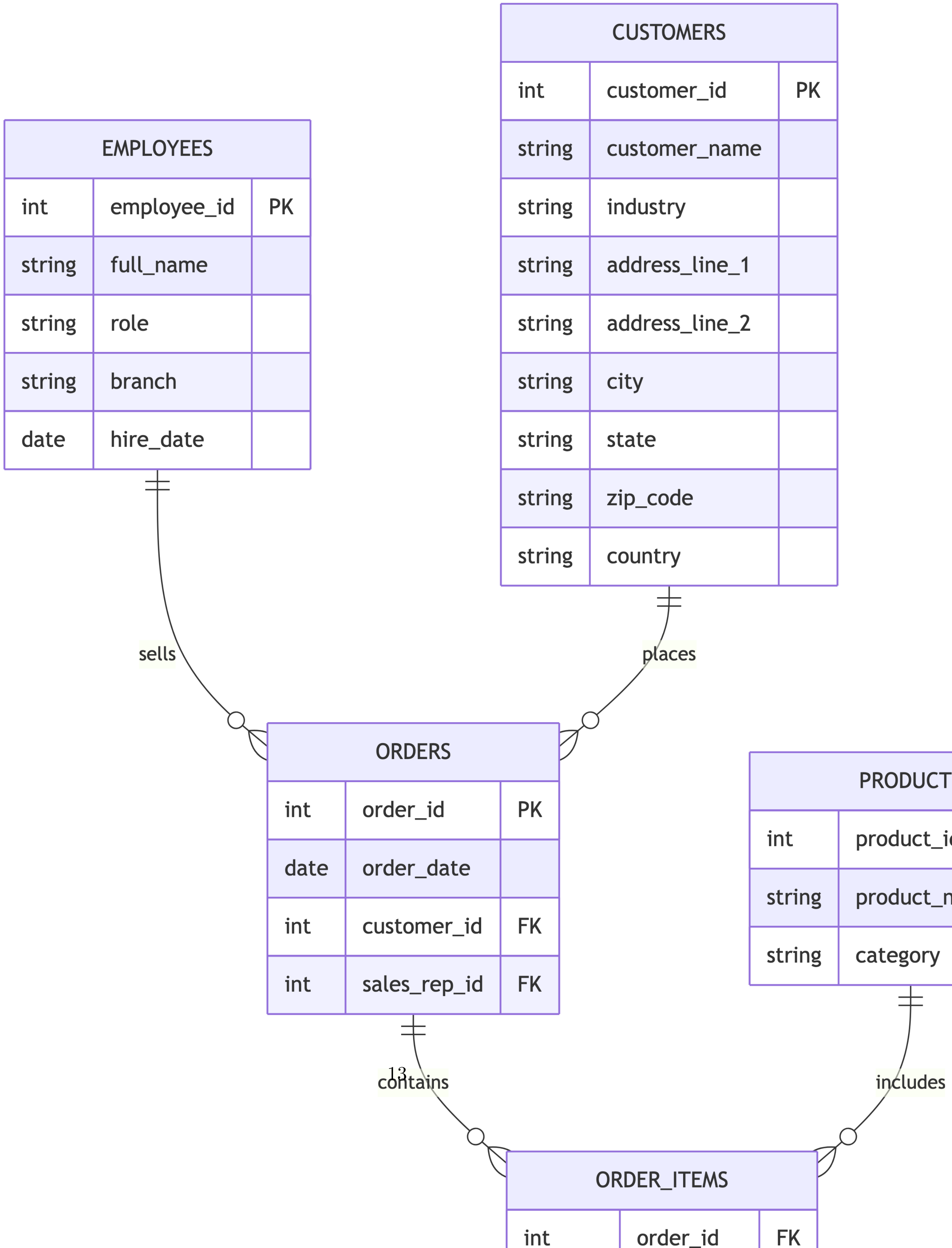
- order_id (PK)
- order_date
- customer_id (FK)
- sales_rep_id (FK)

order_items:

- order_id (FK)
- product_id (FK)
- quantity
- unit_price

> ℹ Note
>
> Foreign Key (FK) is a reference to a primary key in another table. More on this later.

5.8 A simple schema view



EMPLOYEES		
int	employee_id	PK
string	full_name	
string	role	
string	branch	
date	hire_date	

CUSTOMERS		
int	customer_id	PK
string	customer_name	
string	industry	
string	address_line_1	
string	address_line_2	
string	city	
string	state	
string	zip_code	
string	country	

ORDERS		
int	order_id	PK
date	order_date	
int	customer_id	FK
int	sales_rep_id	FK

PRODUCT	
int	product_i
string	product_n
string	category

ORDER_ITEMS		
int	order_id	FK

5.9 Quick question

In what table is the best place to add another address field so that we have both a billing and shipping address?

