

# Data structures

UNDERSTANDING DATA ENGINEERING



**Hadrien Lacroix**

Content Developer at DataCamp

# Structured data

- Easy to search and organize
- Consistent model, rows and columns
- Defined types
- Can be grouped to form relations
- Stored in relational databases
- About 20% of the data is structured
- Created and queried using SQL

# Employee table

index	last_name	first_name	role	team	full_time	office
0	Thien	Vivian	Data Engineer	Data Science	1	Belgium
1	Huong	Julian	Data Scientist	Data Science	1	Belgium
2	Duplantier	Norbert	Software Developer	Infrastructure	1	United Kingdom
3	McColgan	Jeff	Business Developer	Sales	1	United States
4	Sanchez	Rick	Support Agent	Customer Service	0	United States

# Relational database

office	address	number	city	zipcode
Belgium	Martelarenlaan	38	Leuven	3010
UK	Old Street	207	London	EC1V 9NR
USA	5th Ave	350	New York	10118

# Relational database

index	last_name	first_name	office	address	number	city	zipcode
0	Thien	Vivian	Belgium	Martelarenlaan	38	Leuven	3010
1	Huong	Julian	Belgium	Martelarenlaan	38	Leuven	3010
2	Duplantier	Norbert	UK	Old Street	207	London	EC1V 9NR
3	McColgan	Jeff	USA	5th Ave	350	New York	10118
4	Sanchez	Rick	USA	5th Ave	350	New York	10118

# Semi-structured data

- Relatively easy to search and organize
- Consistent model, less-rigid implementation: different observations have different sizes
- Different types
- Can be grouped, but needs more work
- NoSQL databases: JSON, XML, YAML

# Favorite artists JSON file

```
{  
  {"user_1645156":  
    "last_name": "Lacroix",  
    "first_name": "Hadrien",  
    "favorite_artists": ["Fools in Deed", "Gojira", "Pain", "Nanowar of Steel"]},  
  {"user_5913764":  
    "last_name": "Billen",  
    "first_name": "Sara",  
    "favorite_artists": ["Tamino", "Taylor Swift"]},  
  {"user_8436791":  
    "last_name": "Sulmont",  
    "first_name": "Lis",  
    "favorite_artists": ["Arctic Monkeys", "Rihanna", "Nina Simone"]},  
  ...  
}
```

# Unstructured data

- Does not follow a model, can't be contained in rows and columns
- Difficult to search and organize
- Usually text, sound, pictures or videos
- Usually stored in data lakes, can appear in data warehouses or databases
- Most of the data is unstructured
- Can be extremely valuable

Una mattina mi son alzato  
O bella ciao, bella ciao, bella ciao, ciao, ciao  
Una mattina mi son alzato  
E ho trovato l'invasor

