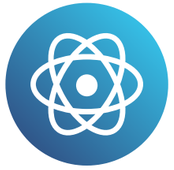
Data engineering and big data

DATA ENGINEERING FOR EVERYONE



Hadrien Lacroix
Content Developer at DataCamp

About the course

- Conceptual course
- No coding involved
- **Objectives**
  - Being able to exchange with data engineers
  - Provide a solid foundation to learn more

Chapter 1

What is data engineering?

1. Data engineering and big data
2. Data engineers vs. data scientists
3. Data pipelines

Chapter 2

How data storage works

1. Structured vs unstructured data
2. SQL
3. Data warehouse and data lakes

Chapter 3

How to move and process data

1. Processing data
2. Scheduling data
3. Parallel computing
4. Cloud computing

Spotflix

Data workflow

Data Collection & Storage

Data workflow

Data Collection & Storage → Data Preparation

Data workflow

Data Collection & Storage → Data Preparation → Exploration & Visualization

Data workflow

Data Collection & Storage

Data Preparation

Exploration & Visualization

Experimentation & Prediction

Data engineers

Data engineers

Data engineers deliver:

- the correct data
- in the right form
- to the right people
- as efficiently as possible

A data engineer's responsibilities

- Ingest data from different sources
- Optimize databases for analysis
- Remove corrupted data
- Develop, construct, test and maintain data architectures

Data engineers and big data

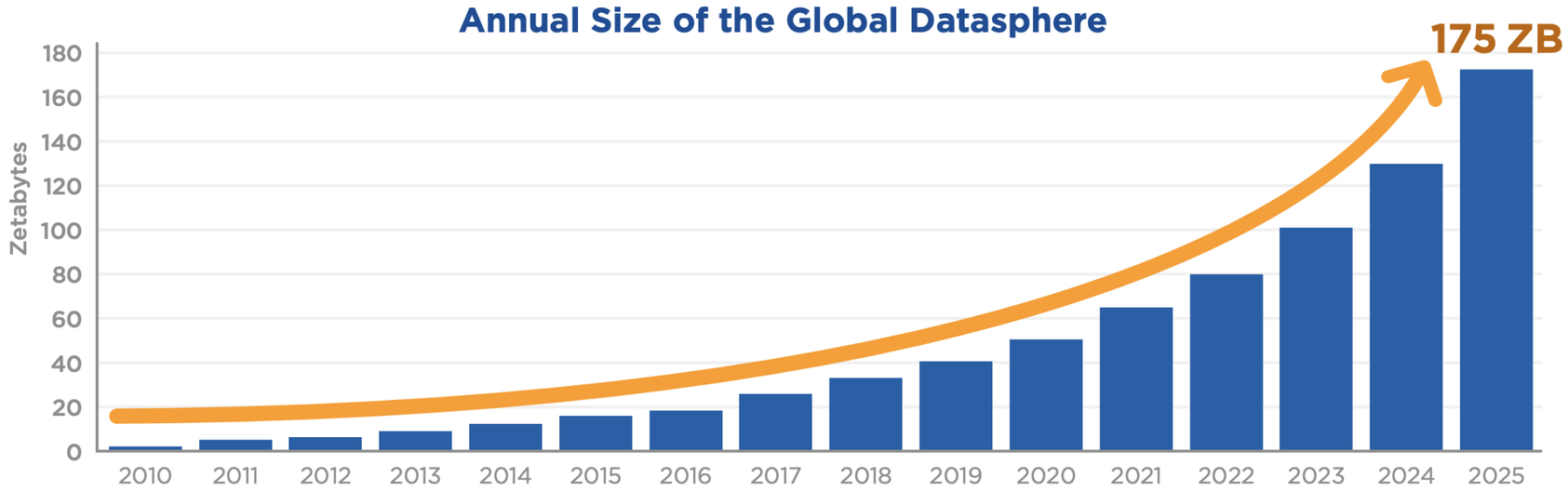
- Big data becomes the norm =>

Data engineers and big data

- Big data becomes the norm => data engineers are more and more needed
- Big data:
  - Have to think about how to deal with its size
  - So large traditional methods don't work anymore

Big data growth

- Sensors and devices
- Social media
- Enterprise data
- VoIP (voice communication, multimedia sessions)



[1] Data Age 2025, Seagate, November 2018

The five Vs

- Volume (how much?)
- Variety (what kind?)
- Velocity (how frequent?)
- Veracity (how accurate?)
- Value (how useful?)

Summary

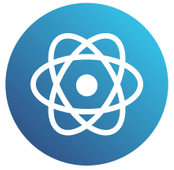
- What's waiting for you
- How data flows through an organization
- When a data engineer intervenes
- What their responsibilities are
- How data engineering relates to big data

Let's practice!

DATA ENGINEERING FOR EVERYONE

Data engineers vs. data scientists

DATA ENGINEERING FOR EVERYONE



Hadrien Lacroix
Content Developer at DataCamp

Data workflow

Data Collection & Storage

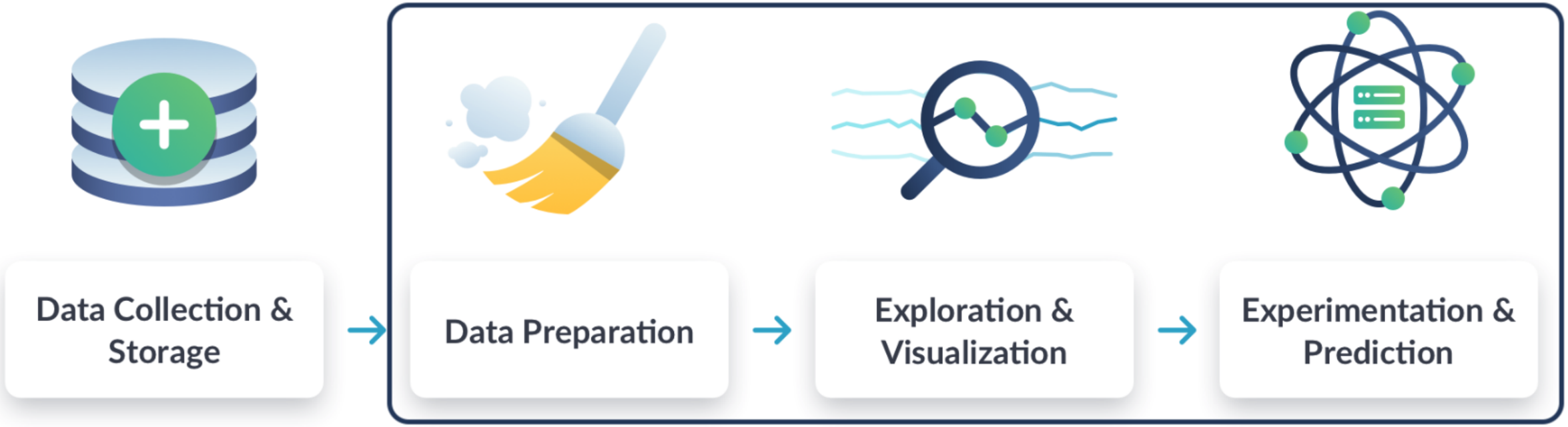
Data Preparation

Exploration & Visualization

Experimentation & Prediction

Data engineers

Data scientists



Data engineers enable data scientists

Data engineer

- Ingest and store data
- Set up databases
- Build data pipelines
- Strong software skills

Data scientist

- Exploit data
- Access databases
- Use pipeline outputs
- Strong analytical skills

Summary

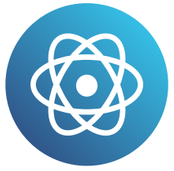
- At which stages data engineers and data scientists intervene
- How data engineers enable data scientists

Let's practice!

DATA ENGINEERING FOR EVERYONE

The data pipeline

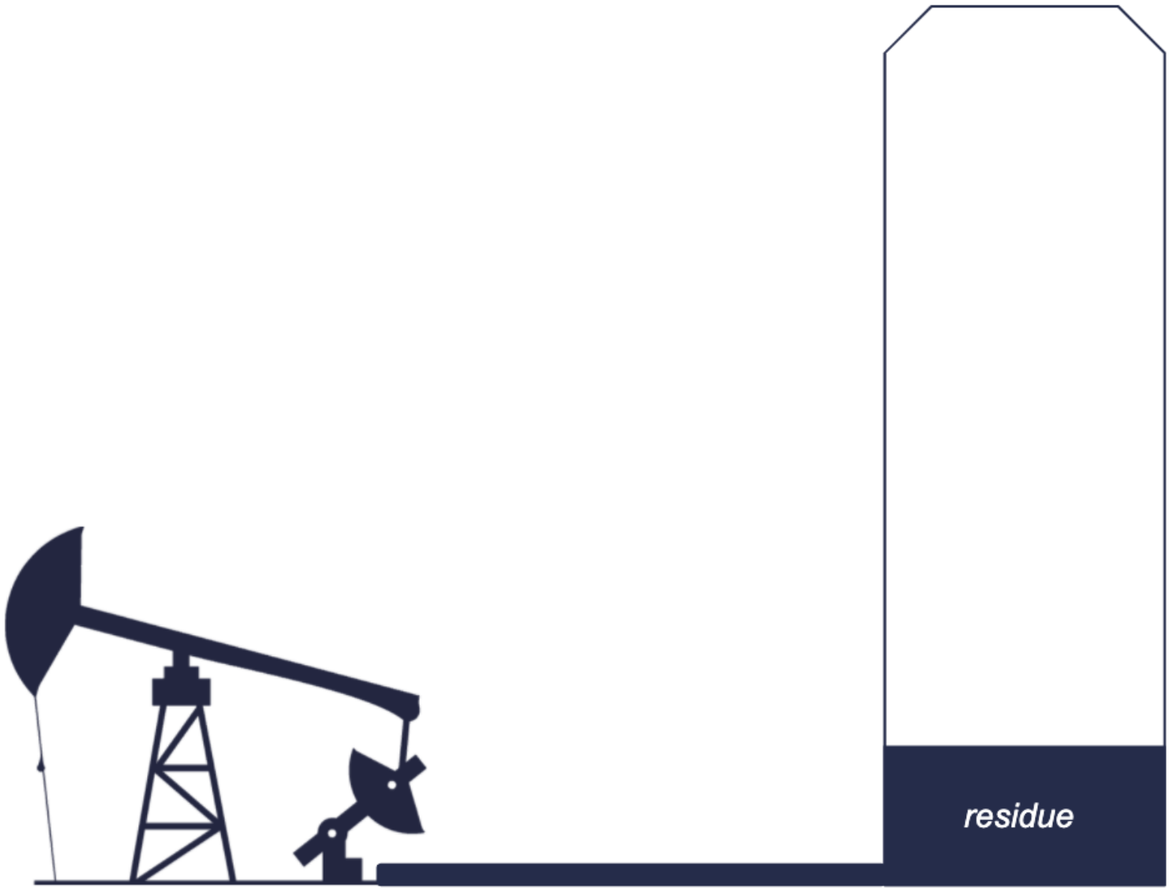
DATA ENGINEERING FOR EVERYONE



Hadrien Lacroix
Content Developer at DataCamp

If data is the new oil...

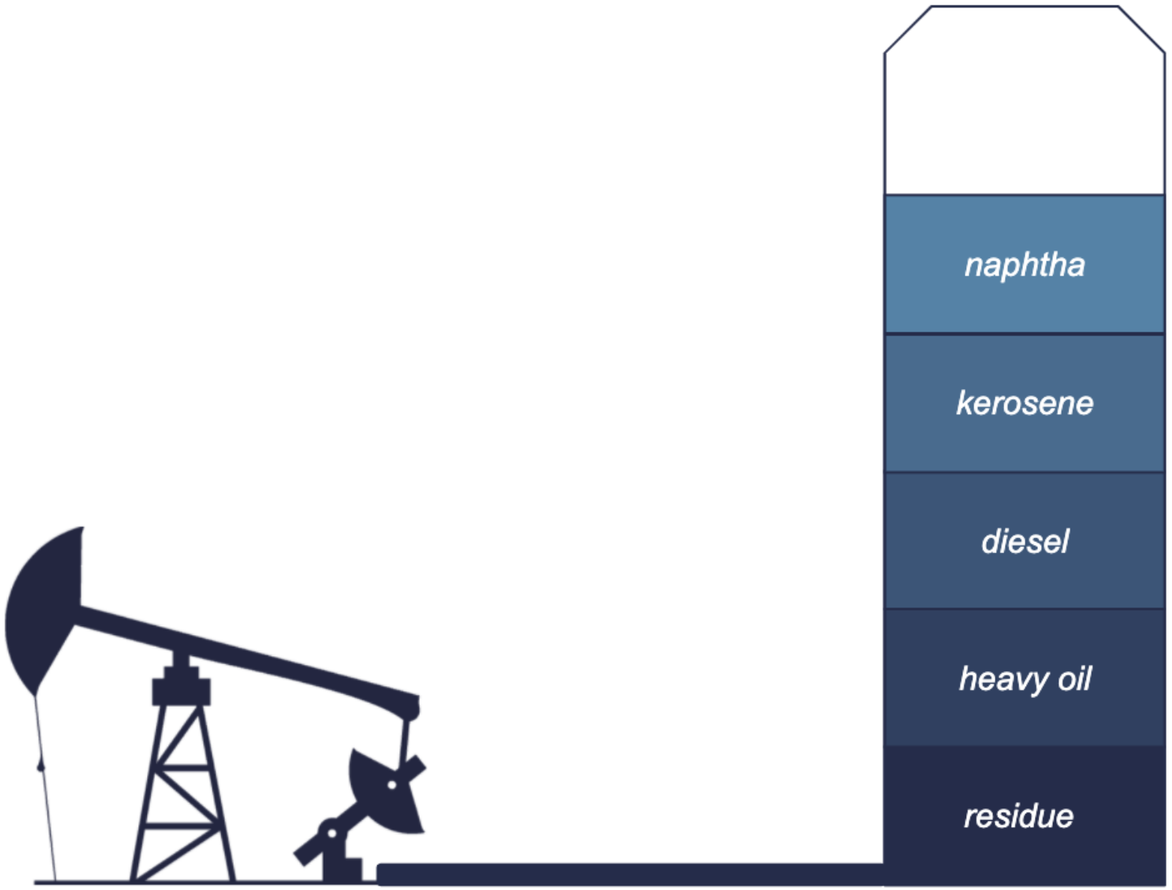
[1] The Economist, 2017-05-06, by David Parkins

