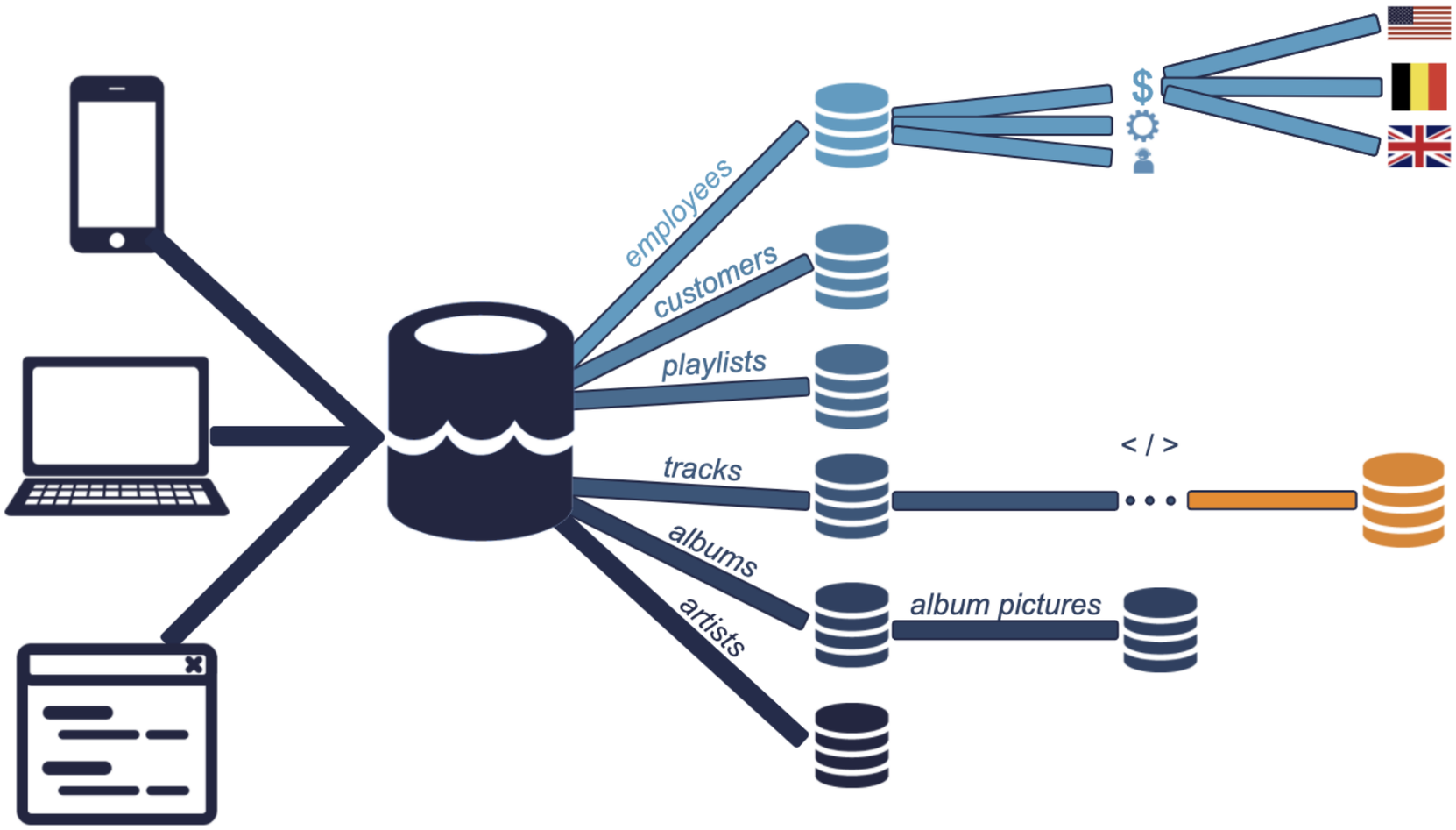
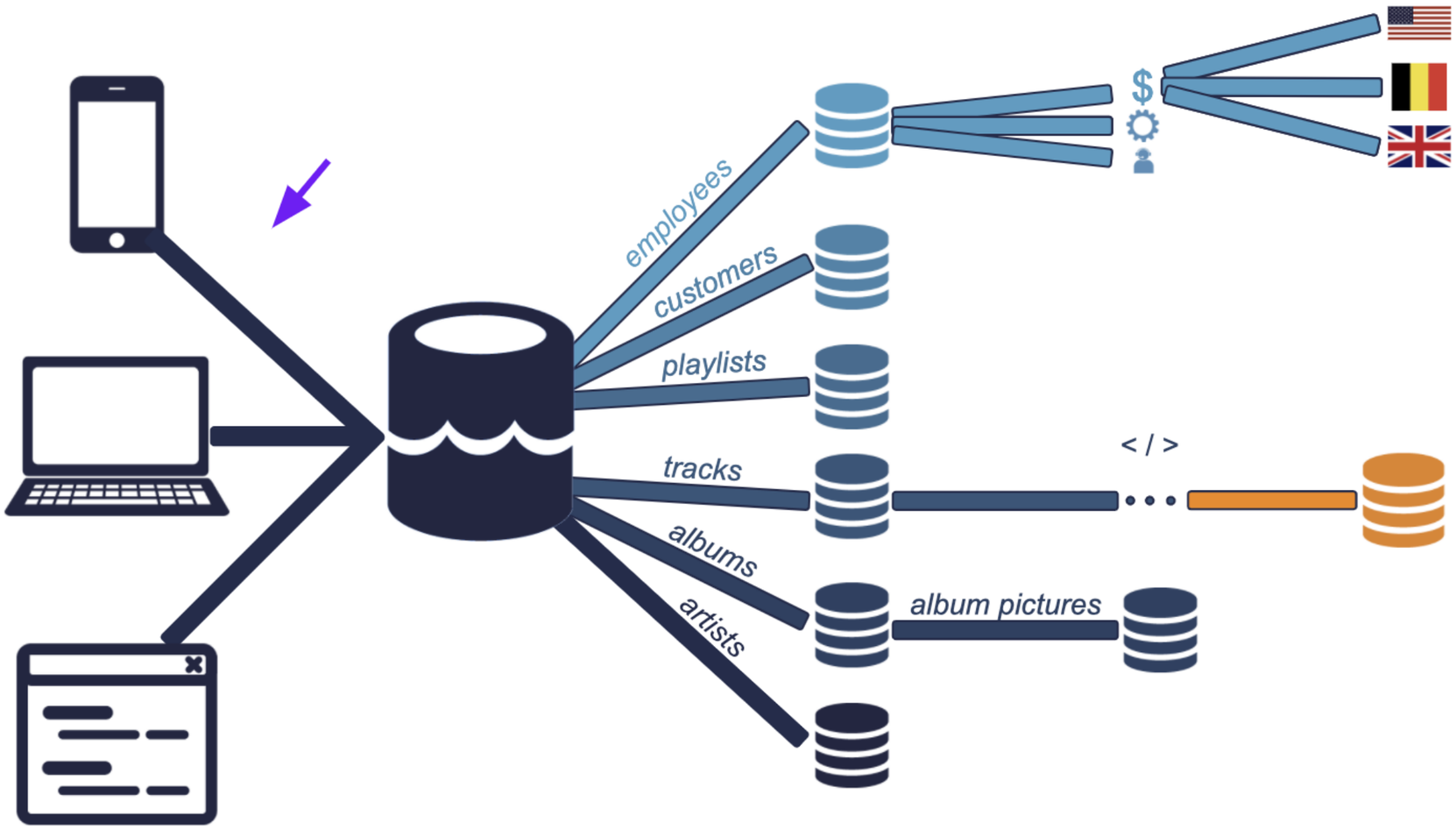
Processing data

