

# Data engineering and big data

UNDERSTANDING DATA ENGINEERING



**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



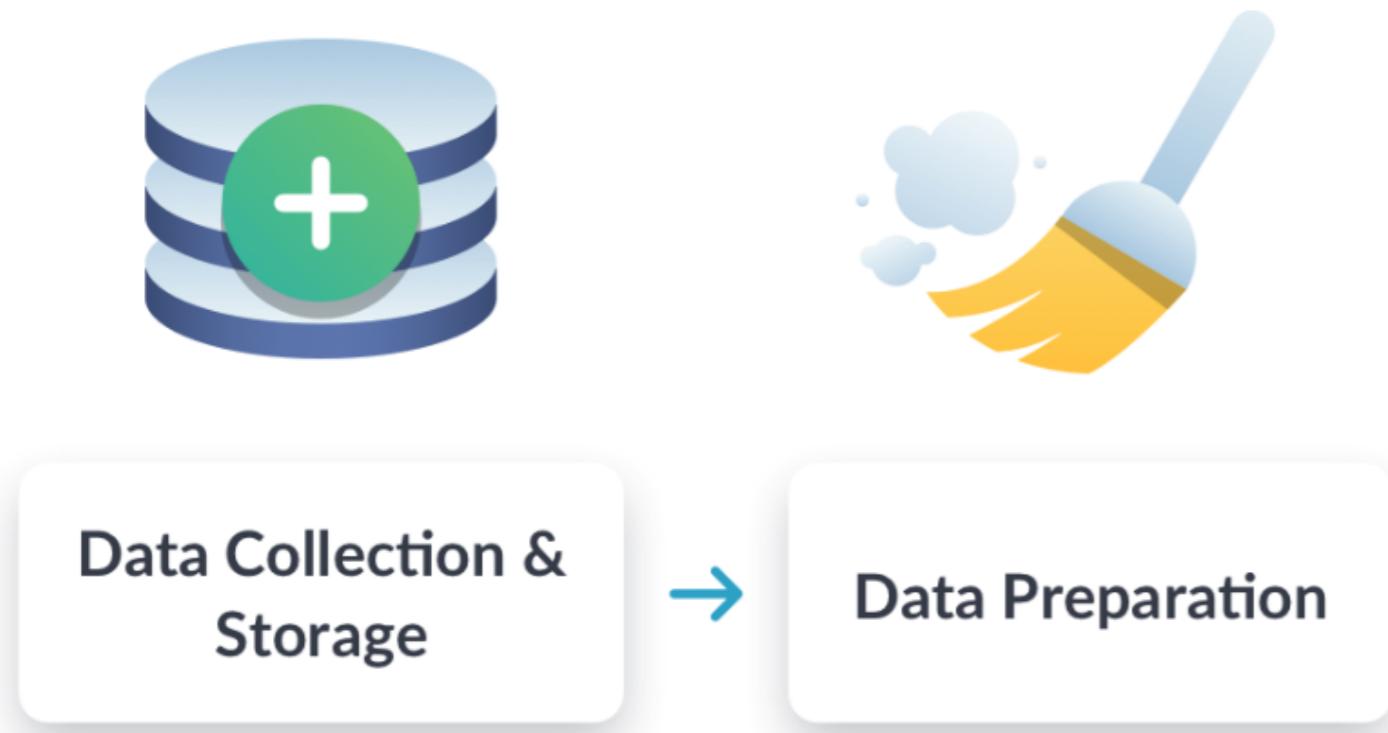
# Spotfliix

# Data workflow

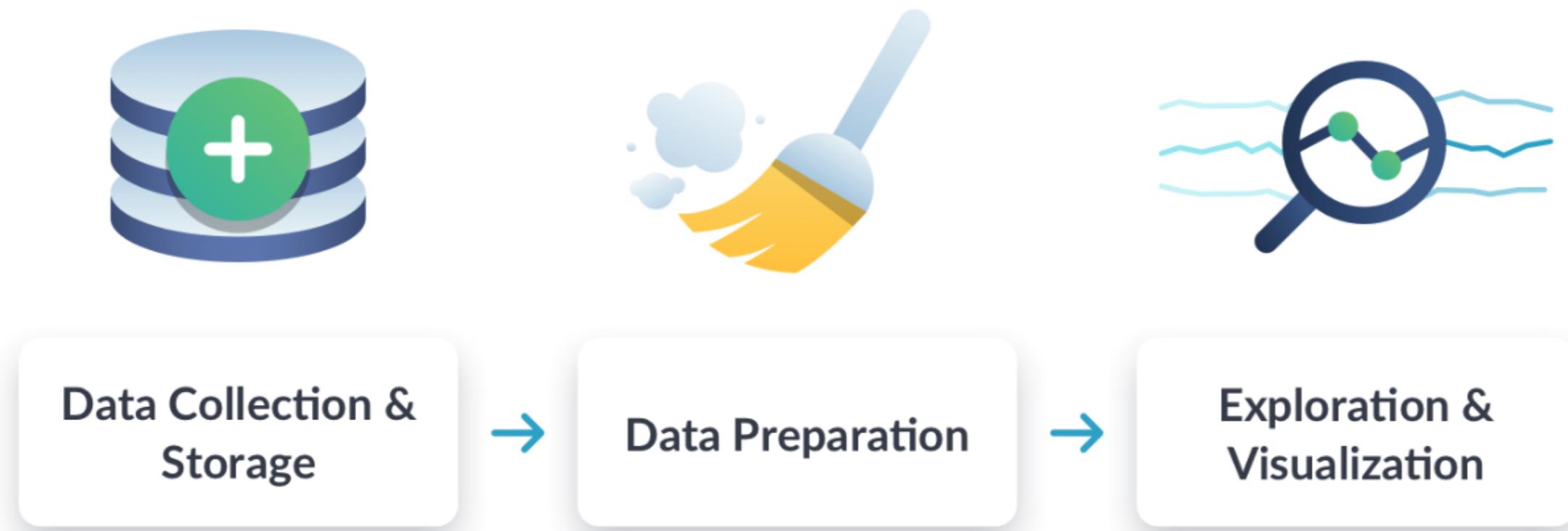