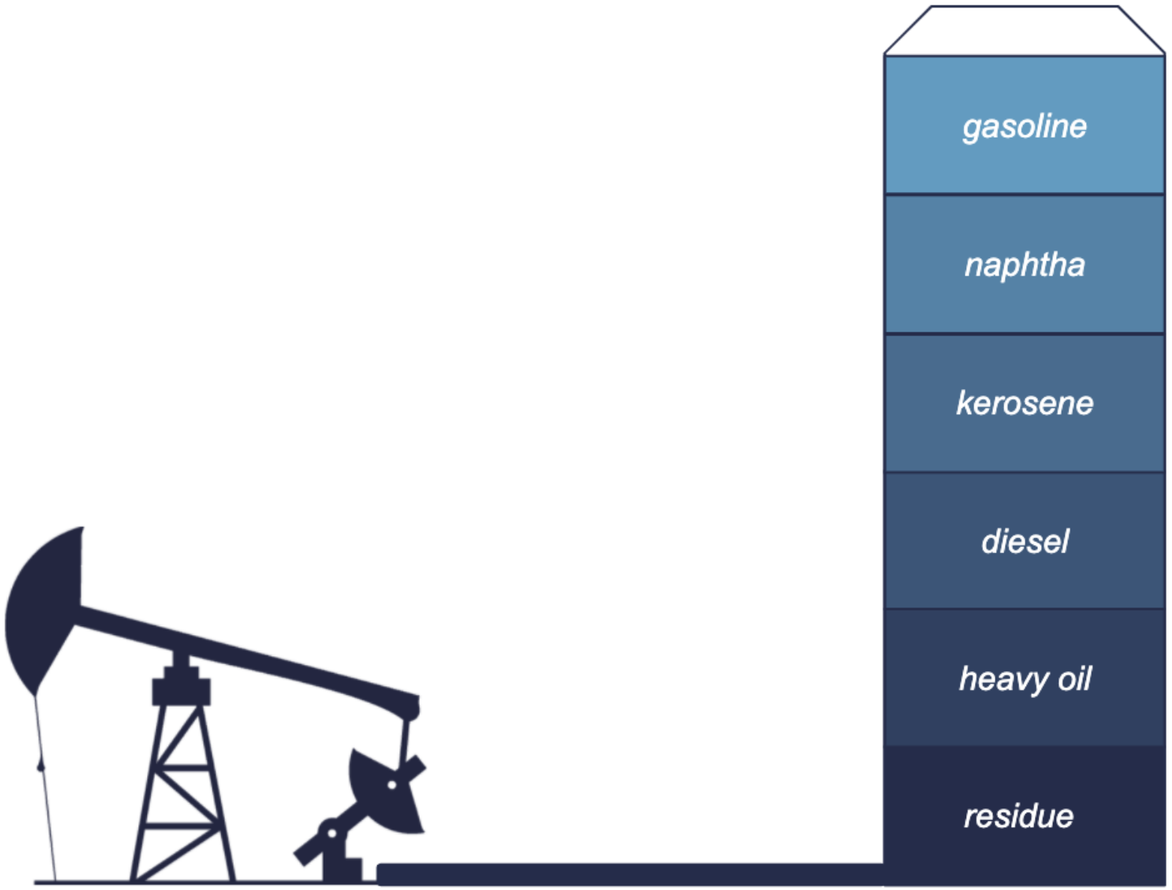

residue

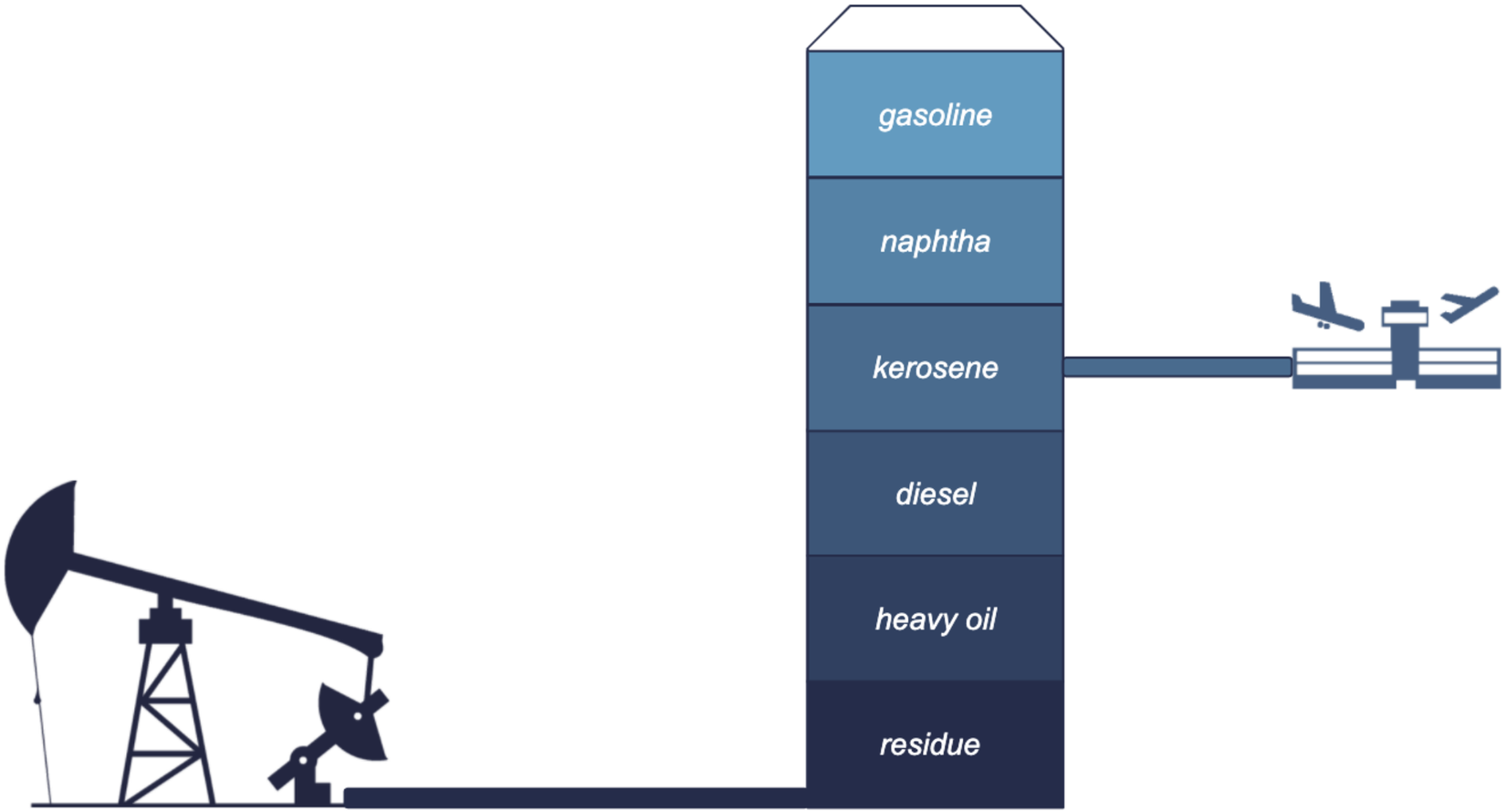

heavy oil
residue

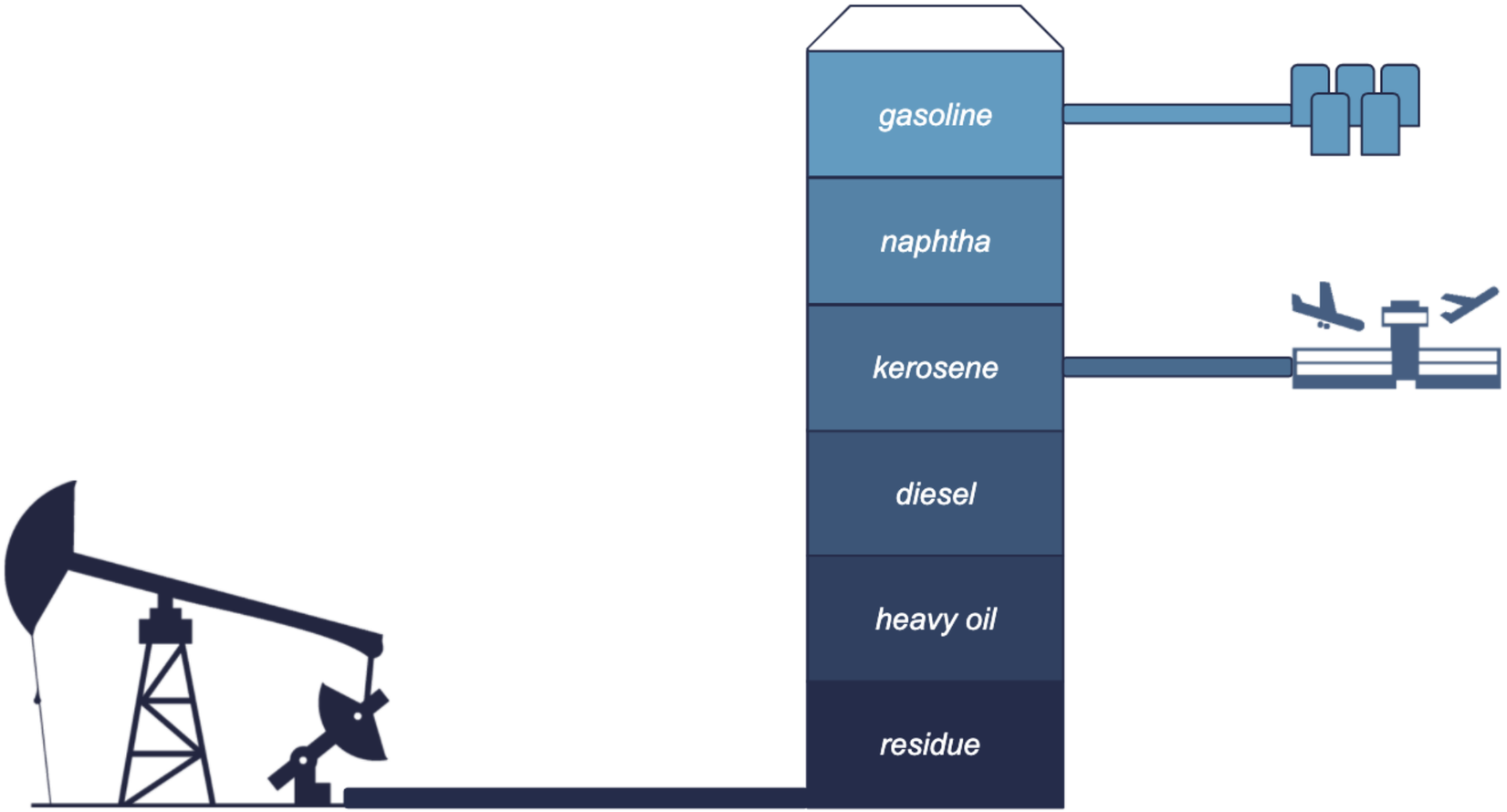

diesel
heavy oil
residue

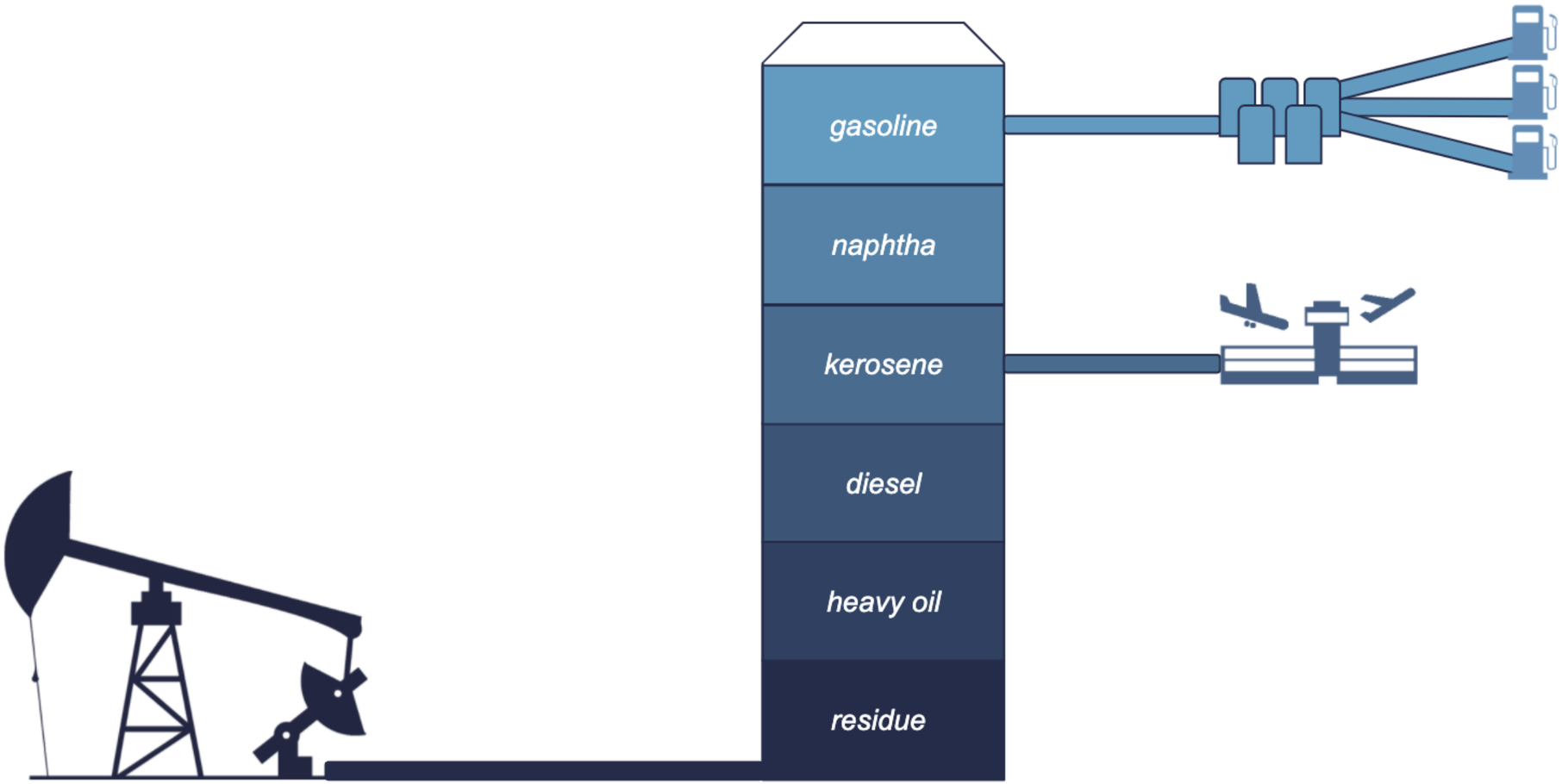

kerosene
diesel
heavy oil
residue

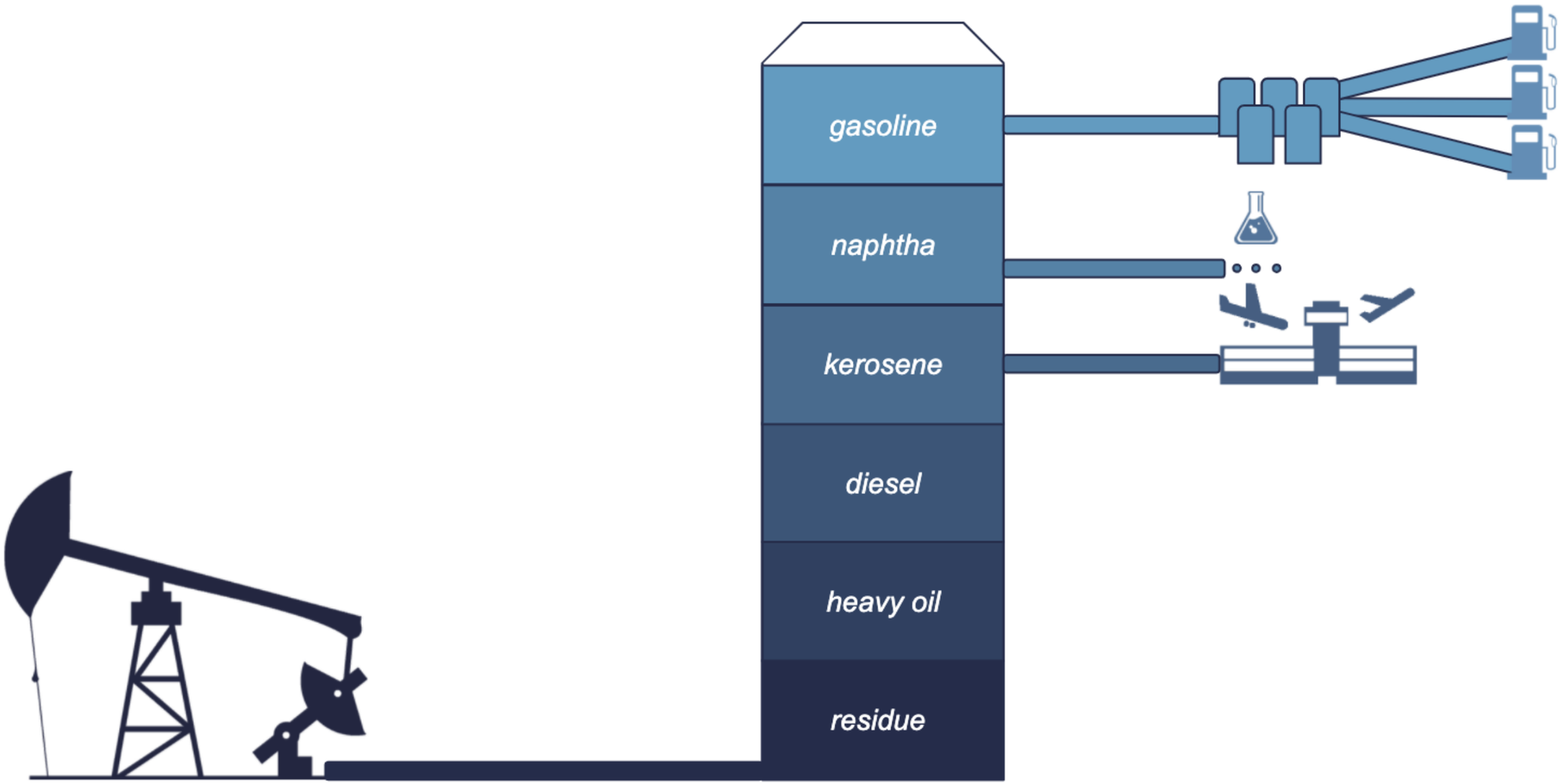

naphtha
kerosene
diesel
heavy oil
residue

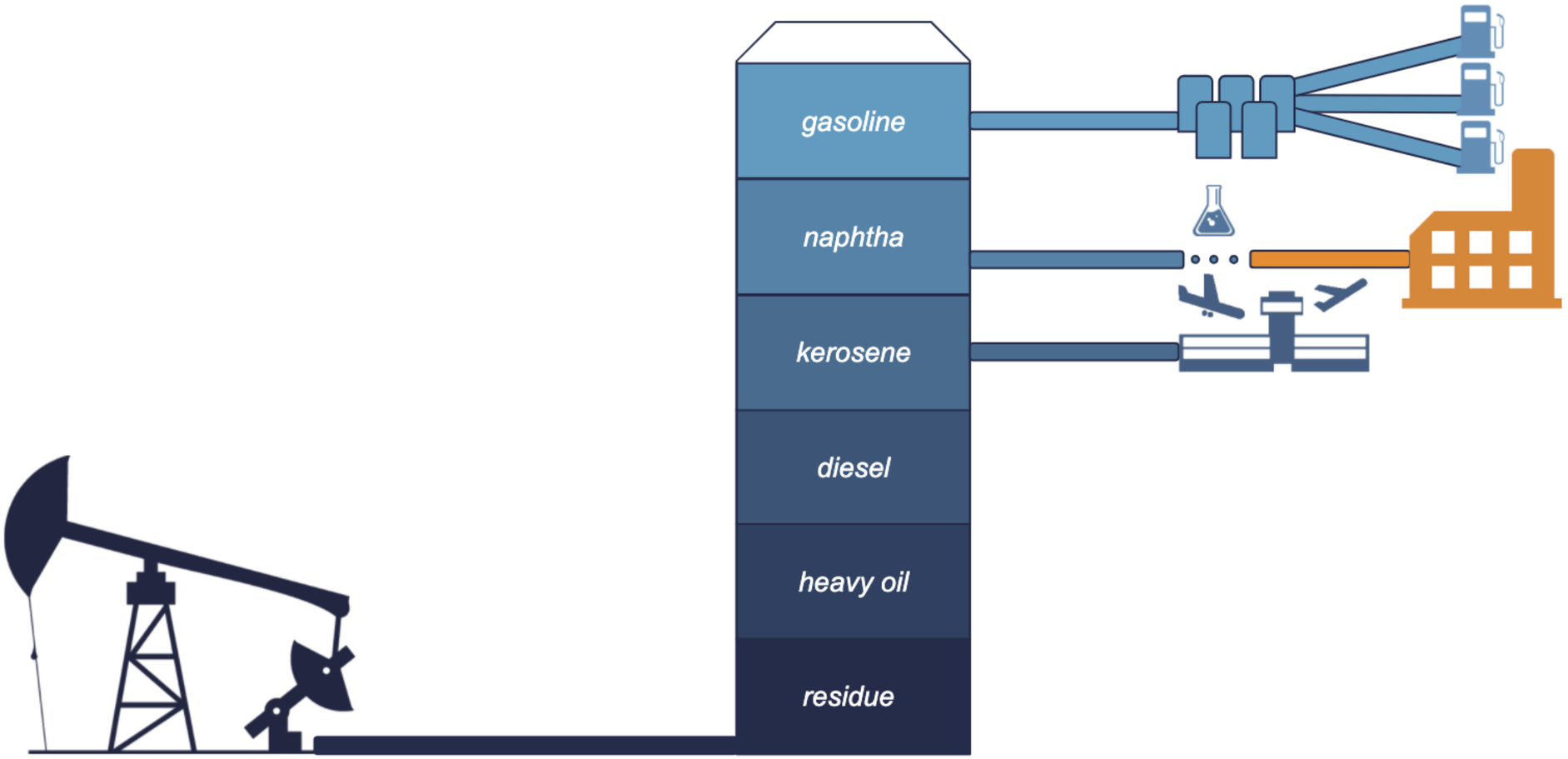

gasoline
naphtha
kerosene
diesel
heavy oil
residue


gasoline
naphtha
kerosene
diesel
heavy oil
residue


gasoline
naphtha
kerosene
diesel
heavy oil
residue

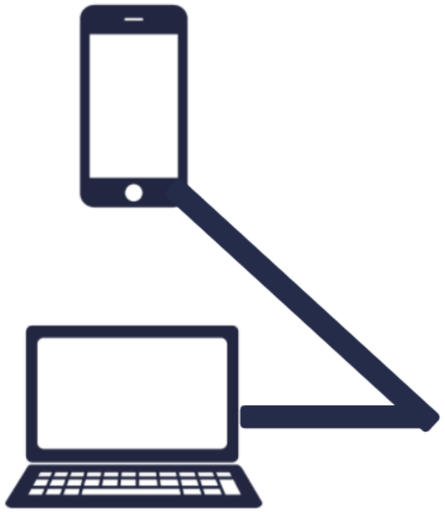

gasoline
naphtha
kerosene
diesel
heavy oil
residue

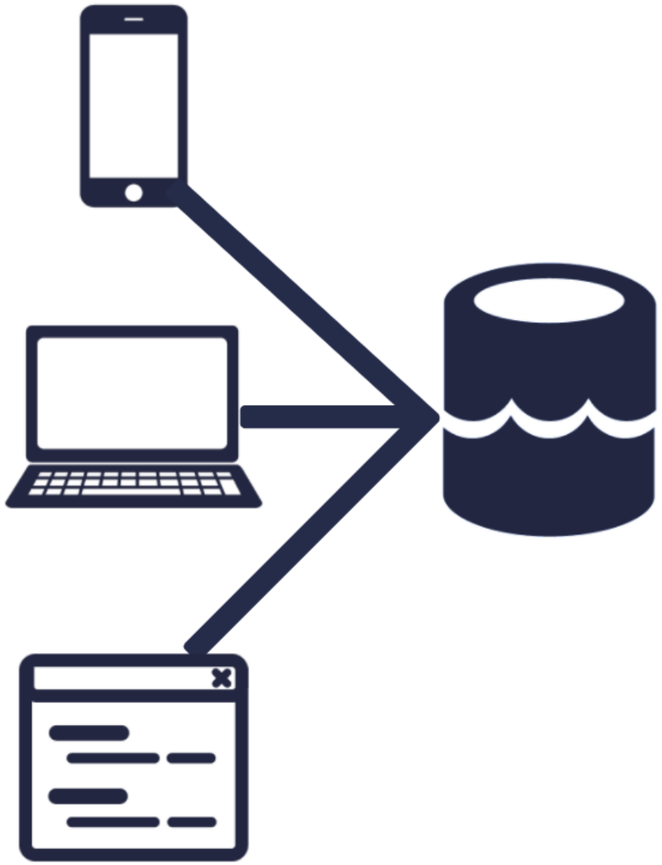

gasoline
naphtha
kerosene
diesel
heavy oil
residue

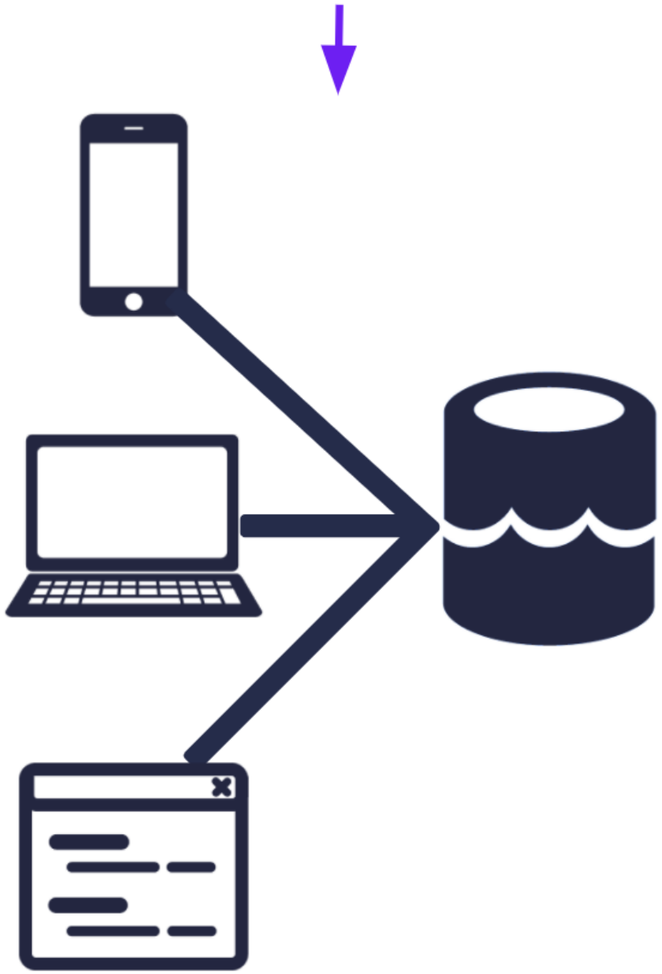

gasoline
naphtha
kerosene
diesel
heavy oil
residue

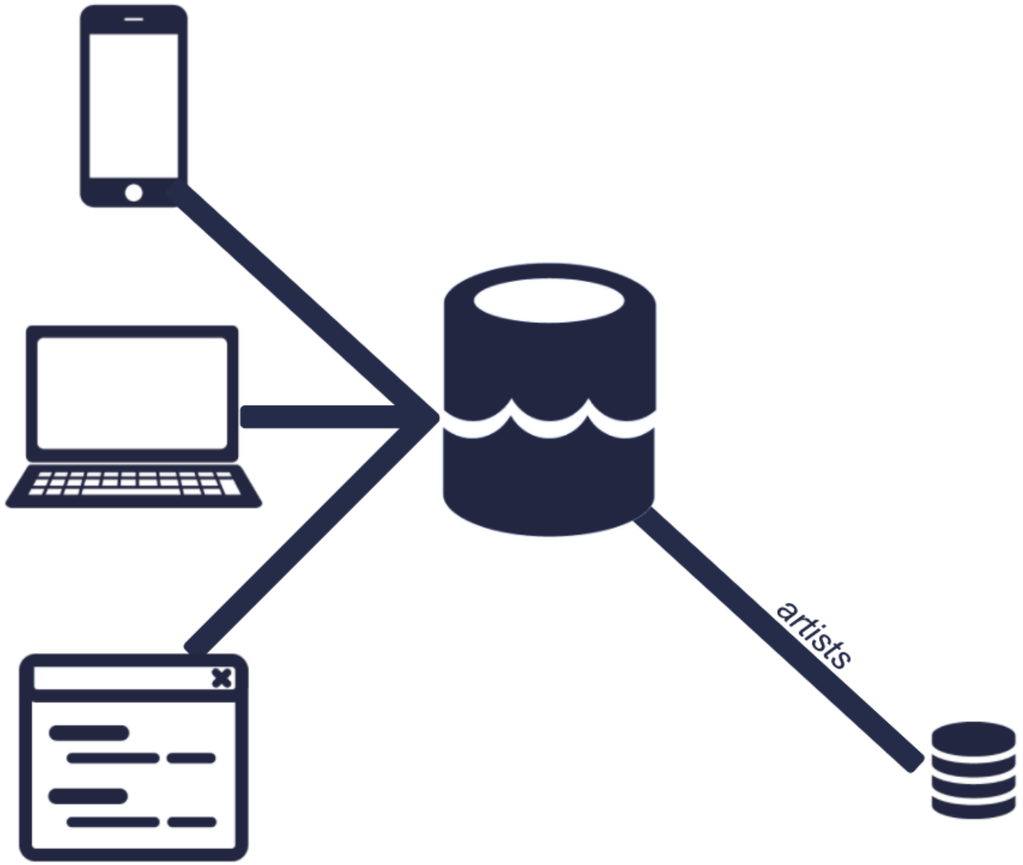
Back to data engineering

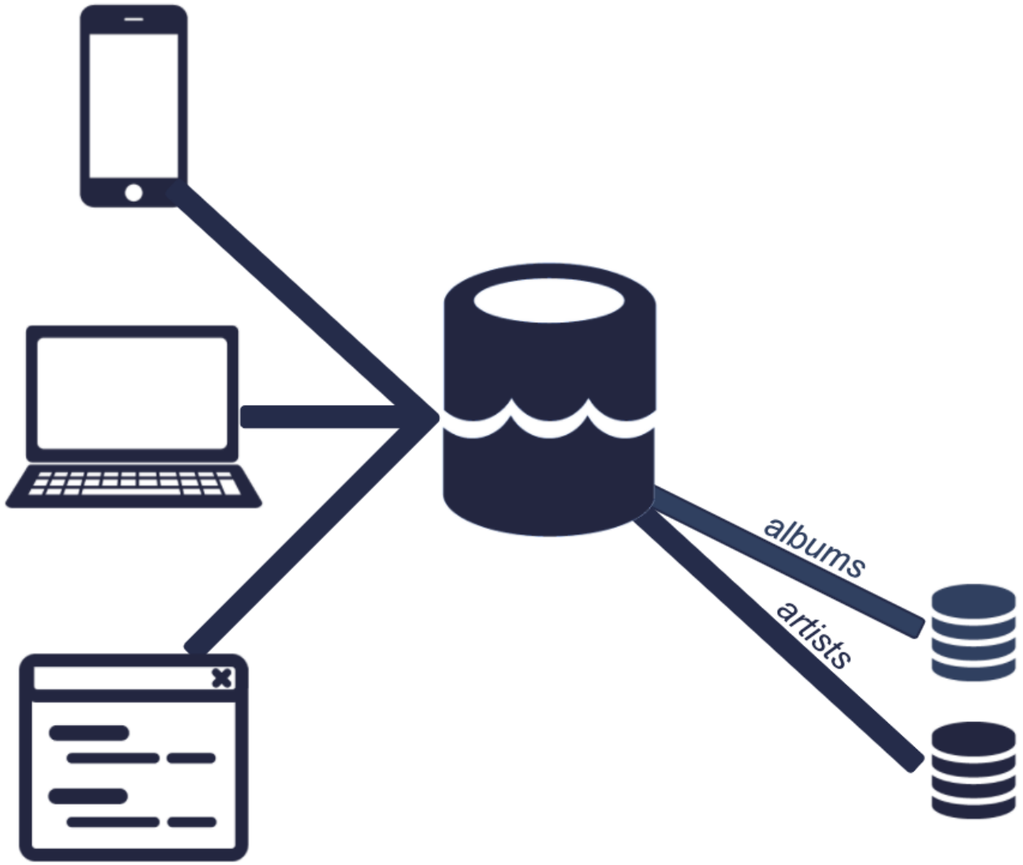
- Ingest
- Process
- Store
- Need pipelines
- Automate flow from one station to the next
- Provide up-to-date, accurate, relevant data