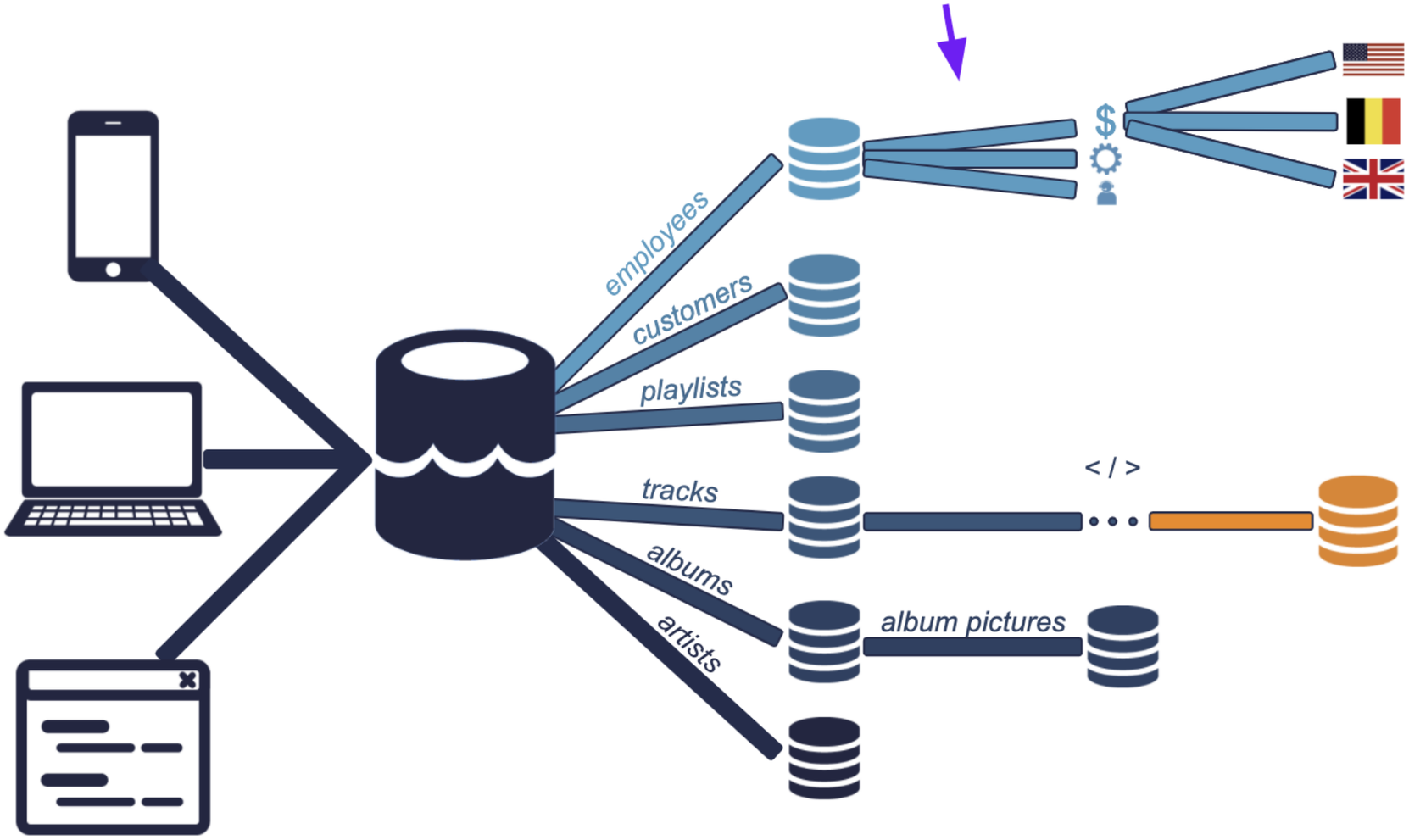
UNDERSTANDING DATA ENGINEERING

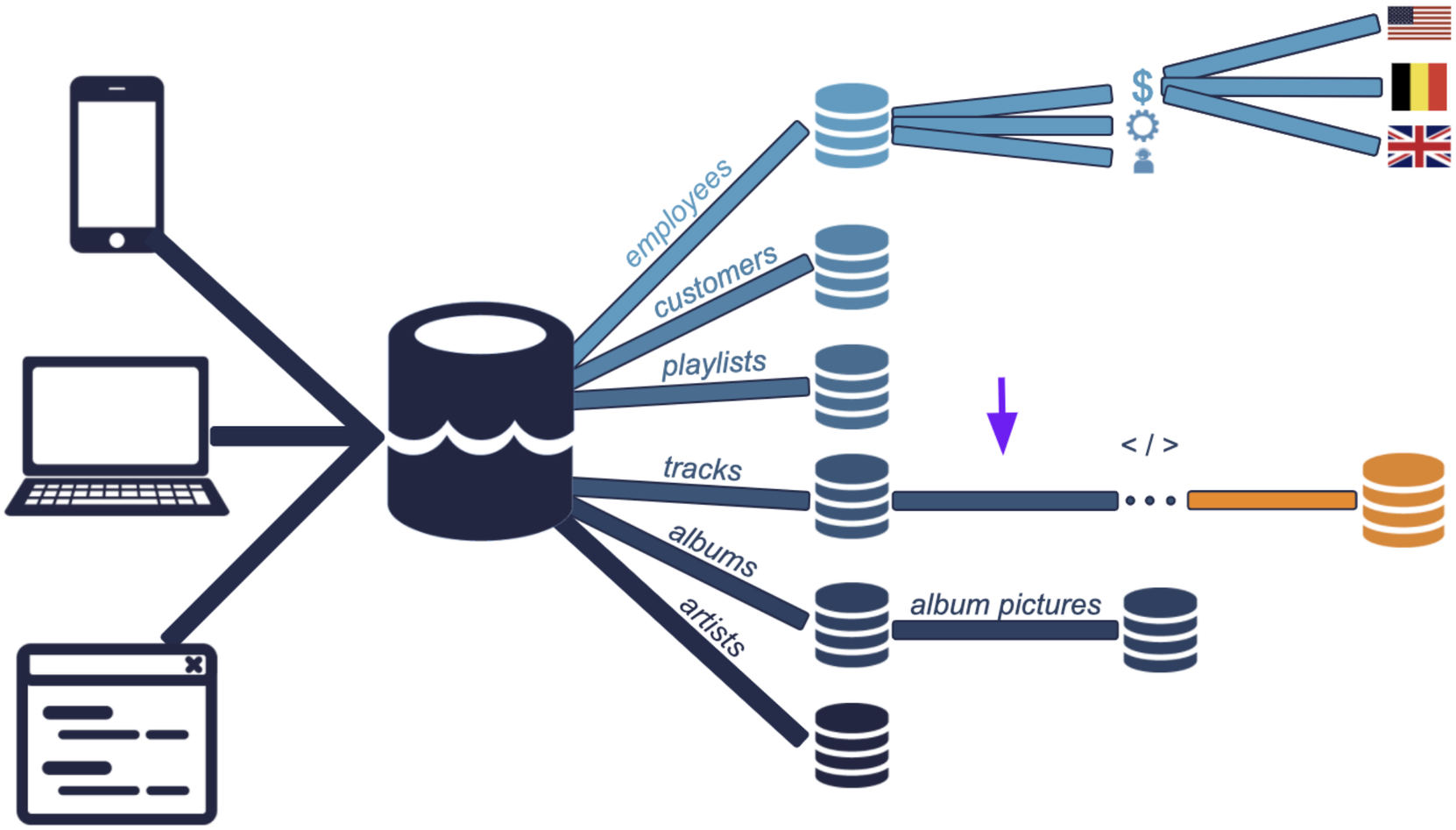
Hadrien Lacroix
Content Developer at DataCamp