**Data Collection &  
Storage**

# Data workflow



# Data workflow



# Data workflow



# 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

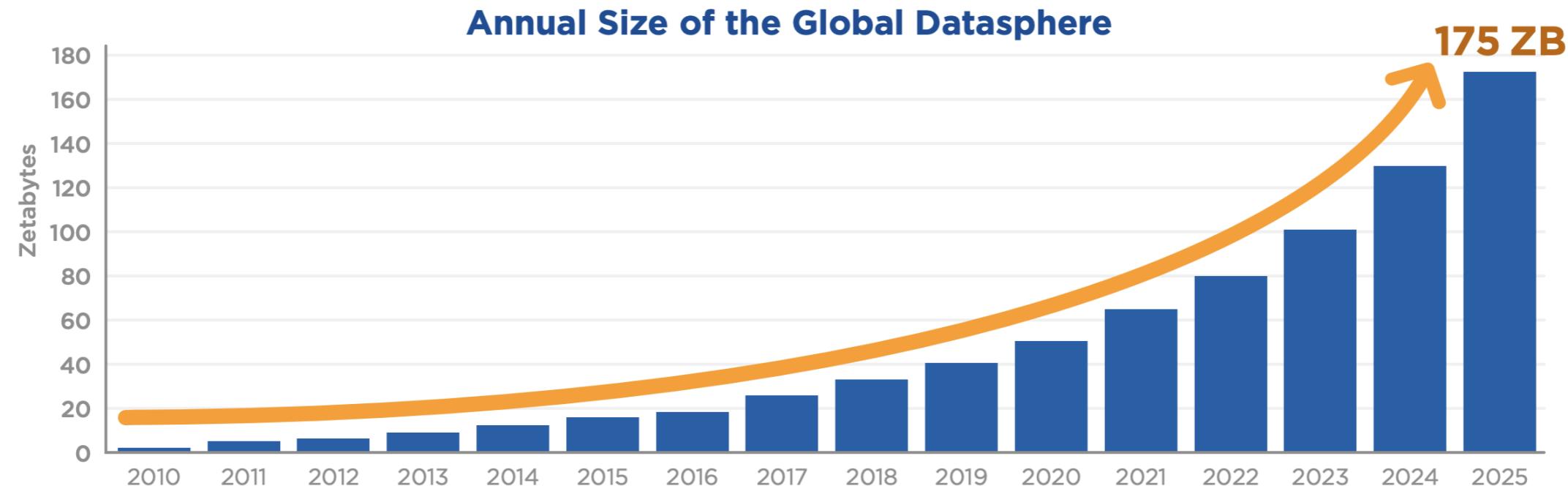
- Big data becomes the norm =>

# 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)