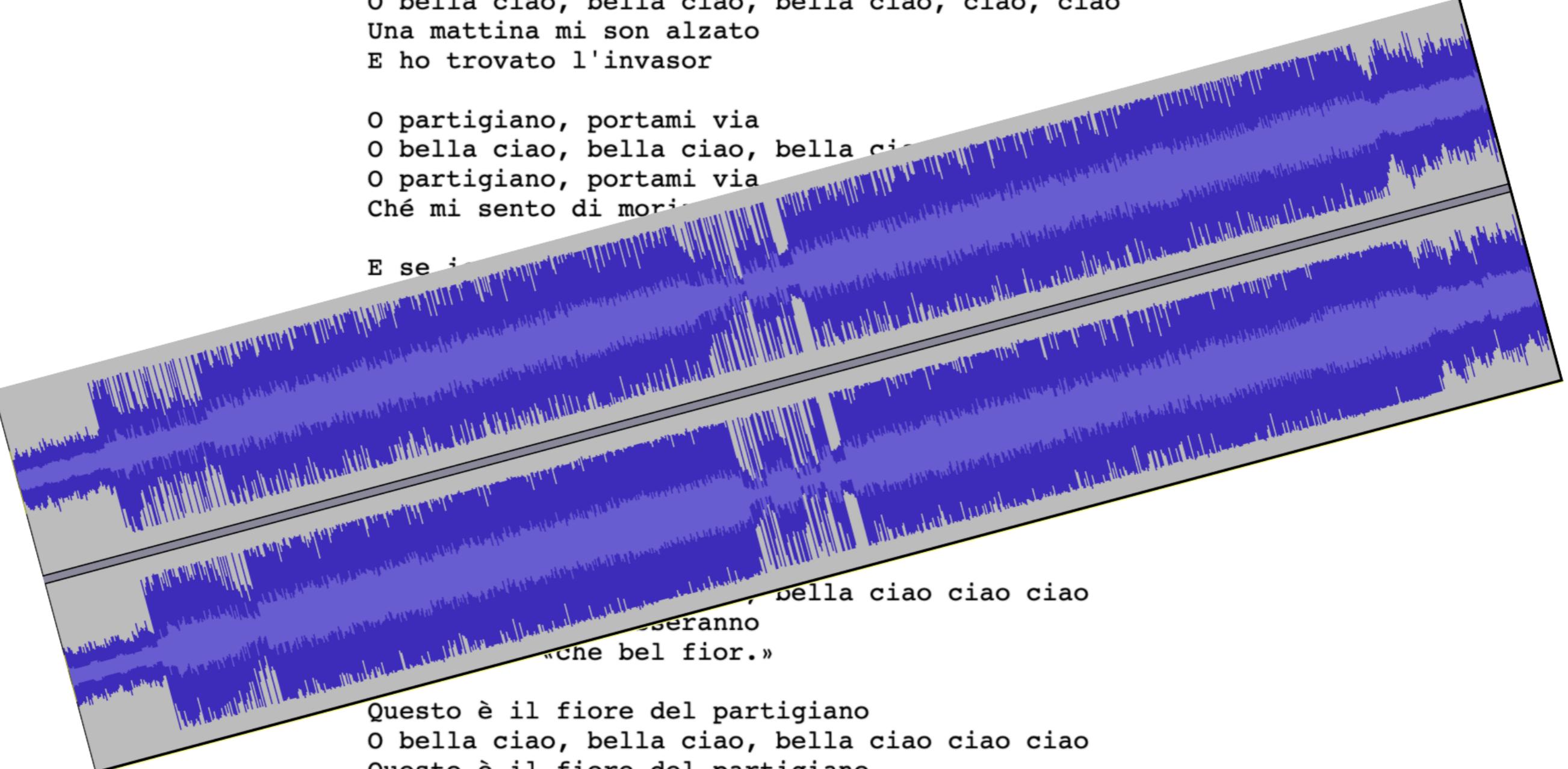
O partigiano, portami via  
O bella ciao, bella ciao, bella ciao, ciao, ciao  
O partigiano, portami via  
Ché mi sento di morir

E se io muoio da partigiano  
O bella ciao, bella ciao, bella ciao, ciao, ciao  
E se io muoio da partigiano  
Tu mi devi seppellir

E seppellire lassù in montagna  
O bella ciao, bella ciao, bella ciao, ciao, ciao  
E seppellire lassù in montagna  
Sotto l'ombra di un bel fior

E le genti che passeranno  
O bella ciao, bella ciao, bella ciao ciao ciao  
E le genti che passeranno  
Mi diranno «che bel fior.»

Questo è il fiore del partigiano  
O bella ciao, bella ciao, bella ciao ciao ciao  
Questo è il fiore del partigiano  
Morto per la libertà



Una mattina mi son alzato  
O bella ciao, bella ciao, bella ciao, ciao, ciao  
Una mattina mi son alzato  
E ho trovato l'invasor

O partigiano, portami via  
O bella ciao, bella ciao, bella ciao  
O partigiano, portami via  
Ché mi sento di morir  
  
E se :

, bella ciao ciao ciao  
seranno  
"che bel fior."