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

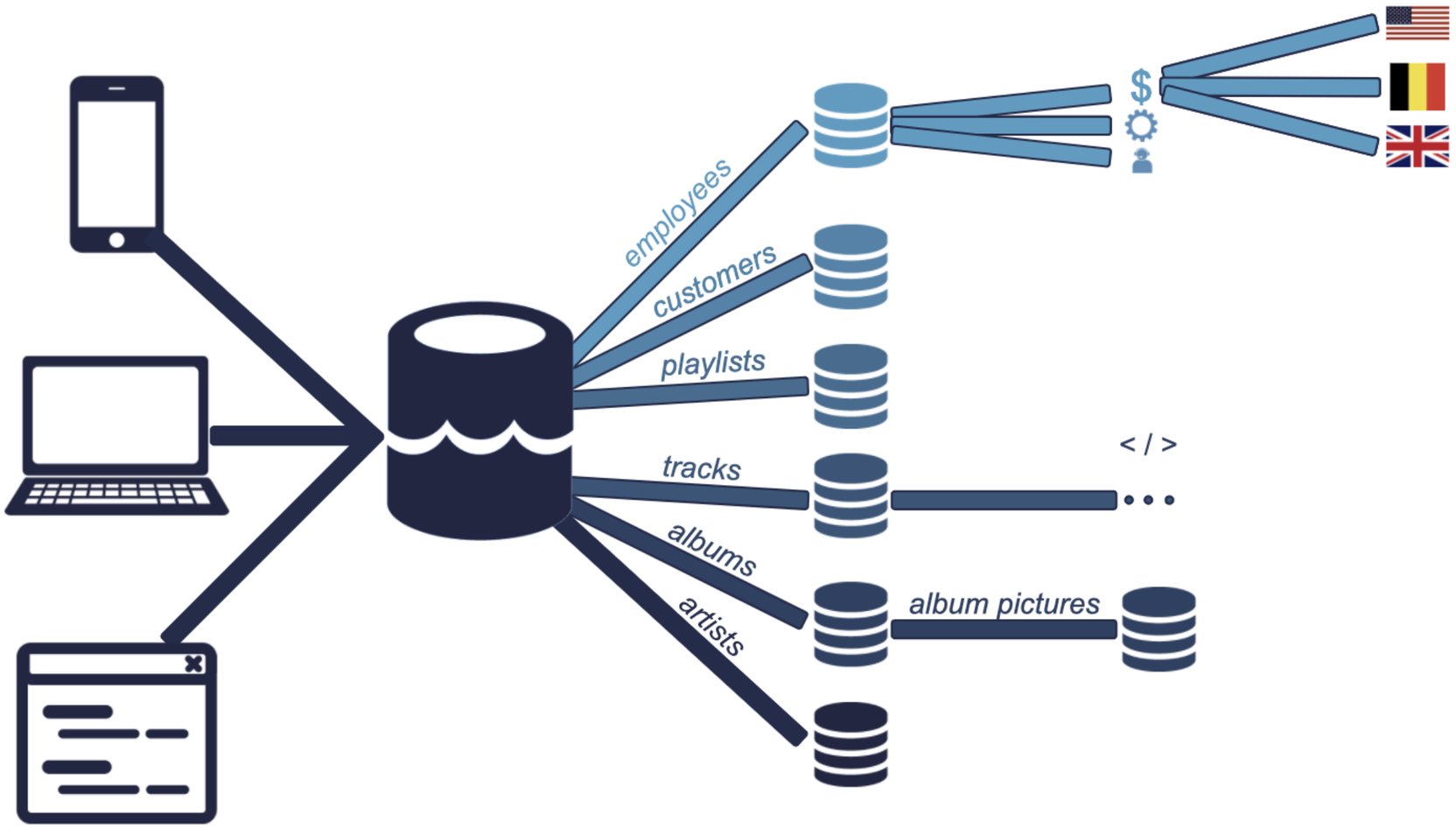
Data processing value

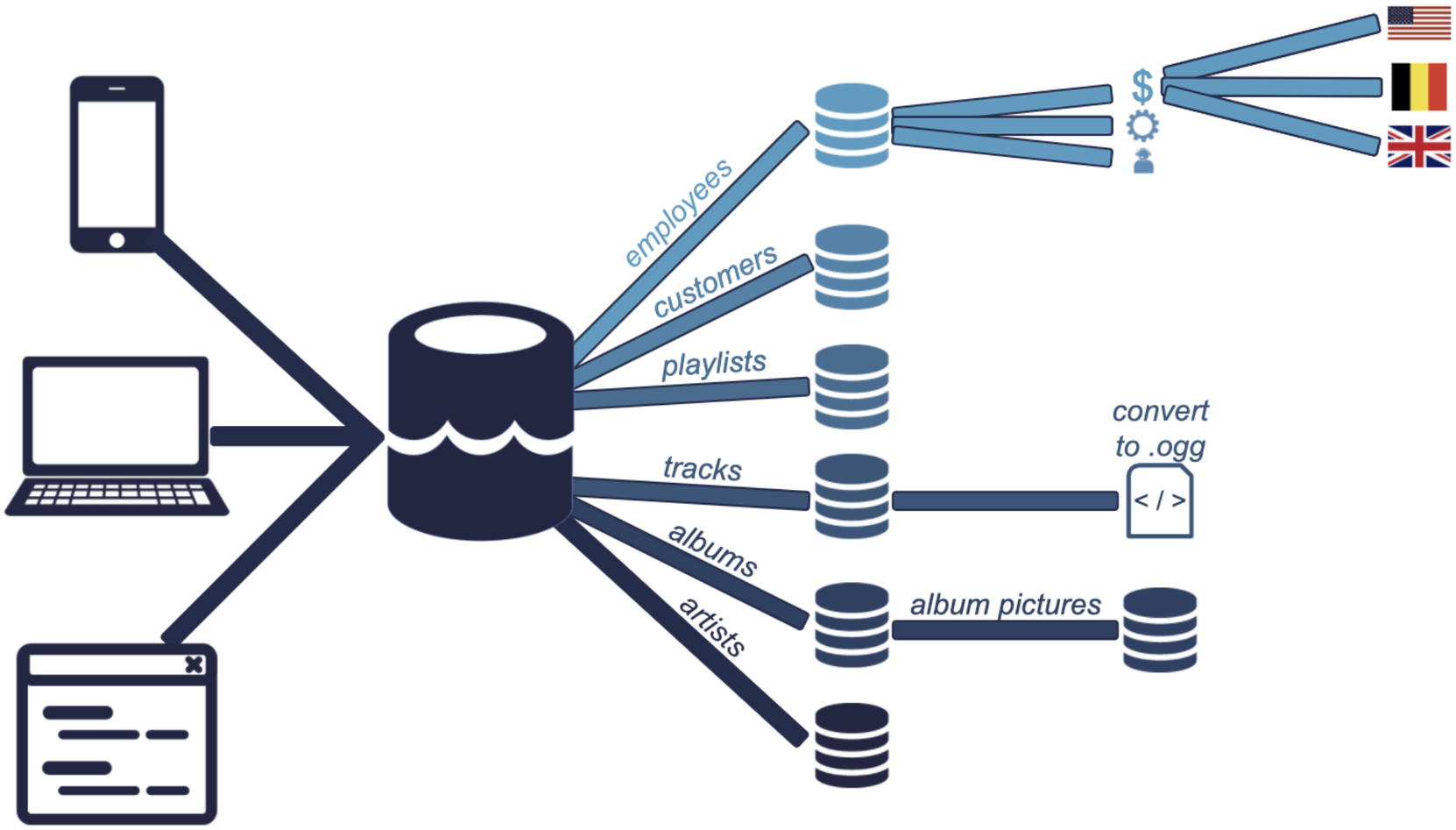
Conceptually

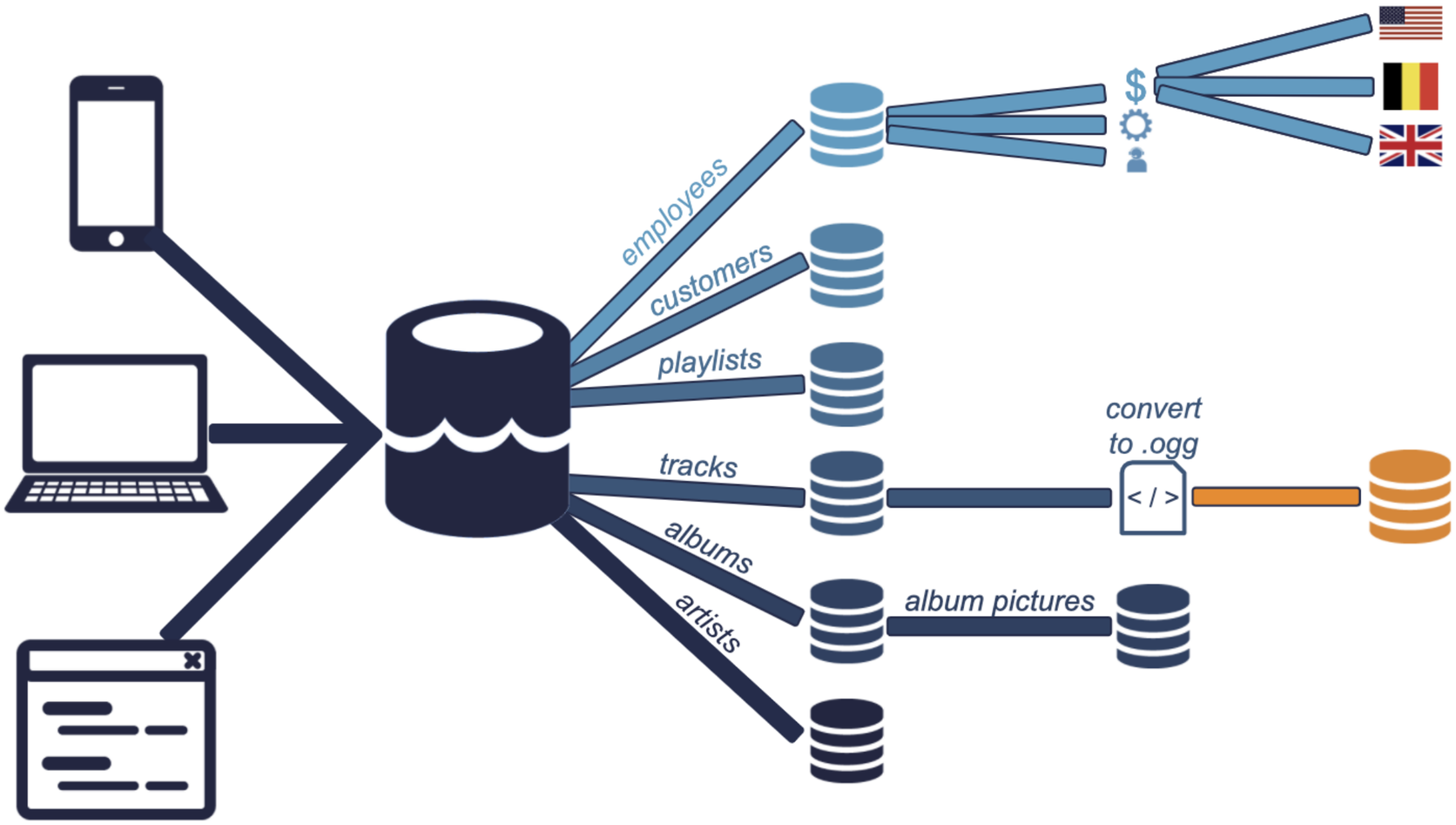
- Remove unwanted data
- Optimize memory, process and network costs
- Convert data from one type to another

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

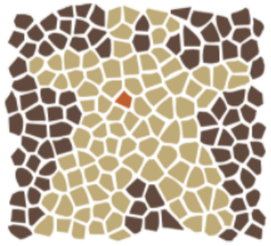
At Spotflix

- 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!

UNDERSTANDING DATA ENGINEERING

Scheduling data

UNDERSTANDING DATA ENGINEERING

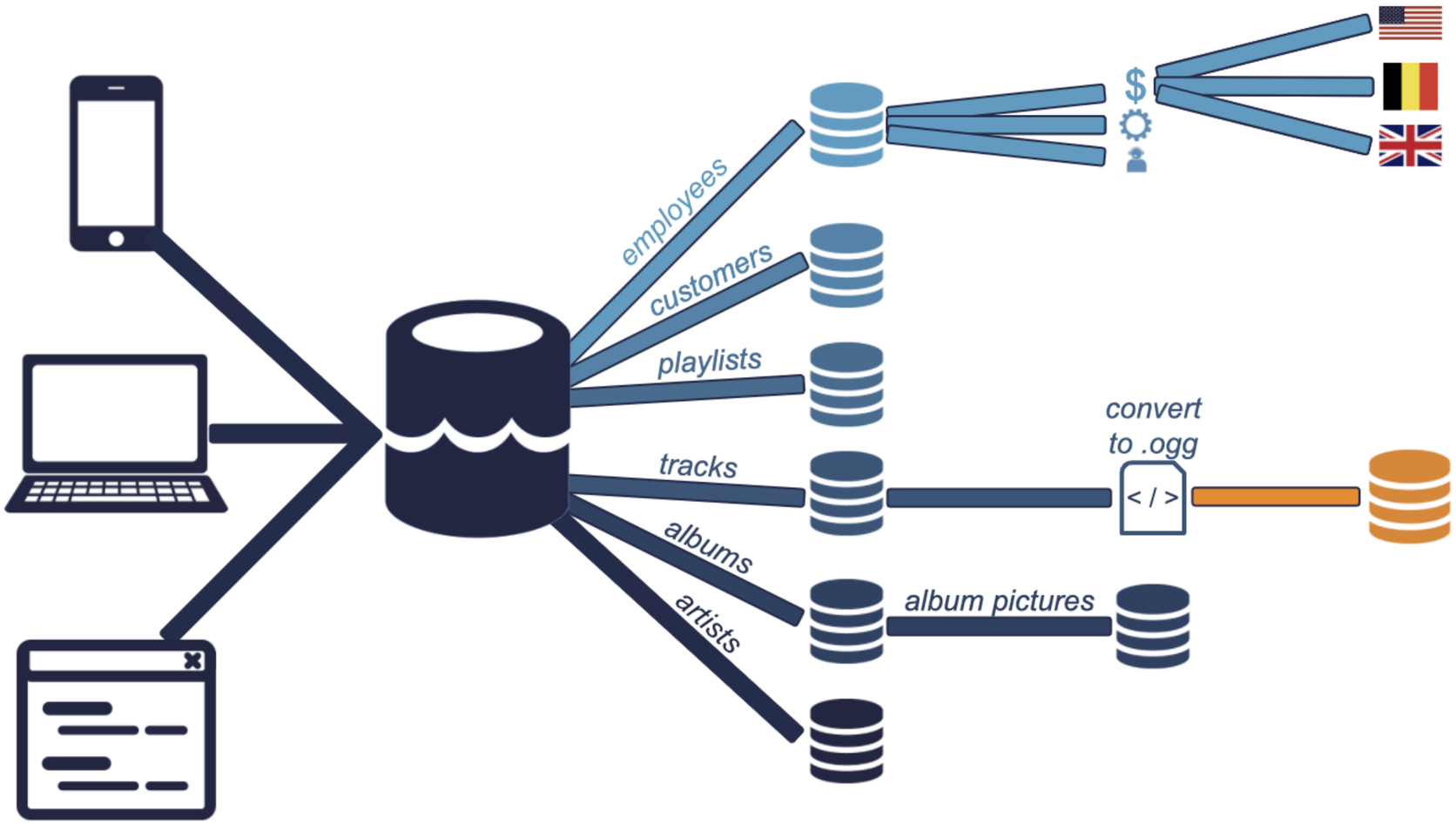
Hadrien Lacroix
Content Developer at DataCamp