<sup>1</sup> 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!**

**UNDERSTANDING DATA ENGINEERING**

# Data engineers vs. data scientists

UNDERSTANDING DATA ENGINEERING



**Hadrien Lacroix**

Content Developer at DataCamp

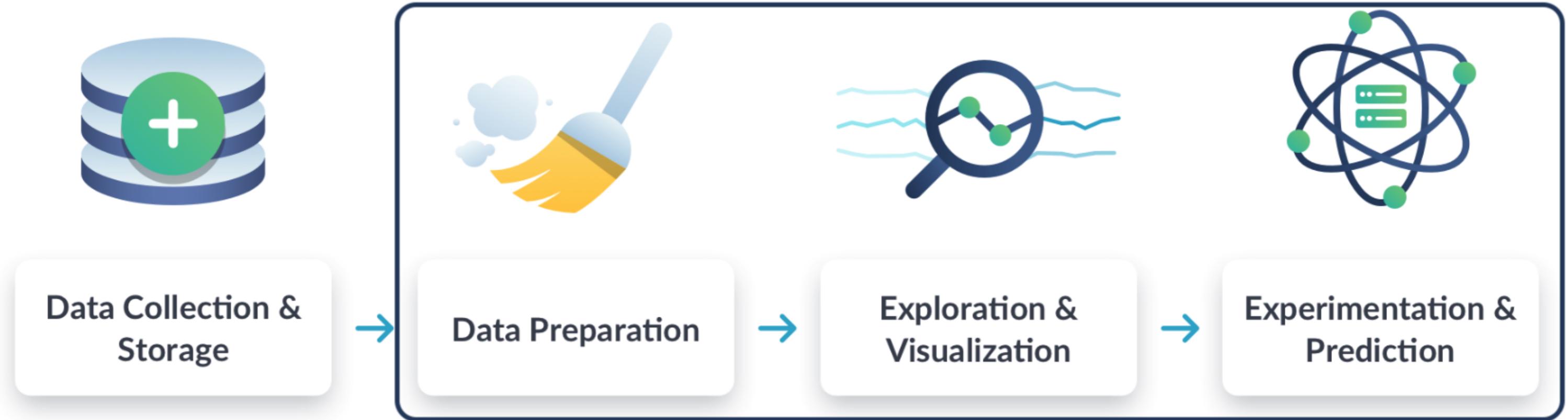