Questo è il fiore del partigiano  
O bella ciao, bella ciao, bella ciao ciao ciao  
Questo è il fiore del partigiano  
Morto per la libertà

Una mattina mi son alzato  
O bella ciao, bella ciao, bella ciao, ciao, ciao

Una mattina mi son alzato  
E ho tro

O parti  
O bella  
O part  
Ché mi

E se

Questo è il fiore dei partigiani  
O bella ciao, bella ciao, bella ciao  
Questo è il fiore del partigiano  
Morto per la libertà





# Adding some structure

- Use AI to search and organize unstructured data
- Add information to make it semi-structured

# Summary

- Structured data
- Semi-structured data
- Unstructured data
- Differences between the three
- Give examples

# **Let's practice!**

**UNDERSTANDING DATA ENGINEERING**

# SQL databases

UNDERSTANDING DATA ENGINEERING



**Hadrien Lacroix**

Content Developer at DataCamp

# SQL

- Structured Query Language
- Industry standard for Relational Database Management System (RDBMS)
- Allows you to access many records at once, and group, filter or aggregate them
- Close to written English, easy to write and understand
- Data engineers use SQL to create and maintain databases
- Data scientists use SQL to query (request information from) databases

# Remember the employees table

index	last_name	first_name	role	team	full_time	office
0	Thien	Vivian	Data Engineer	Data Science	1	Belgium
1	Huong	Julian	Data Scientist	Data Science	1	Belgium
2	Duplantier	Norbert	Software Developer	Infrastructure	1	United Kingdom
3	McColgan	Jeff	Business Developer	Sales	1	United States
4	Sanchez	Rick	Support Agent	Customer Service	0	United States

# SQL for data engineers

- Data engineers use SQL to create, maintain and update tables.

```
CREATE TABLE employees (
    employee_id INT,
    first_name VARCHAR(255),
    last_name VARCHAR(255),
    role VARCHAR(255),
    team VARCHAR(255),
    full_time BOOLEAN,
    office VARCHAR(255)
);
```

# SQL for data scientists

- Data scientist use SQL to query, filter, group and aggregate data in tables.

```
SELECT first_name, last_name  
FROM employees  
WHERE role LIKE '%Data%'
```