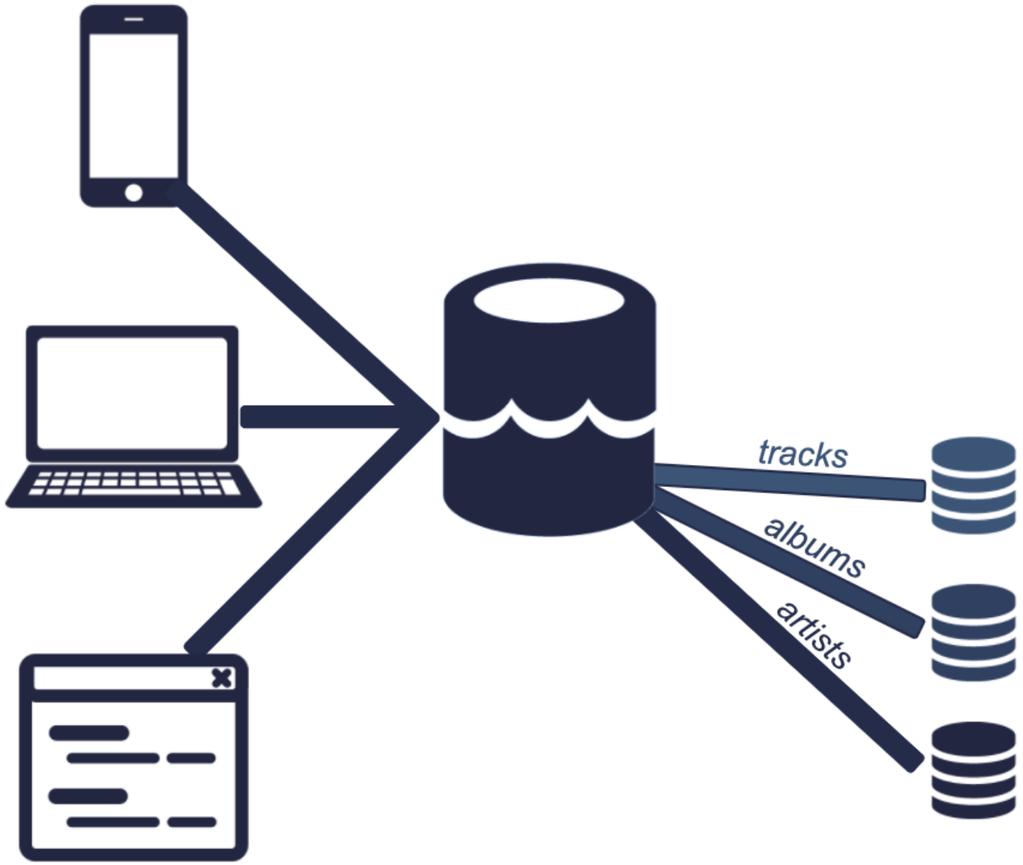


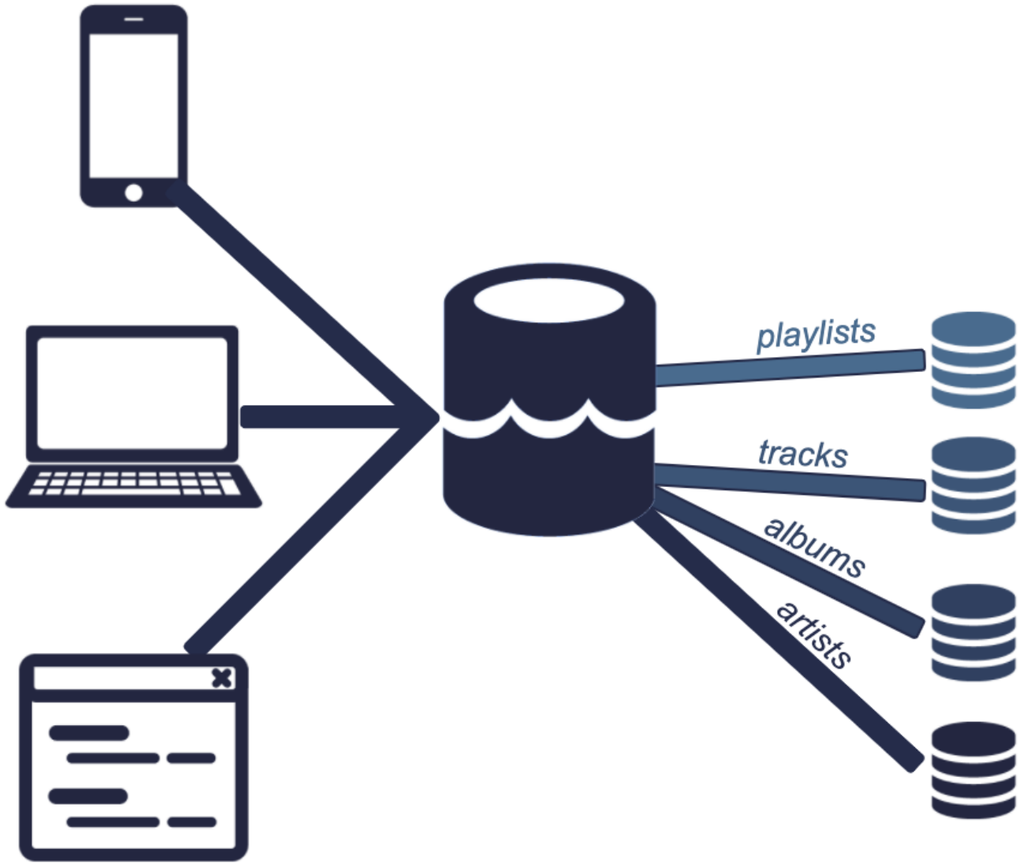


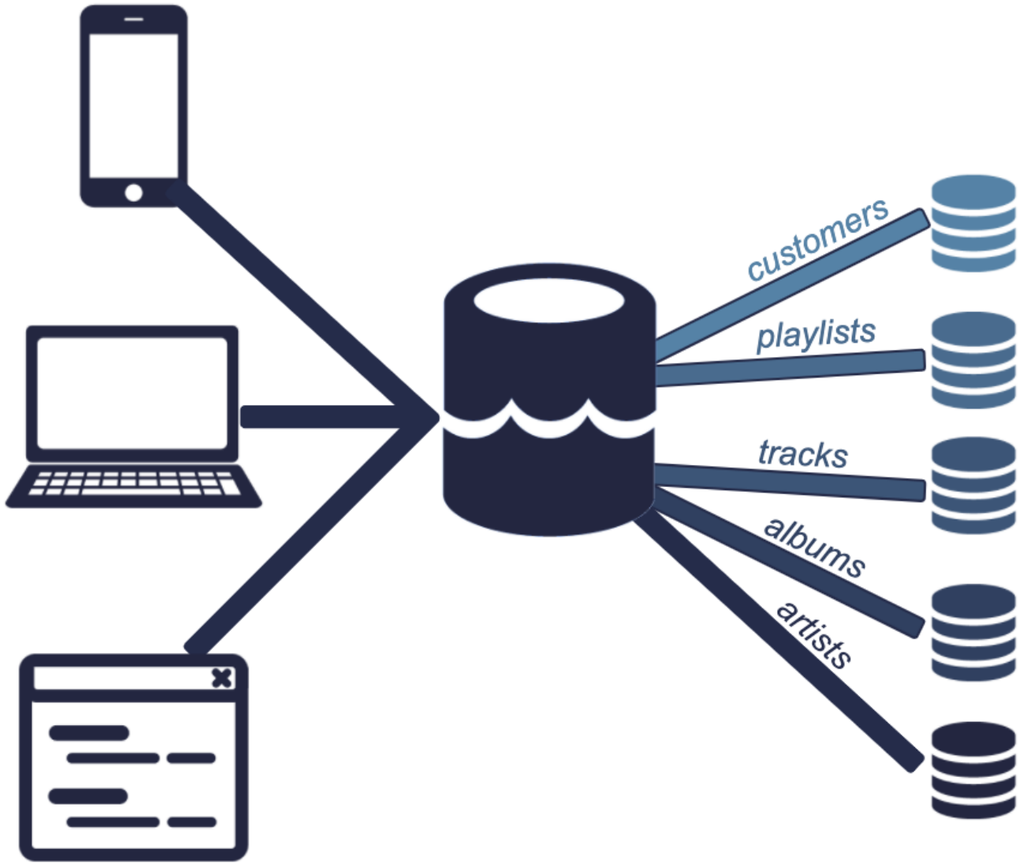


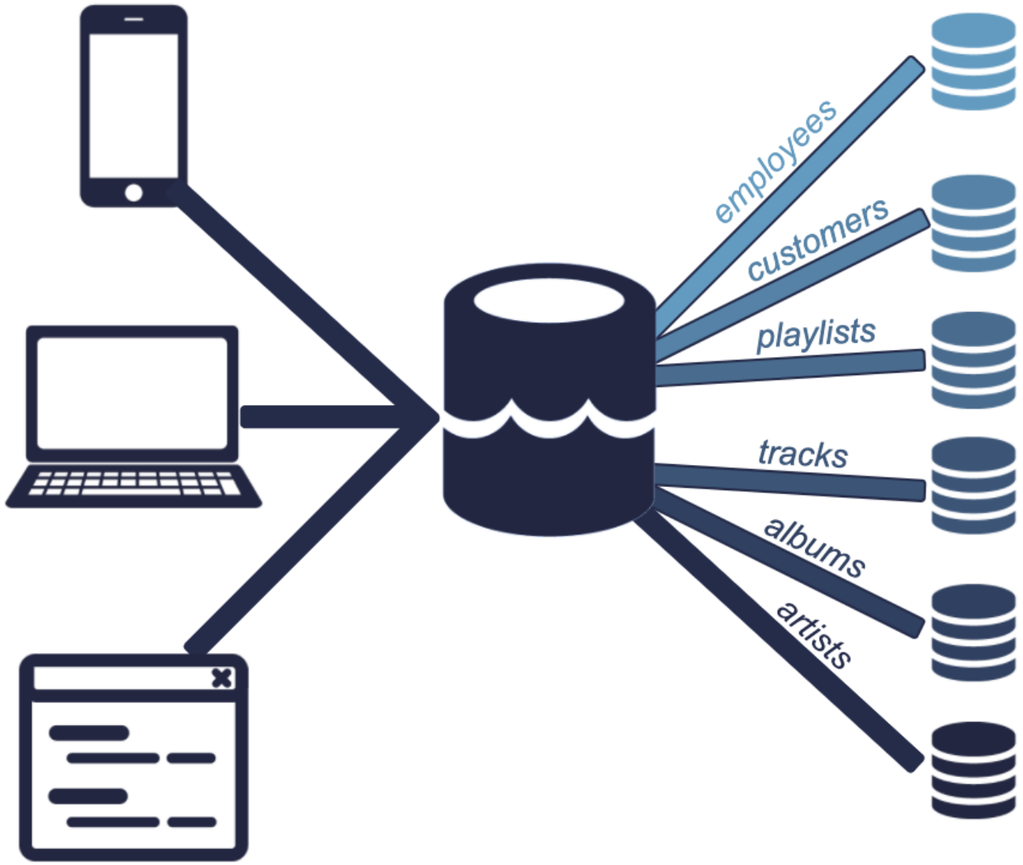

artists

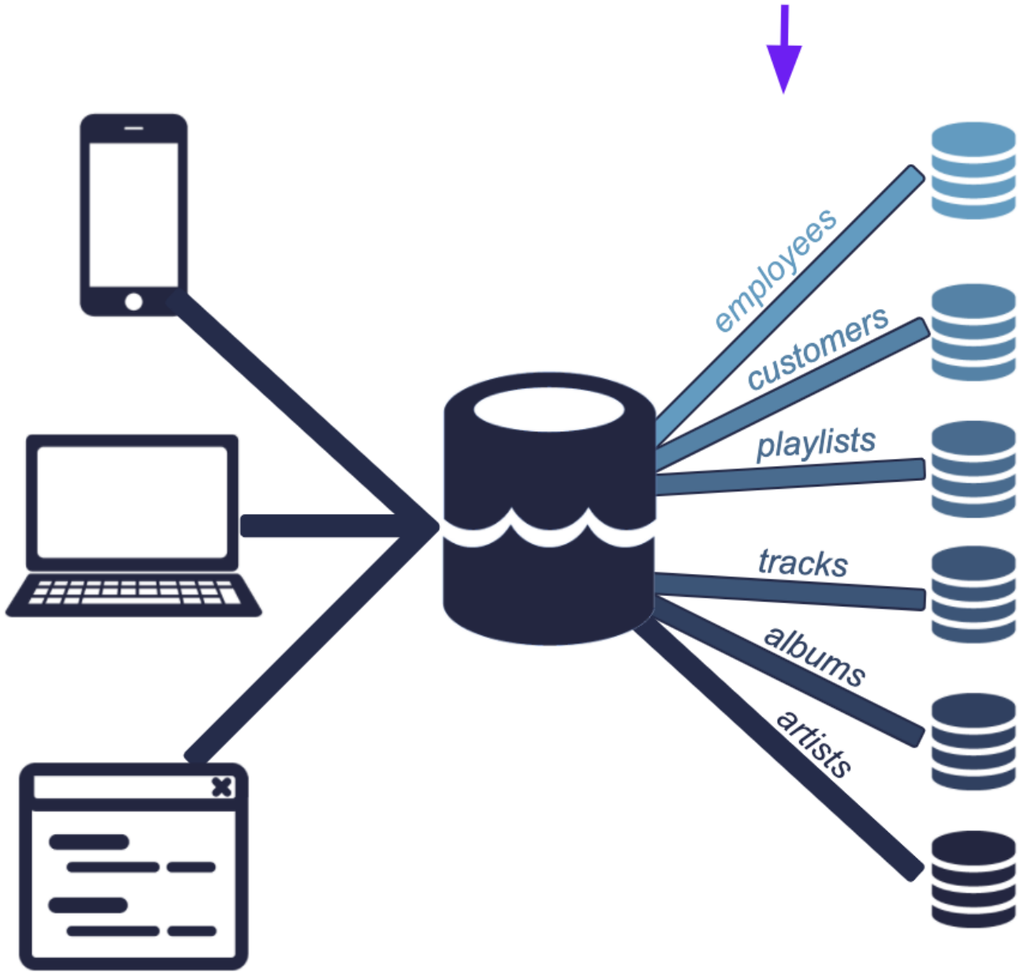

albums
artists

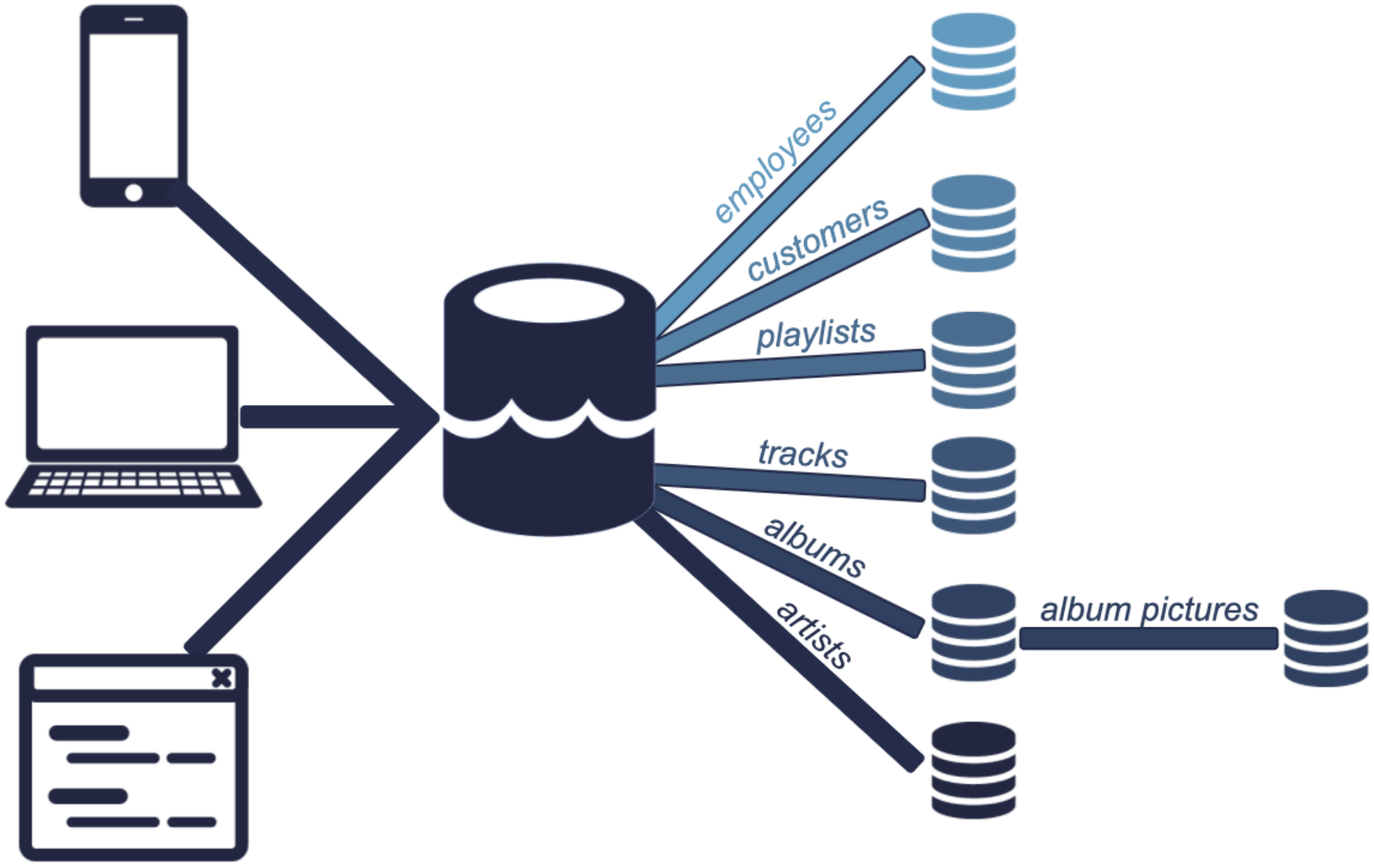

tracks
albums
artists

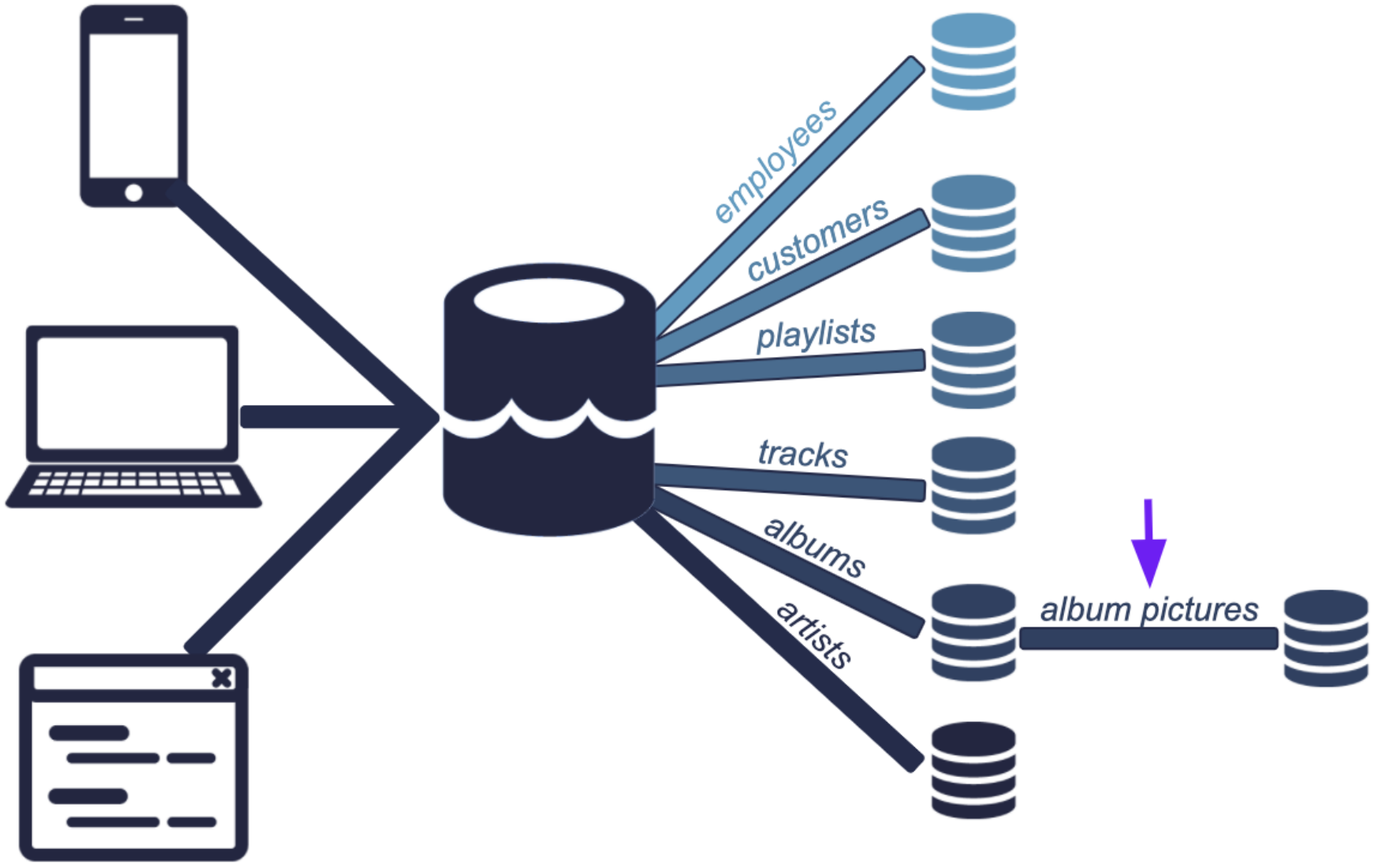

playlists
tracks
albums
artists

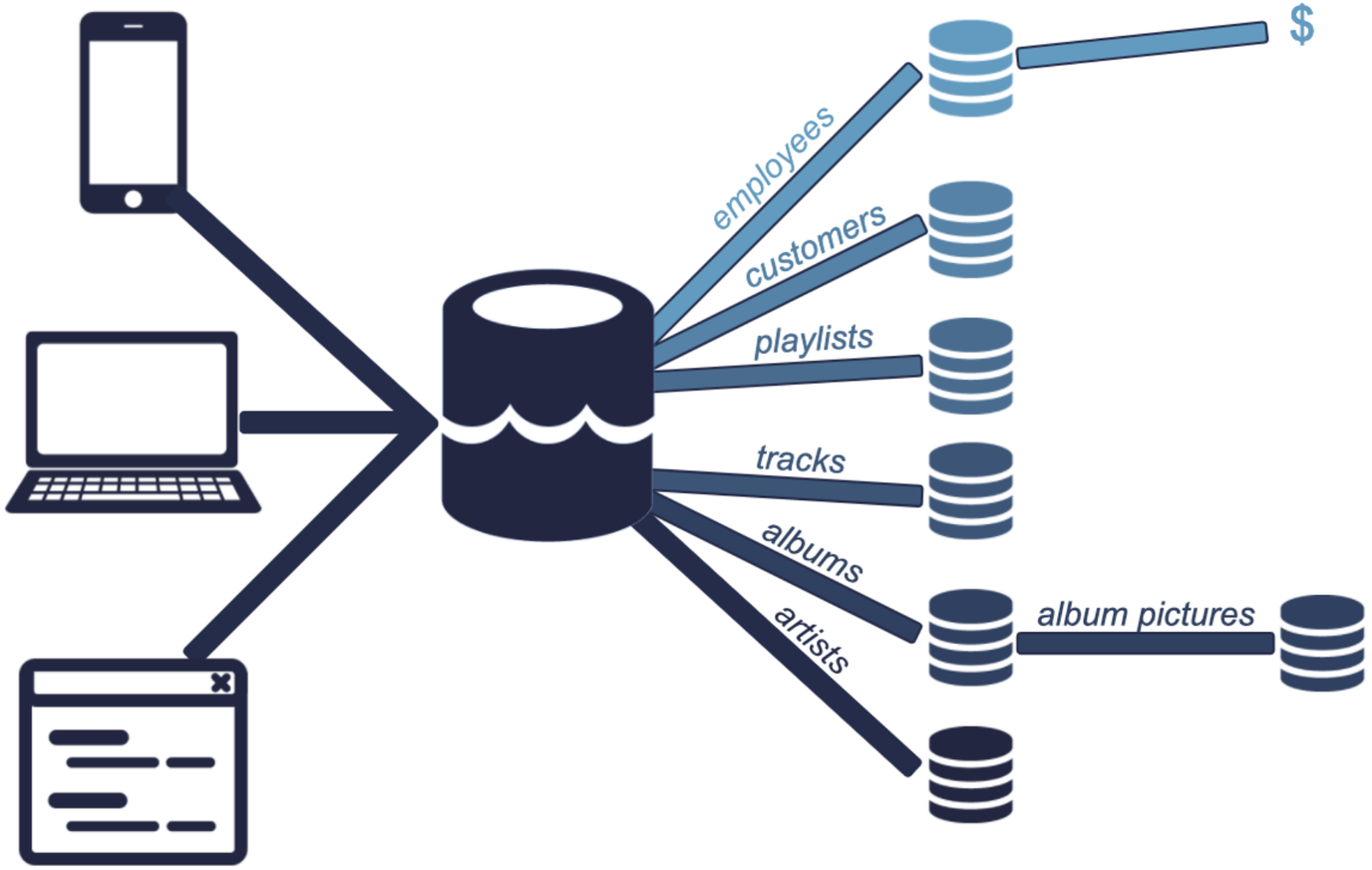

customers
playlists
tracks
albums
artists

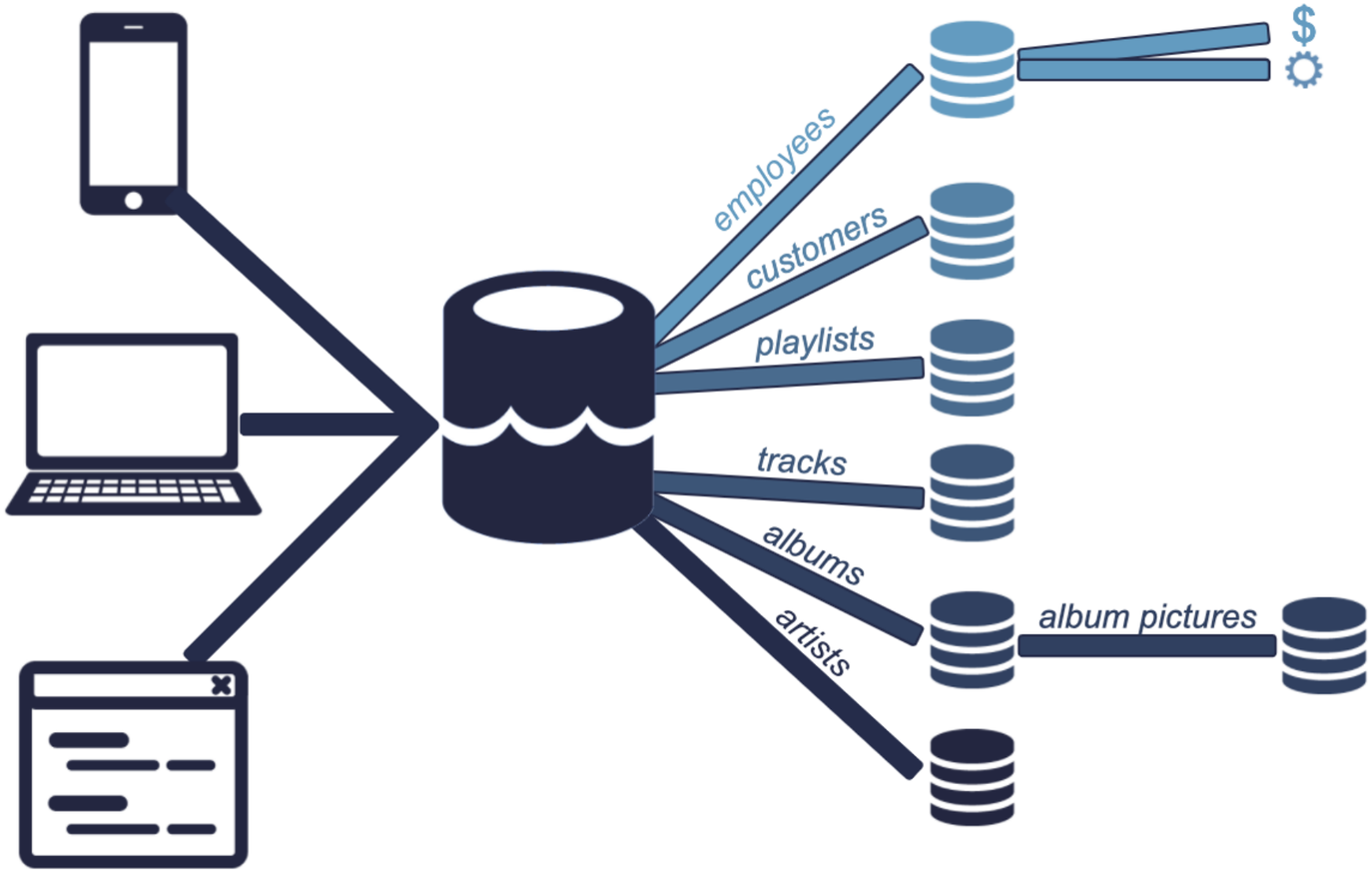

employees
customers
playlists
tracks
albums
artists

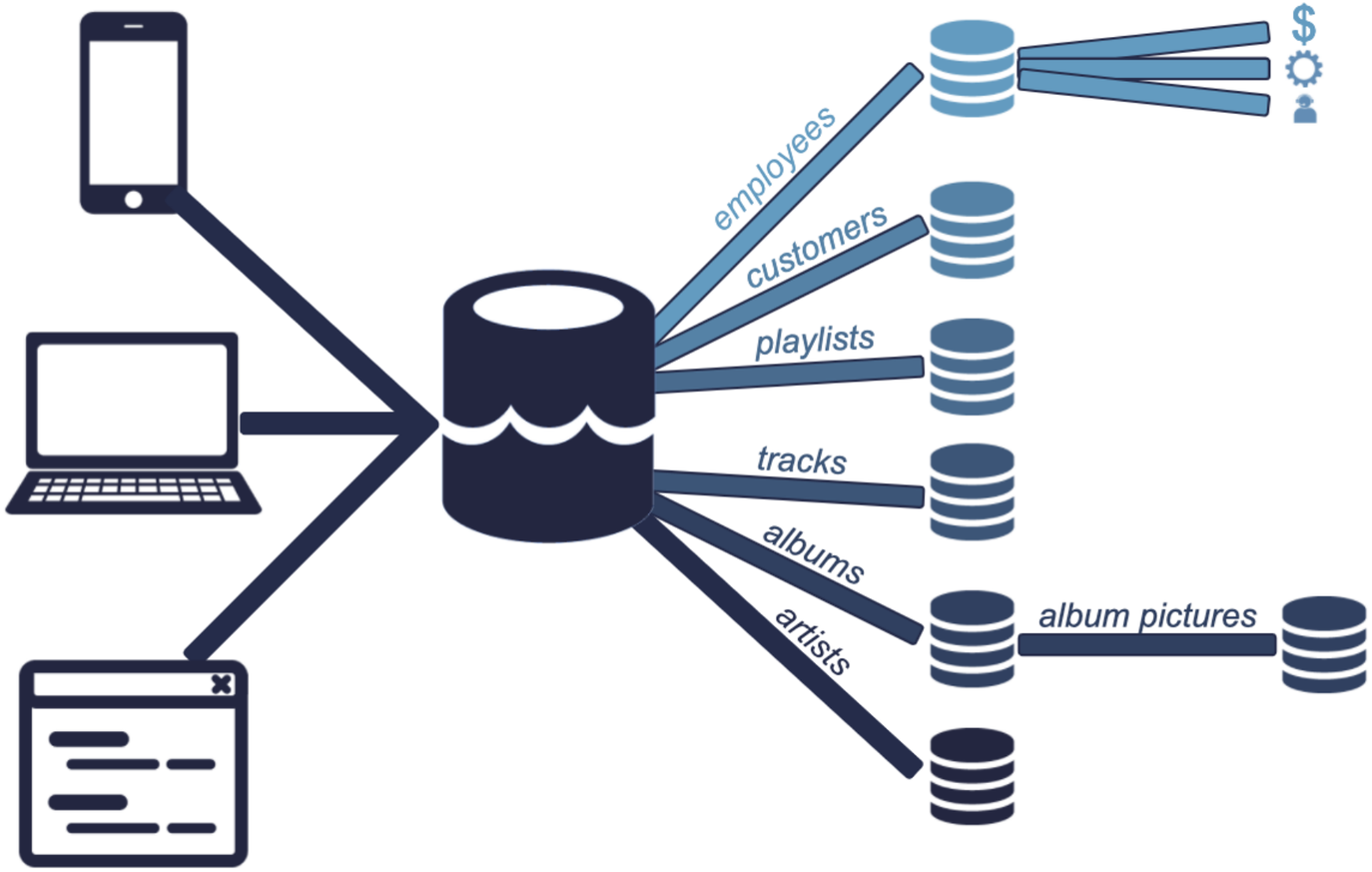

employees
customers
playlists
tracks
albums
artists

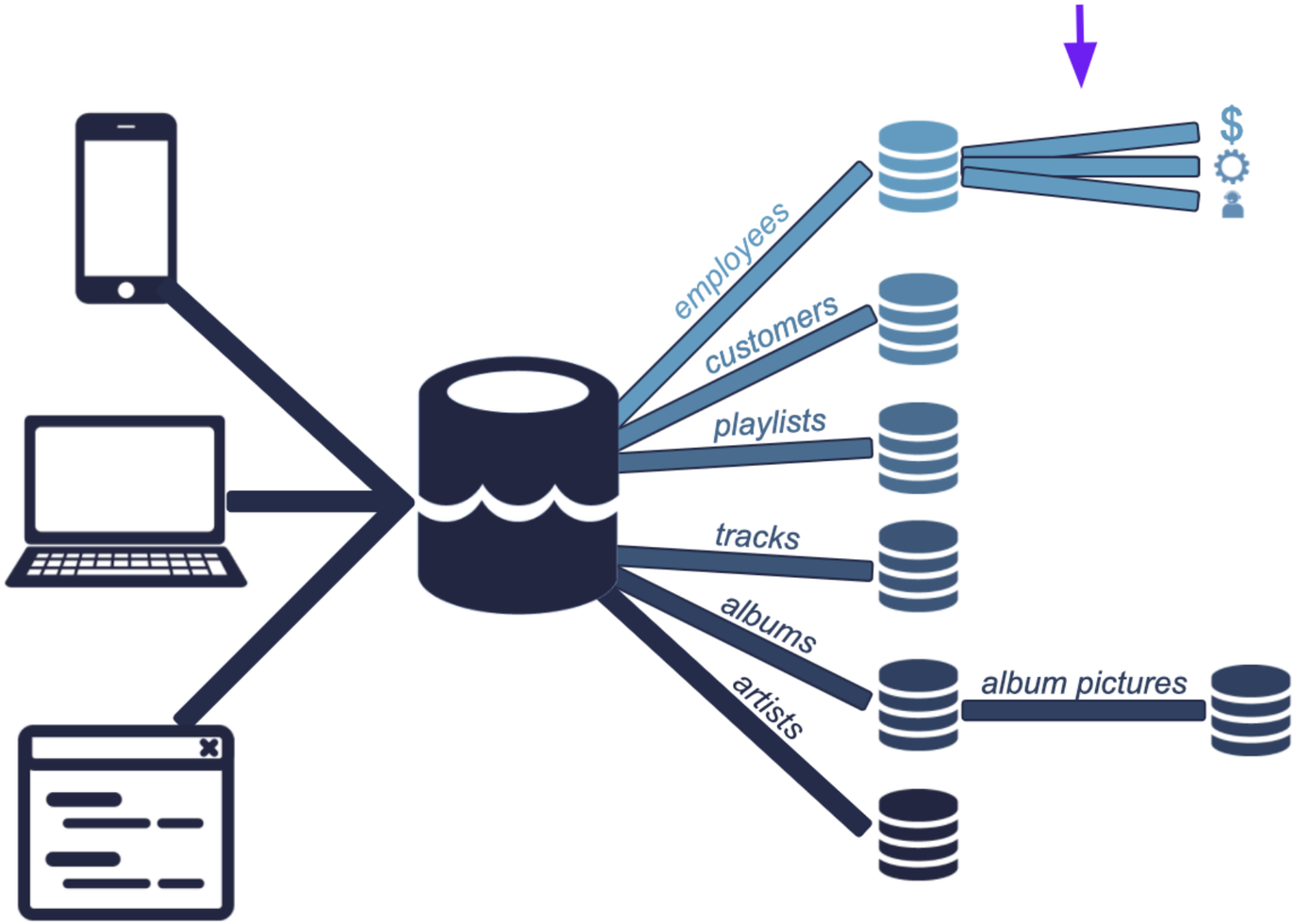

employees
customers
playlists
tracks
albums
artists
album pictures

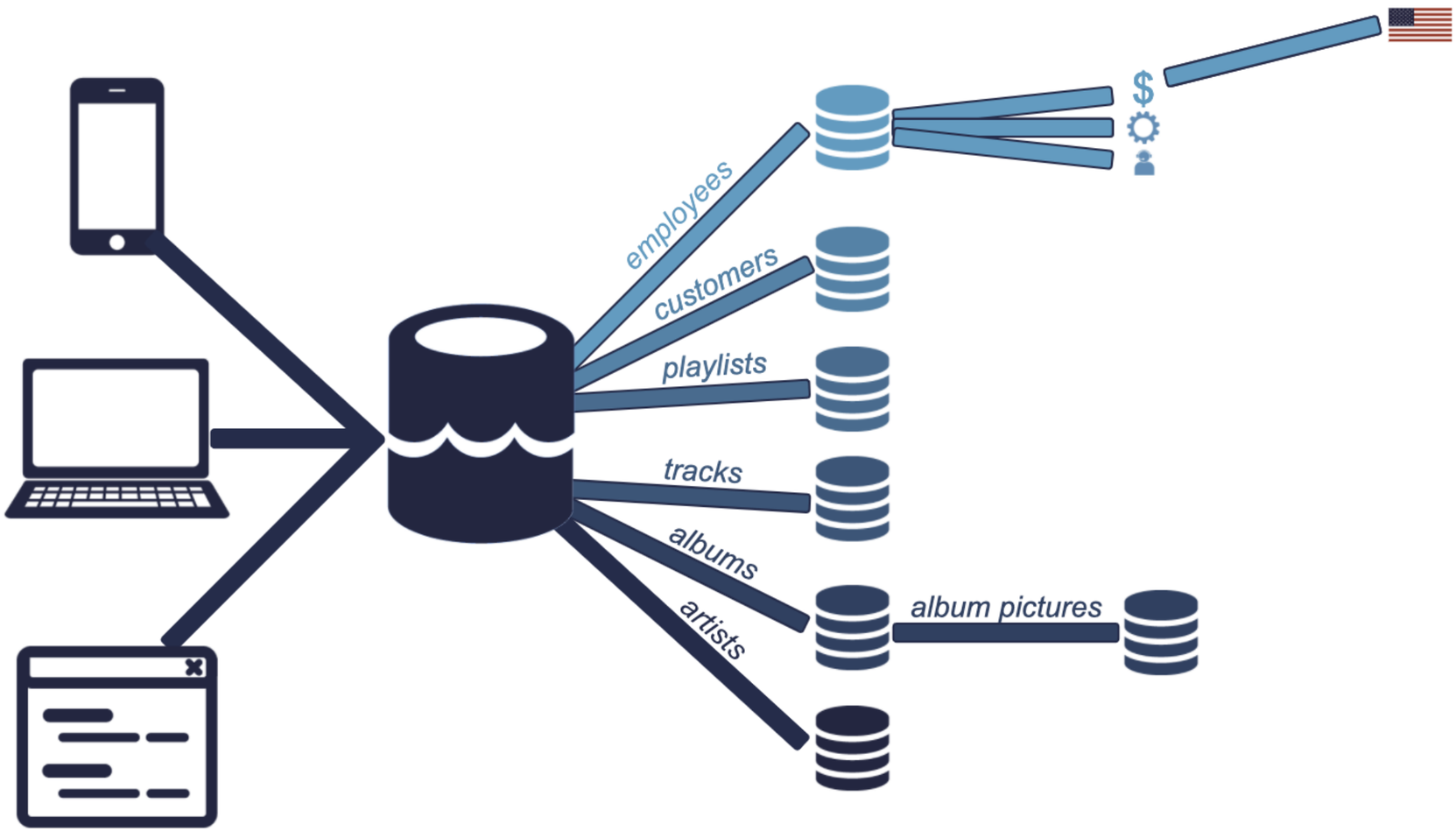

employees
customers
playlists
tracks
albums
artists
album pictures

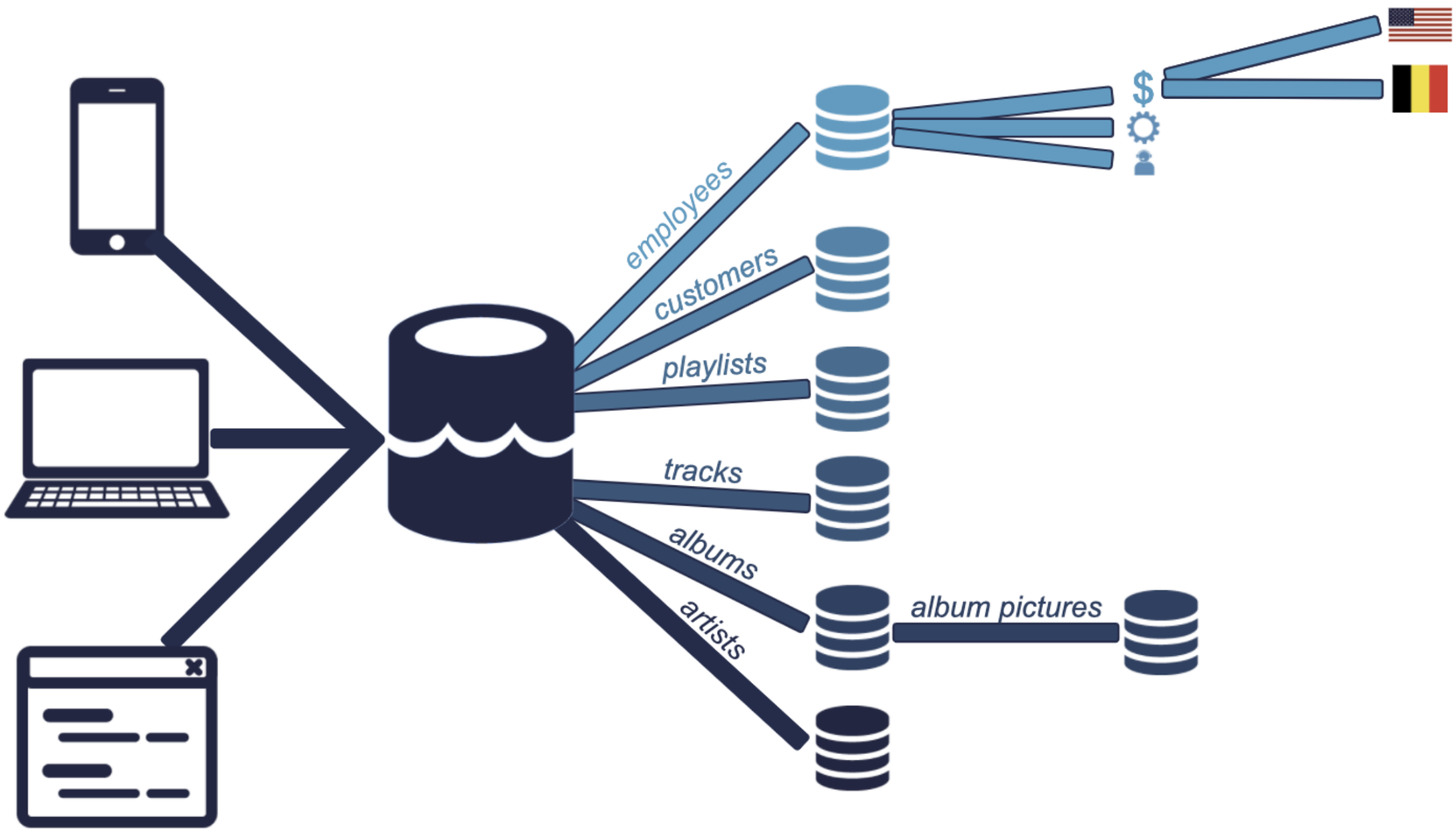

employees
customers
playlists
tracks
albums
artists
$
album pictures

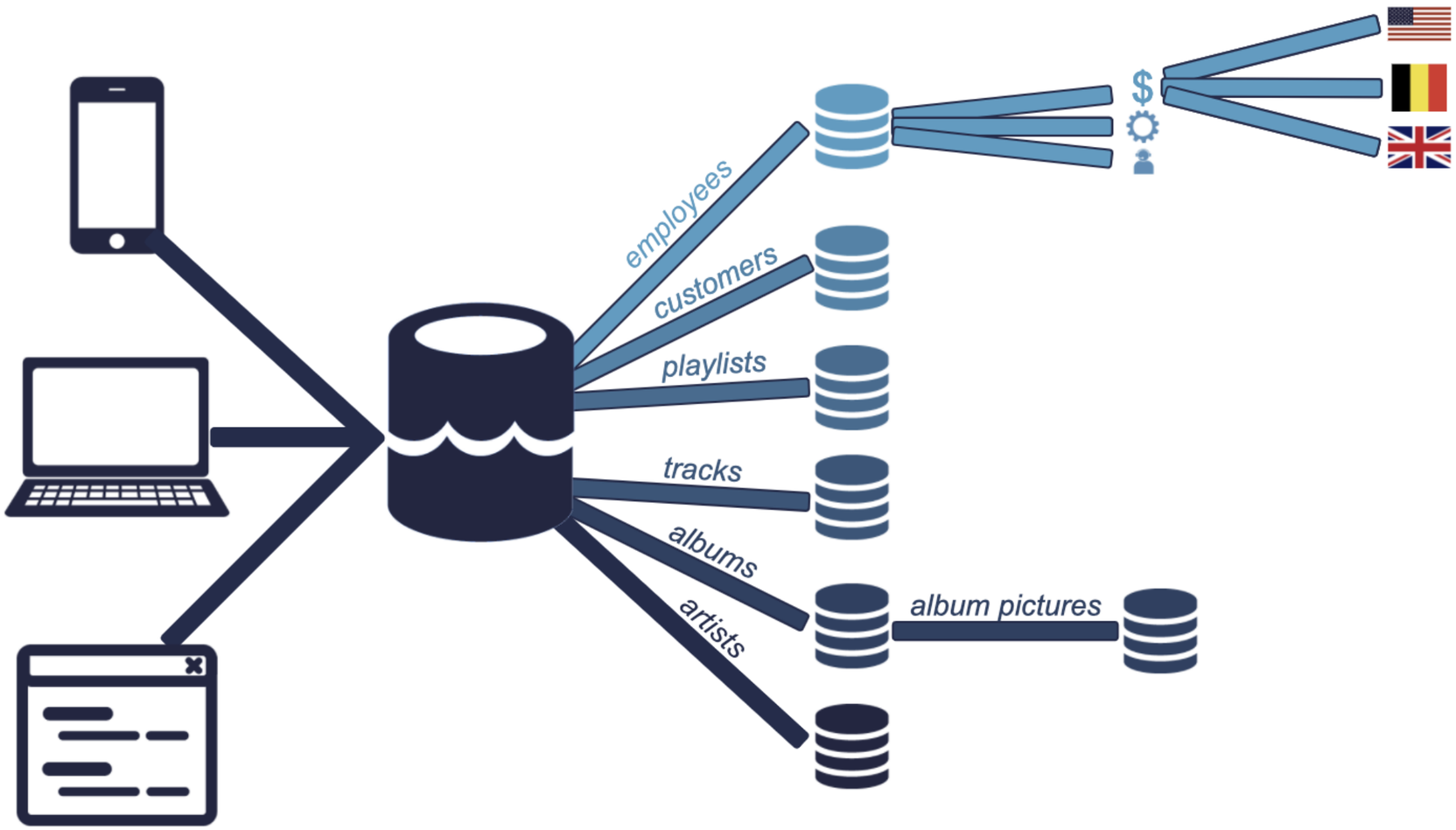

employees
customers
playlists
tracks
albums
artists
album pictures

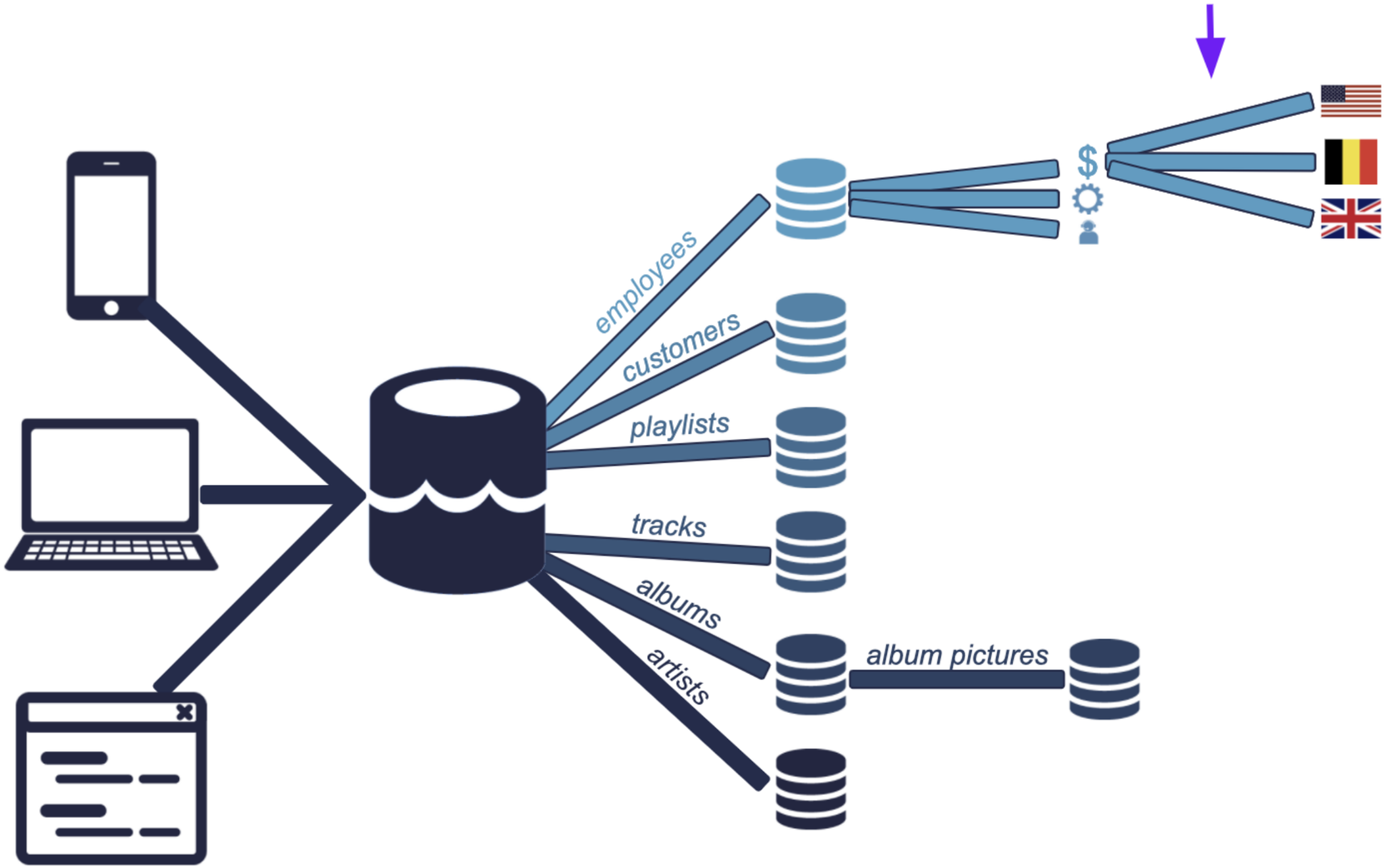

employees
customers
playlists
tracks
albums
artists
album pictures

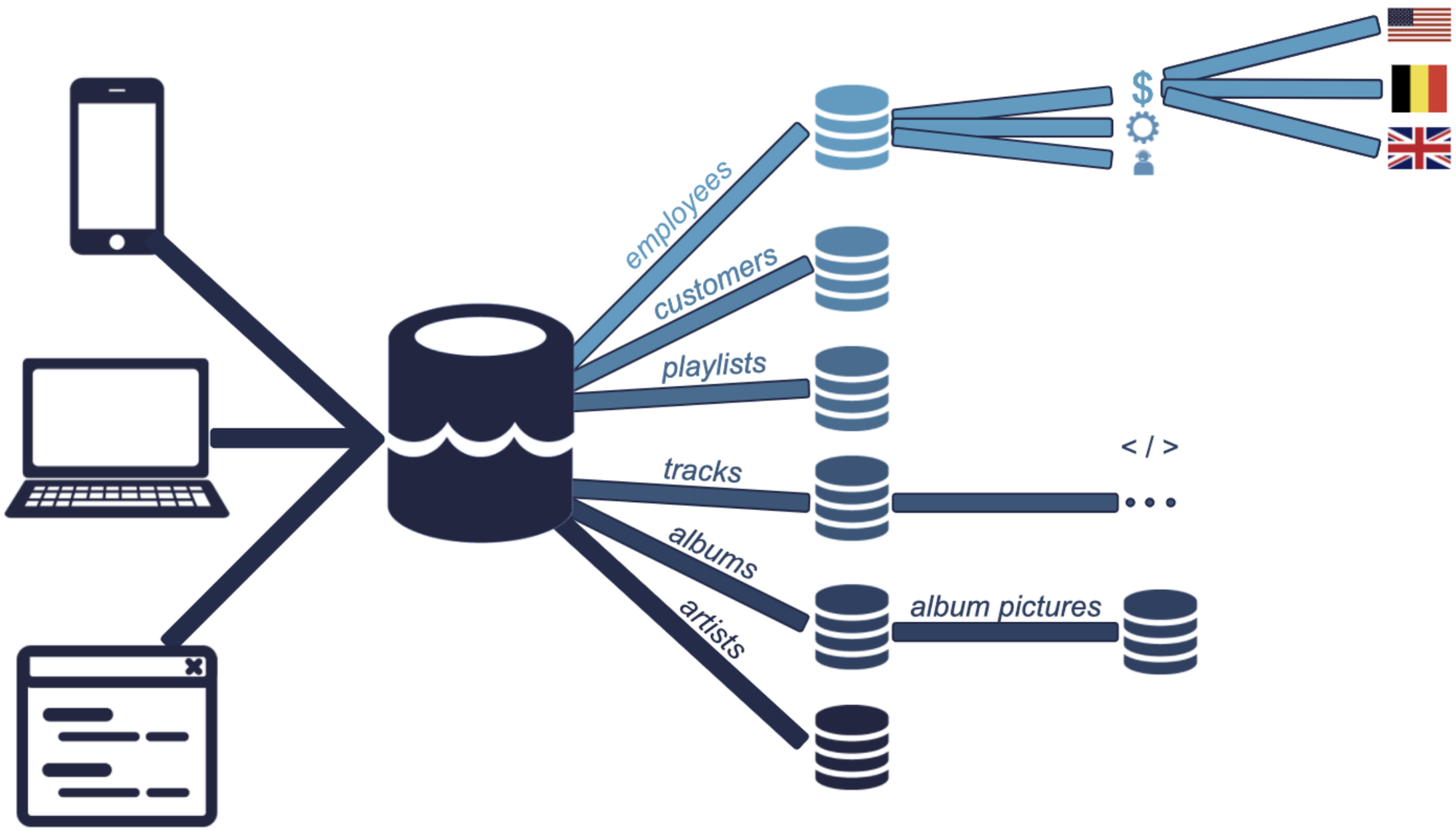

employees
customers
playlists
tracks
albums
artists
album pictures

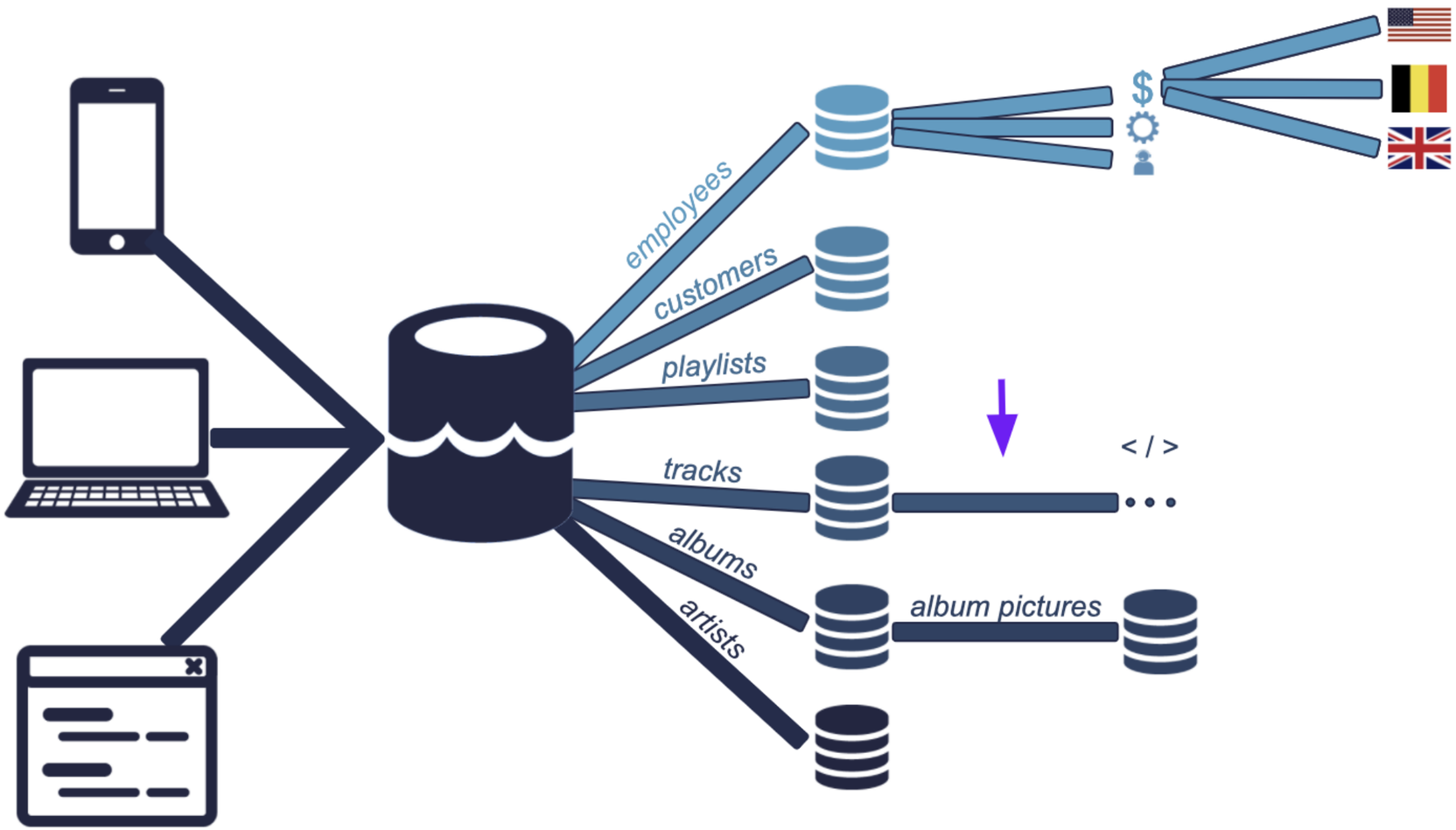

employees
customers
playlists
tracks
albums
artists
album pictures