# Data workflow



# Data engineers



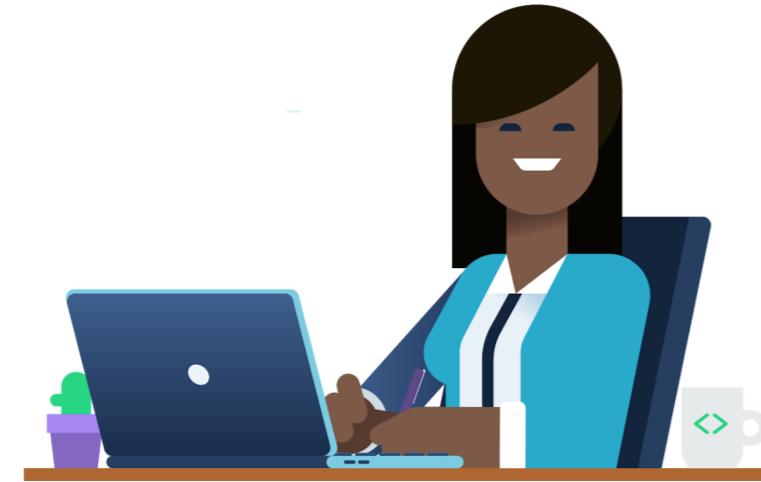
# Data scientists



# 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!**

**UNDERSTANDING DATA ENGINEERING**

# The data pipeline

## UNDERSTANDING DATA