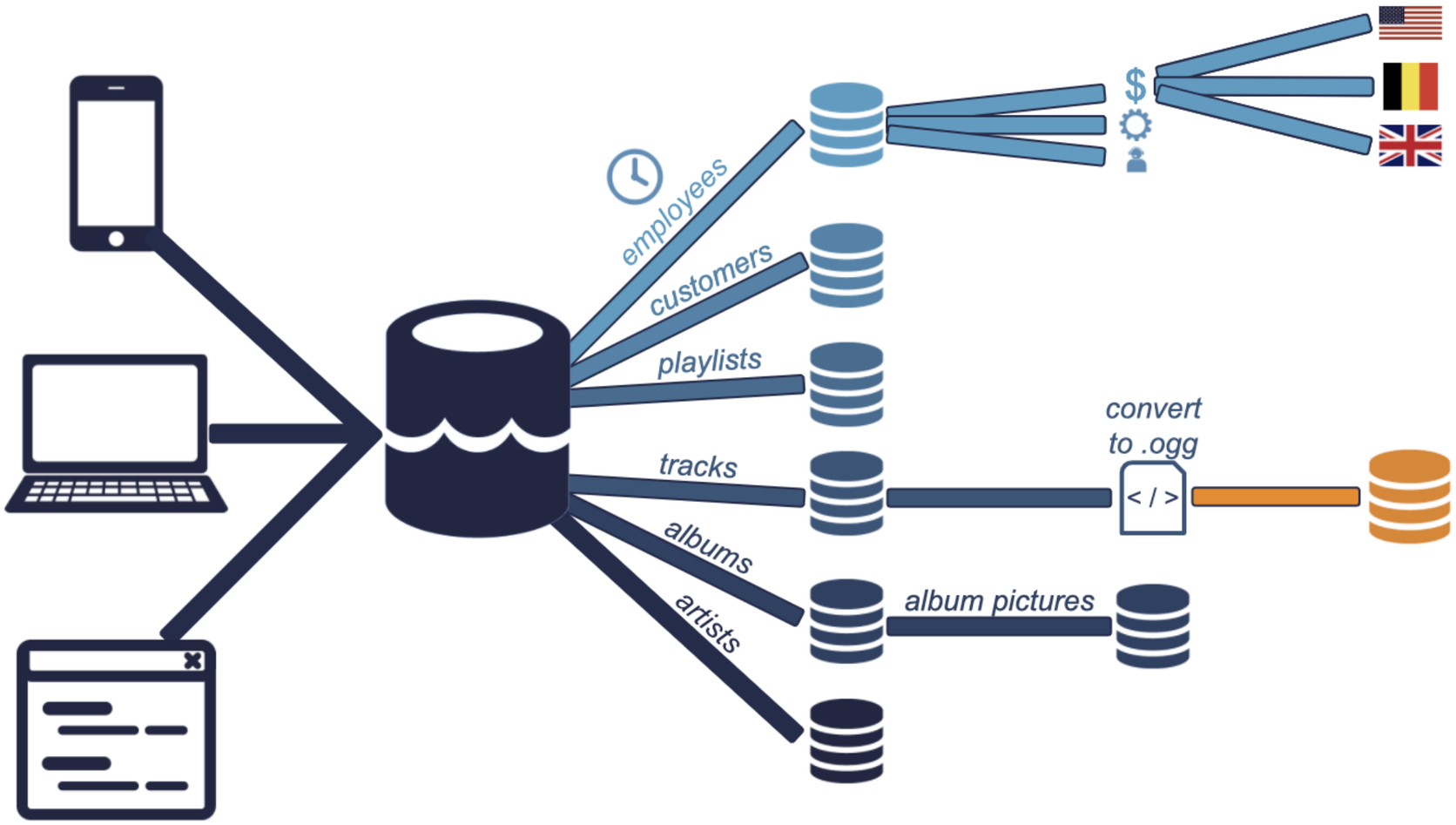
Scheduling

- 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

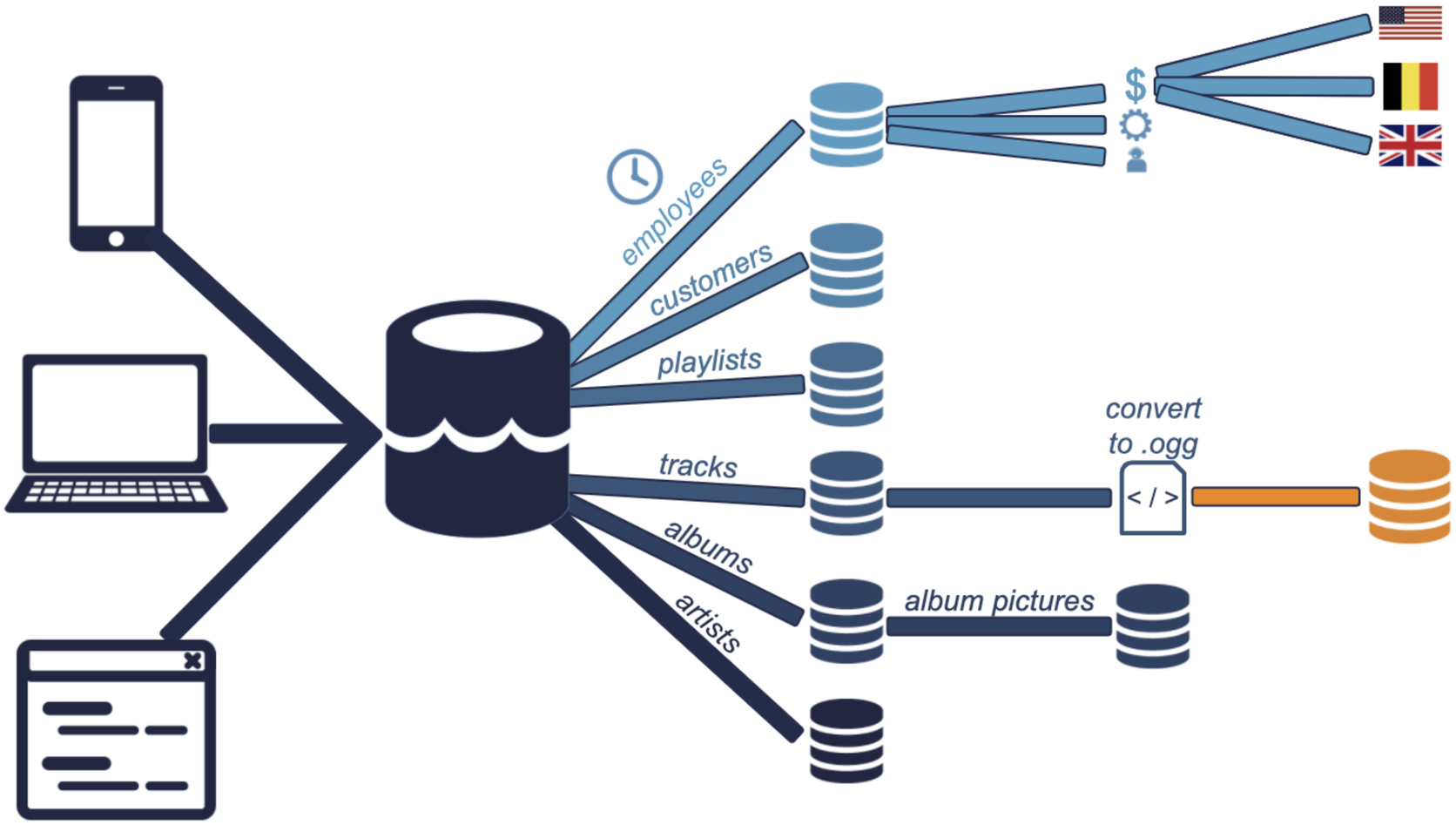

employees
customers
playlists
tracks
albums
artists
convert
to .ogg
< / >
album pictures