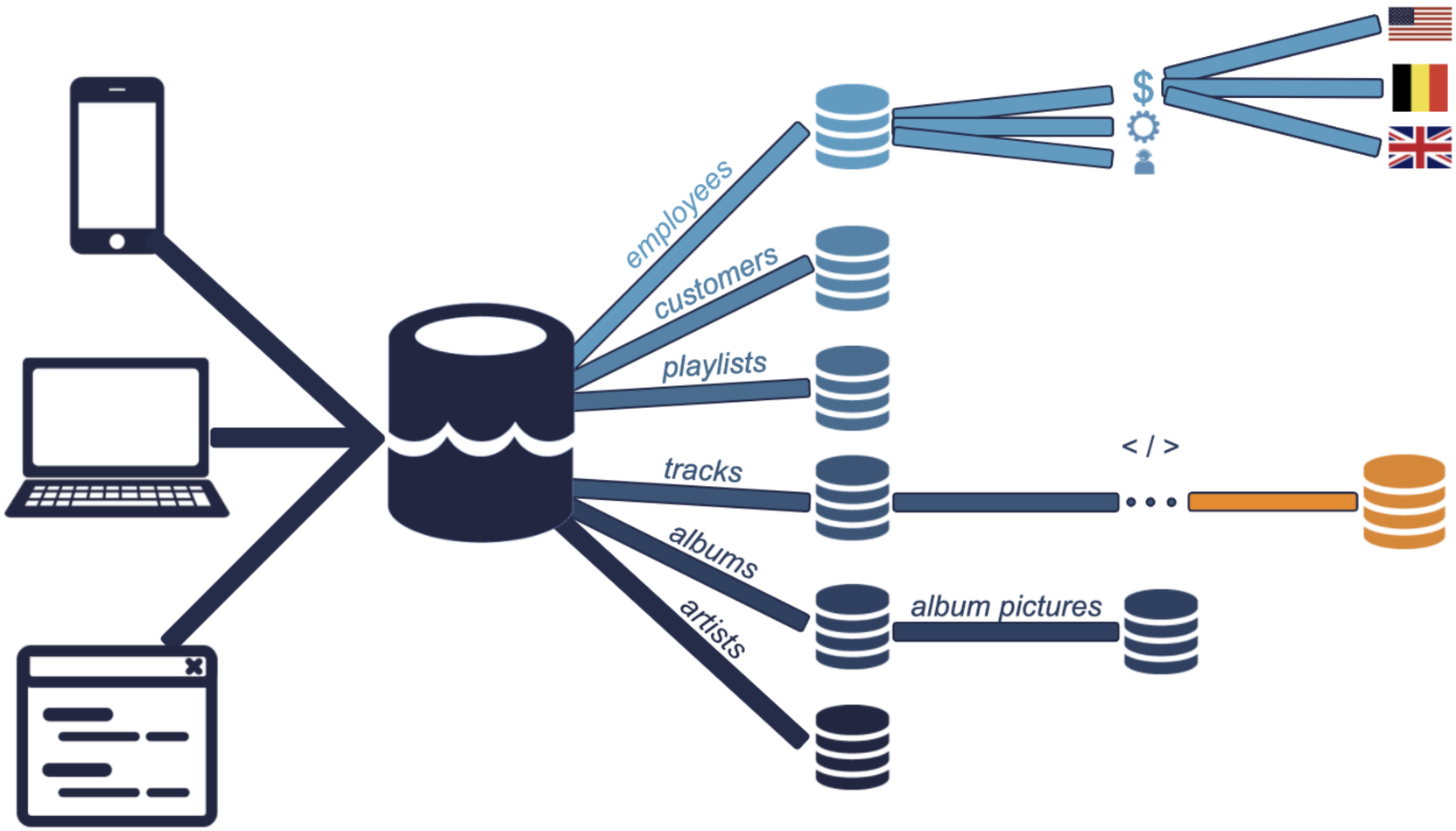

employees
customers
playlists
tracks
albums
artists
album pictures

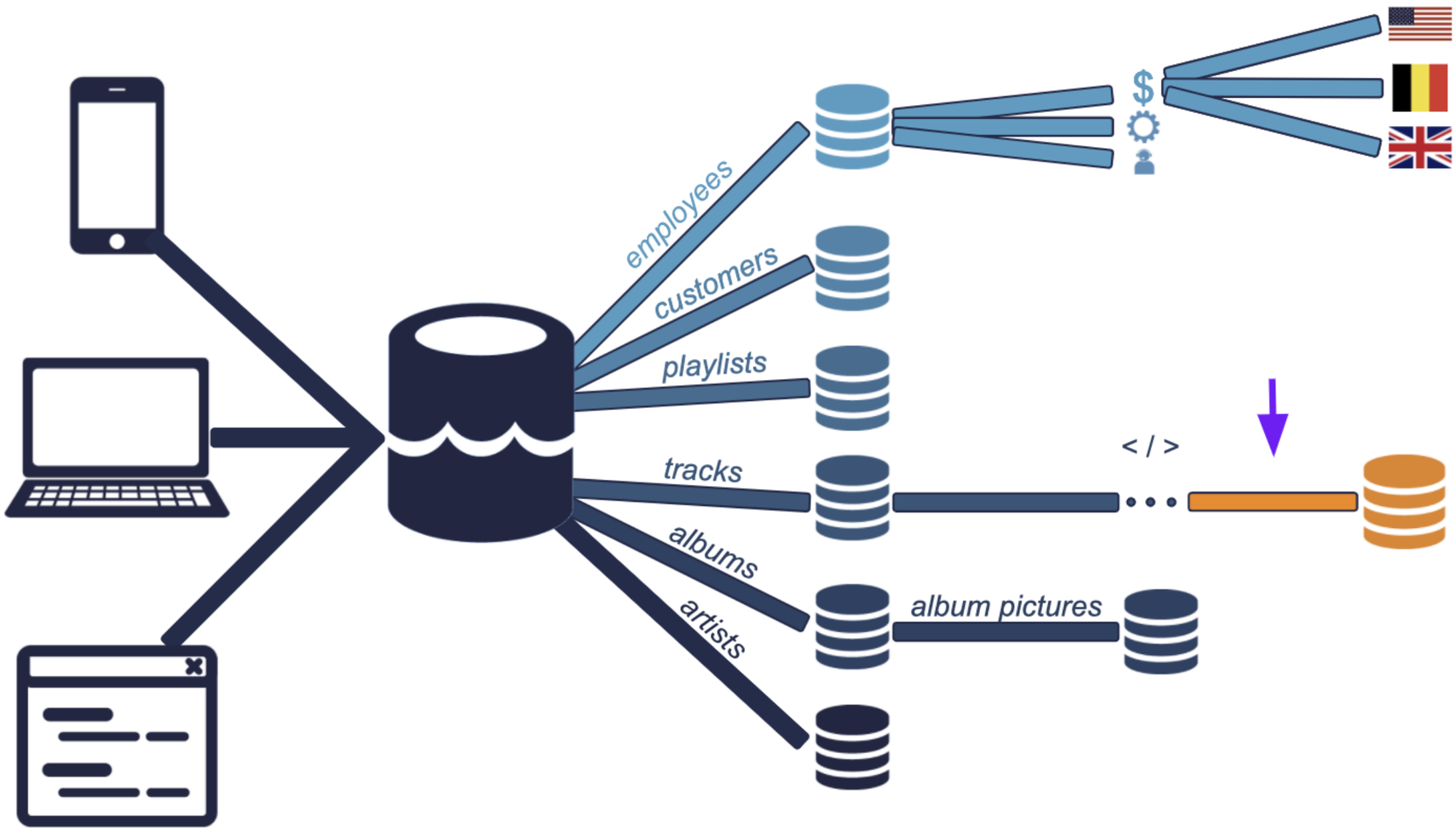

employees
customers
playlists
tracks
albums
artists
album pictures


employees
customers
playlists
tracks
albums
artists
album pictures


employees
customers
playlists
tracks
albums
artists
< / >
album pictures


employees
customers
playlists
tracks
albums
artists
$
< / >
album pictures


employees
customers
playlists
tracks
albums
artists
< / >
album pictures


employees
customers
playlists
tracks
albums
artists
< / >
album pictures


2004
You get a pipeline !


2004
You get a pipeline !


Everybody gets a pipeline !!!
2004

Data pipelines ensure an efficient flow of the data

Automate

- Extracting
- Transforming
- Combining
- Validating
- Loading

Reduce

- Human intervention
- Errors
- Time it takes data to flow

ETL and data pipelines

ETL

- Popular framework for designing data pipelines
- 1) **Extract** data
- 2) **Transform** extracted data
- 3) **Load** transformed data to another database

Data pipelines

- Move data from one system to another
- May follow ETL
- Data may not be transformed
- Data may be directly loaded in applications

Summary

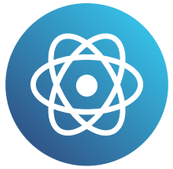
- What a data pipeline is
- What it does
- Why it's important
- How data pipelines are implemented at Spotflix
- What ETL is and its nuances

Let's practice!