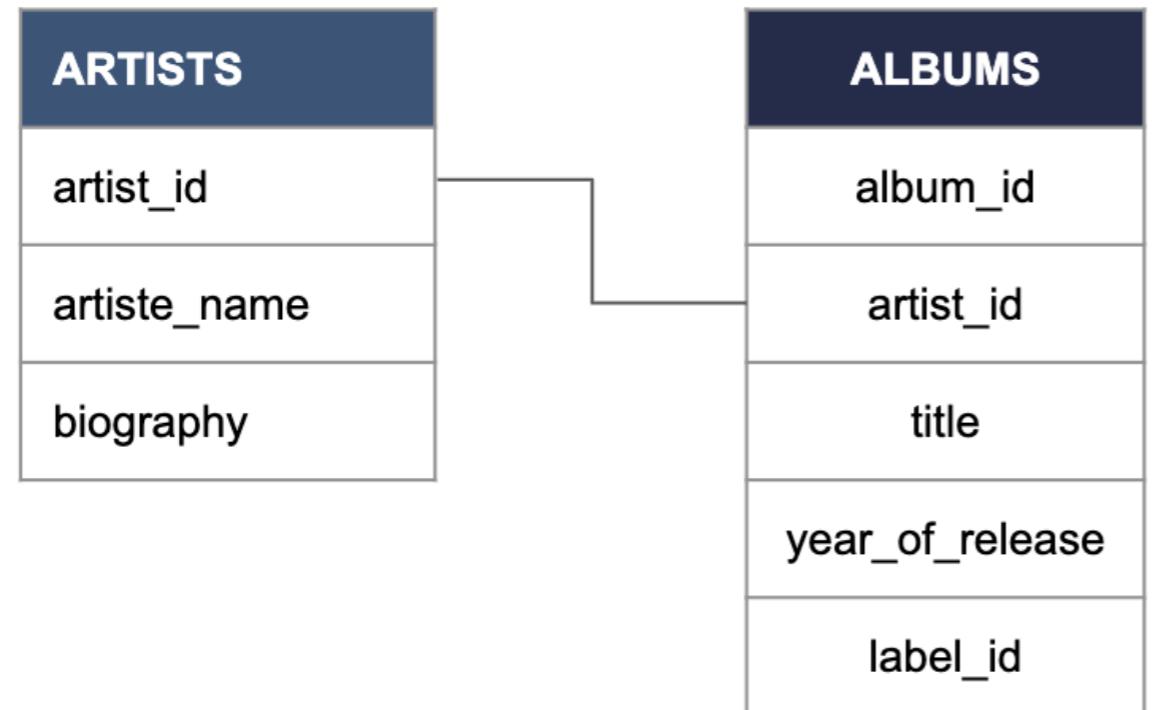
# Database schema

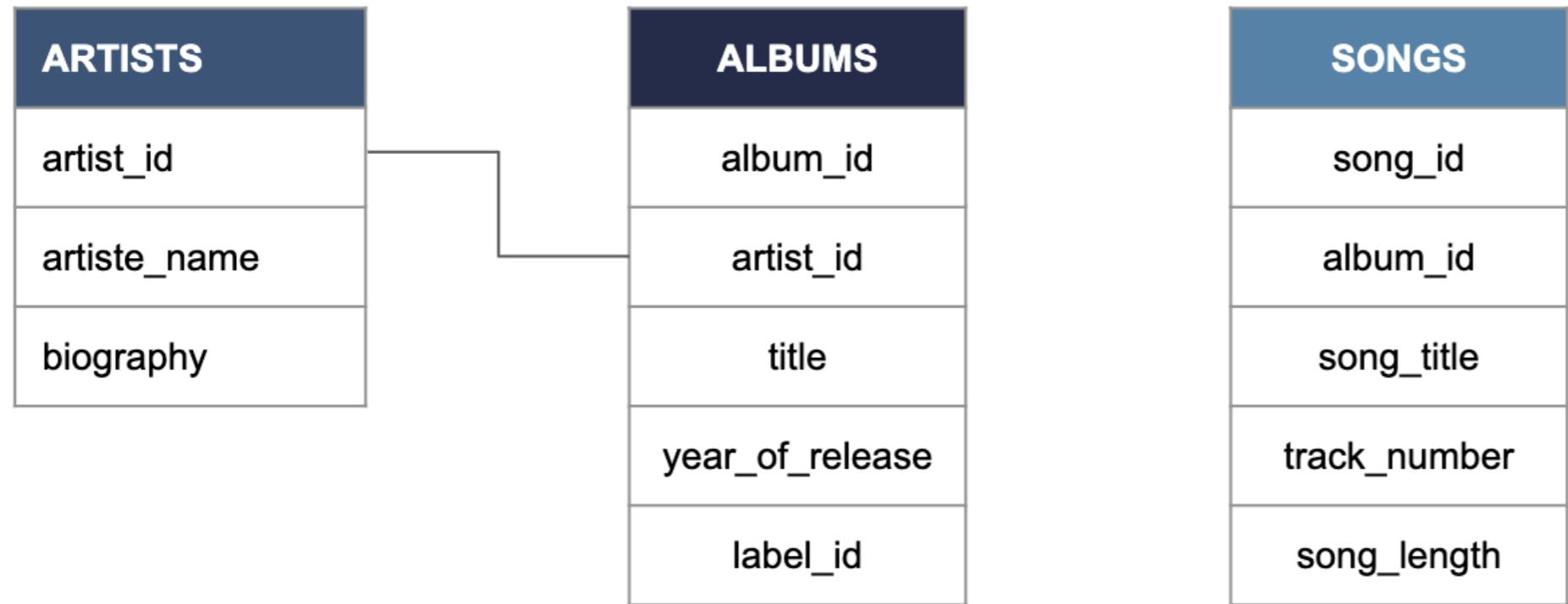
- Databases are made of tables
- The database schema governs how tables are related