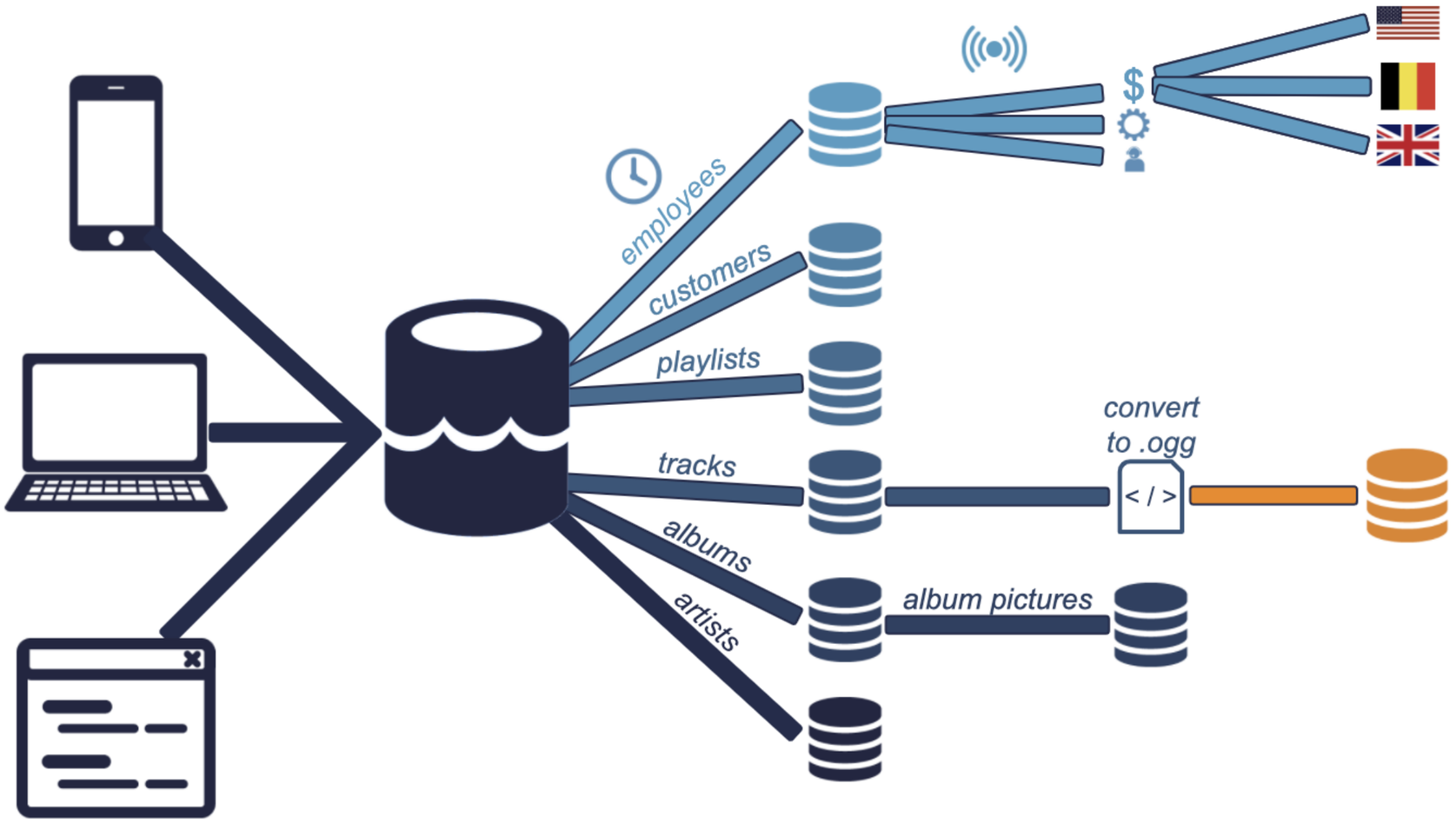

employees
customers
playlists
tracks
albums
artists
convert
to .ogg
< / >
album pictures

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!

UNDERSTANDING DATA ENGINEERING

Parallel computing

UNDERSTANDING DATA ENGINEERING

Hadrien Lacroix
Content Developer at DataCamp

datacamp

Parallel computing

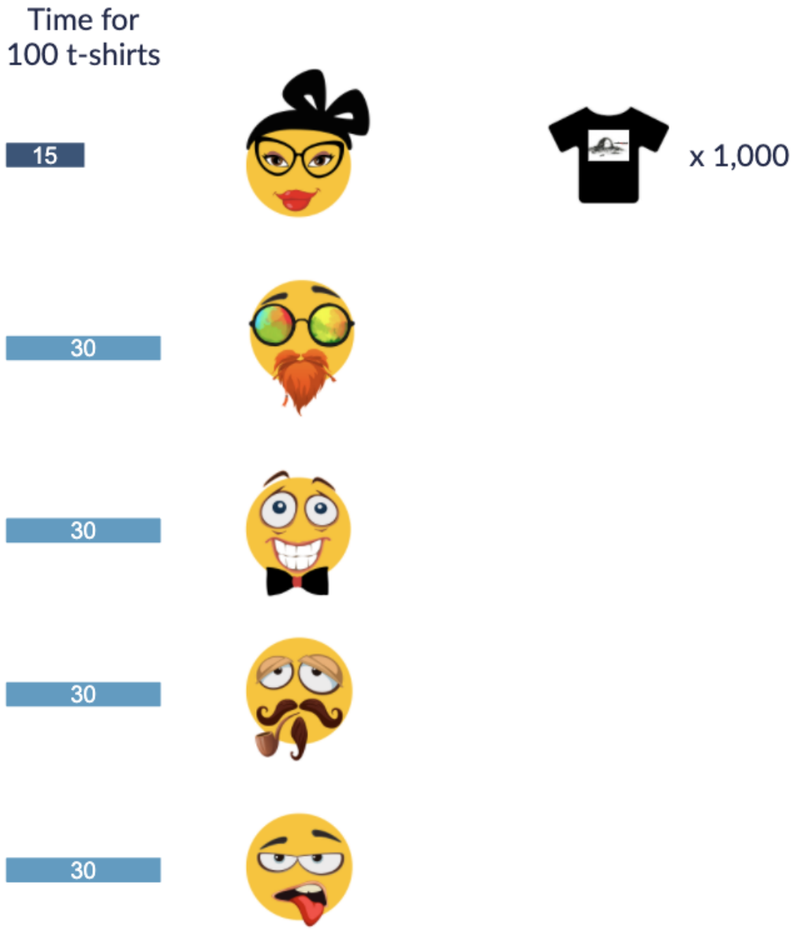
- Basis of modern data processing tools
- Necessary:
  - Mainly because of memory
  - Also for processing power
- How it works:
  - Split tasks up into several smaller subtasks
  - Distribute these subtasks over several computers