ENGINEERING



**Hadrien Lacroix**

Content Developer at DataCamp

# If data is the new oil...



<sup>1</sup> The Economist, 2017-05-06, by David Parkins









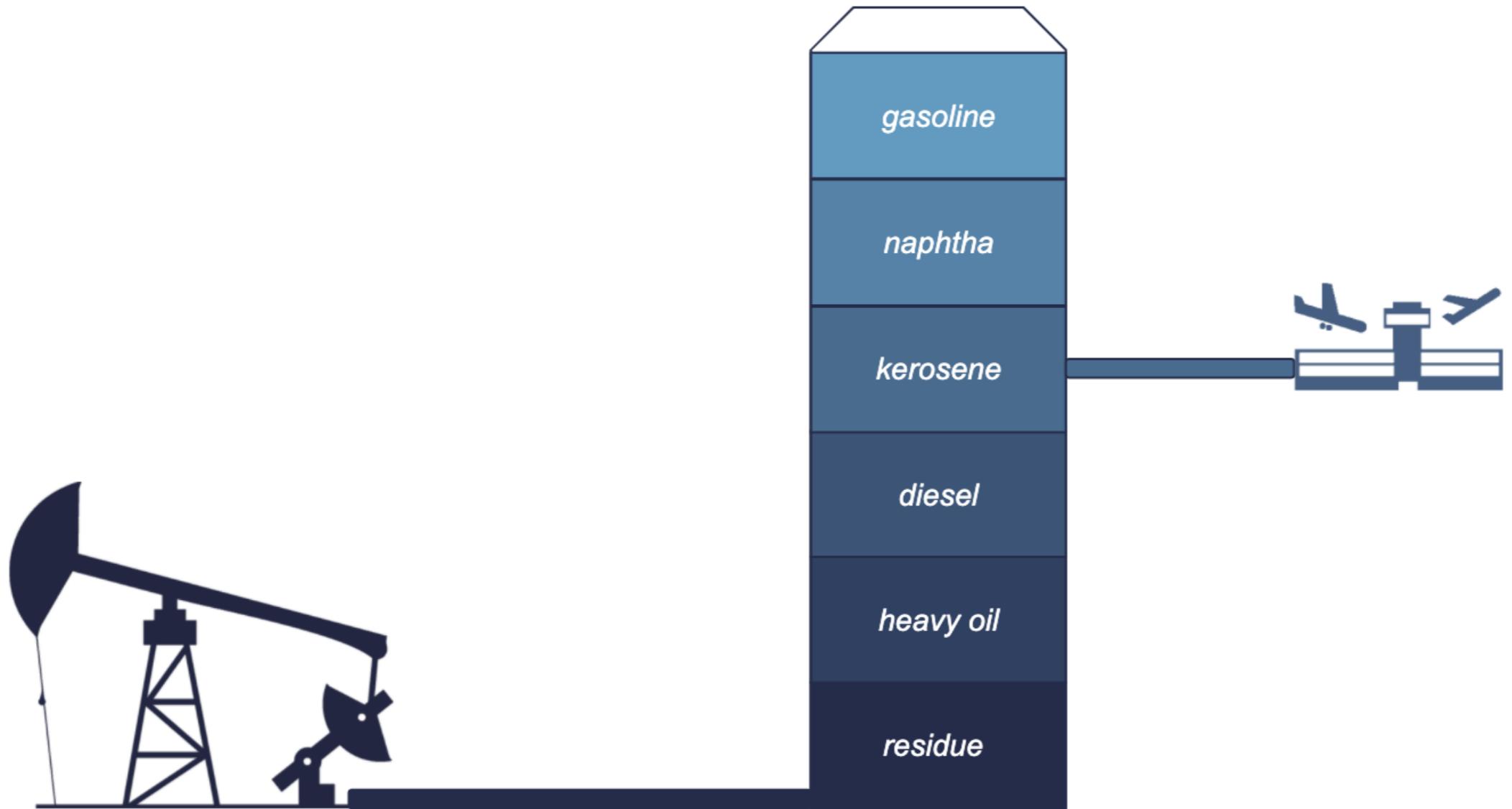


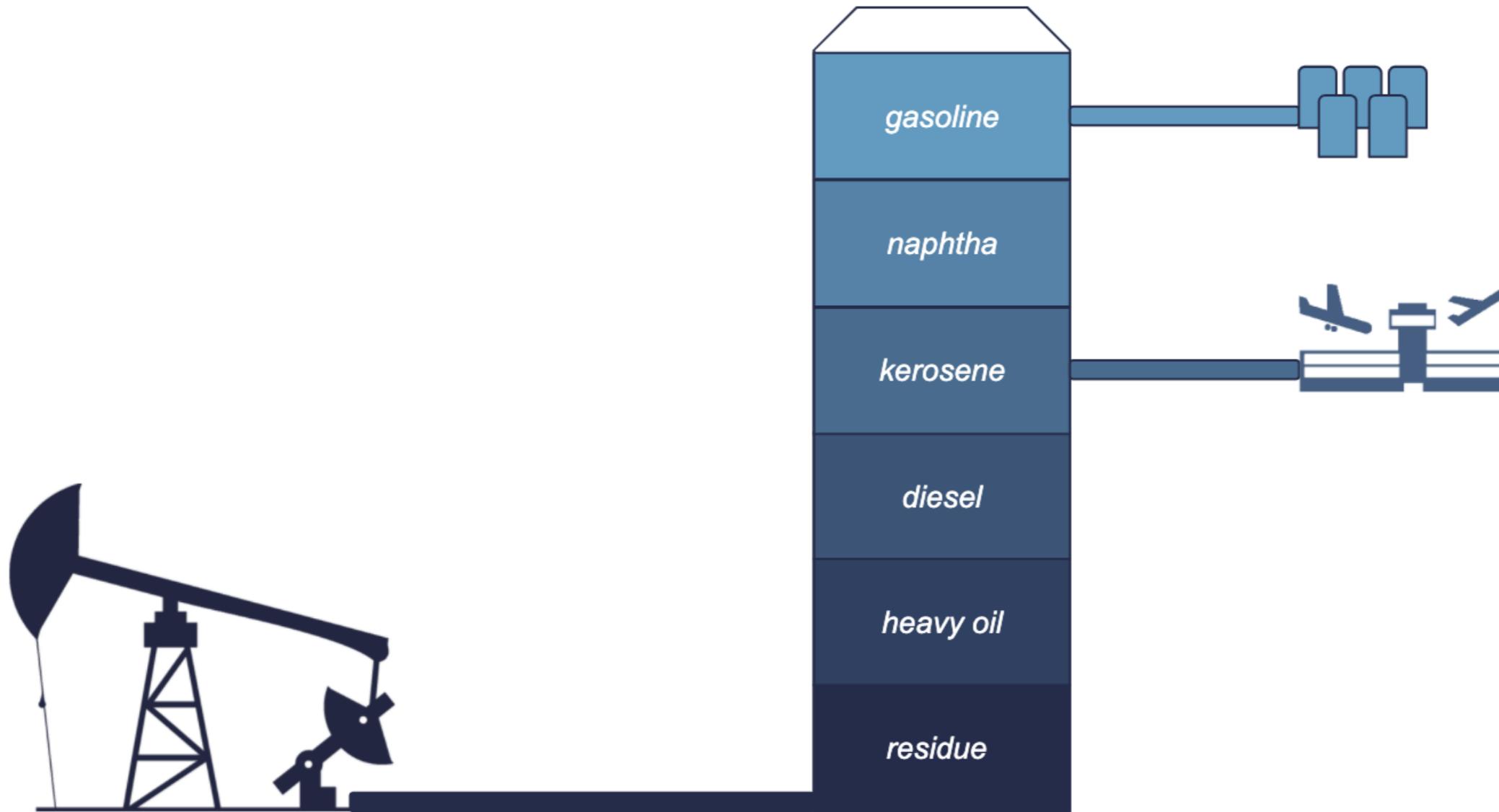