A. customers
B. orders
C. order_items
D. products

Answer: A

> ℹ Note
>
> BUT, we should ask ourselves if there is a better way to approach this problem.

5.10 Normalization in one minute

Normalization is a way to reduce duplication.

- Store the customer names once in customers.
- Reference customers from orders.
- Avoid rewriting customer names on every order.

5.11 Think-pair-share: what is the primary key?

Prompt:

- For "episodes", what could be a reasonable primary key?
- For "orders", why is a single order_id better than (customer_id, date)?

Directions:

- Think (1 minute)
- Pair (2 minutes)
- Share (2 pairs)

6 SQL: asking questions

6.1 SQL is how you ask for answers

SQL lets you:

- Select columns
- Filter rows
- Sort results
- Limit output
- Combine tables with joins
- Aggregate (count, sum, average)

Today we focus on SELECT fundamentals.

6.2 The shape of a SELECT query

```sql
SELECT column_1, column_2
FROM some_table
WHERE some_condition
ORDER BY some_column
LIMIT 10;
```

6.3 Start simple

All employees:

```sql
SELECT *
FROM employees;
```

employee_id	full_name	role	branch	hire_date
1	Michael Scott	Regional Manager	Scranton	1992-03-15
2	Dwight Schrute	Assistant Regional Manager	Scranton	1995-04-01
3	Jim Halpert	Sales Representative	Scranton	1999-08-01
4	Pam Beesly	Receptionist	Scranton	2000-01-03

employee_id	full_name	role	branch	hire_date
5	Stanley Hudson	Sales Representative	Scranton	1990-09-10
6	Phyllis Vance	Sales Representative	Scranton	2000-02-14
7	Kevin Malone	Accountant	Scranton	1998-06-15
8	Oscar Martinez	Accountant	Scranton	1996-11-20
9	Angela Martin	Head of Accounting	Scranton	1994-05-05
10	Creed Bratton	Quality Assurance	Scranton	1993-12-01

> ℹ Note
>
> SELECT * is a wildcard that selects all columns. It is not a good practice to use * in production queries. Instead, you should list the columns you need.

6.4 Choose columns

Only names and roles:

```sql
SELECT full_name, role
FROM employees;
```

full_name	role
Michael Scott	Regional Manager
Dwight Schrute	Assistant Regional Manager
Jim Halpert	Sales Representative
Pam Beesly	Receptionist
Stanley Hudson	Sales Representative
Phyllis Vance	Sales Representative
Kevin Malone	Accountant
Oscar Martinez	Accountant
Angela Martin	Head of Accounting
Creed Bratton	Quality Assurance

full_name	role

6.5 DISTINCT

Unique branches:

```
SELECT DISTINCT branch
FROM employees;
```

branch
Scranton

6.6 WHERE

All Scranton employees:

```
SELECT full_name, role
FROM employees
WHERE branch = 'Scranton';
```

full_name	role
Michael Scott	Regional Manager
Dwight Schrute	Assistant Regional Manager
Jim Halpert	Sales Representative
Pam Beesly	Receptionist
Stanley Hudson	Sales Representative
Phyllis Vance	Sales Representative
Kevin Malone	Accountant
Oscar Martinez	Accountant
Angela Martin	Head of Accounting
Creed Bratton	Quality Assurance

6.7 ORDER BY

Newest hires first:

```
SELECT full_name, hire_date
FROM employees
ORDER BY hire_date DESC;
```

full_name	hire_date
Phyllis Vance	2000-02-14
Pam Beesly	2000-01-03
Jim Halpert	1999-08-01
Kevin Malone	1998-06-15
Oscar Martinez	1996-11-20
Dwight Schrute	1995-04-01
Angela Martin	1994-05-05
Creed Bratton	1993-12-01
Michael Scott	1992-03-15
Stanley Hudson	1990-09-10

6.8 LIMIT

Top 5 newest hires:

```
SELECT full_name, hire_date
FROM employees
ORDER BY hire_date DESC
LIMIT 5;
```

full_name	hire_date
Phyllis Vance	2000-02-14
Pam Beesly	2000-01-03
Jim Halpert	1999-08-01
Kevin Malone	1998-06-15
Oscar Martinez	1996-11-20

6.9 Building a query step by step

Question:

- "Show the 5 largest order line items by total line value."