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

15

x 1,000

30

x 250

30

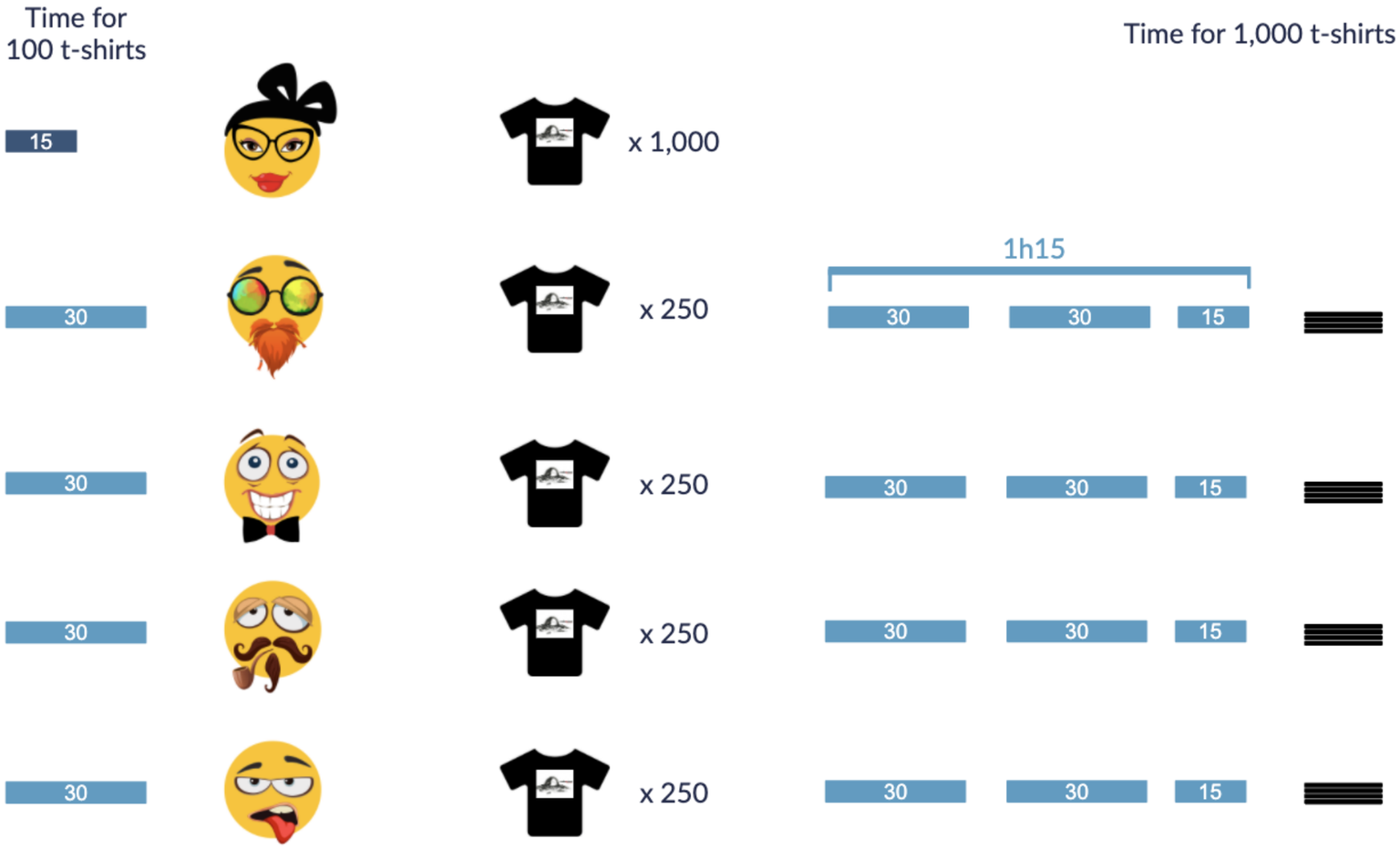
x 250

30

x 250

30

x 250


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

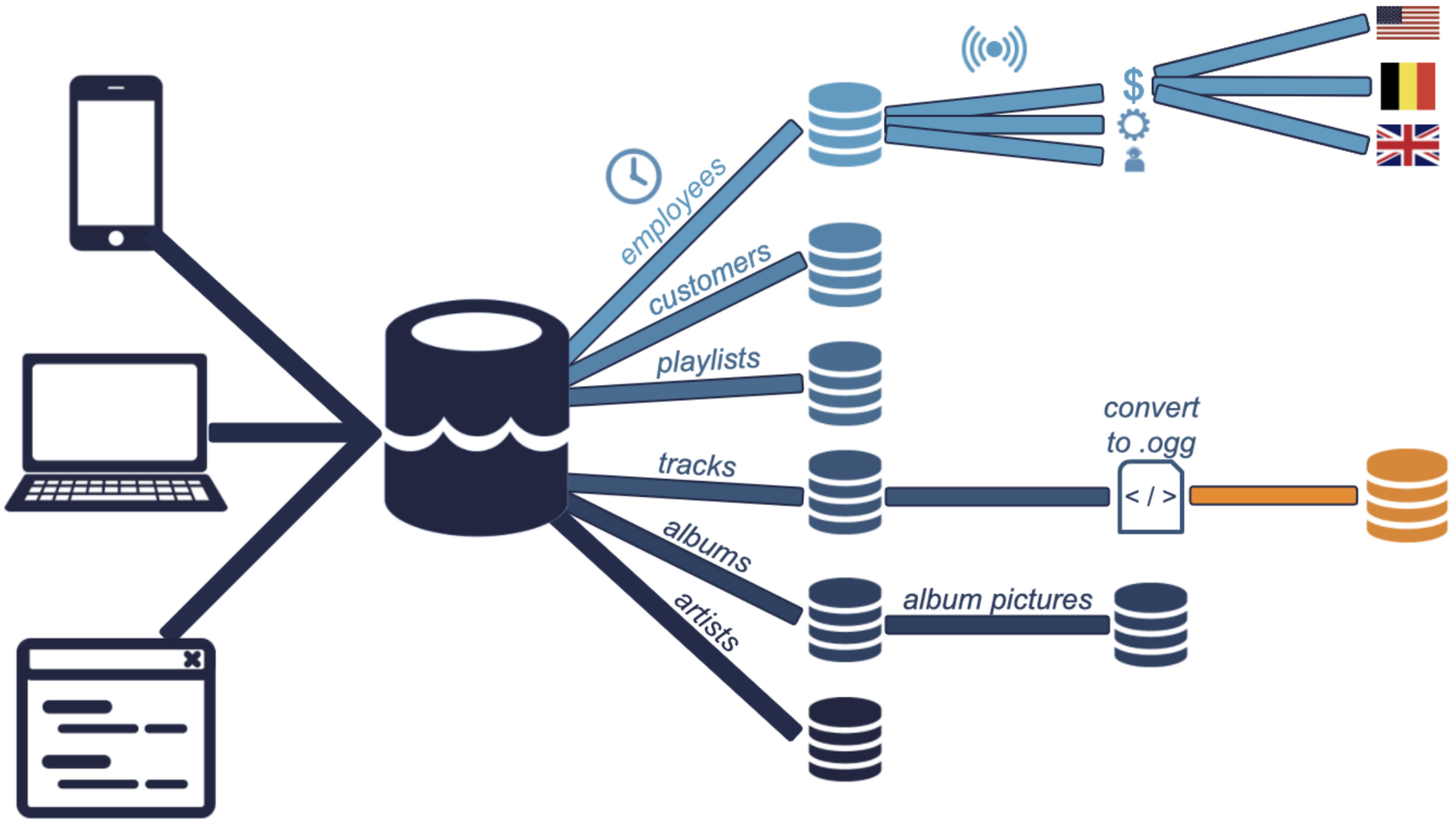
Benefits and risks of parallel computing

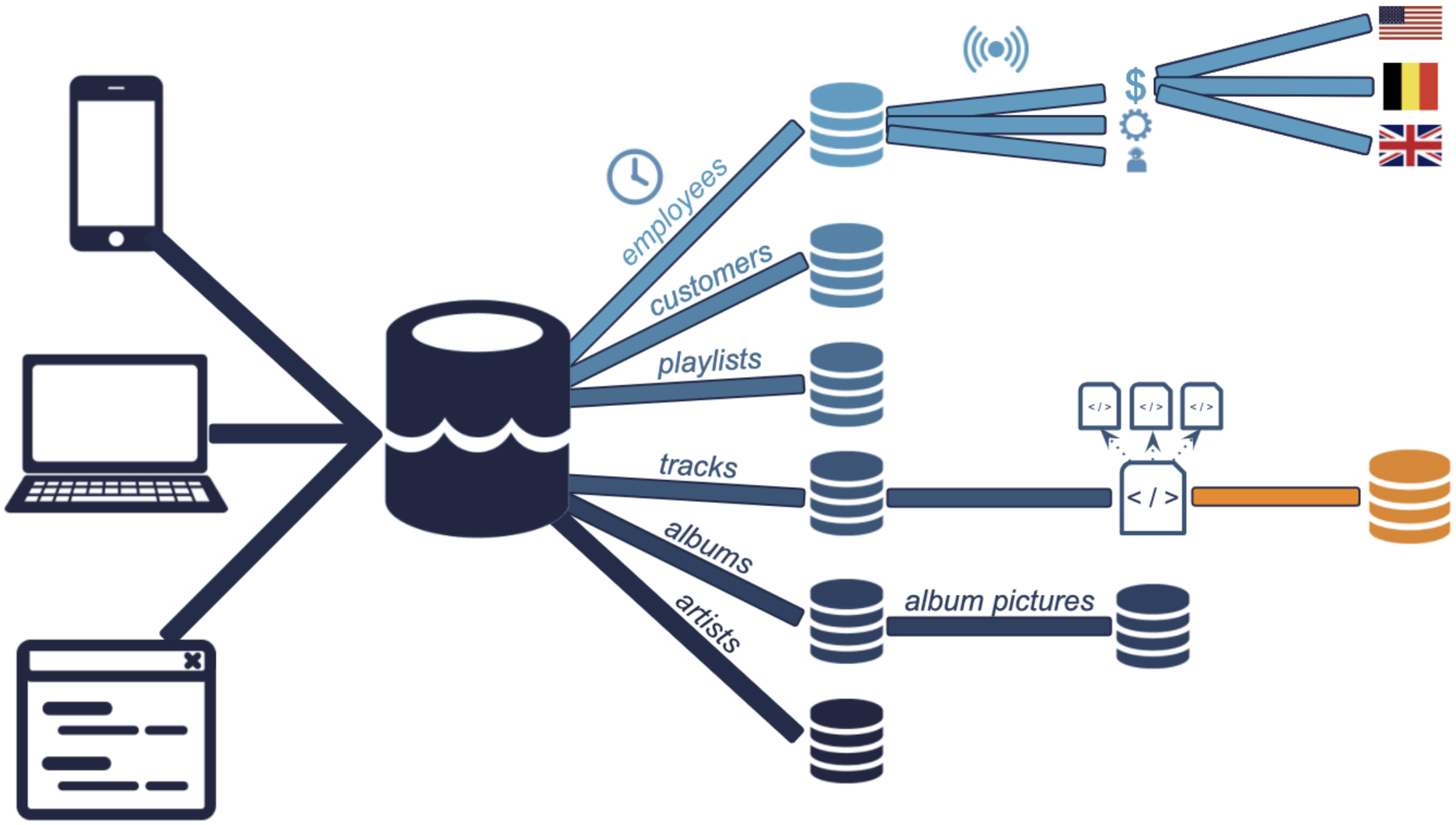
- Employees = processing units
- Advantages
  - Extra processing power
  - Reduced memory footprint
- Disadvantages
  - Moving data incurs a cost
  - Communication time


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


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!