DATA ENGINEERING FOR EVERYONE

Data structures

DATA ENGINEERING FOR EVERYONE



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

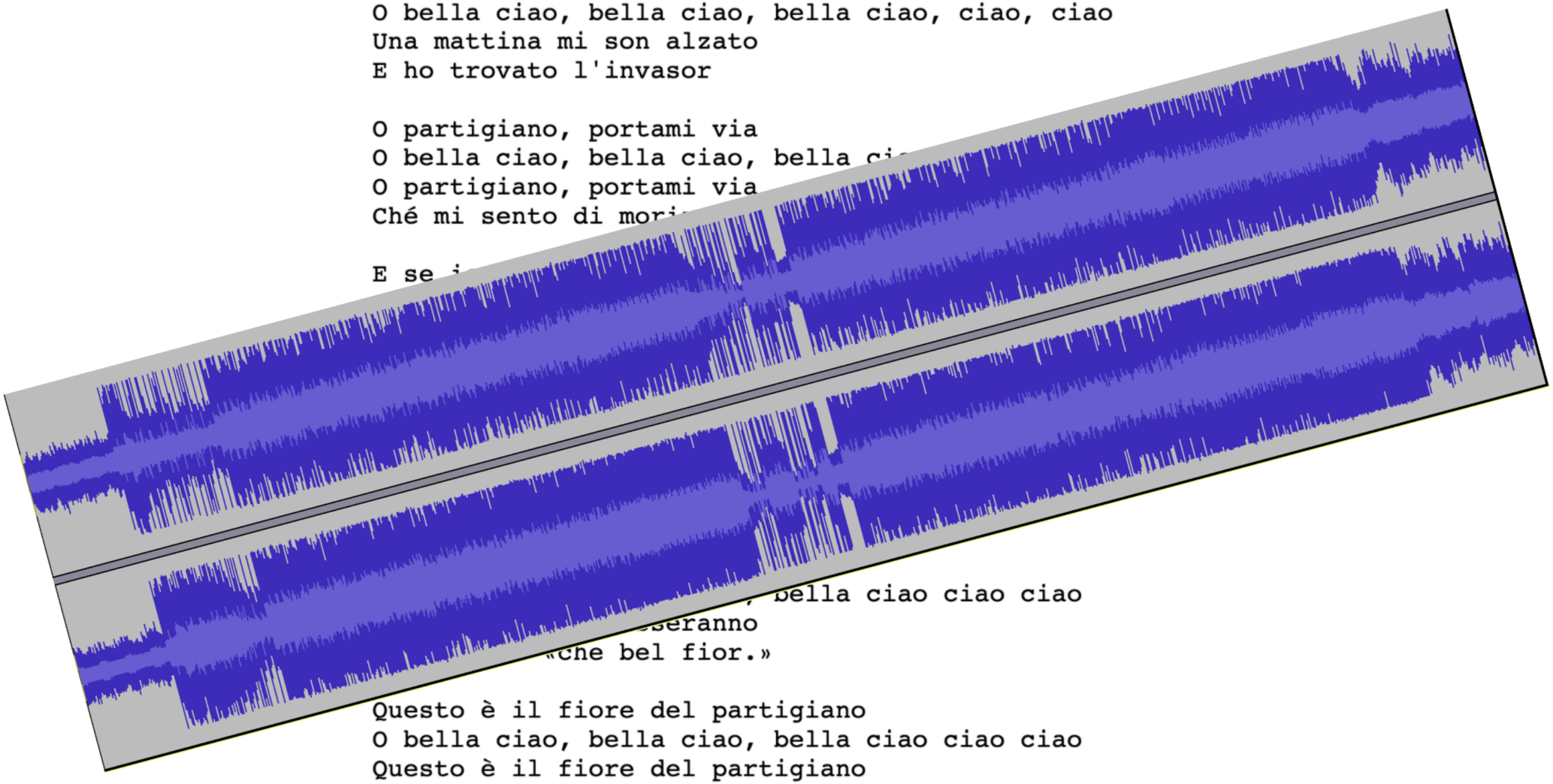
- Does not follow a model, can't be contained in rows and columns
- Difficult to search and organize
- Usually text, sound, pictures or videos
- Usually stored in data lakes, can appear in data warehouses or databases
- Most of the data is unstructured
- Can be extremely valuable

```
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 ci
O partigiano, portami via
Ché mi sento di mori

E se i



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 matti alzato
E ho tro
O parti
O bella
O part
Ché mi
E se
Questo è il fiore del
O bella ciao, bella ciao, be
Questo è il fiore del partigiano
Morto per la libertà
roolsindeed
Meander

Una mattina mi son alzato
O bella ciao, bella ciao, bella ciao, ciao, ciao
Una matti[illegible]n alzato
E ho tro[illegible]

[illegible]ciao, bella ciao, be[illegible]
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

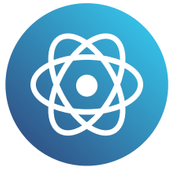
- Structured data
- Semi-structured data
- Unstructured data
- Differences between the three
- Give examples

Let's practice!

DATA ENGINEERING FOR EVERYONE

SQL databases

DATA ENGINEERING FOR EVERYONE



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%'
```

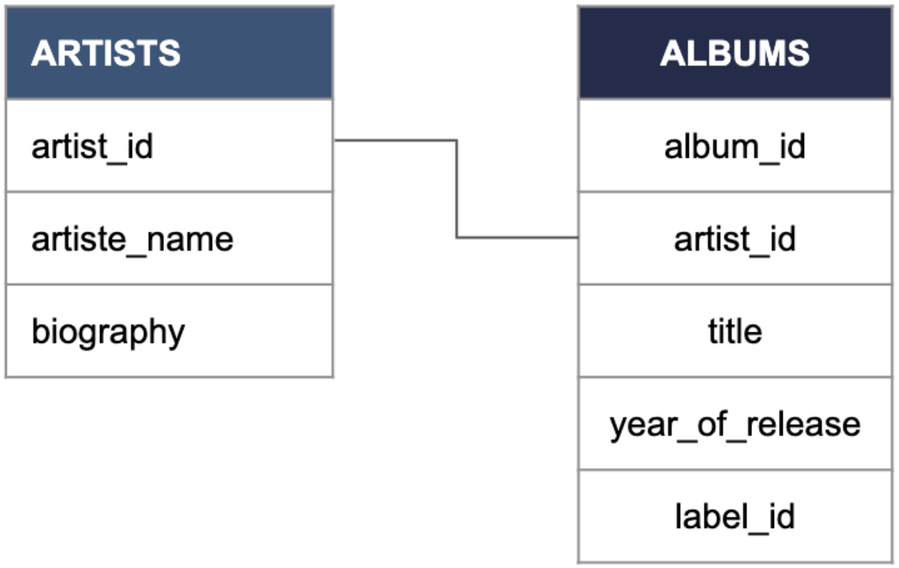
Database schema

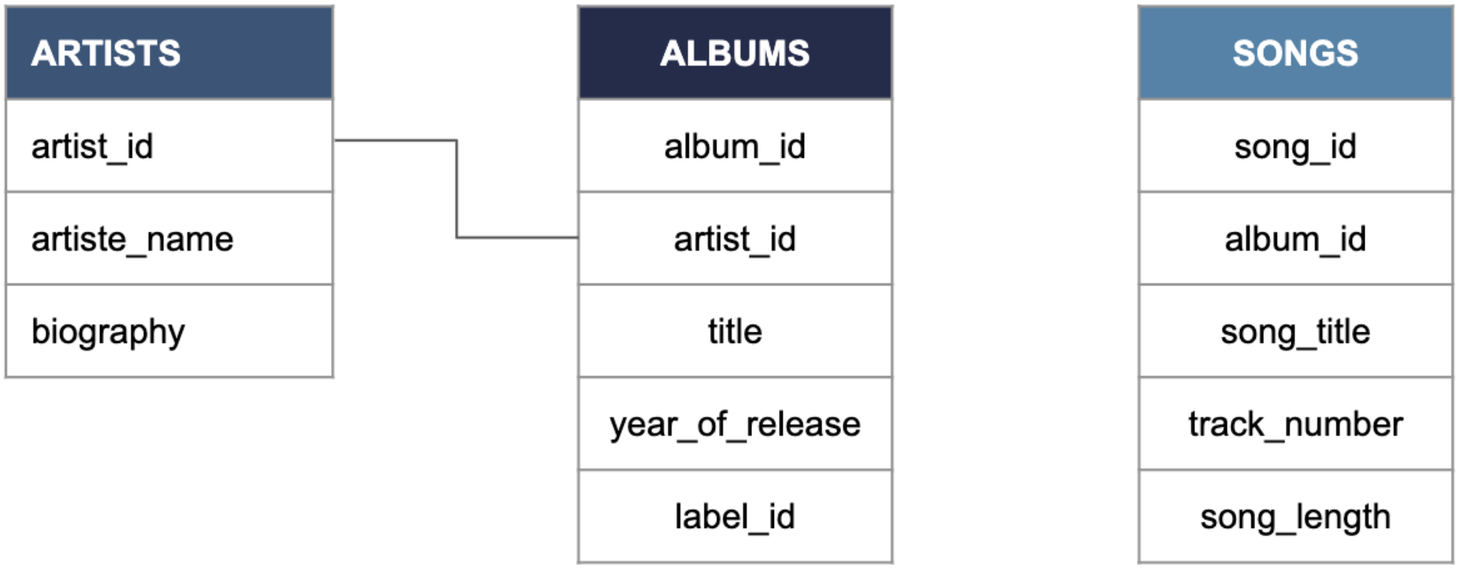
- Databases are made of tables
- The database schema governs how tables are related

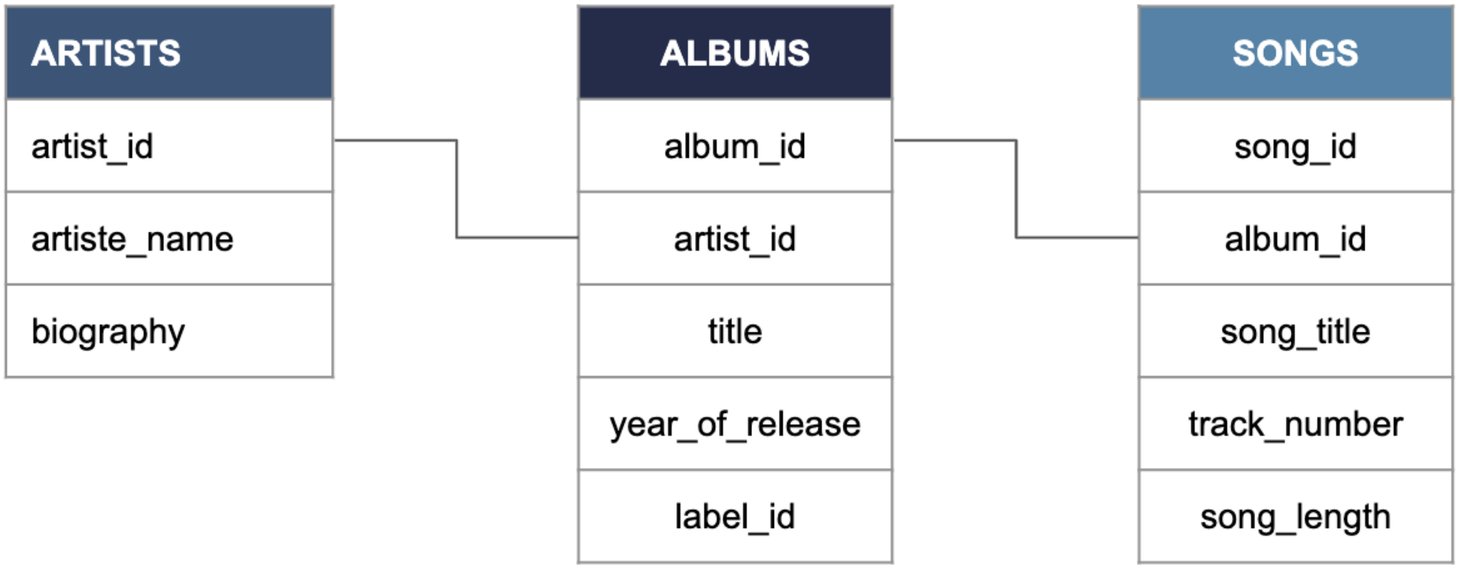
ALBUMS
album_id
artist_id
title
year_of_release
label_id

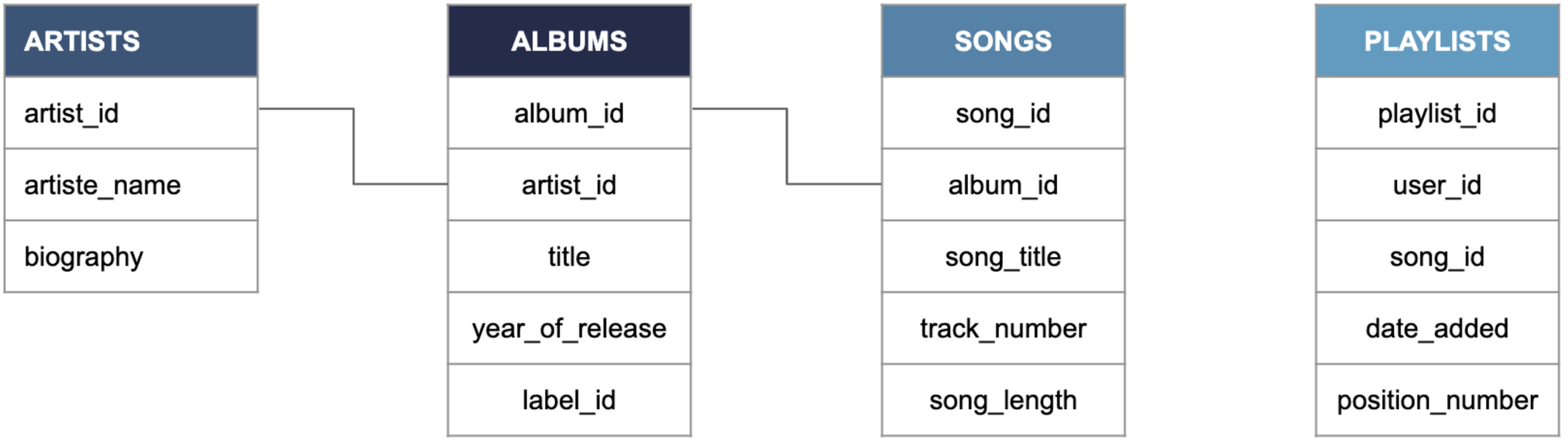
ARTISTS
artist_id
artiste_name
biography

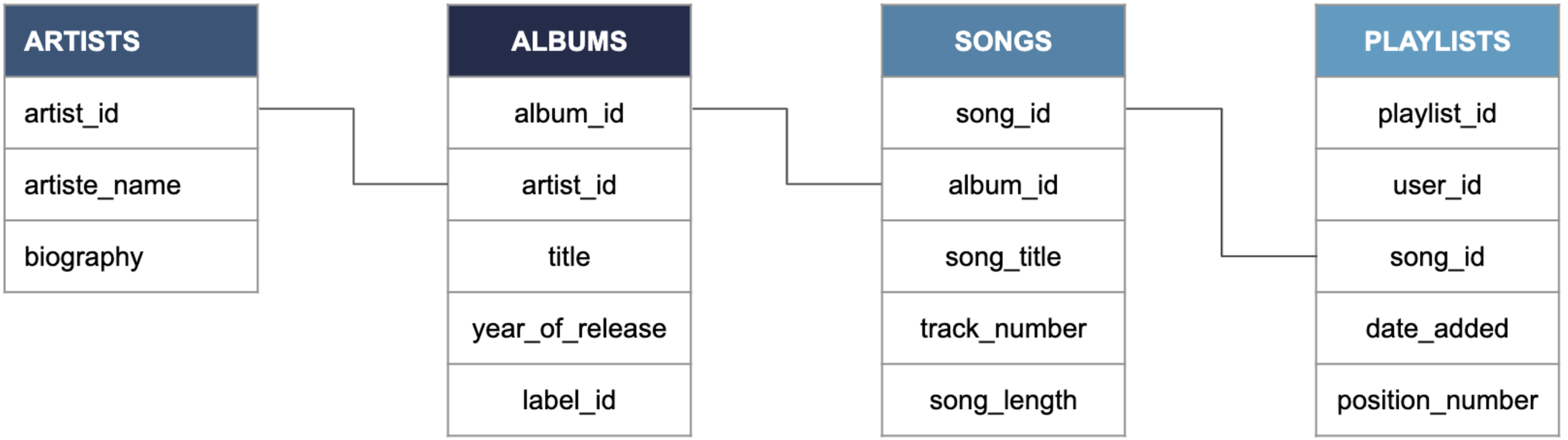
ALBUMS
album_id
artist_id
title
year_of_release
label_id


ARTISTS
artist_id
artiste_name
biography
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
SONGS
song_id
album_id
song_title
track_number
song_length


ARTISTS
artist_id
artiste_name
biography
ALBUMS
album_id
artist_id
title
year_of_release
label_id
SONGS
song_id
album_id
song_title
track_number
song_length


ARTISTS
artist_id
artiste_name
biography
ALBUMS
album_id
artist_id
title
year_of_release
label_id
SONGS
song_id
album_id
song_title
track_number
song_length
PLAYLISTS
playlist_id
user_id
song_id
date_added
position_number


ARTISTS
artist_id
artiste_name
biography
ALBUMS
album_id
artist_id
title
year_of_release
label_id
SONGS
song_id
album_id
song_title
track_number
song_length
PLAYLISTS
playlist_id
user_id
song_id
date_added
position_number

Several implementations

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

Summary

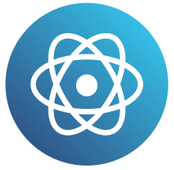
- SQL = industry standard
- Explain how Data engineers and Data scientists use it differently
- Database schema
- SQL implementations

Let's practice!

DATA ENGINEERING FOR EVERYONE

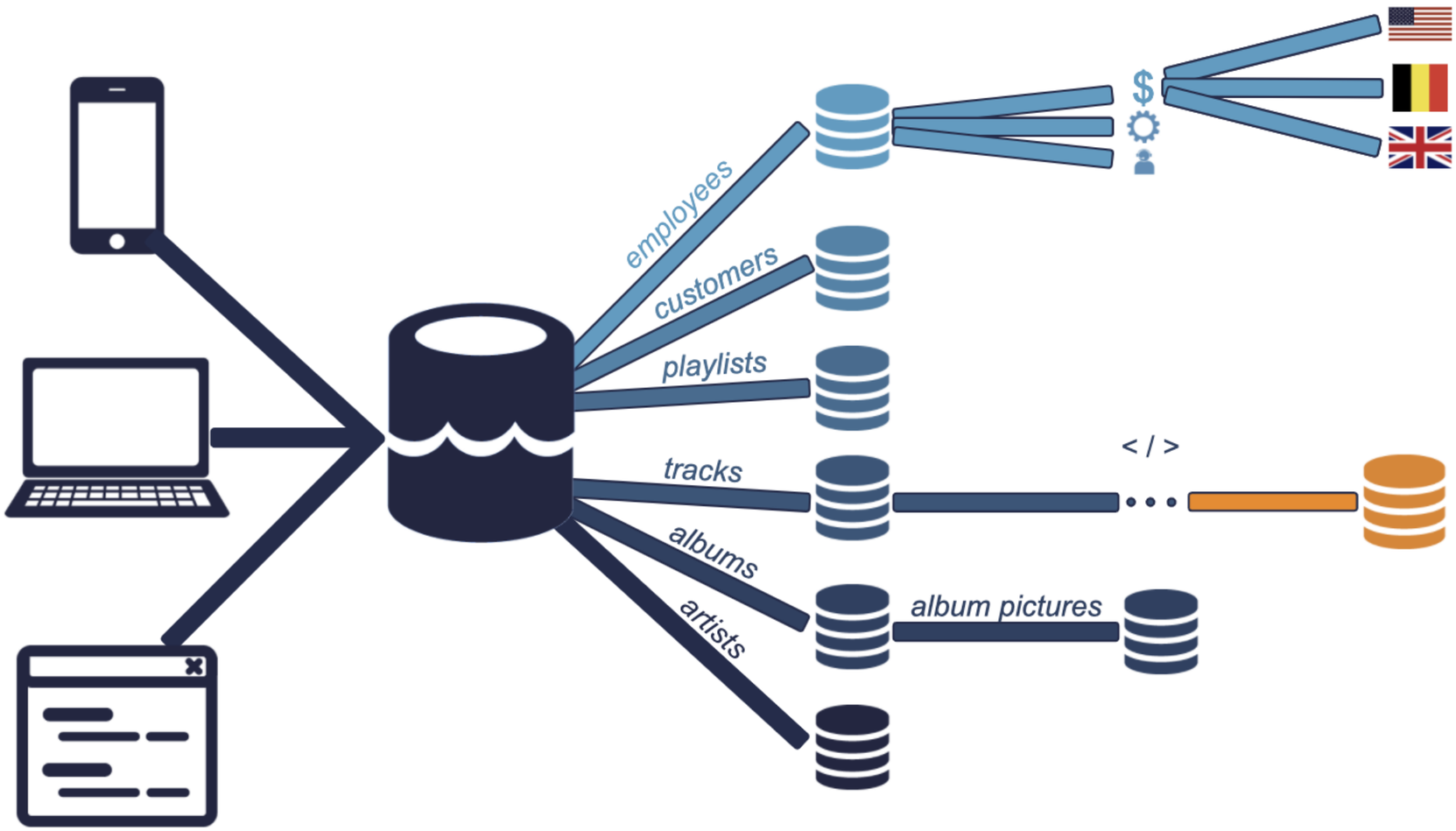
Data warehouses and data lakes

DATA ENGINEERING FOR EVERYONE



Hadrien Lacroix
Content Developer

Warehouses with stunning view on the lake


employees
customers
playlists
tracks
albums
artists
< / >
album pictures

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

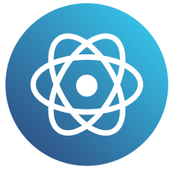
- Data lakes
- Data warehouses
- Databases
- Data catalog

Let's practice!