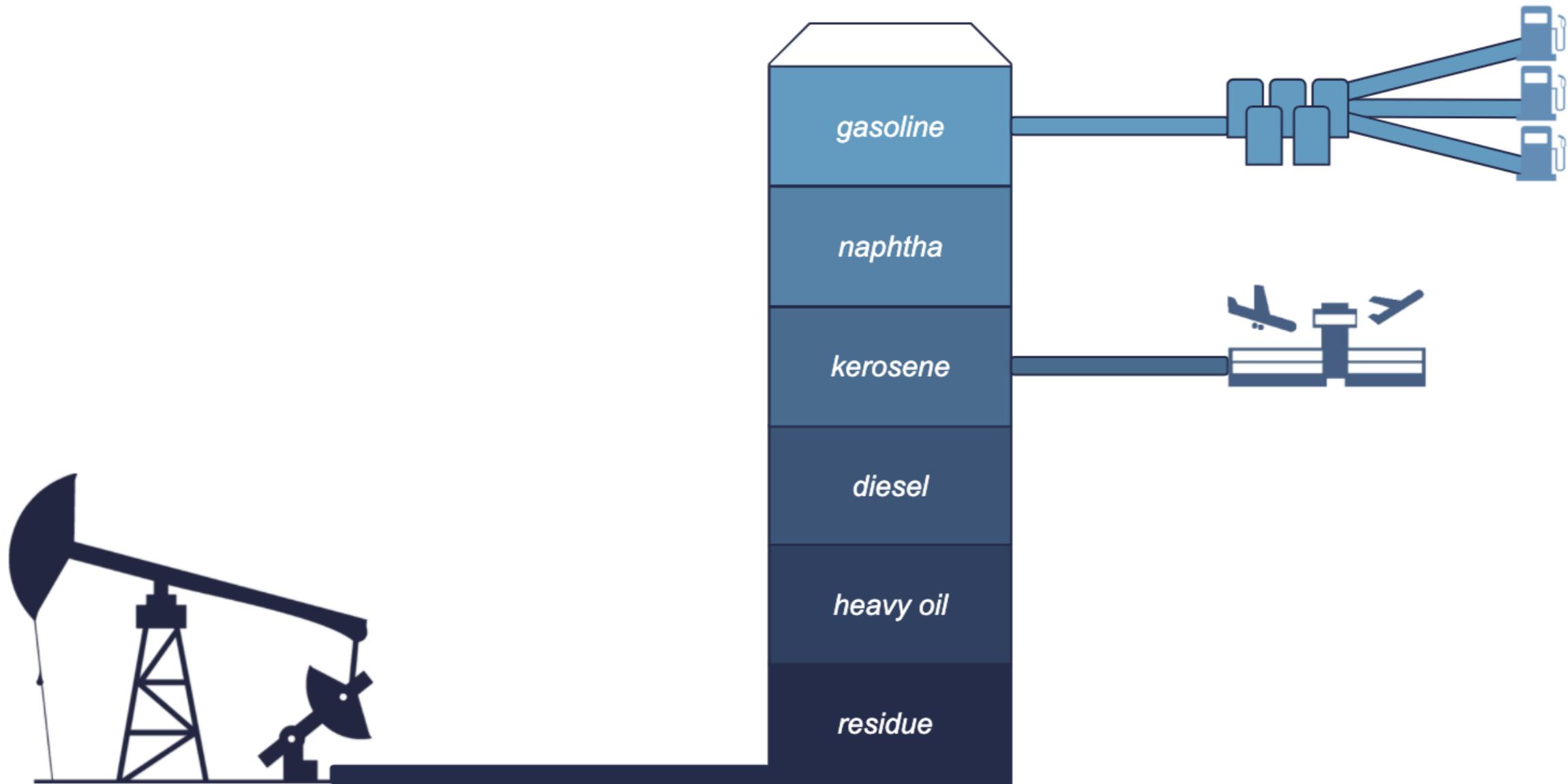


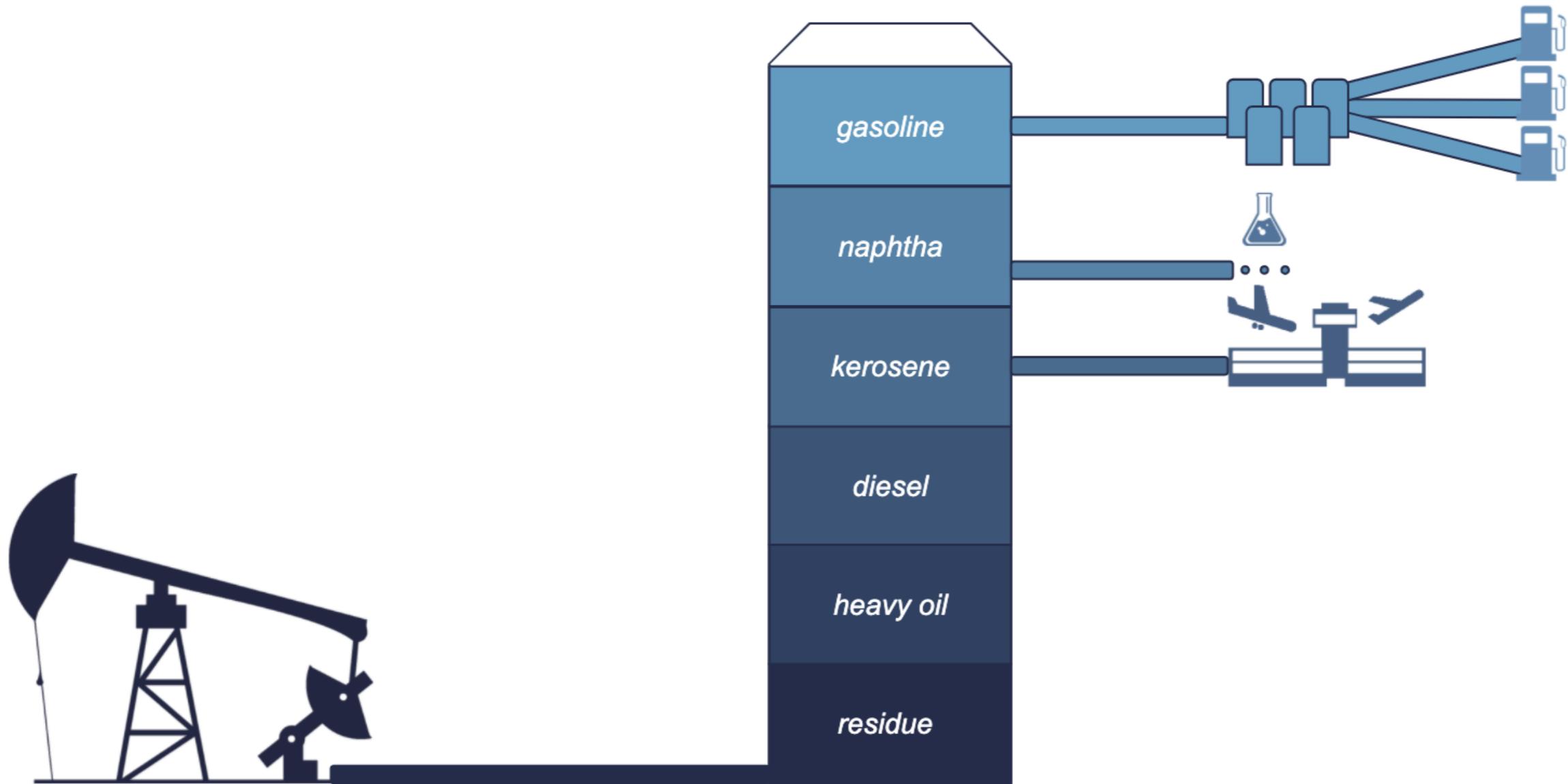


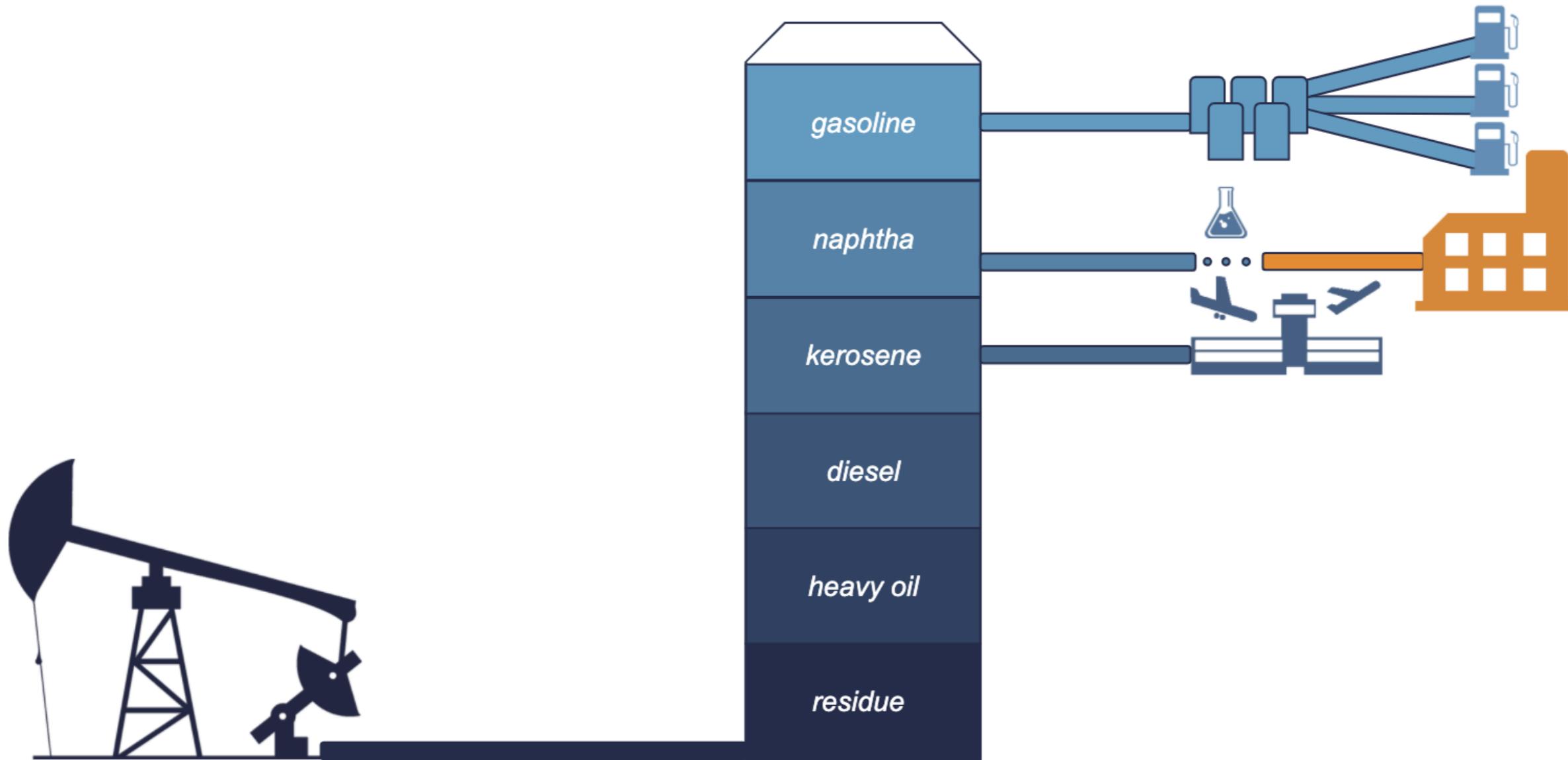












# Back to data engineering

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



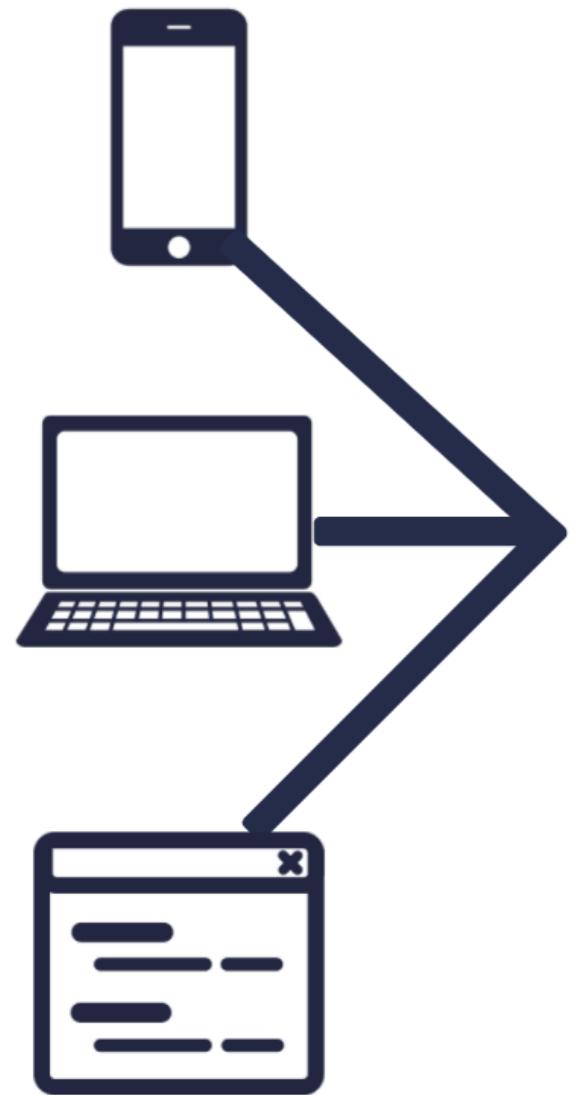