We define line value as:

- quantity * unit_price

6.10 Step 1: pick columns

```sql
SELECT order_id, product_id, quantity, unit_price
FROM order_items;
```

6.11 Step 2: add a computed column

```sql
SELECT
  order_id,
  product_id,
  quantity,
  unit_price,
  quantity * unit_price AS line_value
FROM order_items;
```

6.12 Step 3: sort and limit

```sql
SELECT
  order_id,
  product_id,
  quantity,
  unit_price,
  quantity * unit_price AS line_value
FROM order_items
ORDER BY line_value DESC
LIMIT 5;
```

6.13 Quick question

If you filter rows, which clause do you use?

A. FROM
B. WHERE
C. ORDER BY
D. LIMIT

Answer: B

7 Informal exercise: build a SELECT

7.1 The exercise (individual then pair)

We are going to build a single SELECT statement for a given table.

Table:

- episodes

Columns:

- episode_id
- season
- episode_number
- title
- air_date
- imdb_rating

7.2 Task 1

Write a query to list:

- season
- episode_number
- title
- imdb_rating

Conditions:

- only season 2
- only ratings 8.5 or higher

Output:

- highest rated first

Limit:

- top 5

7.3 Hint: start from the skeleton

```sql
SELECT
  -- columns
FROM episodes
WHERE
  -- conditions
ORDER BY
  -- sorting
LIMIT
  -- number
;
```

7.4 Think-pair-share: compare solutions

Directions:

- Think (2 minutes): write your query.
- Pair (3 minutes): compare with a neighbor.
- Share (3 minutes): we will build the “class version” together.

7.5 One possible solution

```sql
SELECT
  season,
  episode_number,
  title,
  imdb_rating
FROM episodes
WHERE season = 2
  AND imdb_rating >= 8.5
ORDER BY imdb_rating DESC
LIMIT 5;
```

7.6 Task 2

Modify your query to break ties by episode_number ascending.

7.7 One possible solution

```sql
SELECT
  season,
  episode_number,
  title,
  imdb_rating
FROM episodes
WHERE season = 2
  AND imdb_rating >= 8.5
ORDER BY imdb_rating DESC, episode_number ASC
LIMIT 5;
```

8 A peek ahead: joins

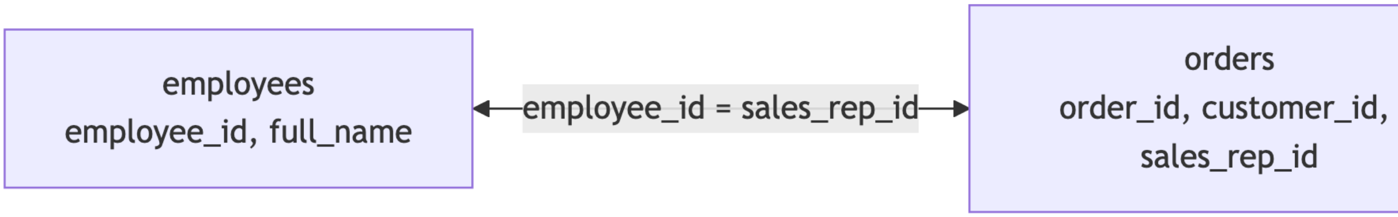
8.1 Why joins exist

Most real questions require combining tables.

Example:

- "Which customers did Jim sell to last month?"
- That information is split across:
  - employees
  - orders
  - customers

8.2 Conceptual join



8.3 One join teaser

```sql
SELECT
  e.full_name AS sales_rep,
  c.customer_name,
  o.order_date
FROM orders o
JOIN employees e
  ON o.sales_rep_id = e.employee_id
JOIN customers c
  ON o.customer_id = c.customer_id
WHERE e.full_name = 'Jim Halpert'
ORDER BY o.order_date DESC
LIMIT 10;
```

8.4 Quick question

What is the main purpose of a foreign key?

A. Make queries faster
B. Guarantee a relationship points to an existing row
C. Store text efficiently
D. Replace the need for indexes

Answer: B

9 Wrap-up

9.1 What you should leave with

- A clear definition of what data engineering is.
- A mental model of a pipeline.
- A sense of why relational databases matter.
- The ability to write basic SELECT queries with:
  - WHERE
  - ORDER BY
  - LIMIT
  - DISTINCT

9.2 Exit ticket

Write down:

- One concept that felt clear.
- One concept that felt fuzzy.
- One question you want answered next lecture.

Send me your answers on Canvas on the Week 1 Participation Activity.

9.3 Vibe check

- If you had to explain “ETL vs ELT” to Michael in two sentences, what would you say?
- If you had to explain “foreign key” to Dwight in two sentences, what would you say?