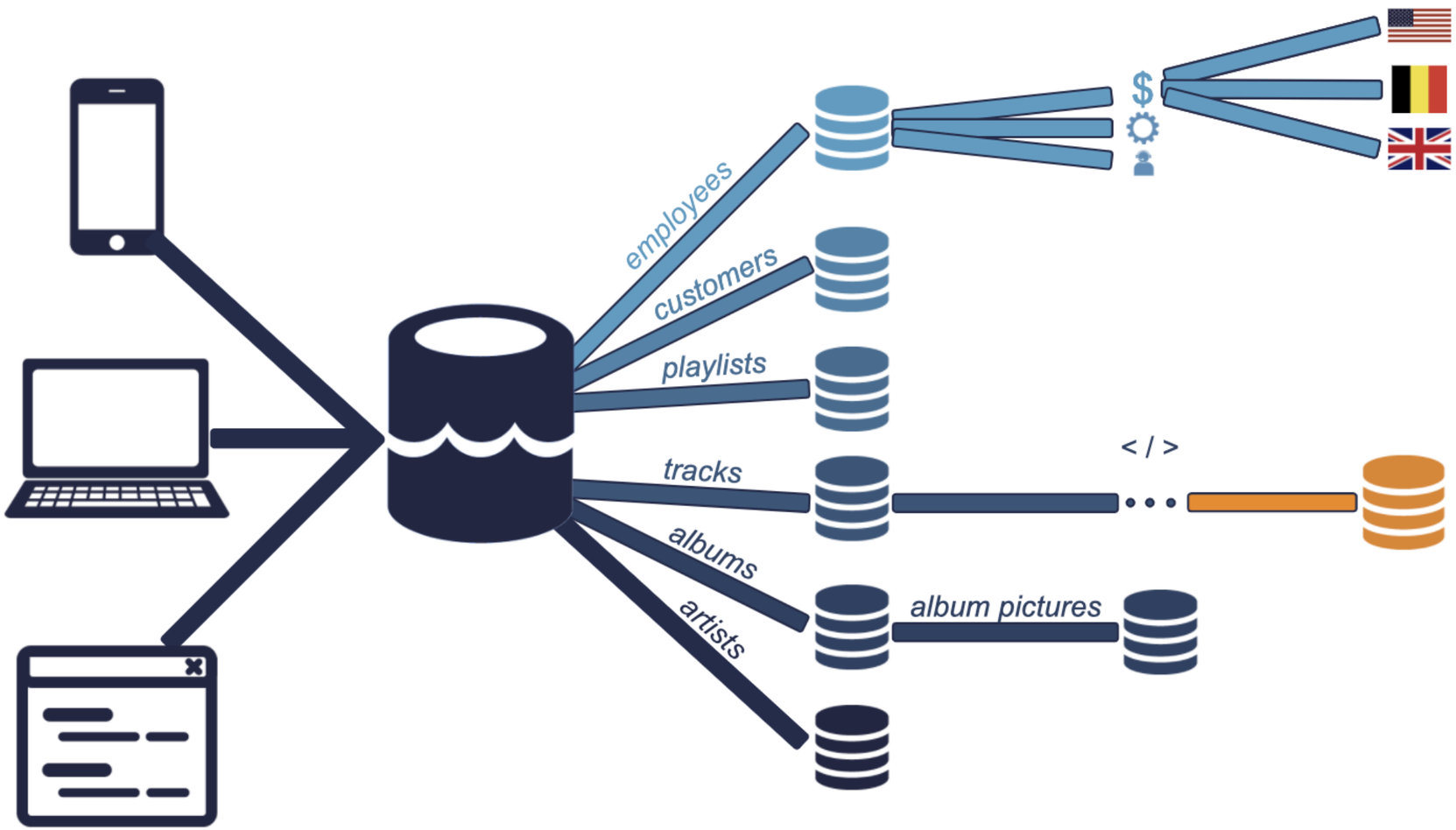
DATA ENGINEERING FOR EVERYONE

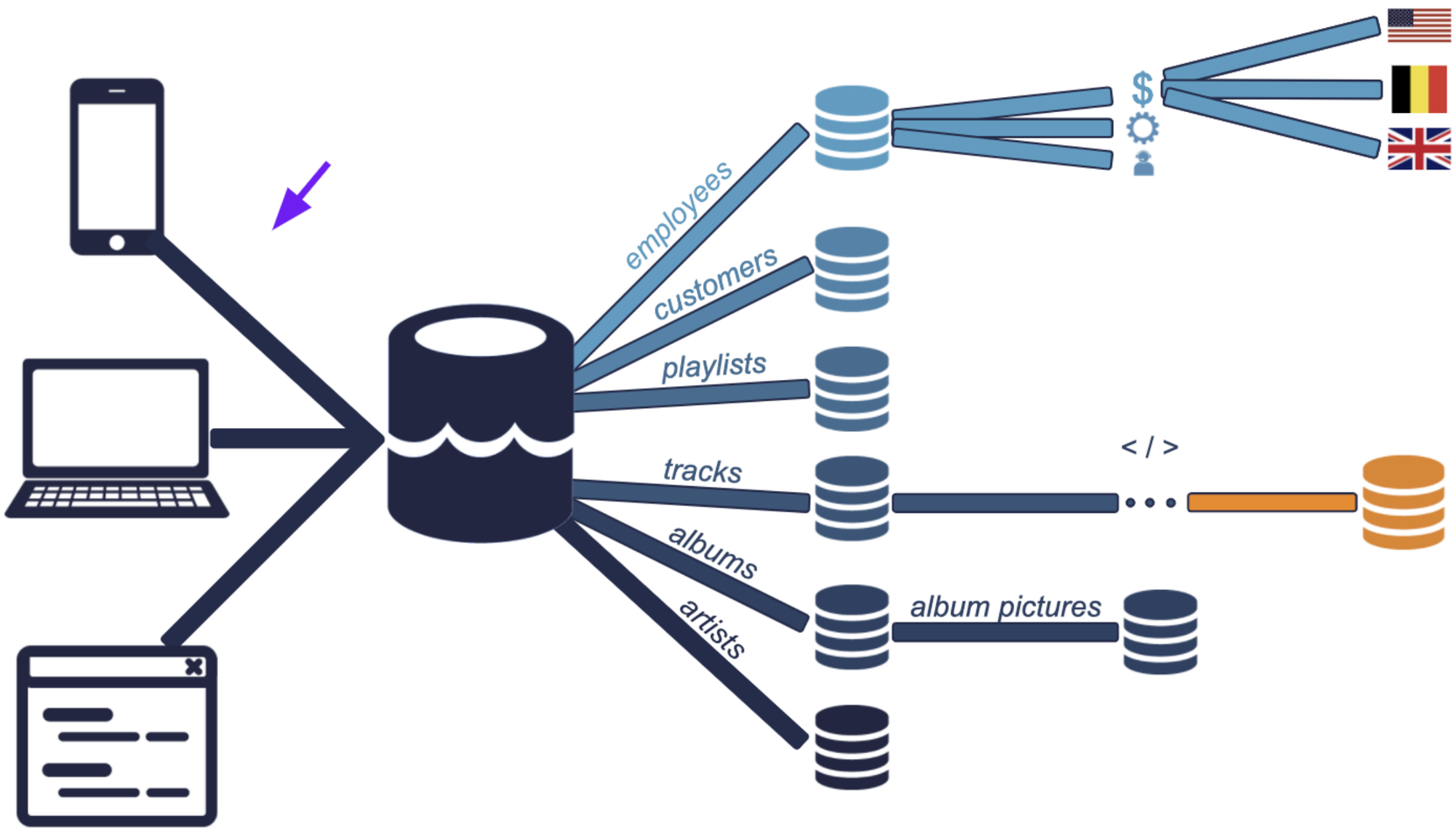
Processing data

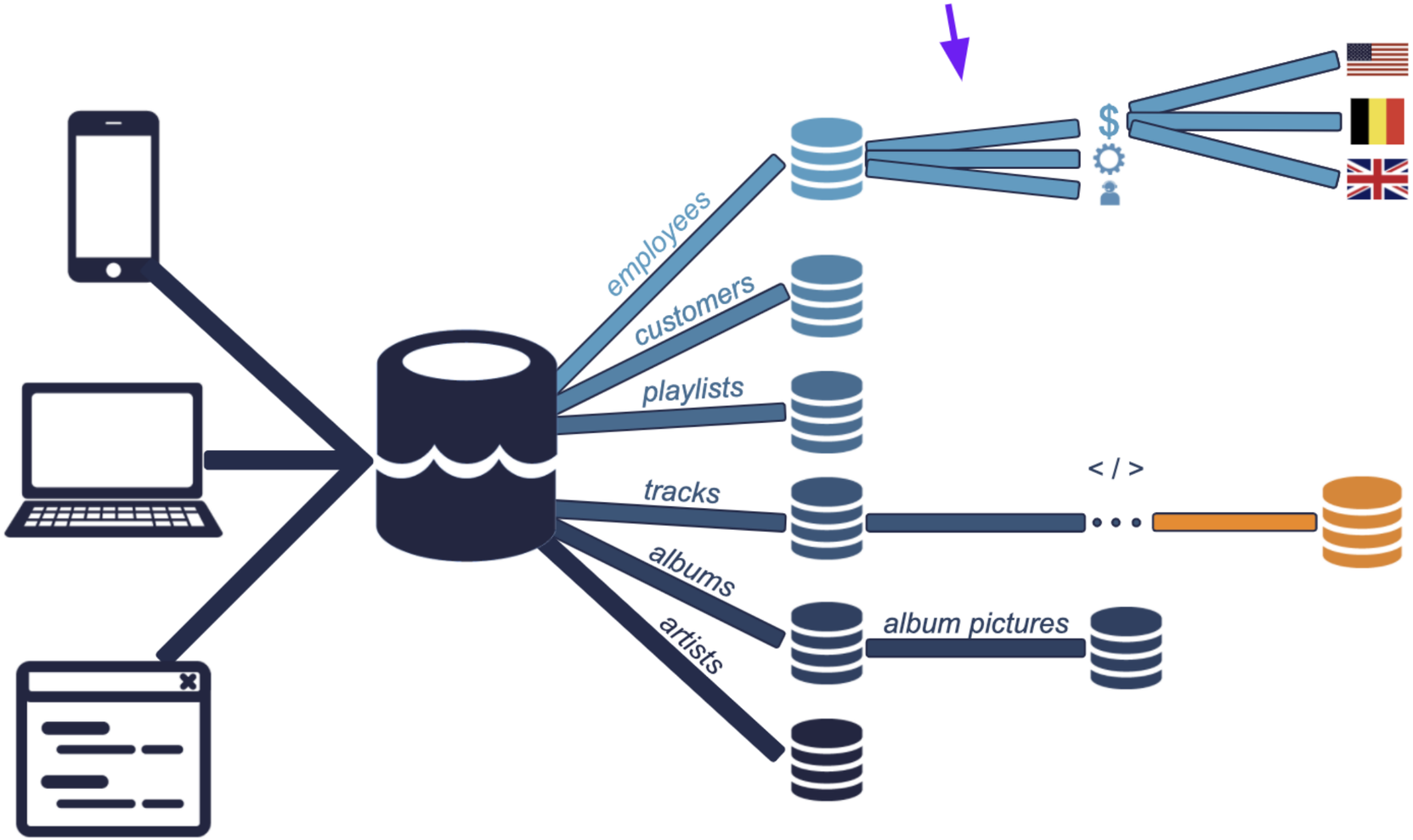
DATA ENGINEERING FOR EVERYONE

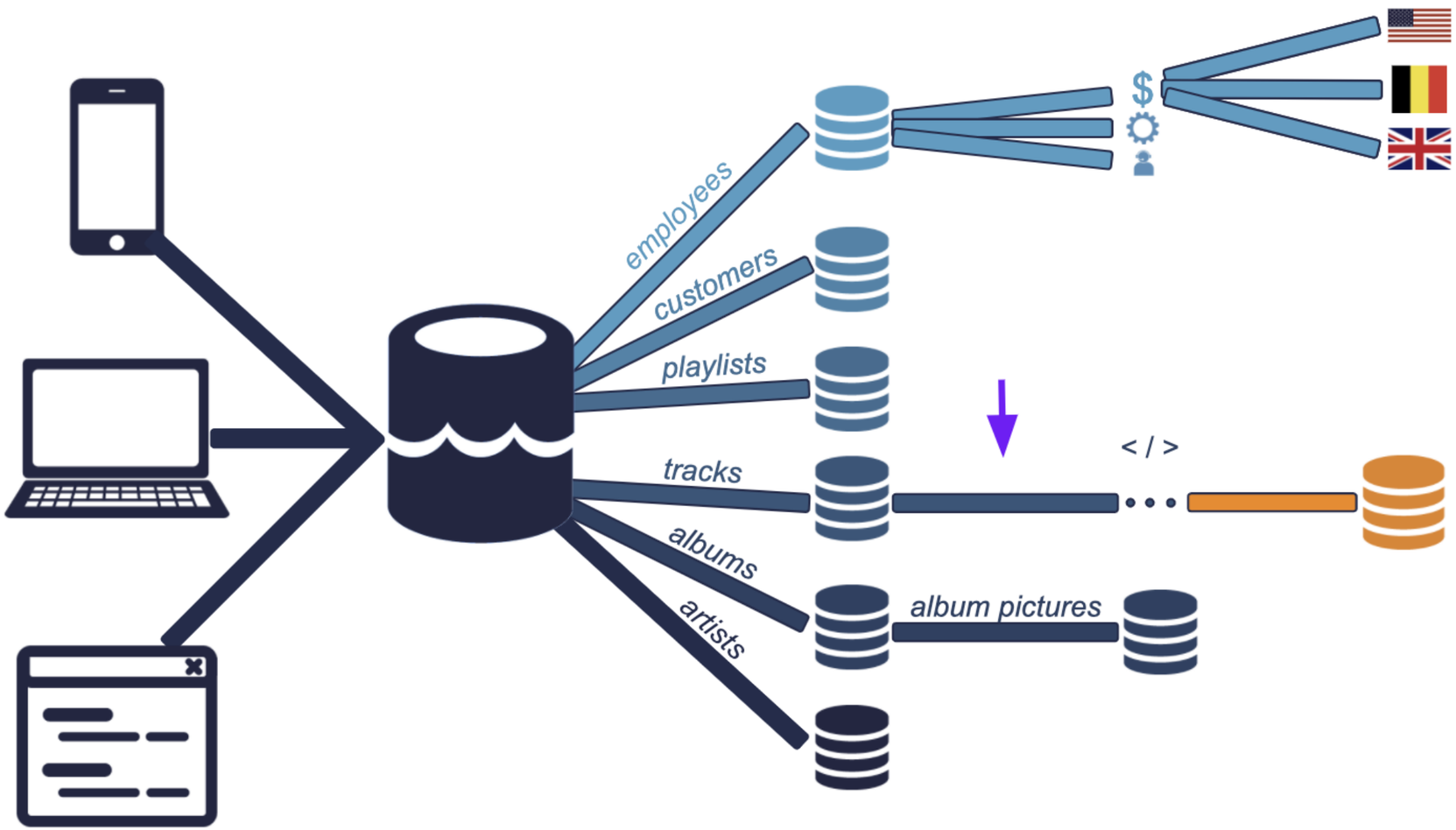


Hadrien Lacroix
Content Developer at DataCamp


employees
customers
playlists
tracks
albums
artists
< / >
album pictures


employees
customers
playlists
tracks
albums
artists
< / >
album pictures


employees
customers
playlists
tracks
albums
artists
< / >
album pictures


employees
customers
playlists
tracks
albums
artists
< / >
album pictures

A general definition

- Data processing: converting **raw** data into **meaningful** information

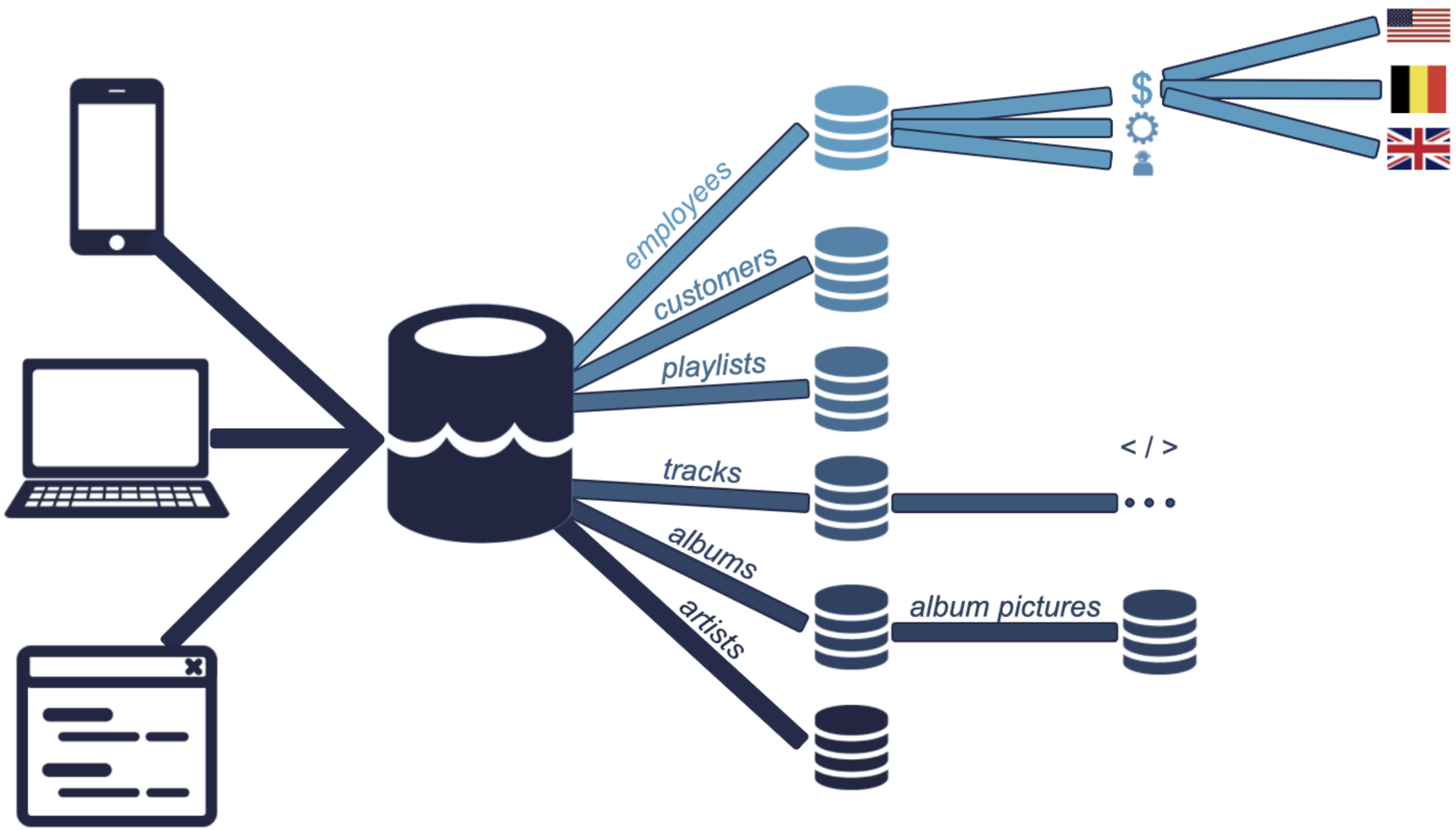
Data processing value

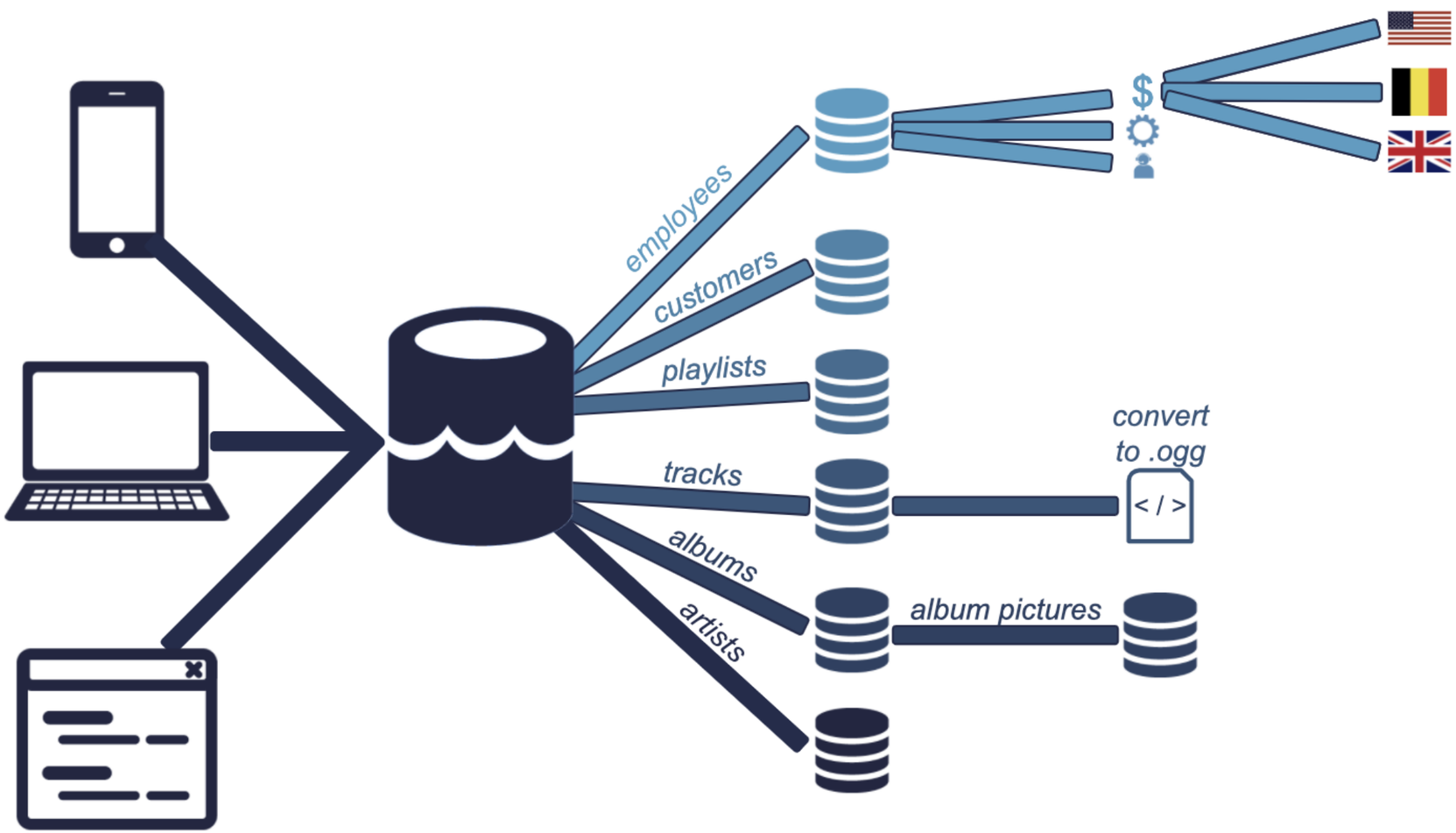
Conceptually

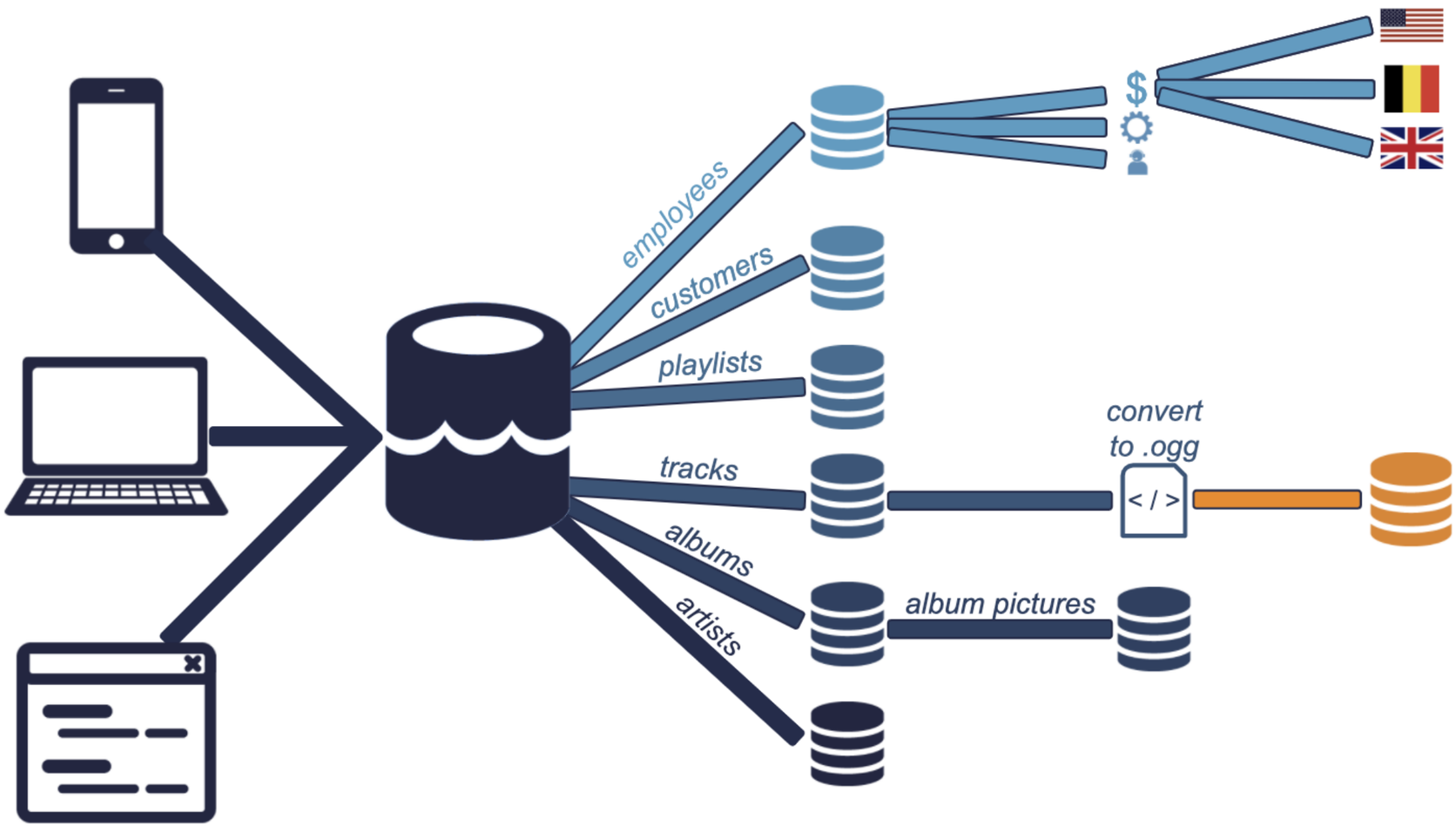
- Remove unwanted data
- Optimize memory, process and network costs
- Convert data from one type to another

At Spotflix

- No long term need for testing feature data
- Can't afford to store and stream files this big


employees
customers
playlists
tracks
albums
artists
< / >
album pictures


employees
customers
playlists
tracks
albums
artists
convert
to .ogg
< / >
album pictures


employees
customers
playlists
tracks
albums
artists
convert
to .ogg
< / >
album pictures

Data processing value

Conceptually

- Remove unwanted data
- To save memory
- Convert data from one type to another
- Organize data
- To fit into a schema/structure
- Increase productivity

At Spotflix

- No need for lossless format
- Can't afford to store files this big
- Convert songs from `.flac` to `.ogg`
- Reorganize data from the data lake to data warehouses
- Employee table example
- Enable data scientists

How data engineers process data

- Data manipulation, cleaning, and tidying tasks
  - that can be automated
  - that will always need to be done
- Store data in a sanely structured database
- Create views on top of the database tables
- Optimizing the performance of the database
- Rejecting corrupt song files
- Deciding what happens with missing metadata
- Separate artists and albums tables...
- ...but provide view combining them
- Indexing

Batch processing

Stream processing

[1] The difference between batch and stream will be explained in the next lesson!


APACHE
Spark
TM

Summary

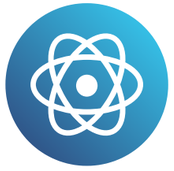
- What data processing is
- Why it's necessary
- What it consists in
- How we process data at Spotflix

Let's practice!