UNDERSTANDING DATA ENGINEERING

Cloud computing

UNDERSTANDING DATA ENGINEERING

Hadrien Lacroix
Content Developer

datacamp

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* 	

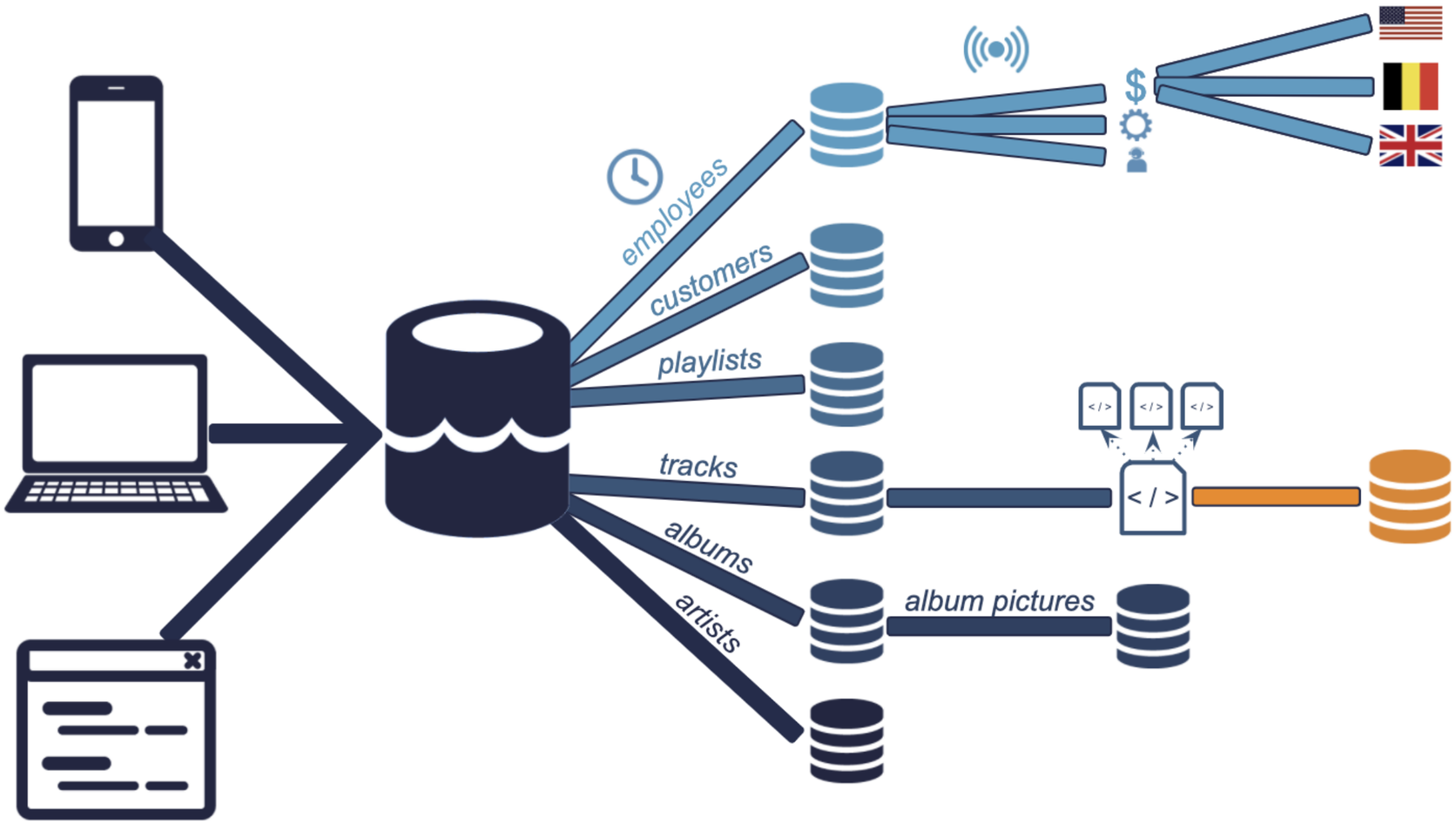
	File storage	Computation
aws	*AWS S3* 	*AWS EC2* 
Microsoft Azure	*Azure Blob Storage* 	*Azure Virtual Machines* 
Google Cloud	*Google Cloud Storage* 	*Google Compute Engine* 

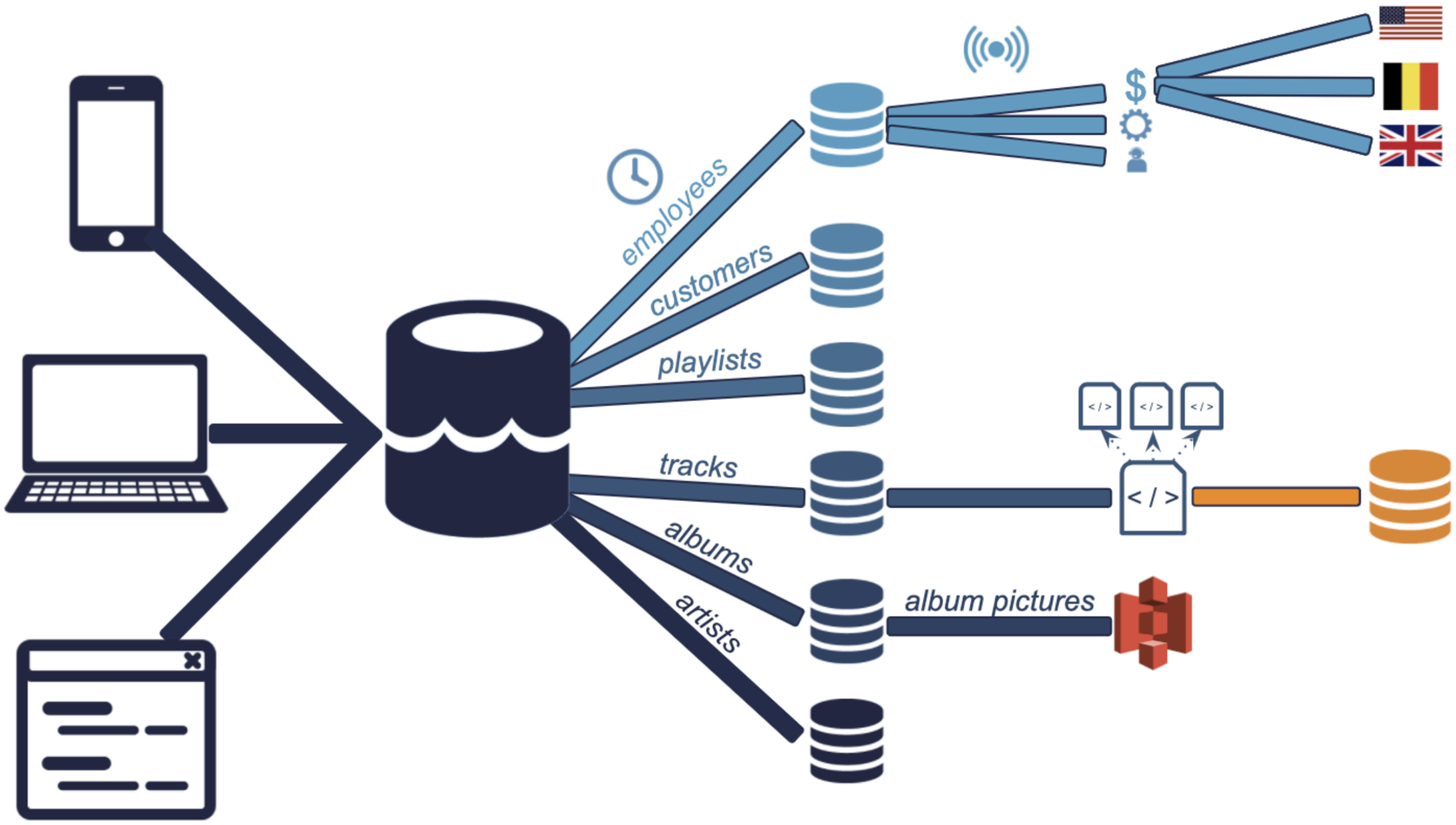
	File storage	Computation	Databases
aws	*AWS S3* 	*AWS EC2* 	
Microsoft Azure	*Azure Blob Storage* 	*Azure Virtual Machines* 	
Google Cloud	*Google Cloud Storage* 	*Google Compute Engine* 	

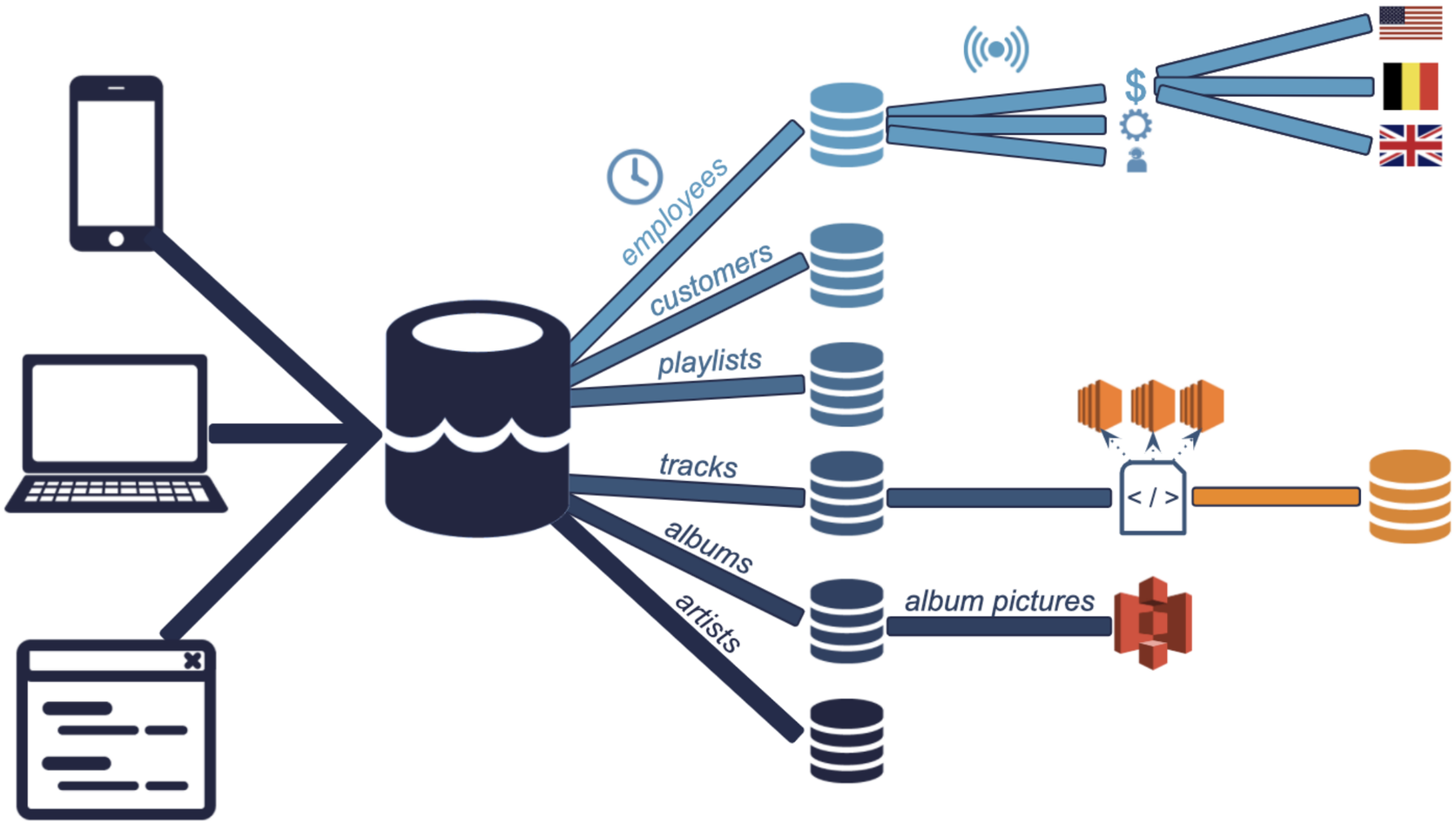
	File storage	Computation	Databases
aws	*AWS S3* 	*AWS EC2* 	*AWS RDS* 
Microsoft Azure	*Azure Blob Storage* 	*Azure Virtual Machines* 	
Google Cloud	*Google Cloud Storage* 	*Google Compute Engine* 	