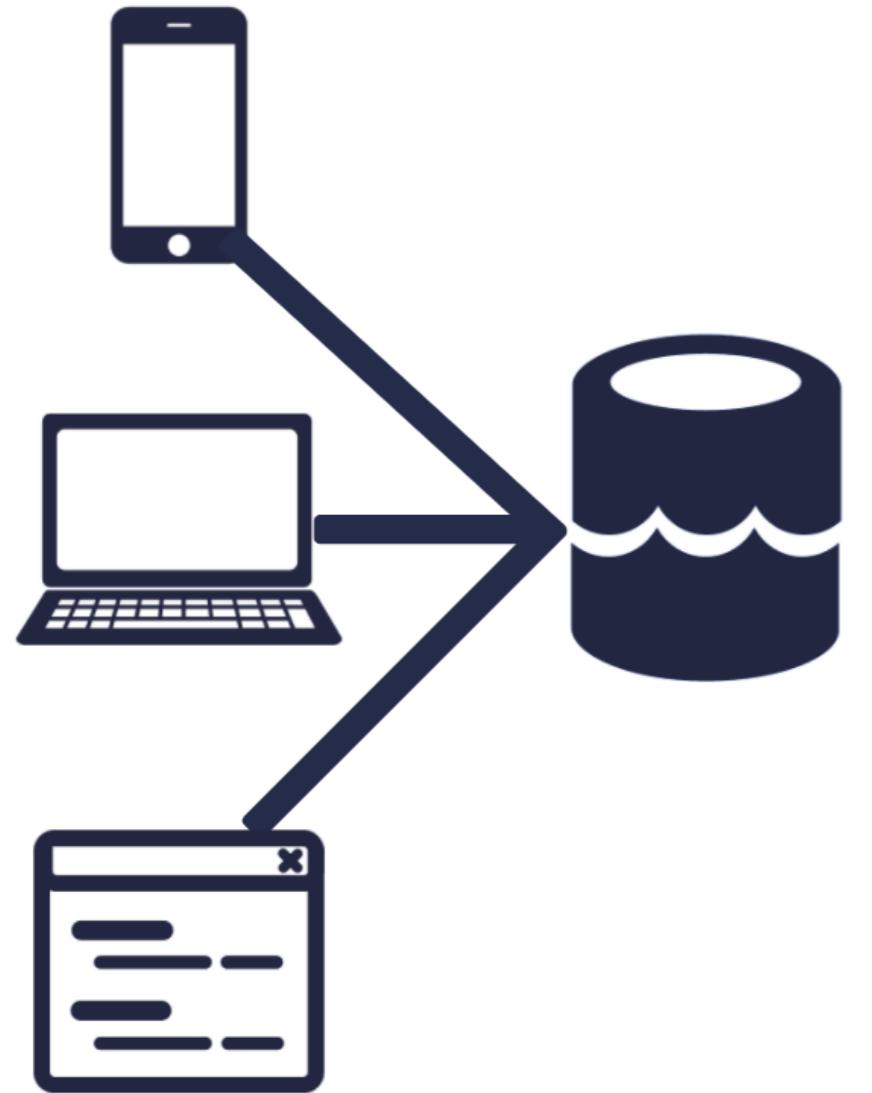


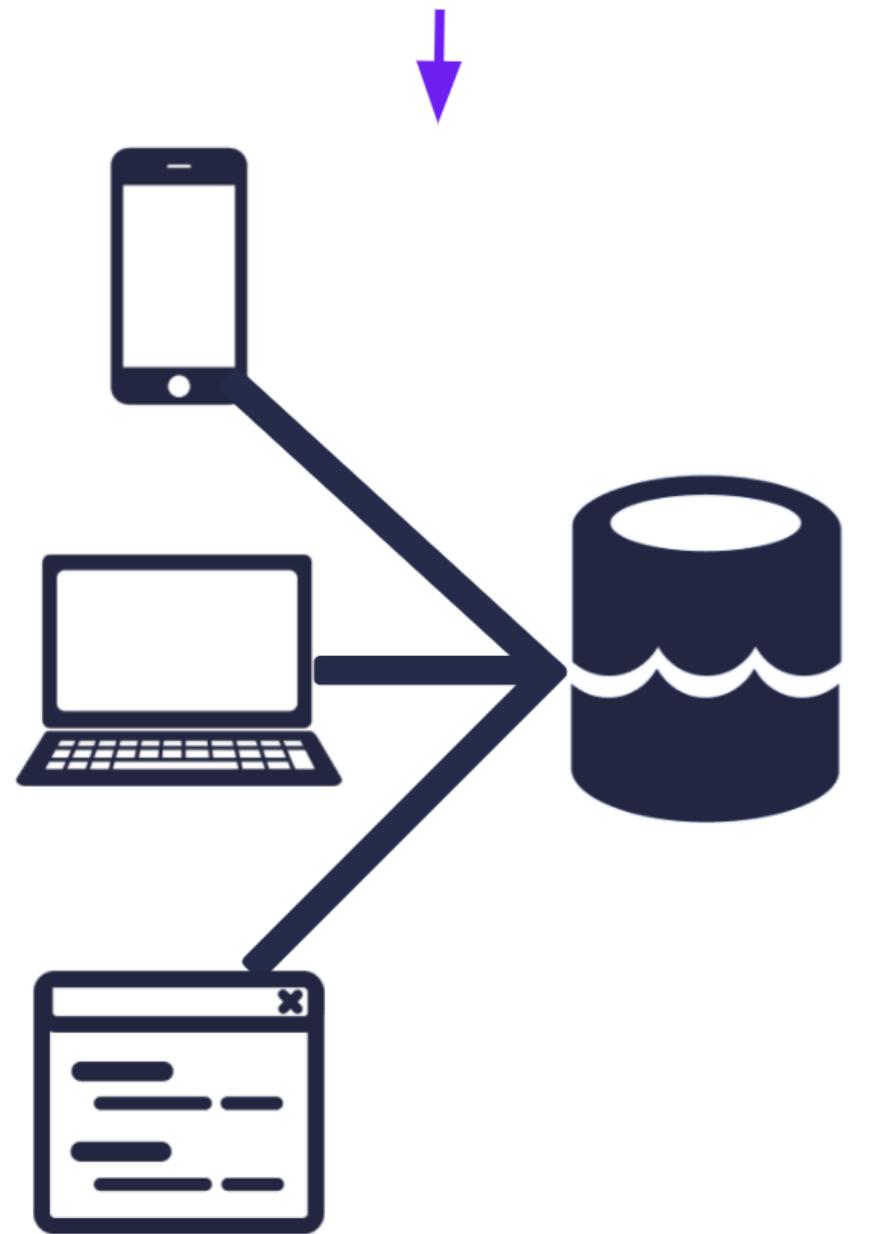


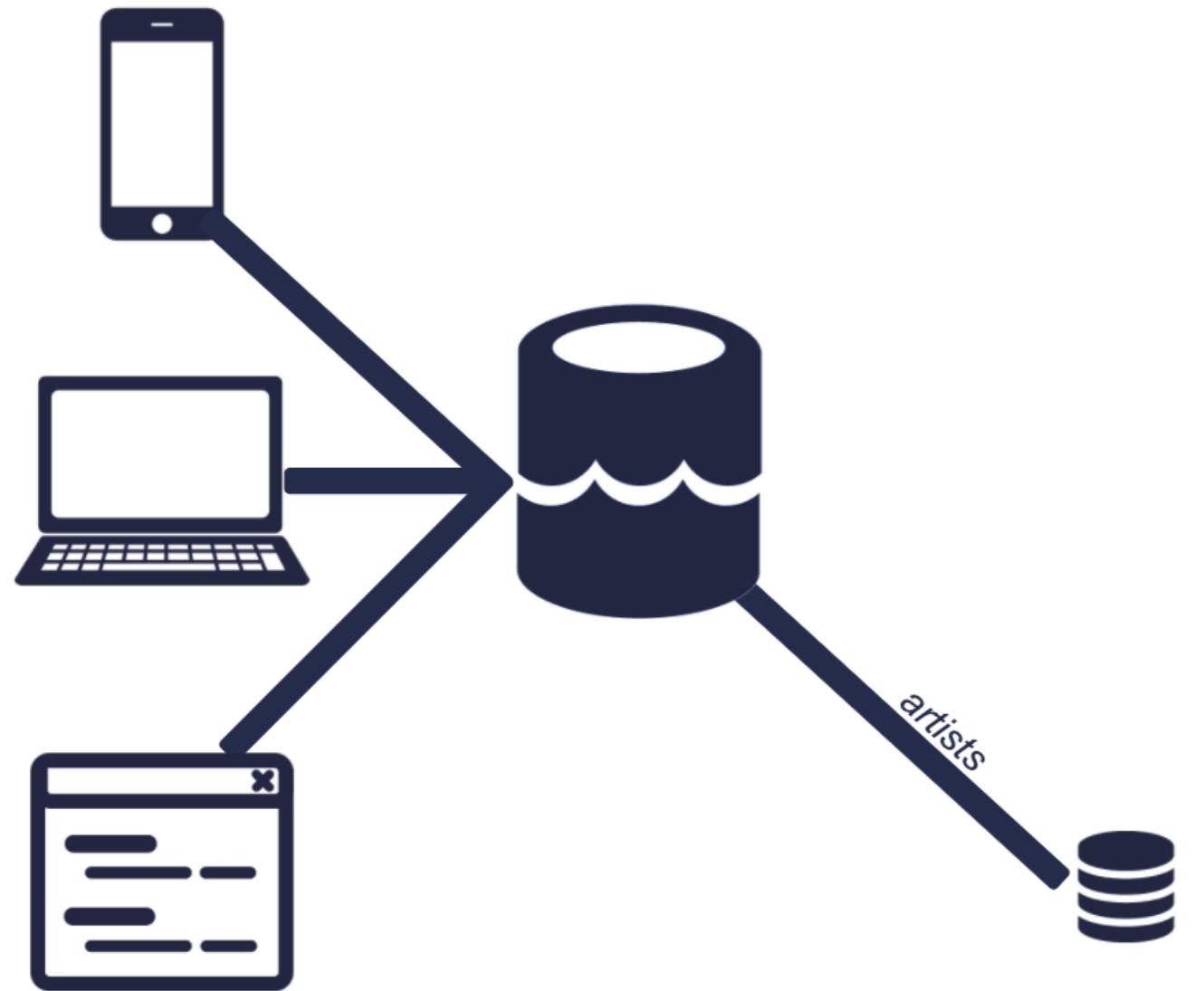


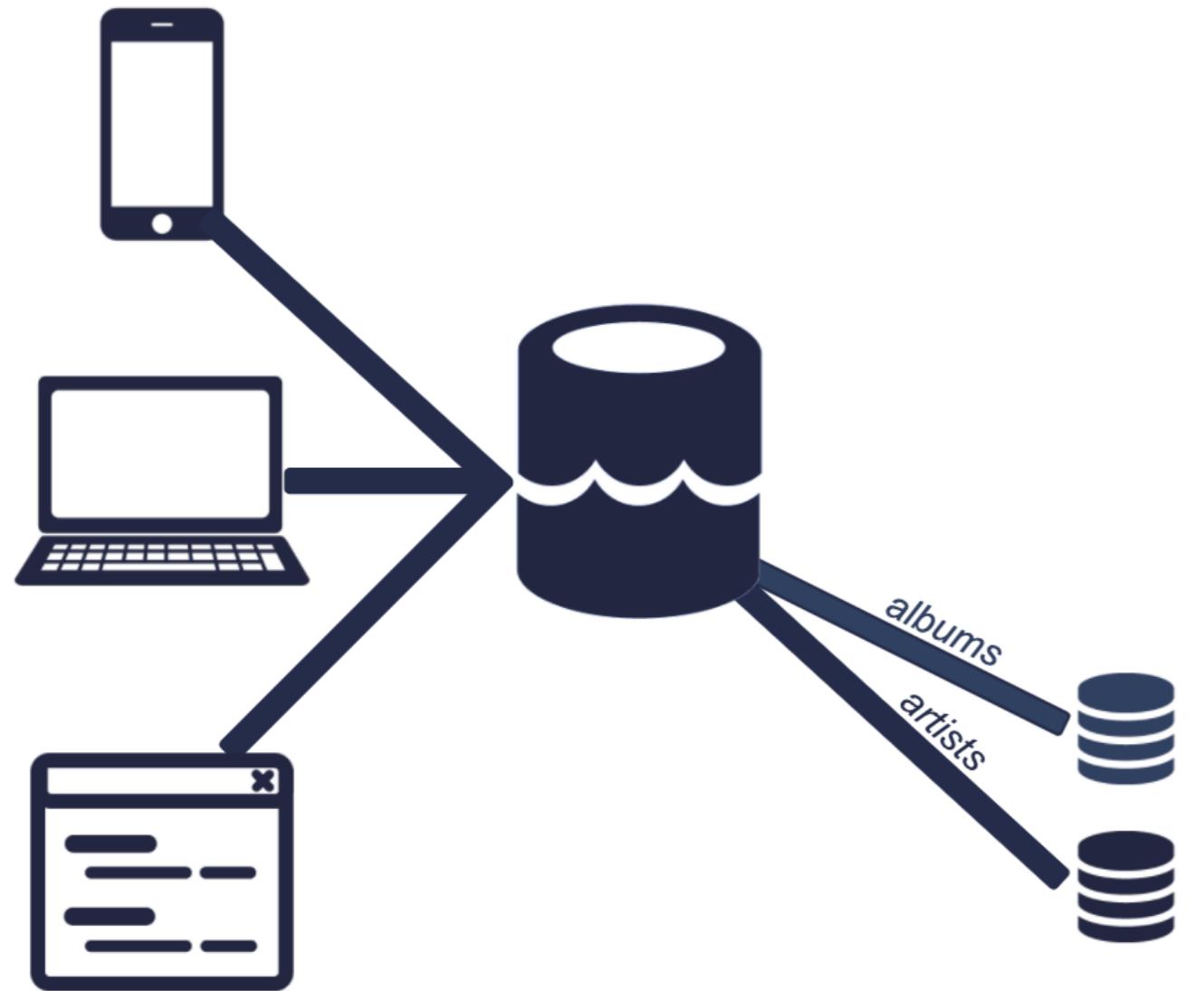


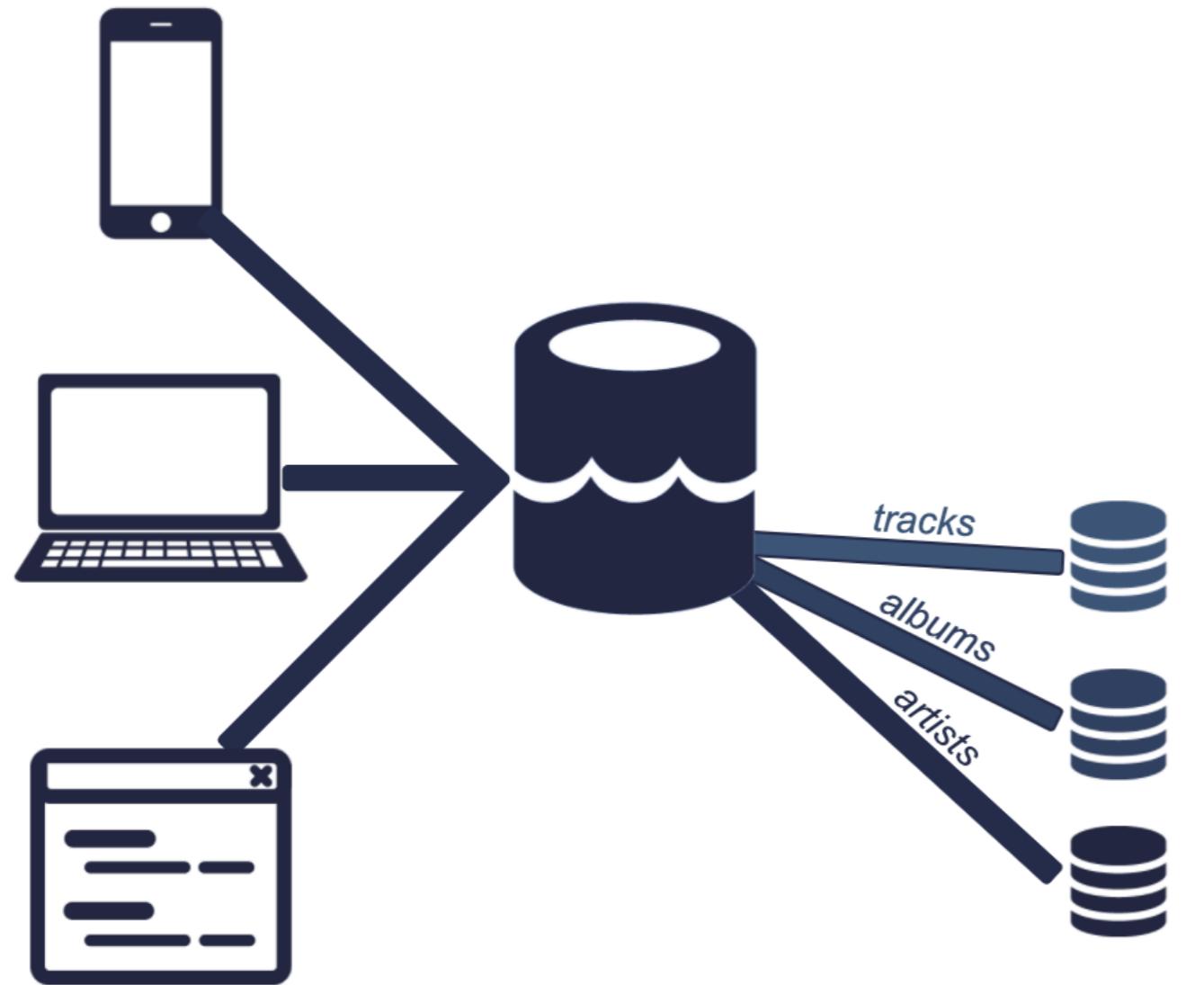


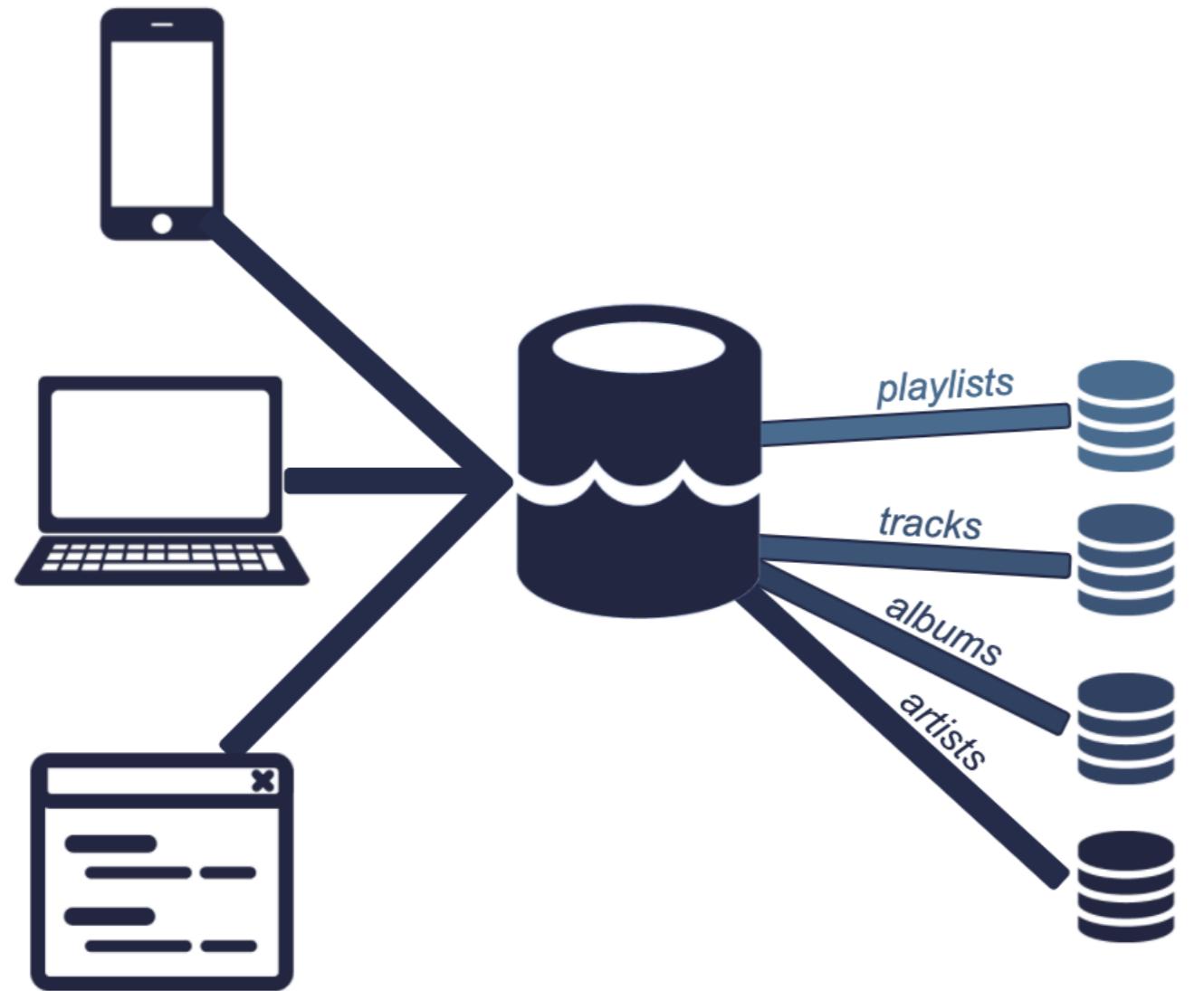