DATA ENGINEERING FOR EVERYONE

Scheduling data

DATA ENGINEERING FOR EVERYONE



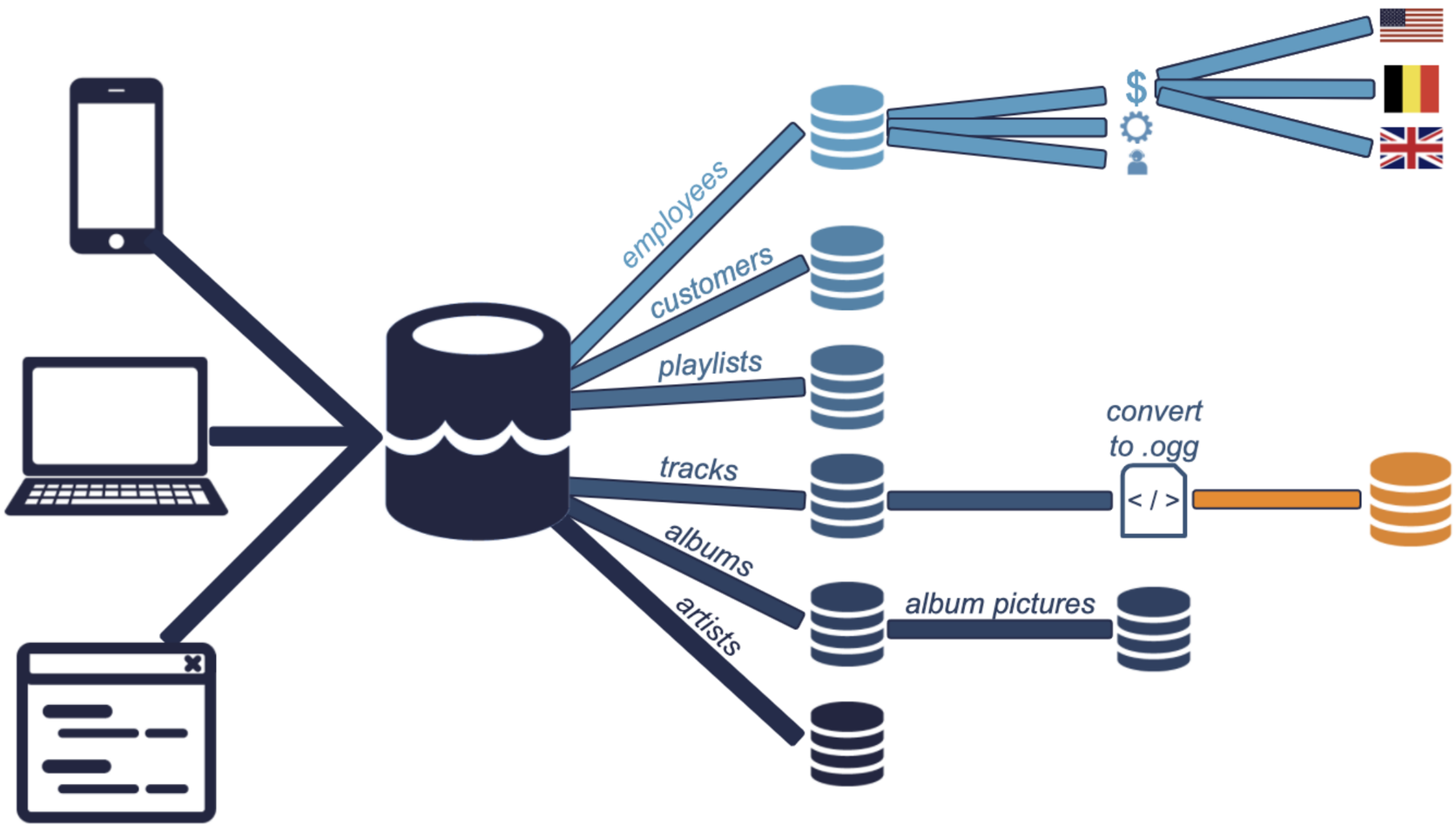
Hadrien Lacroix
Content Developer at DataCamp

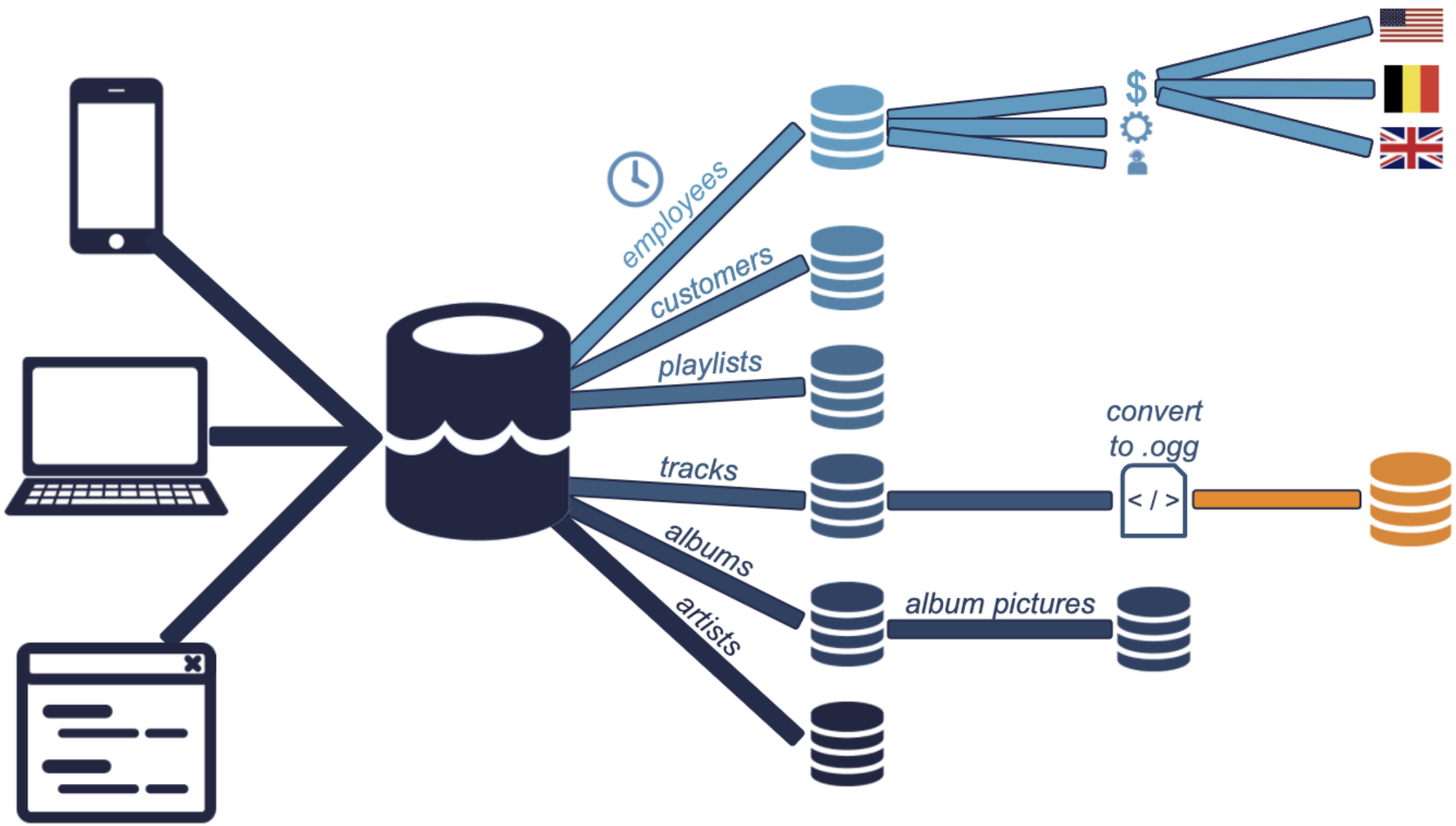
Scheduling

- Can apply to any task listed in data processing
- Scheduling is the glue of your system
- Holds each piece and organize how they work together
- Runs tasks in a specific order and resolves all dependencies

Manual, time and sensor scheduling

- Manually
- Manually update the employee table

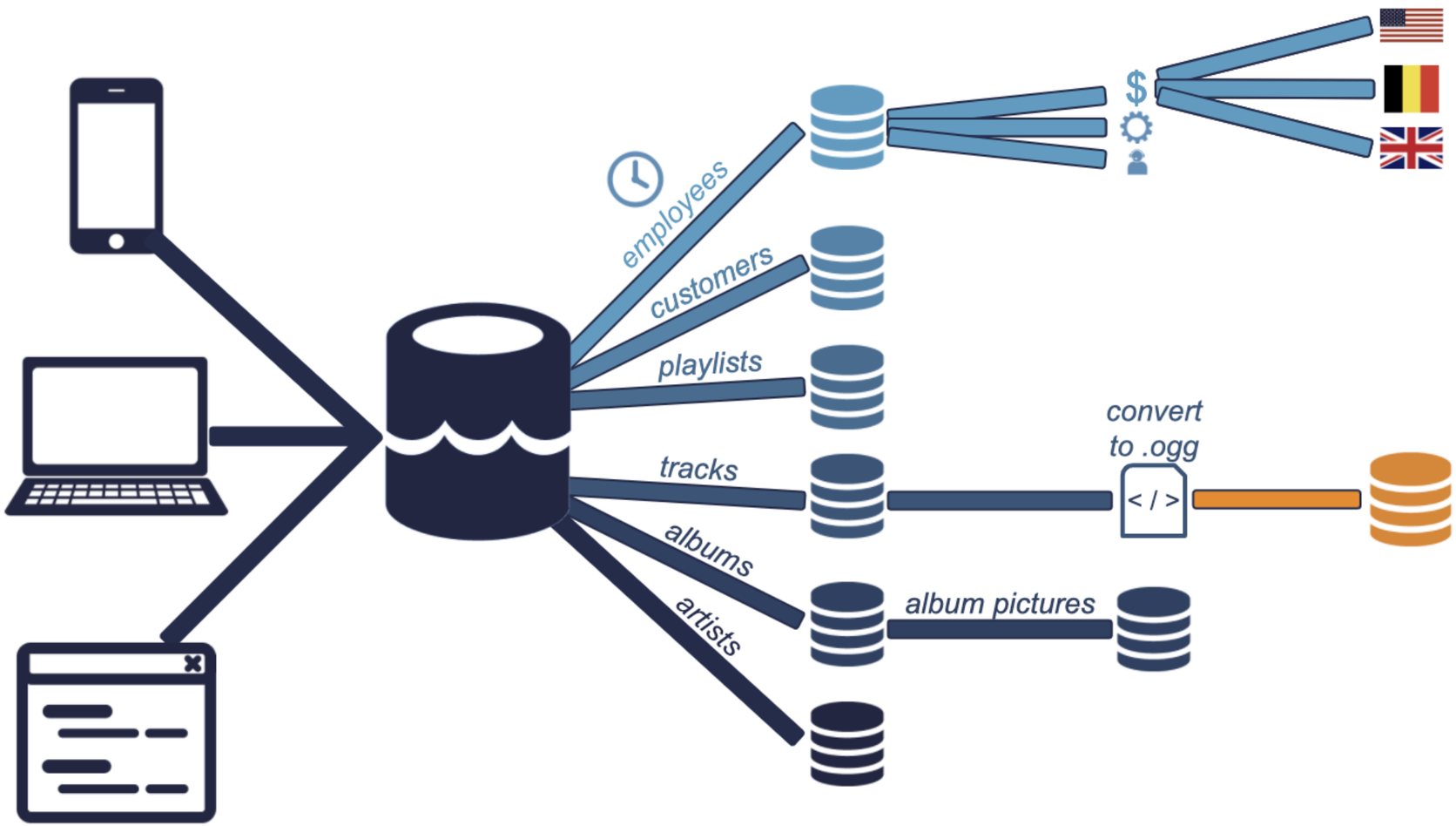

employees
customers
playlists
tracks
albums
artists
convert
to .ogg
< / >
album pictures

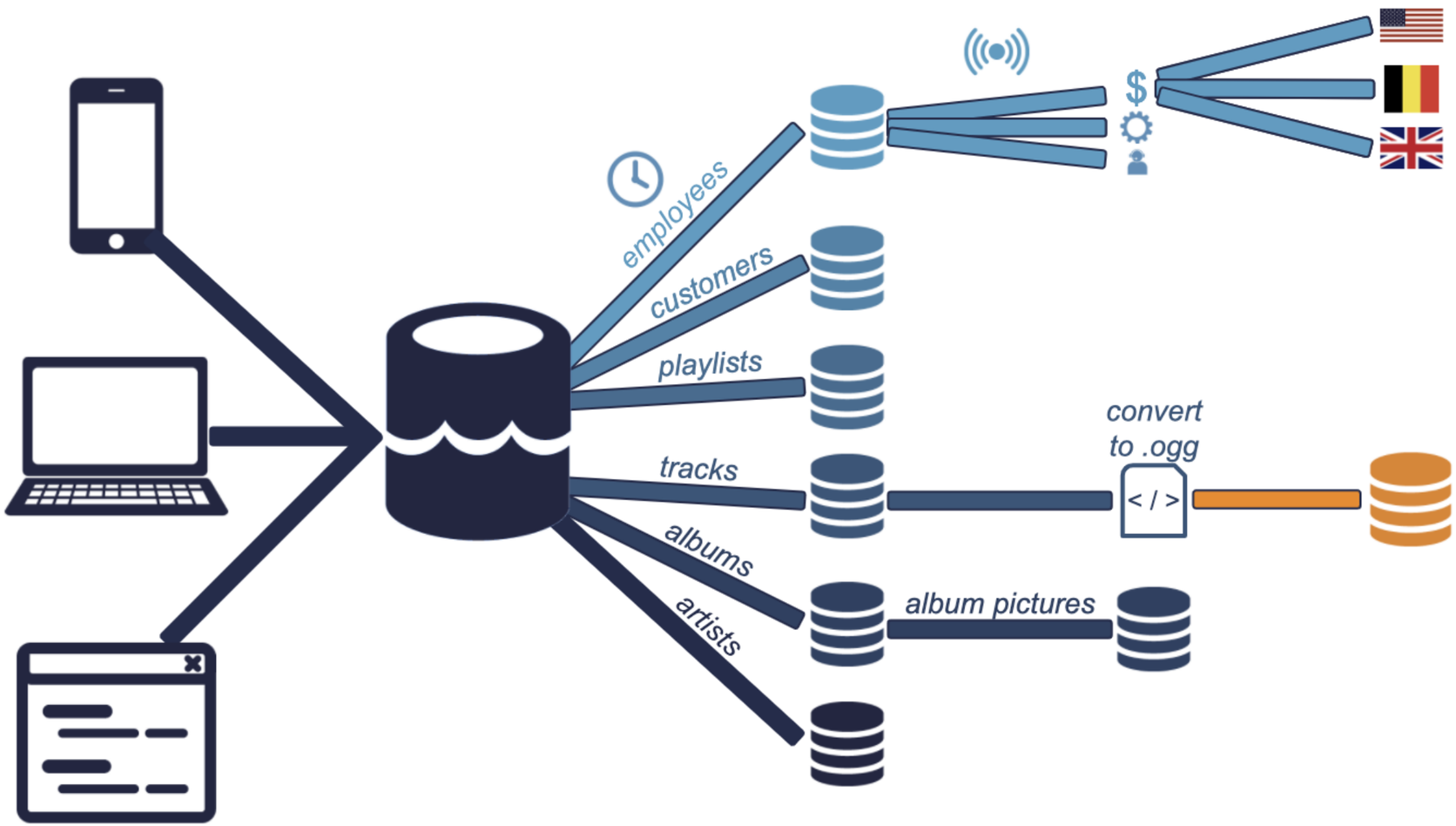

employees
customers
playlists
tracks
albums
artists
convert
to .ogg
< / >
album pictures

Manual, time and sensor scheduling

- Manually
- Automatically run at a specific time
- Automatically run if a specific condition is met
  - Sensor scheduling

- Manually update the employee table
- Update the employee table at 6 AM


employees
customers
playlists
tracks
albums
artists
convert
to .ogg
< / >
album pictures


employees
customers
playlists
tracks
albums
artists
convert
to .ogg
< / >
album pictures

Manual, time, and sensor scheduling

- Manually
- Automatically run at a specific time
- Automatically run if a specific condition is met
  - Sensor scheduling

- Manually update the employee table
- Update the employee table at 6 AM
- Update the department tables if a new employee was added

Batches and streams

- Batches
  - Group records at intervals
  - Often cheaper
- Streams
  - Send individual records right away

- Songs uploaded by artists
- Employee table
- Revenue table
- New users signing in
- Another example: online vs. offline listening

Scheduling tools

Apache Airflow

Summary

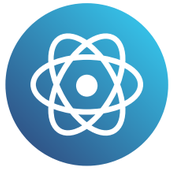
- What scheduling is
- Different ways to set it up
- Difference between batches and streams
- How scheduling is implemented at Spotflix
- Airflow, Luigi

Let's practice!

DATA ENGINEERING FOR EVERYONE

Parallel computing

DATA ENGINEERING FOR EVERYONE



Hadrien Lacroix
Content Developer at DataCamp

Parallel computing

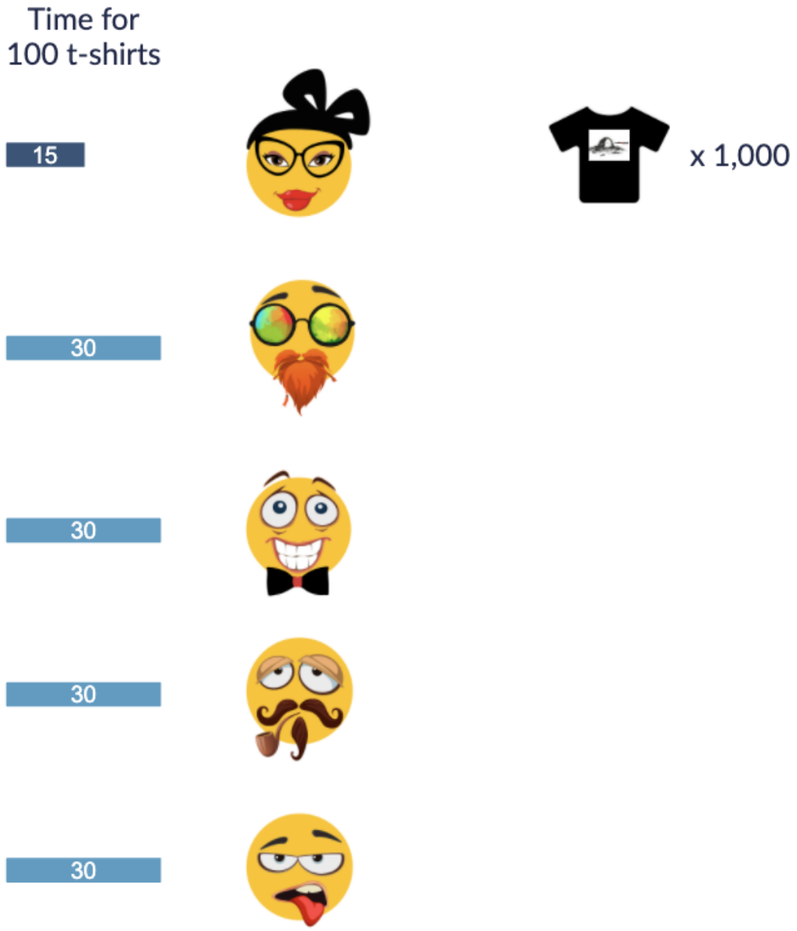
- Basis of modern data processing tools
- Necessary:
  - Mainly because of memory
  - Also for processing power
- How it works:
  - Split tasks up into several smaller subtasks
  - Distribute these subtasks over several computers

x 1,000

Time for
100 t-shirts

15

x 1,000



[1] Emojis by Mohamed Hassan

Time for
100 t-shirts

15

x 1,000

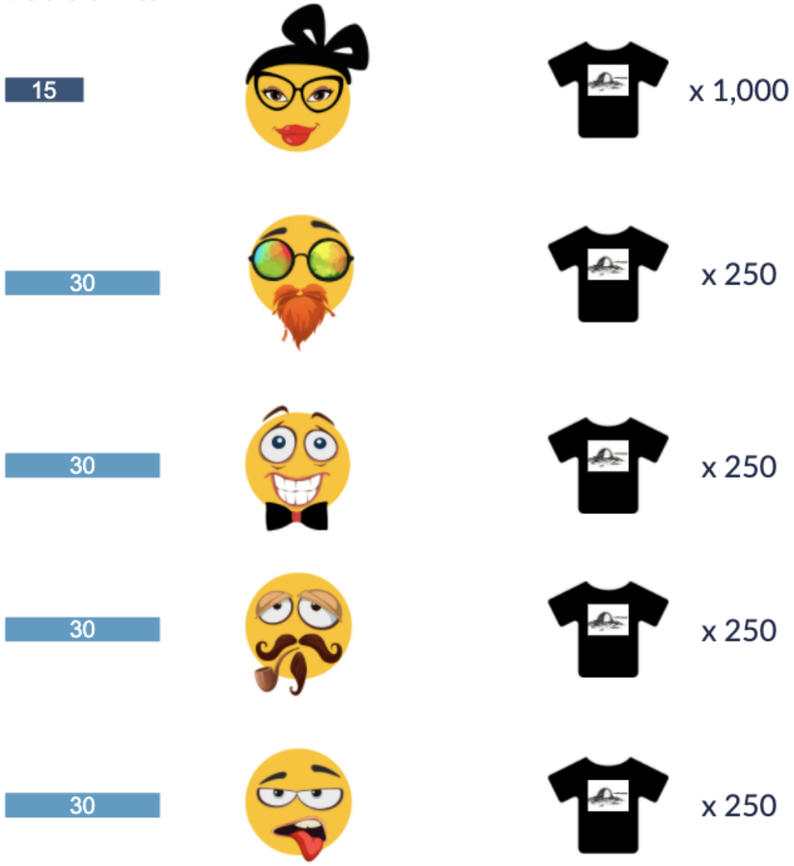
30

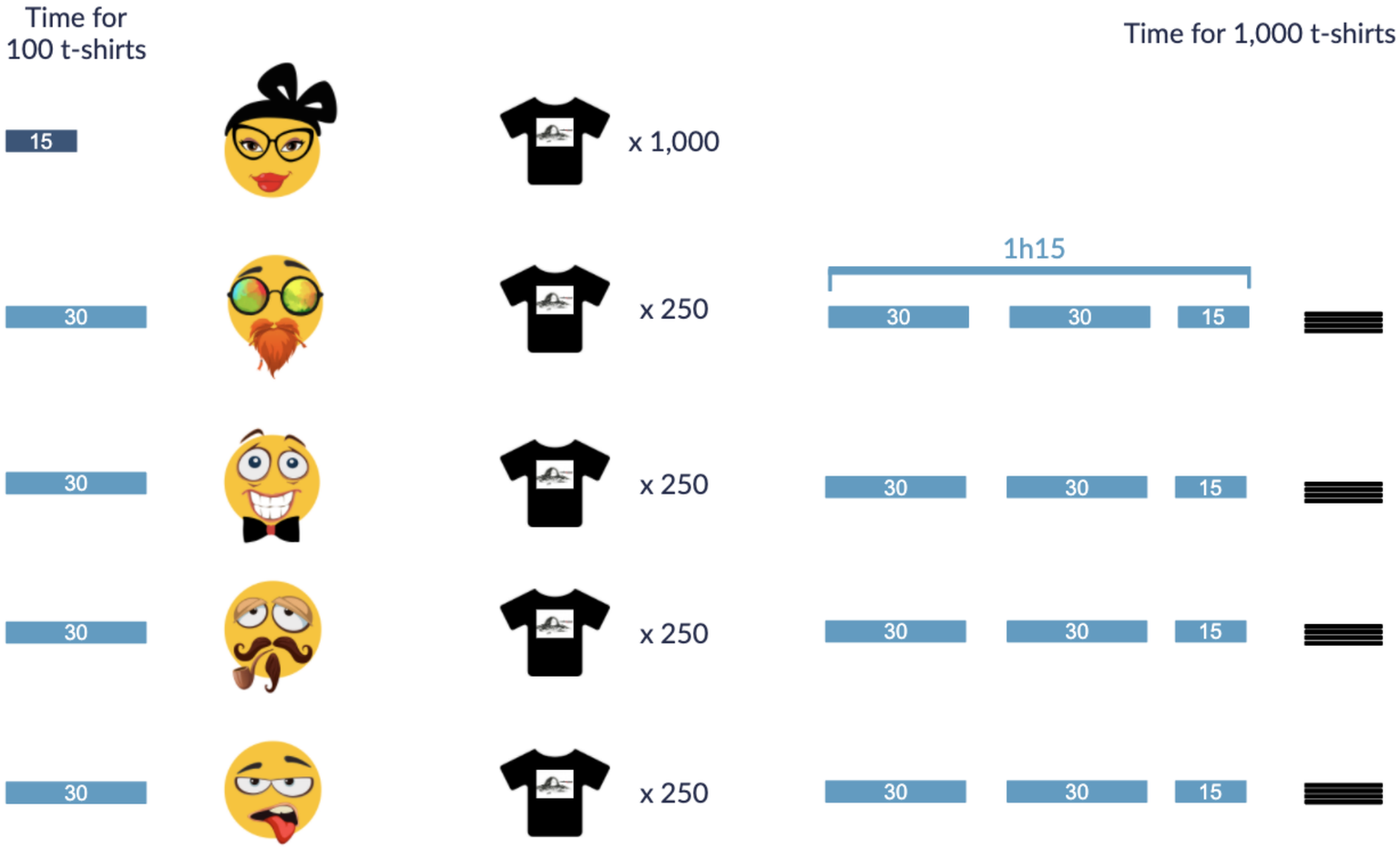
30

30

30

Time for
100 t-shirts




Time for
100 t-shirts
Time for 1,000 t-shirts
15
x 1,000
1h15
30
30
15
x 250
30
30
30
15
x 250
30
30
30
15
x 250
30
30
30
15
x 250


Time for
100 t-shirts
Time for 1,000 t-shirts
2h30
15
x 1,000
15 15 15 15 15 15 15 15 15 15
30
x 250
1h15
30 30 15
30
x 250
30 30 15
30
x 250
30 30 15
30
x 250
30 30 15

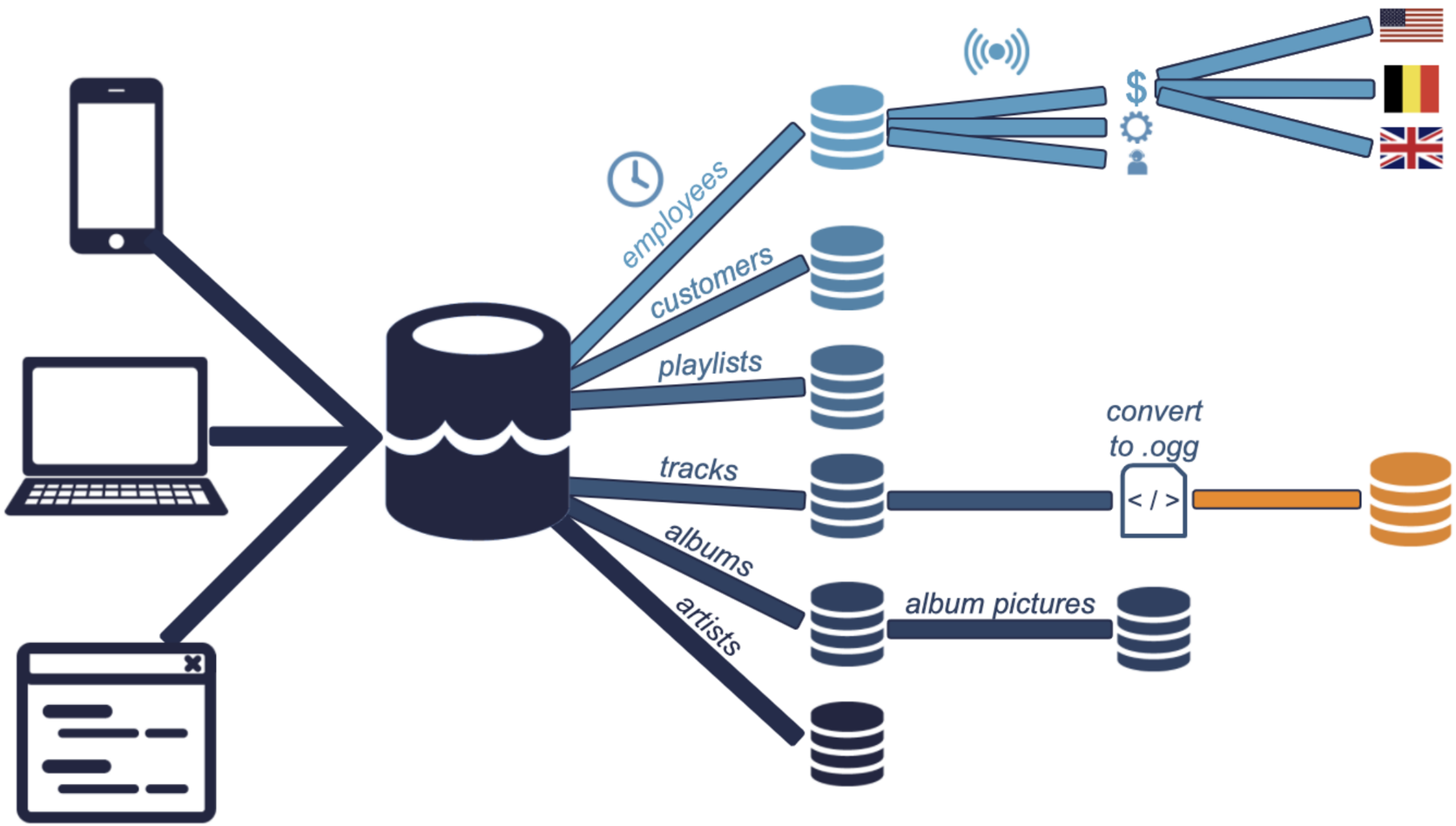
Benefits and risks of parallel computing

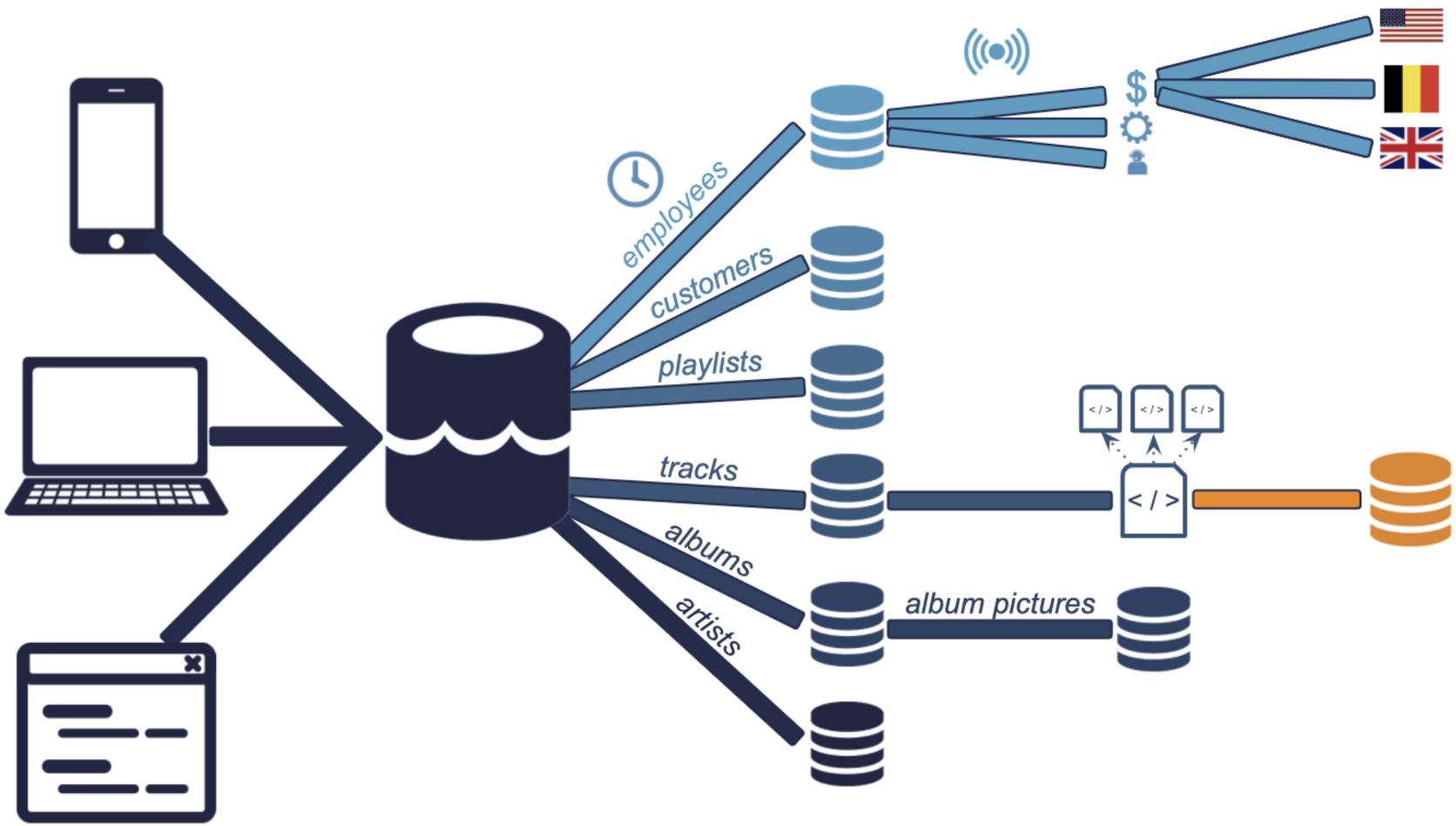
- Employees = processing units
- Advantages
  - Extra processing power
  - Reduced memory footprint
- Disadvantages
  - Moving data incurs a cost
  - Communication time


Time for
100 t-shirts
Time for 1,000 t-shirts
2h30
15
x 1,000
15 15 15 15 15 15 15 15 15 15
30
x 250
1h15
30 30 15
30
x 250
30 30 15
30
x 250
30 30 15
30
x 250
30 30 15


Time for
100 t-shirts
Time for 1,000 t-shirts
2h30
15
x 1,000
0h10
1h15
30
x 250
x 250
x 250
x 250


Time for
100 t-shirts
15
30
30
30
30
x 1,000
x 250
x 250
x 250
x 250
Time for 1,000 t-shirts
2h30
15 15 15 15 15 15 15 15 15 15
0h10
1h15
0h05
30 30 15
30 30 15
30 30 15
30 30 15


employees
customers
playlists
tracks
albums
artists
convert
to .ogg
< / >
album pictures


employees
customers
playlists
tracks
albums
artists
album pictures
< / >
$

Summary

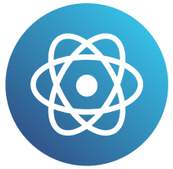
- Benefits and risks
- How it's implemented at Spotflix

Let's practice!