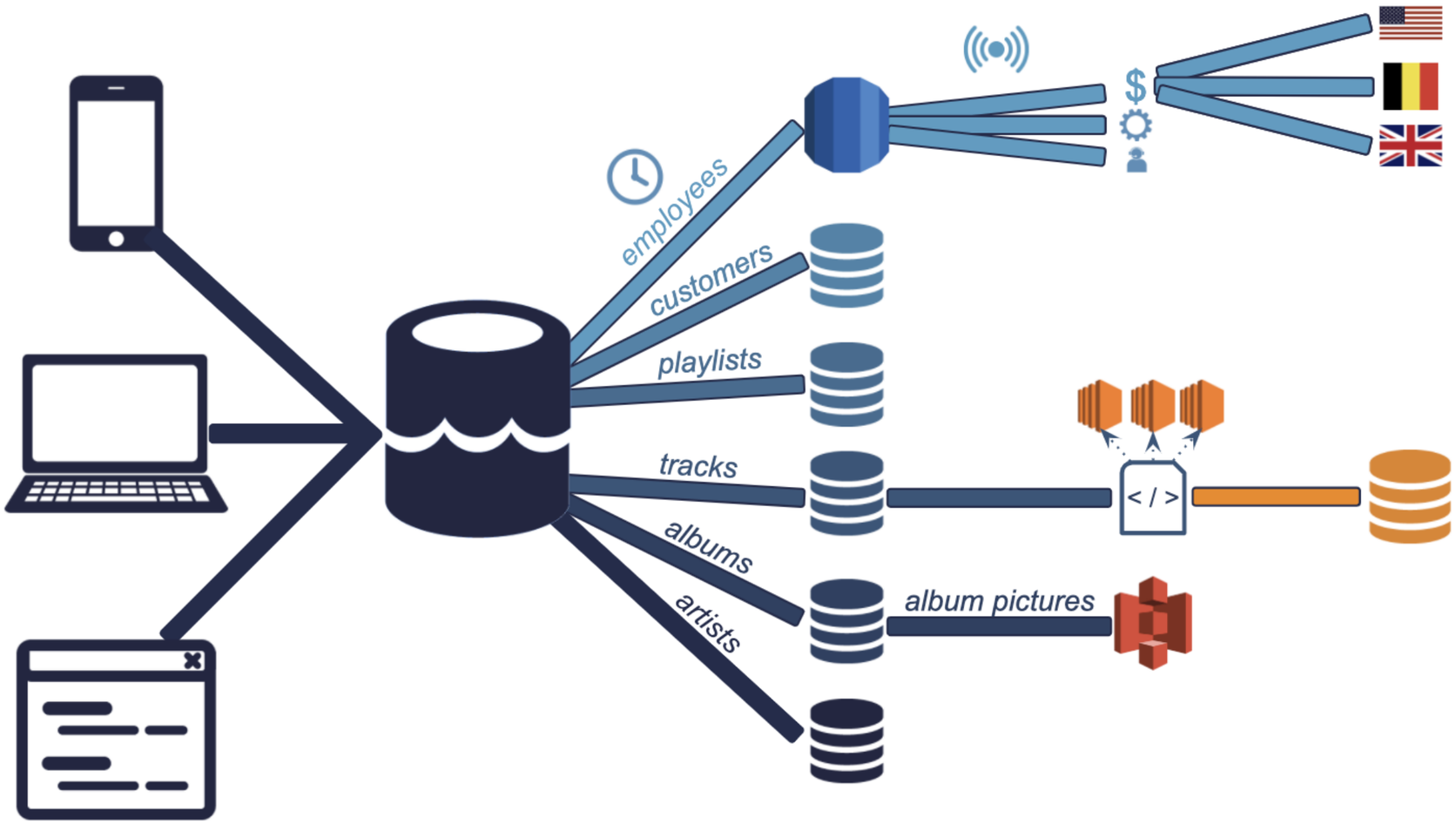
	File storage	Computation	Databases
aws	*AWS S3* 	*AWS EC2* 	*AWS RDS* 
Microsoft Azure	*Azure Blob Storage* 	*Azure Virtual Machines* 	*Azure SQL Database* 
Google Cloud	*Google Cloud Storage* 	*Google Compute Engine* 	

	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
- Mitigating 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!

UNDERSTANDING DATA ENGINEERING

We are the champions

UNDERSTANDING DATA ENGINEERING

Hadrien Lacroix
Content Developer at DataCamp

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

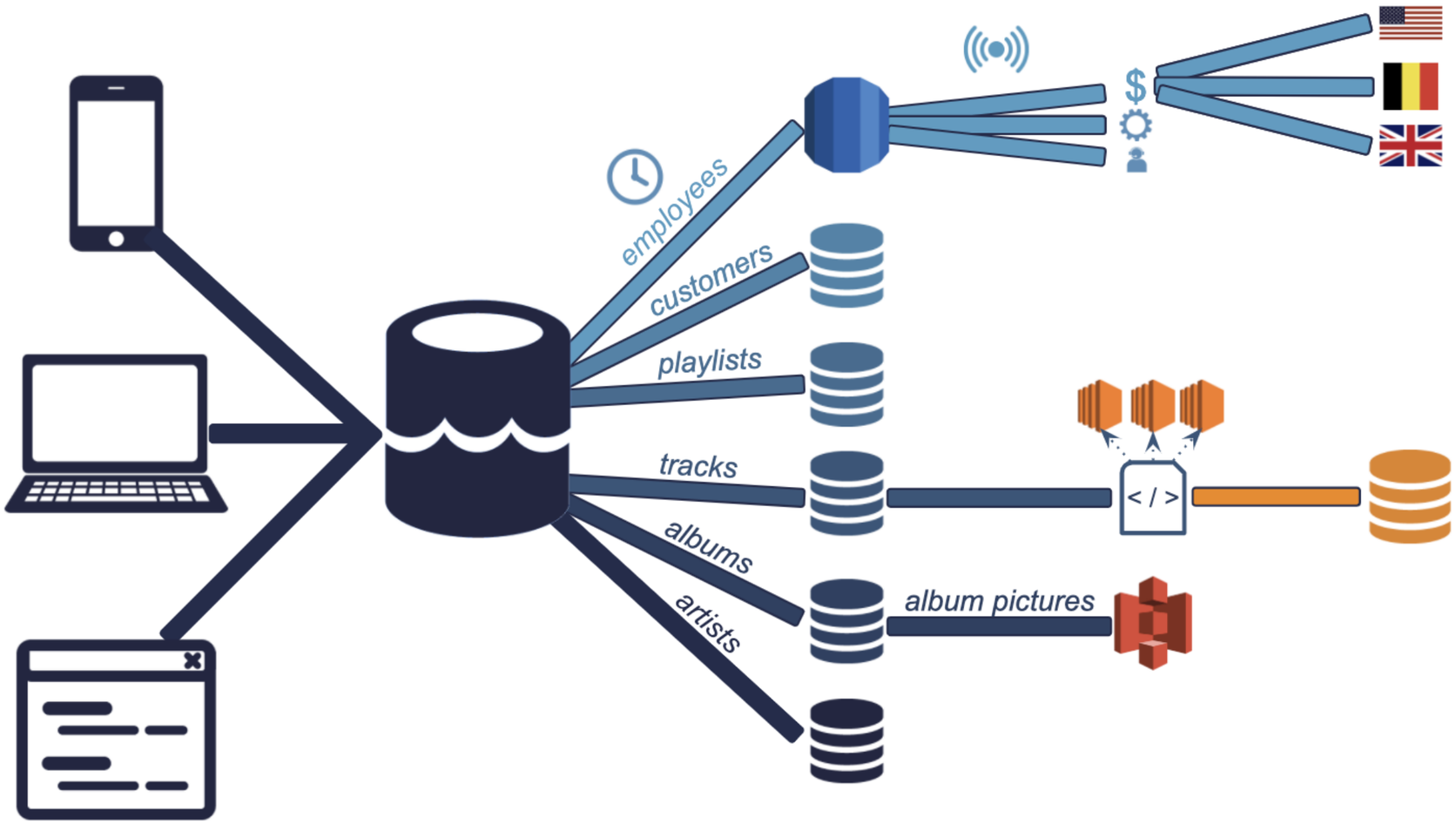
- The different structures data can take
- How fundamental SQL is
- The differences between data lakes, data warehouses and databases

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

Data Engineering for Everyone - Lexicon

Data Engineering for Everyone - Lexicon

- **Airflow**: an open-source workflow management platform used to schedule data engineering tasks. Started at Airbnb, now maintained by the Apache foundation.
- **AWS**: Amazon Web Services. Amazon's cloud computing services.
- **Azure**: Microsoft's cloud services.
- **Big data**: the systematic storage, management and analysis of datasets that are too large or complex to be dealt with by traditional data-processing application software. Big Data revolves around 4 Vs: volume, variety, velocity, and veracity.
- **Cloud computing**: the use of a network of remote servers hosted on the Internet to store, manage, and process data, rather than a local server or a personal computer.

A promise is a promise, DataChamps!

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

Congratulations!

UNDERSTANDING DATA ENGINEERING