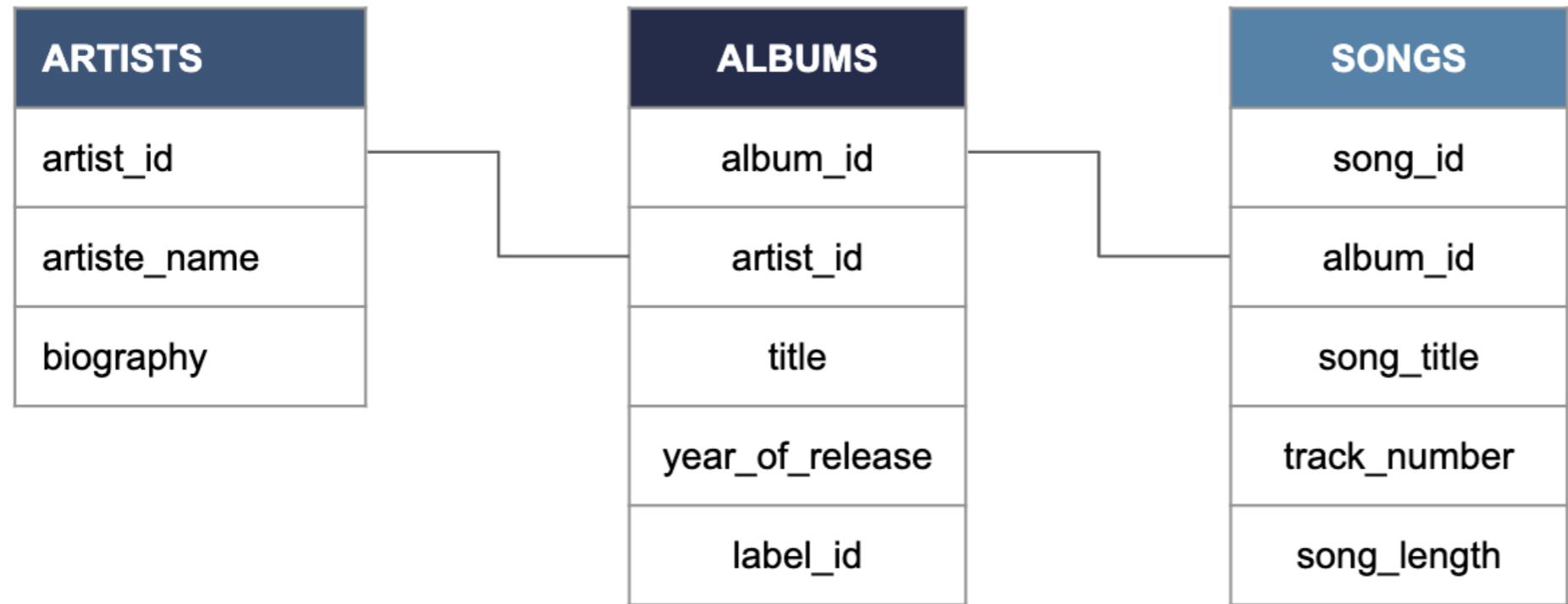
ALBUMS
album_id
artist_id
title
year_of_release
label_id

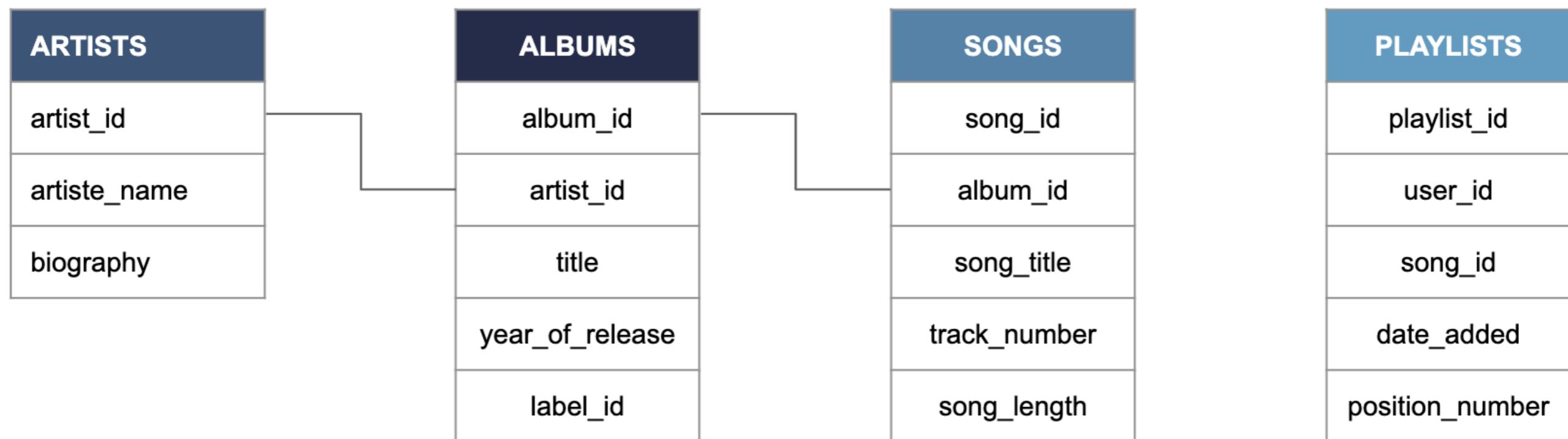
ARTISTS
artist_id
artiste_name
biography