DATA ENGINEERING FOR EVERYONE

Cloud computing

DATA ENGINEERING FOR EVERYONE



Hadrien Lacroix
Content Developer

Cloud computing for data processing

Servers on premises

- Bought
- Need space
- Electrical and maintenance cost
- Enough power for peak moments
- Processing power unused at quieter times

Servers on the cloud

- Rented
- Don't need space
- Use just the resources we need
- When we need them
- The closer to the user the better

Cloud computing for data storage

- Database reliability: data replication
- Risk with sensitive data

32.4%

32.4%

17.6%

32.4%

17.6%

6%

File storage

File storage

AWS S3

File storage

AWS S3

Azure Blob Storage

File storage

AWS S3

Azure Blob Storage

Google Cloud Storage

	File storage	Computation
aws	*AWS S3* 	
Microsoft Azure	*Azure Blob Storage* 	
Google Cloud	*Google Cloud Storage* 	

	File storage	Computation
aws	*AWS S3* 	*AWS EC2* 
Microsoft Azure	*Azure Blob Storage* 	
Google Cloud	*Google Cloud Storage* 	

	File storage	Computation
aws	*AWS S3* 	*AWS EC2* 
Microsoft Azure	*Azure Blob Storage* 	*Azure Virtual Machines* 
Google Cloud	*Google Cloud Storage* 	

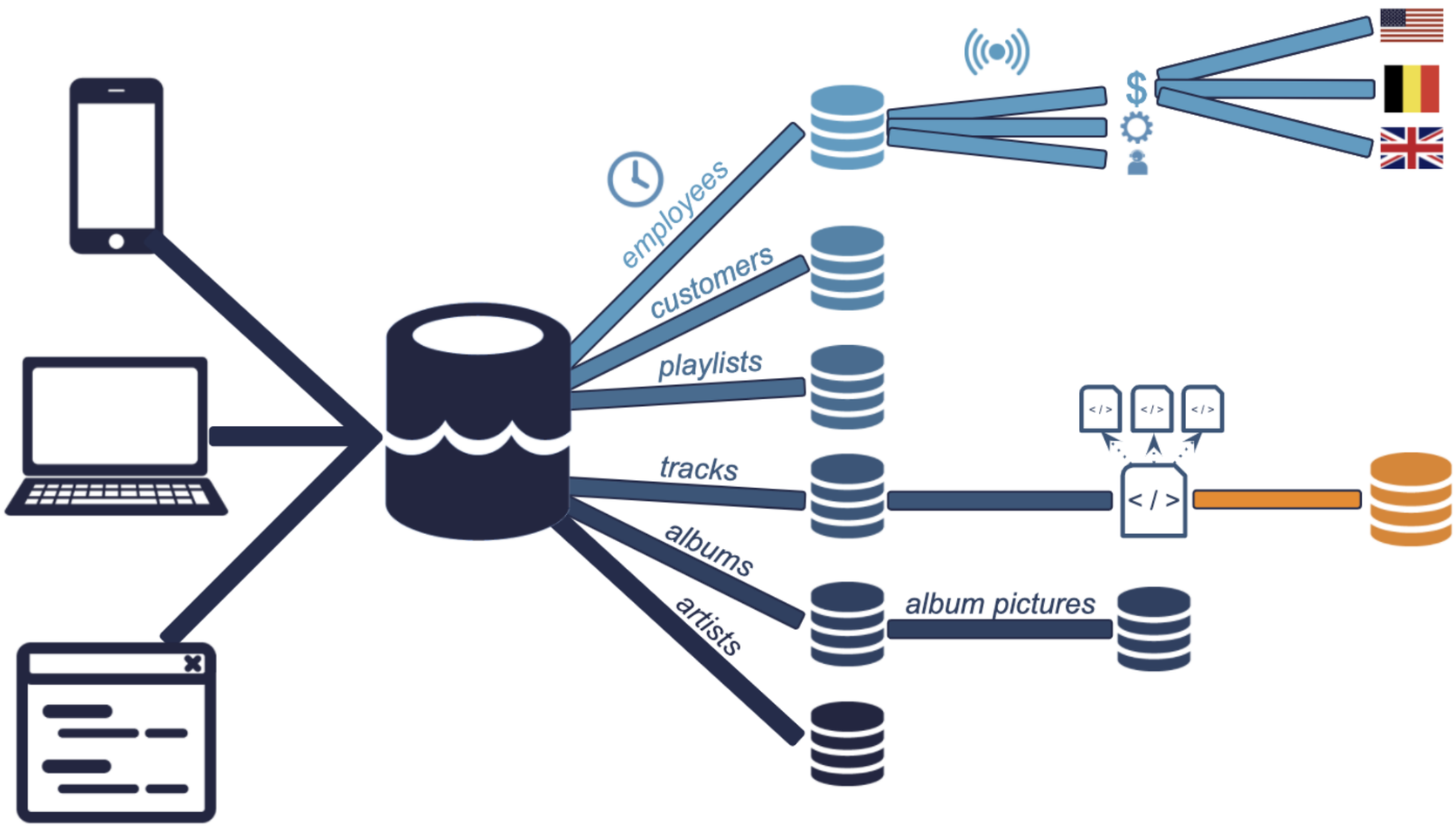
	File storage	Computation
aws	*AWS S3* 	*AWS EC2* 
Microsoft Azure	*Azure Blob Storage* 	*Azure Virtual Machines* 
Google Cloud	*Google Cloud Storage* 	*Google Compute Engine* 

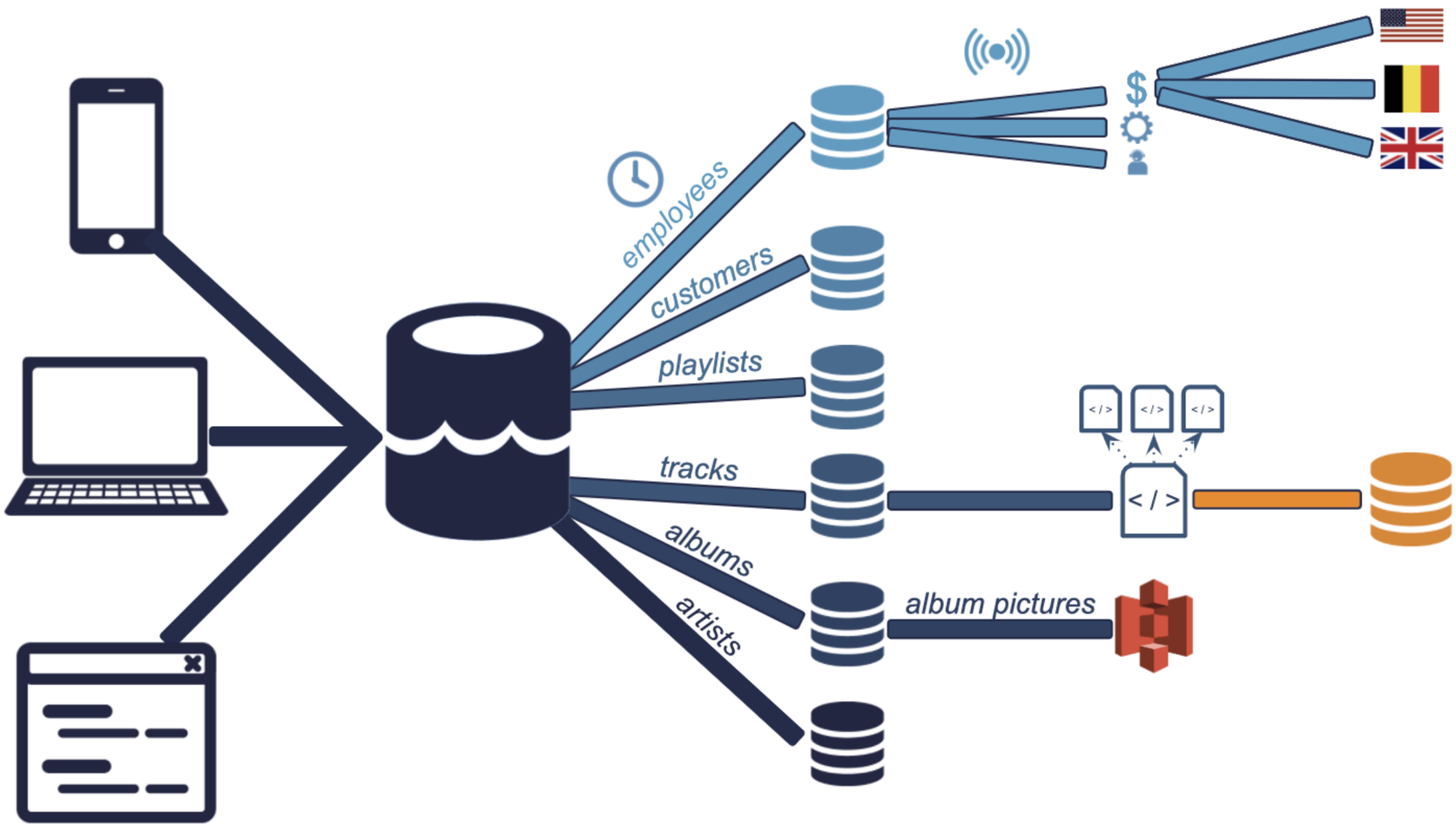
	File storage	Computation	Databases
aws	*AWS S3* 	*AWS EC2* 	
Microsoft Azure	*Azure Blob Storage* 	*Azure Virtual Machines* 	
Google Cloud	*Google Cloud Storage* 	*Google Compute Engine* 	

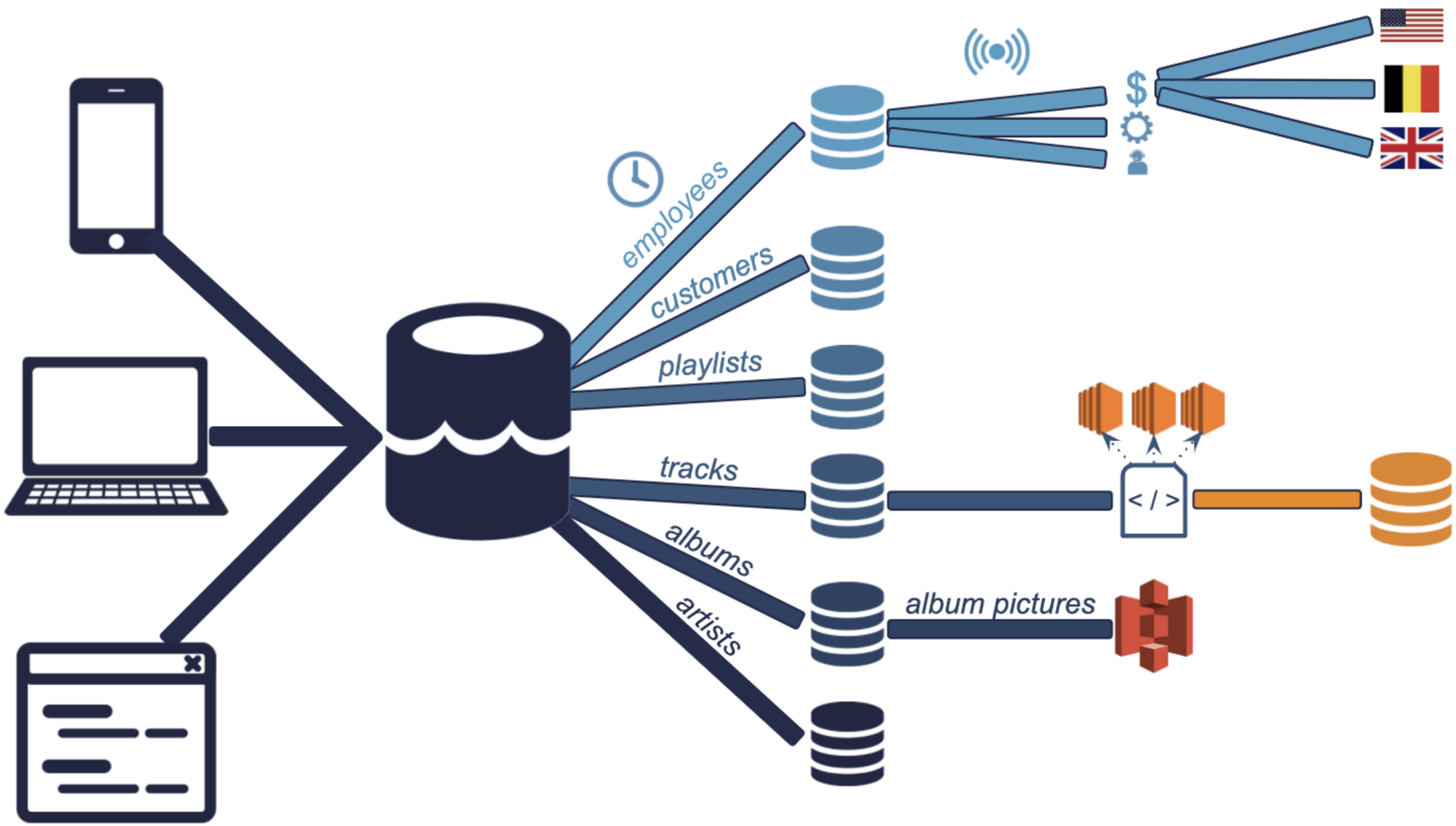
	File storage	Computation	Databases
aws	*AWS S3* 	*AWS EC2* 	*AWS RDS* 
Microsoft Azure	*Azure Blob Storage* 	*Azure Virtual Machines* 	
Google Cloud	*Google Cloud Storage* 	*Google Compute Engine* 	

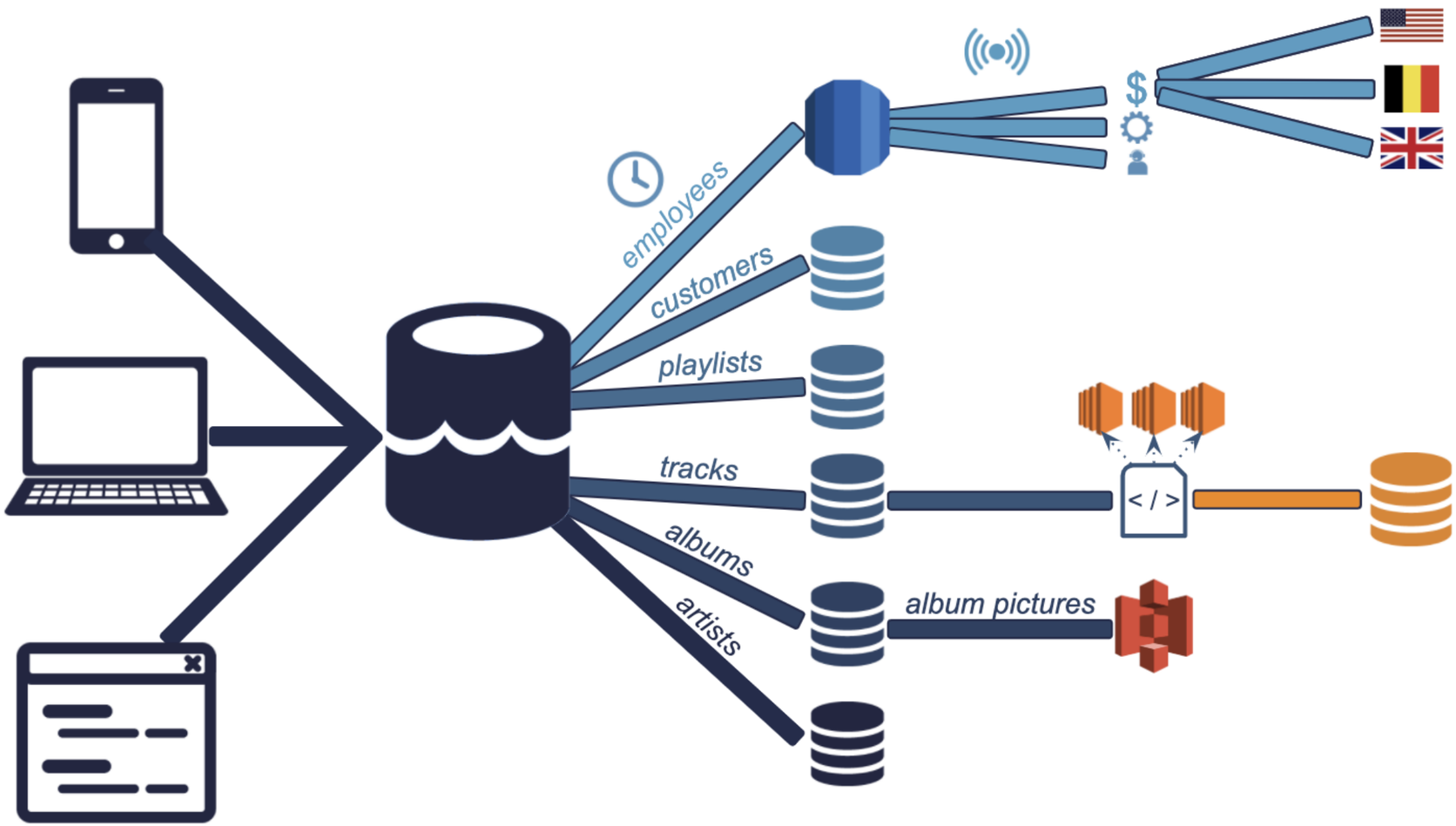
	File storage	Computation	Databases
aws	*AWS S3* 	*AWS EC2* 	*AWS RDS* 
Microsoft Azure	*Azure Blob Storage* 	*Azure Virtual Machines* 	*Azure SQL Database* 
Google Cloud	*Google Cloud Storage* 	*Google Compute Engine* 	

	File storage	Computation	Databases
aws	*AWS S3* 	*AWS EC2* 	*AWS RDS* 
Microsoft Azure	*Azure Blob Storage* 	*Azure Virtual Machines* 	*Azure SQL Database* 
Google Cloud	*Google Cloud Storage* 	*Google Compute Engine* 	*Google Cloud SQL* 


employees
customers
playlists
tracks
albums
artists
album pictures
< / >
$


employees
customers
playlists
tracks
albums
artists
album pictures
< / >
$


employees
customers
playlists
tracks
albums
artists
album pictures
< / >


employees
customers
playlists
tracks
albums
artists
album pictures
< / >

Multicloud

Pros

- Reducing reliance on a single vendor
- Cost-efficiencies
- Local laws requiring certain data to be physically present within the country
- Militating against disasters

Cons

- Cloud providers try to lock in consumers
- Incompatibility
- Security and governance

Summary

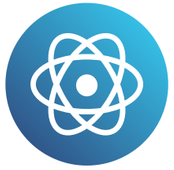
- Benefits and risks of cloud computing
- How it is implemented at Spotflix
- Can cite the main cloud providers and their services

Let's practice!

DATA ENGINEERING FOR EVERYONE

We are the champions

DATA ENGINEERING FOR EVERYONE



Hadrien Lacroix
Content Developer at DataCamp

Actually, YOU are the champion!

What you learned - chapter 1

- What Data Engineering is
- How important it is
- How data engineers differ from data scientists
- What a data pipeline is and how it works

What you learned - chapter 2

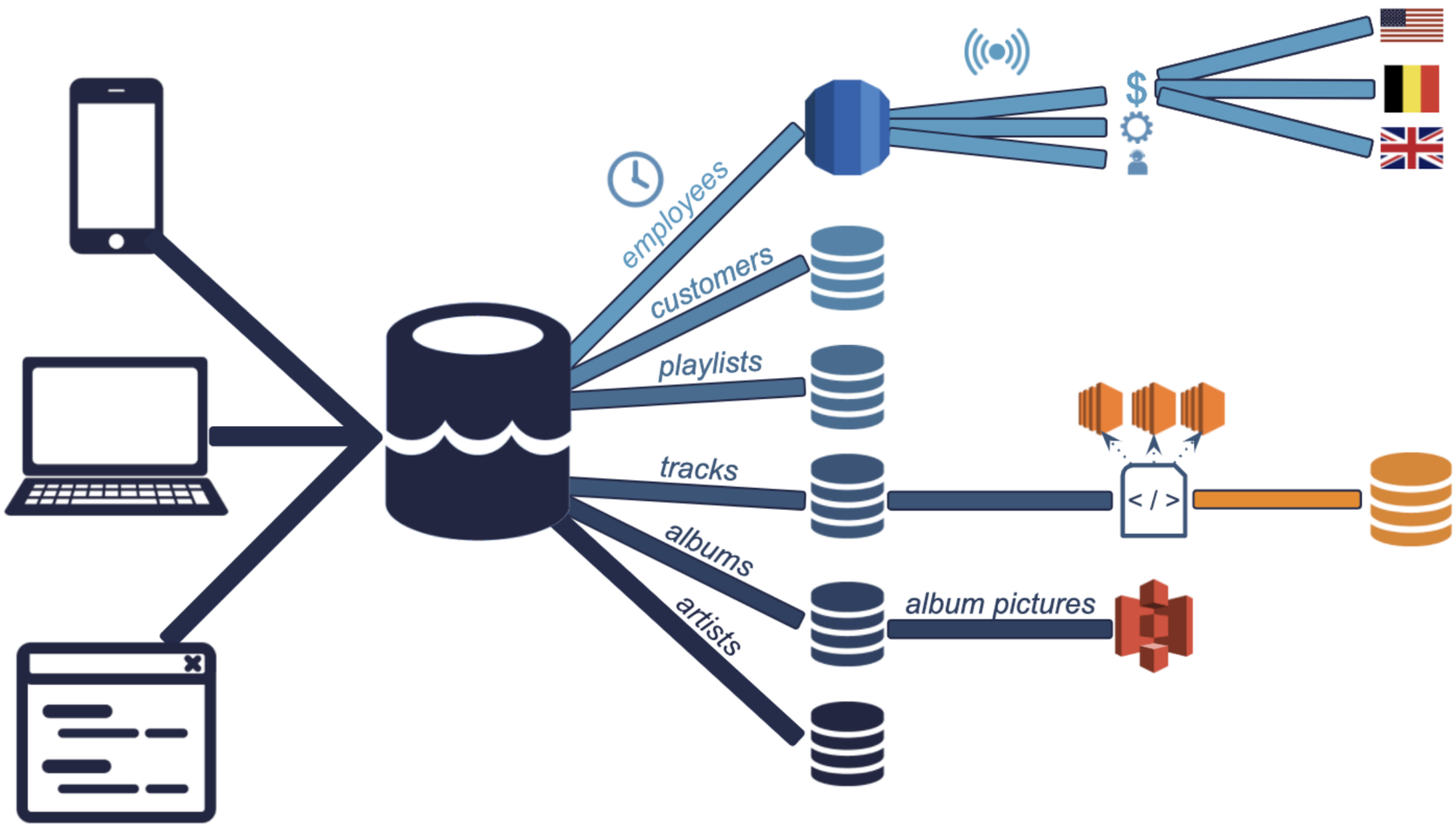
- The different structures data can take
- How fundamental SQL is
- The differences between data lakes, data warehouses and databases

What you learned - chapter 3

- How data is processed
- How scheduling holds it all together
- Parallel computing
- Cloud computing

And some more

- What SQL code actually looks like
- Main tools and technologies used in data engineering
- And some more


employees
customers
playlists
tracks
albums
artists
album pictures
< / >

Lexicon

DataCamp

Search Catalog

My Account

- My Progress
- My Bookmarks
- Organizations
- Custom Tracks
- Career Tracks
- Skill Tracks
- Courses
- Practice
- Projects
- Assessments

In 2019, data engineers overtook data scientists in terms of salary. How did that happen? As eager as companies are of turning their data into gold, if the mine hasn't been build, there isn't much that data scientists can surface. Data engineers lay the groundwork that makes data science possible: no wonder they are in high demand! In this conceptual course, you will get acquainted with their responsibilities, see how they differ from and enable data scientists, and how they manage the flow of data through an organization. Throughout the course, you will understand how data engineering is implemented at a fictional company named Spotflix. At the end of the course, you will be ready to have a conversation with a data engineer, understand what your company's data engineers do, or have a solid foundation to start your journey to becoming one yourself!

1 **What is Data Engineering** FREE 0%

This chapter starts with an explanation of what data engineering is and why there is an increasing need for it. Equipped with this foundational knowledge, you will then understand where Data Engineering stands in the realm of Data Science, and how a Data Engineer differs from a Data Scientist. You will finish this chapter with a first exposure to a complete data pipeline.

VIEW CHAPTER DETAILS

Continue Chapter

Hadrien Lacroix

Content Developer at DataCamp

Hadrien is a Content Developer at DataCamp. He's helping instructors daily build the best Data Science and Machine Learning courses possible.

See More

COLLABORATOR(S)

Lis Sulmont

A promise is a promise, DataChamps!

- All the exercises are song titles
- Search for "DataChamps" on Spotify

Congratulations!

DATA ENGINEERING FOR EVERYONE