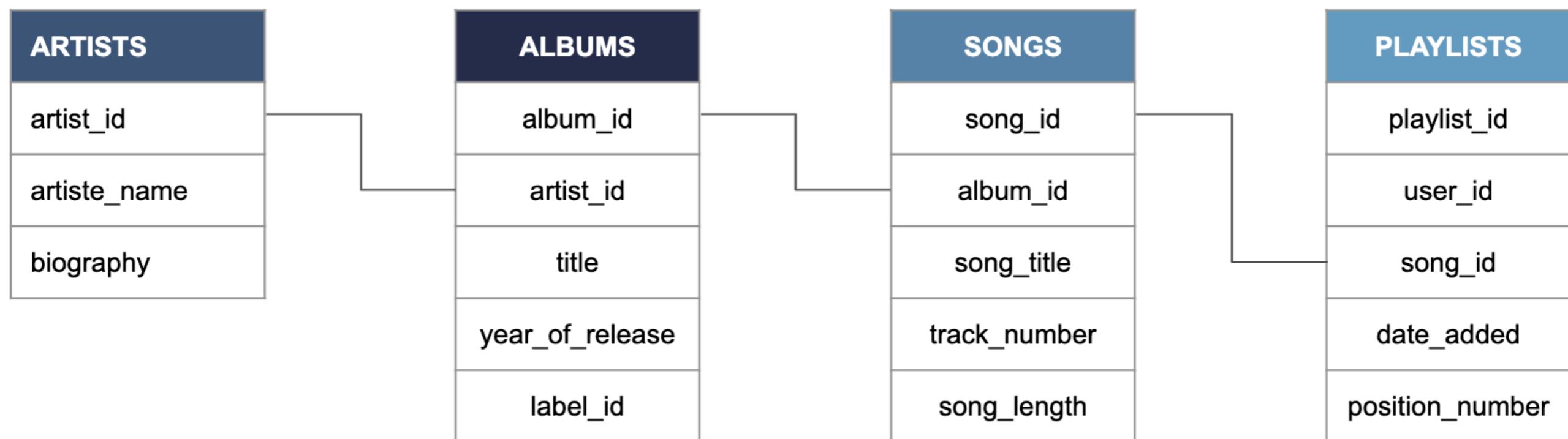
ALBUMS
album_id
artist_id
title
year_of_release
label_id











# Several implementations

- SQLite
- MySQL
- PostgreSQL
- Oracle SQL
- SQL Server

# Summary

- SQL = industry standard
- Explain how Data engineers and Data scientists use it differently
- Database schema
- SQL implementations

# **Let's practice!**

**UNDERSTANDING DATA ENGINEERING**

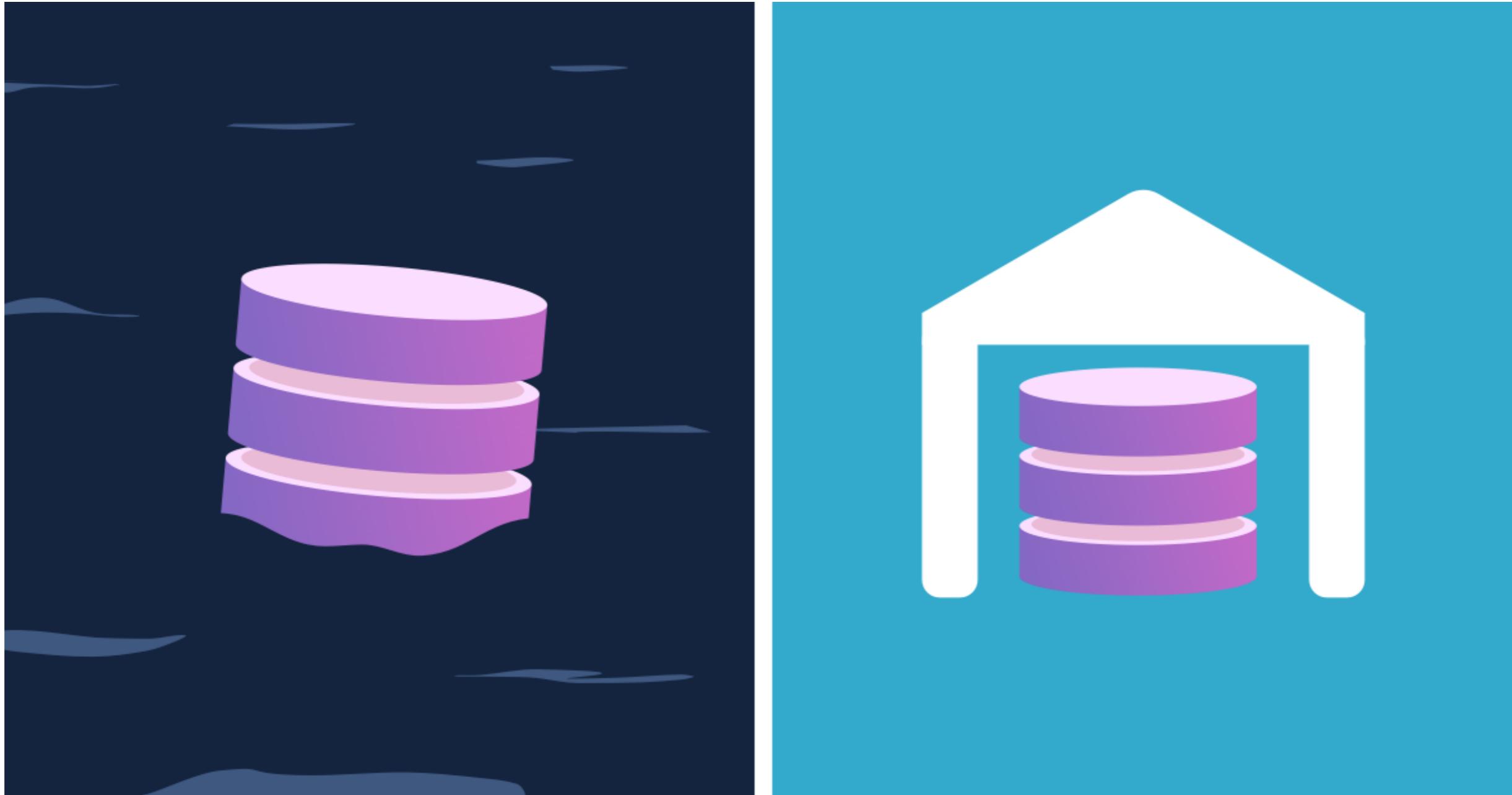
# Data warehouses and data lakes

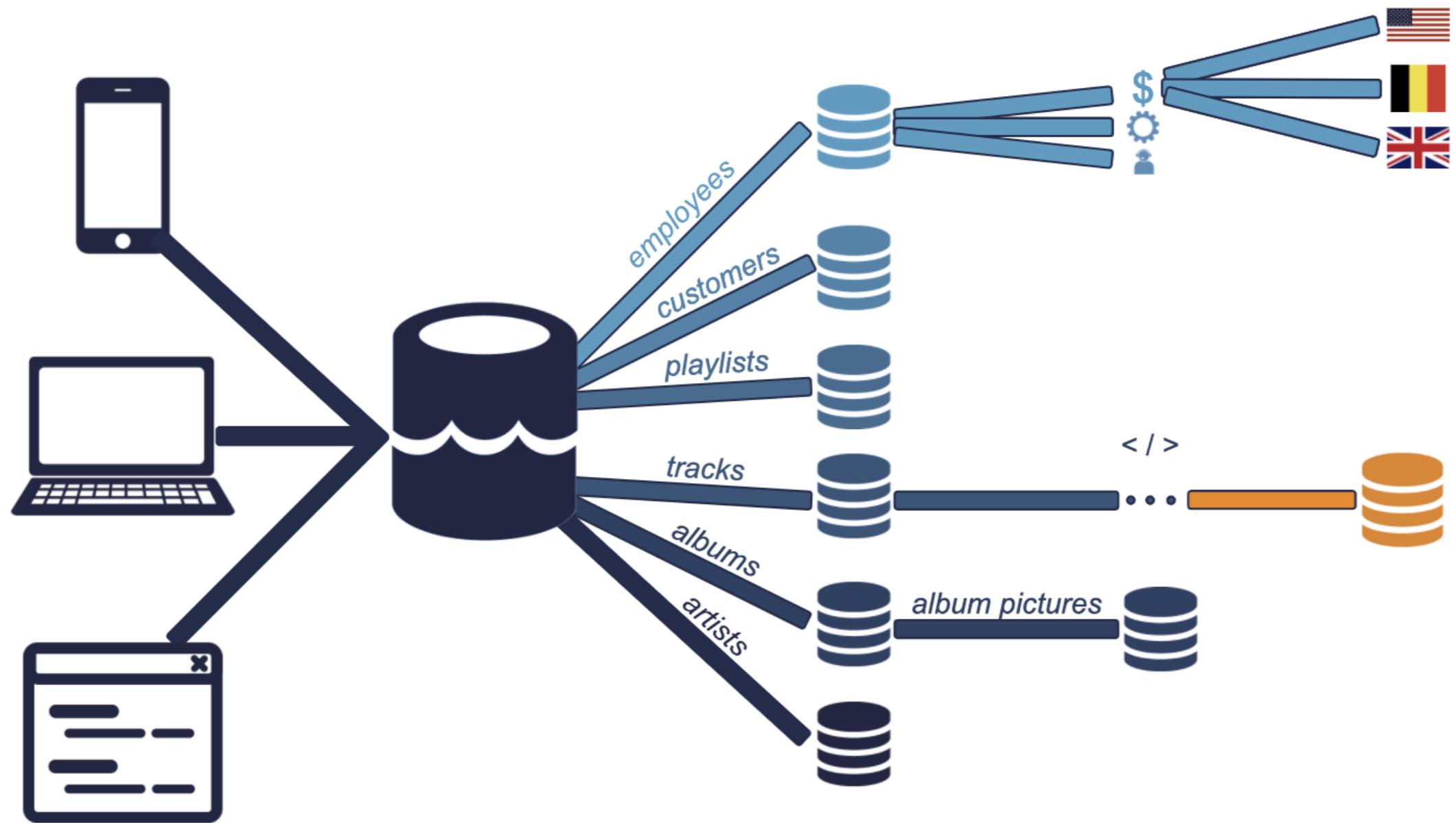
UNDERSTANDING DATA ENGINEERING



**Hadrien Lacroix**  
Content Developer

# Warehouses with stunning view on the lake





# Data lakes and data warehouses

## Data lake

- Stores all the raw data
- Can be petabytes (1 million GBs)
- Stores all data structures
- Cost-effective
- Difficult to analyze
- Requires an up-to-date data catalog
- Used by data scientists
- Big data, real-time analytics

## Data warehouse

- Specific data for specific use
- Relatively small
- Stores mainly structured data
- More costly to update
- Optimized for data analysis
- Also used by data analysts and business analysts
- Ad-hoc, read-only queries

# Data catalog for data lakes

- What is the source of this data?
- Where is this data used?
- Who is the owner of the data?
- How often is this data updated?
- Good practice in terms of data governance
- Ensures reproducibility
- No catalog --> data swamp
- **Good practice for any data storage solution**
  - Reliability
  - Autonomy
  - Scalability
  - Speed

# Database vs. data warehouse

- Database:
  - General term
  - Loosely defined as *organized data stored and accessed on a computer*
- Data warehouse is a type of database

# Summary

- Data lakes
- Data warehouses
- Databases
- Data catalog

# **Let's practice!**

**UNDERSTANDING DATA ENGINEERING**