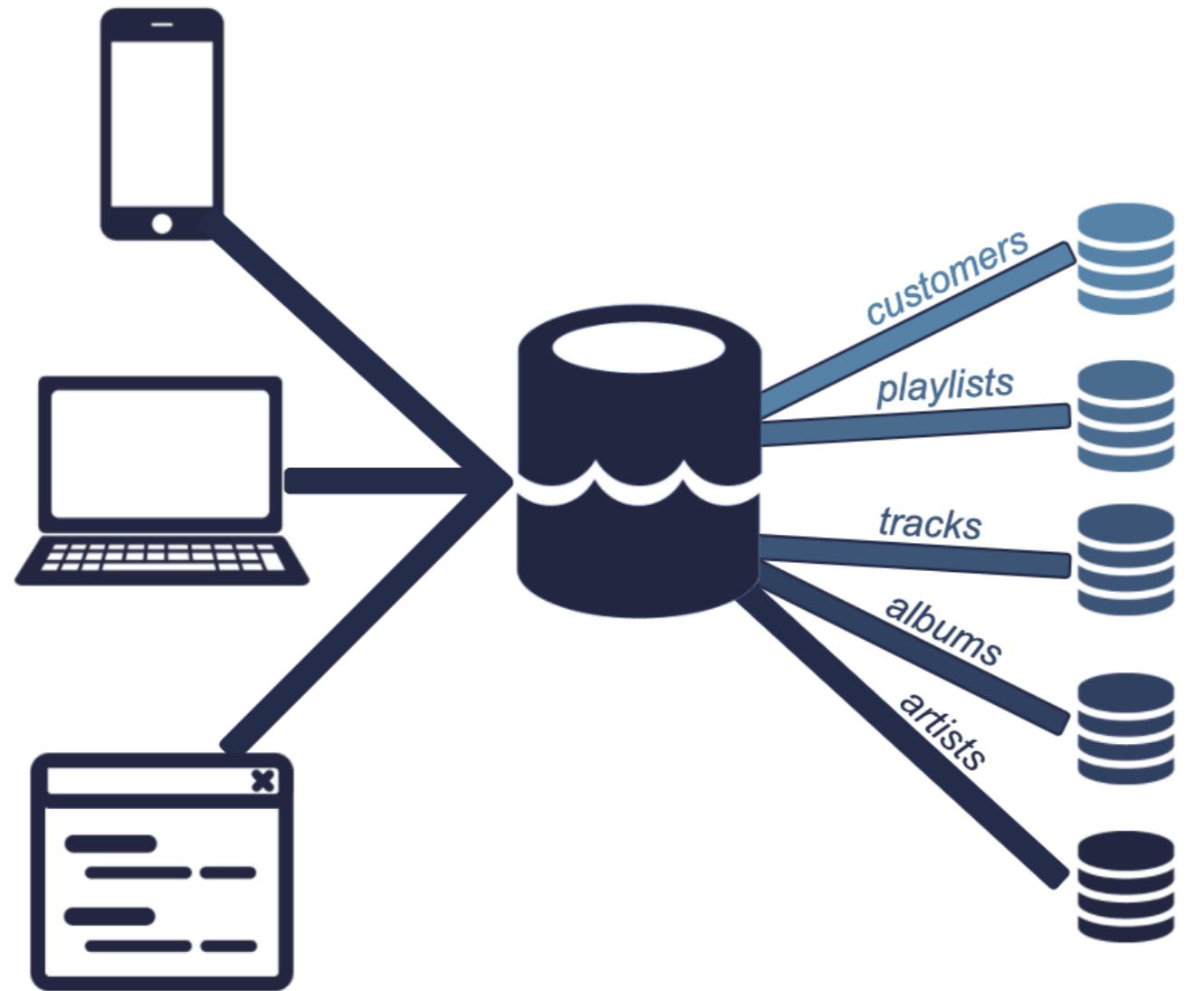


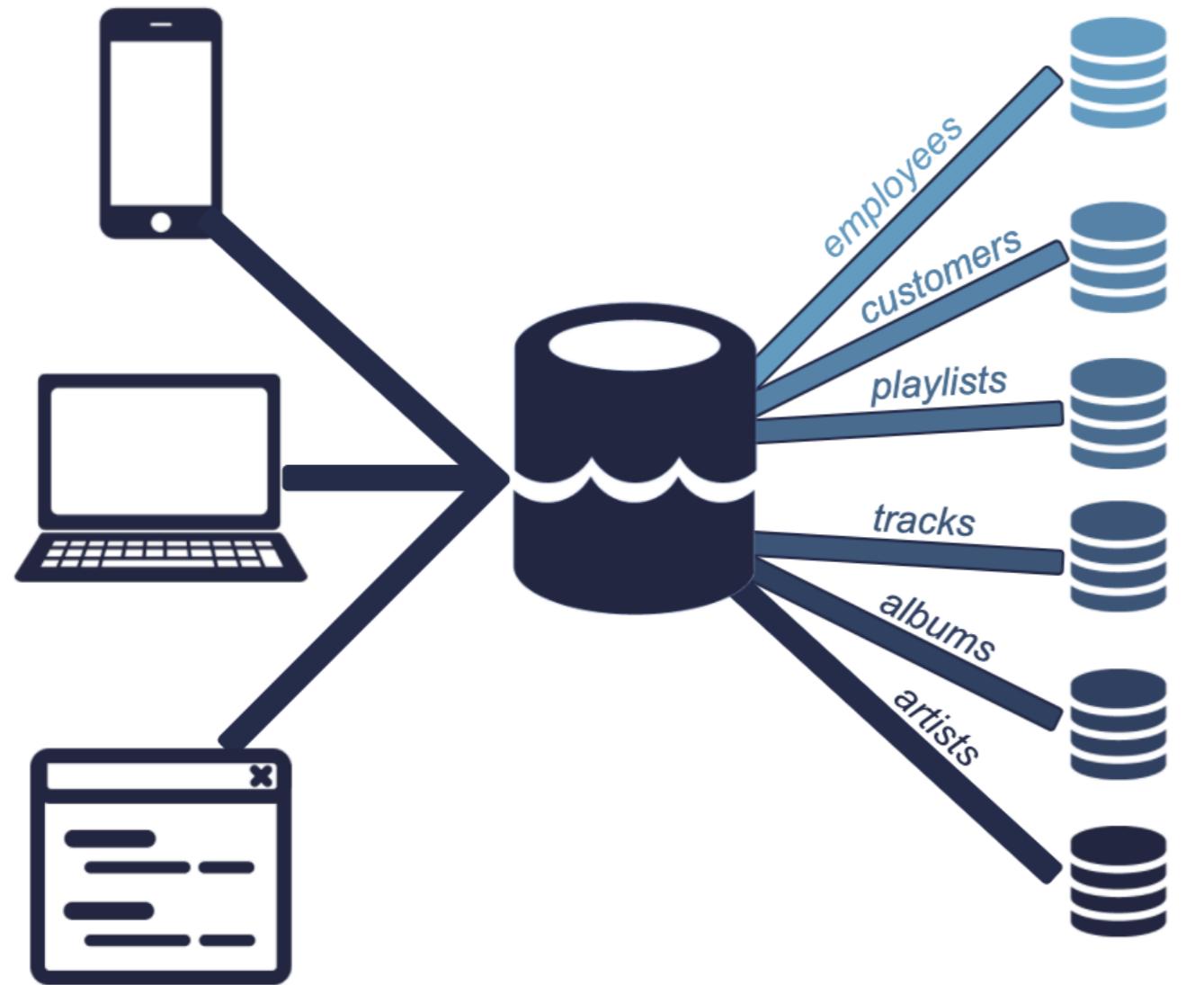


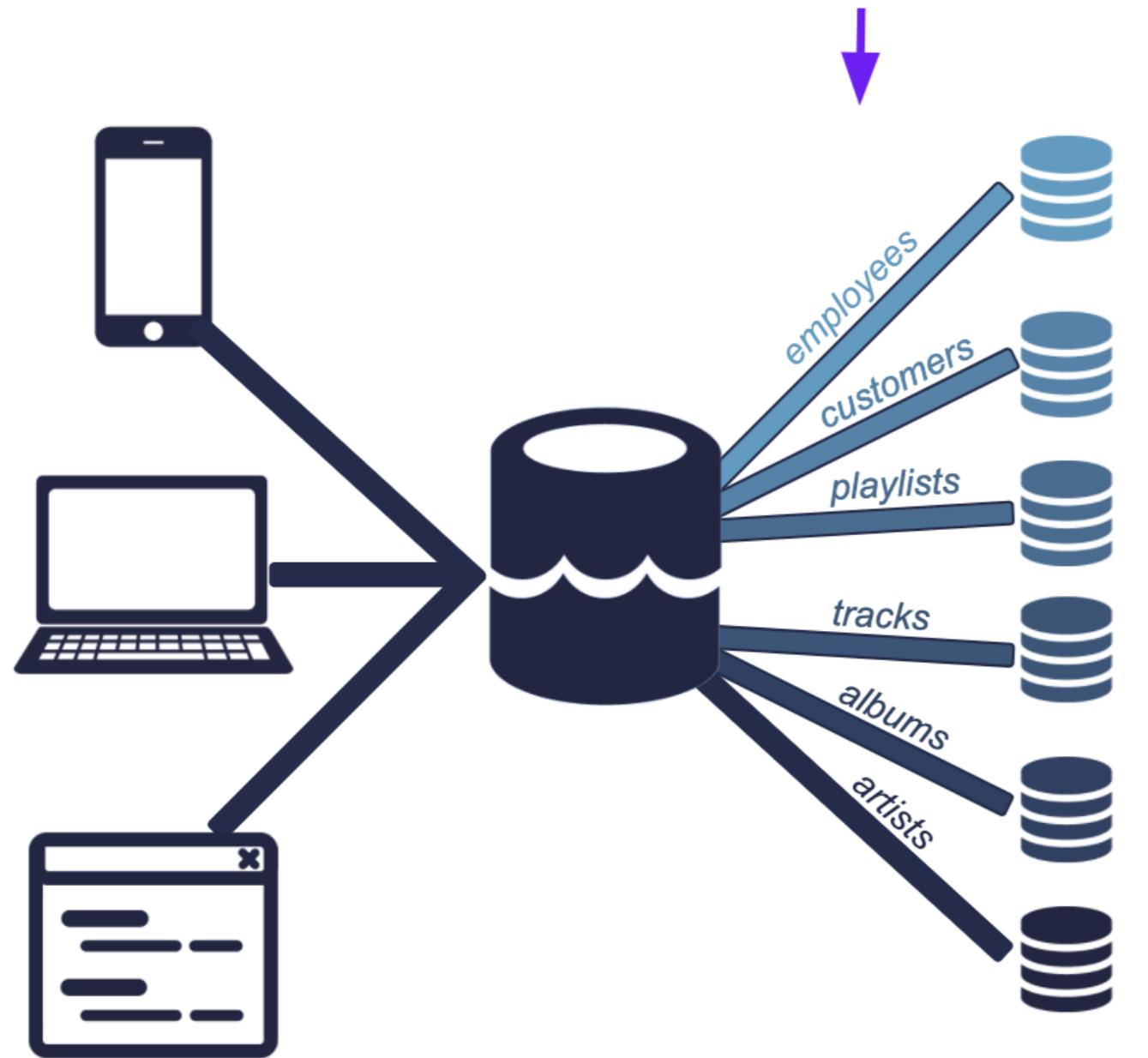


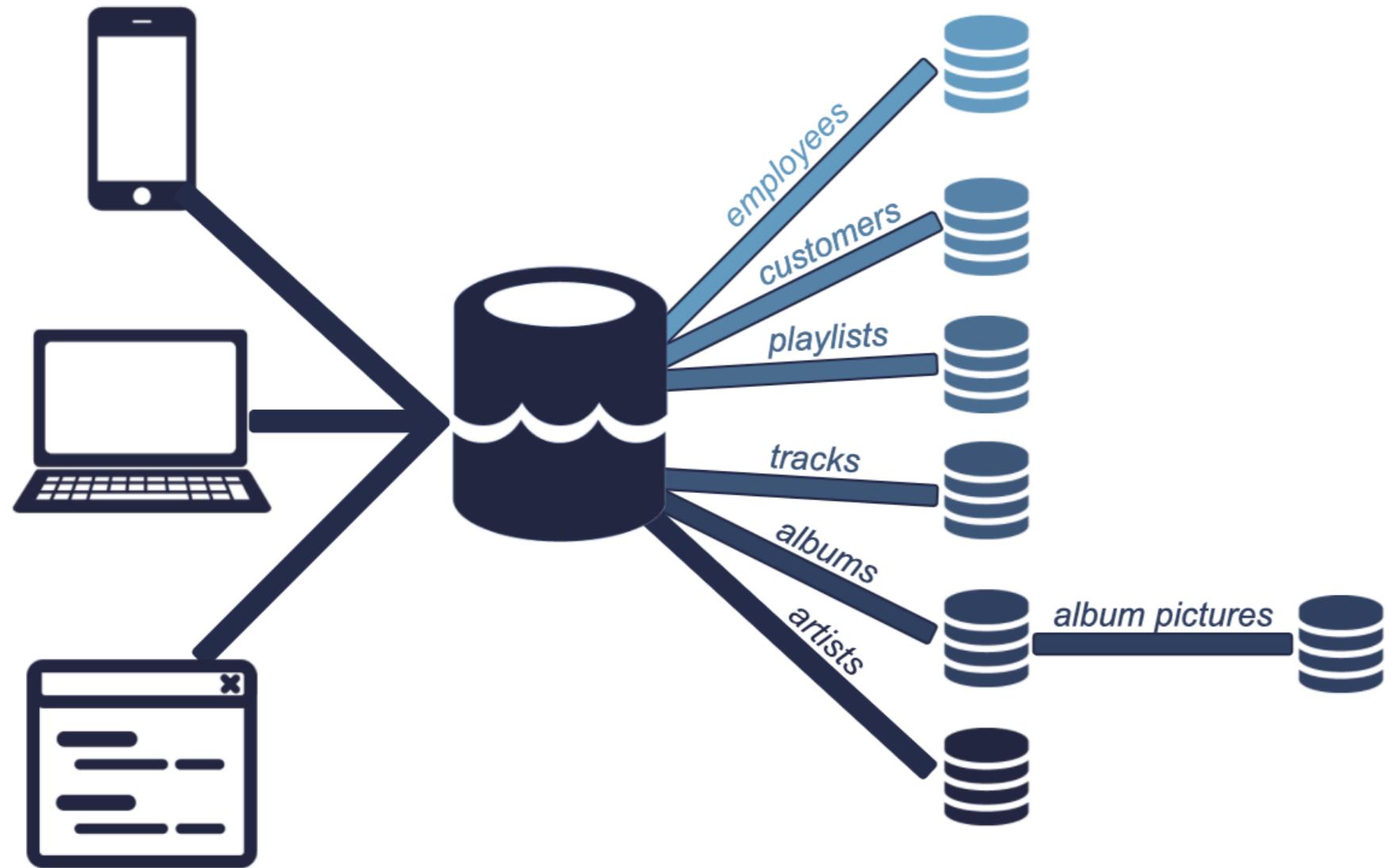


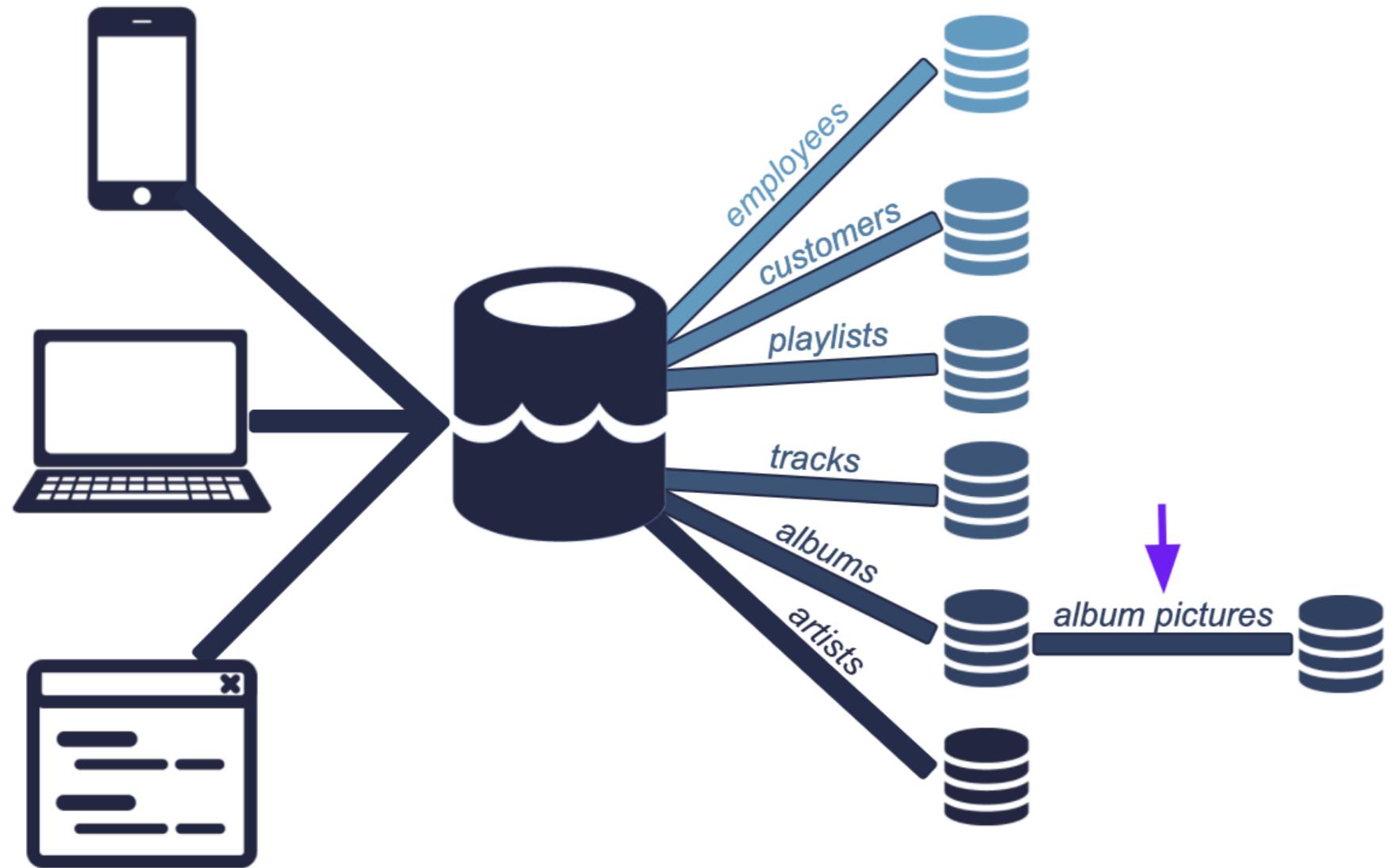


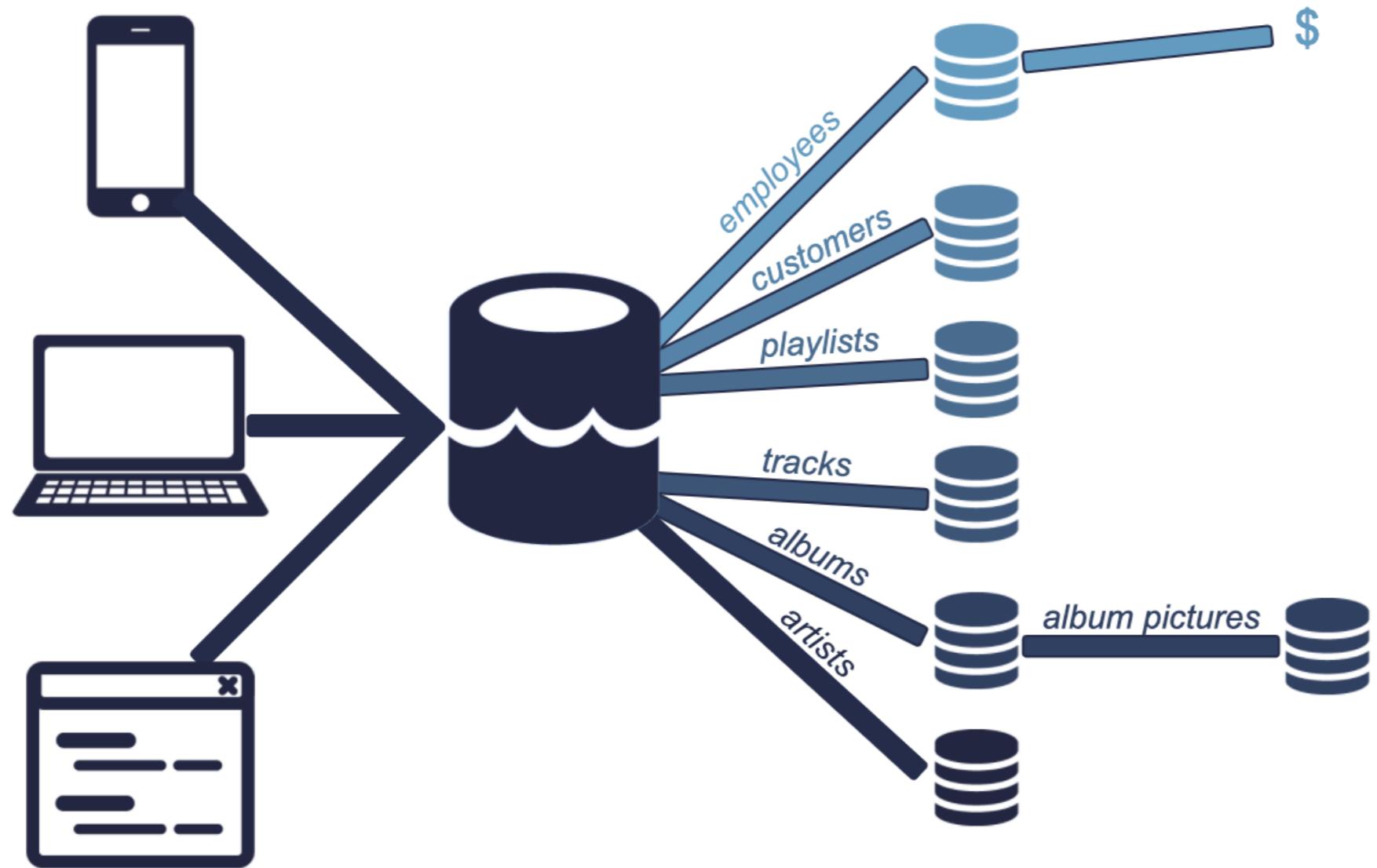


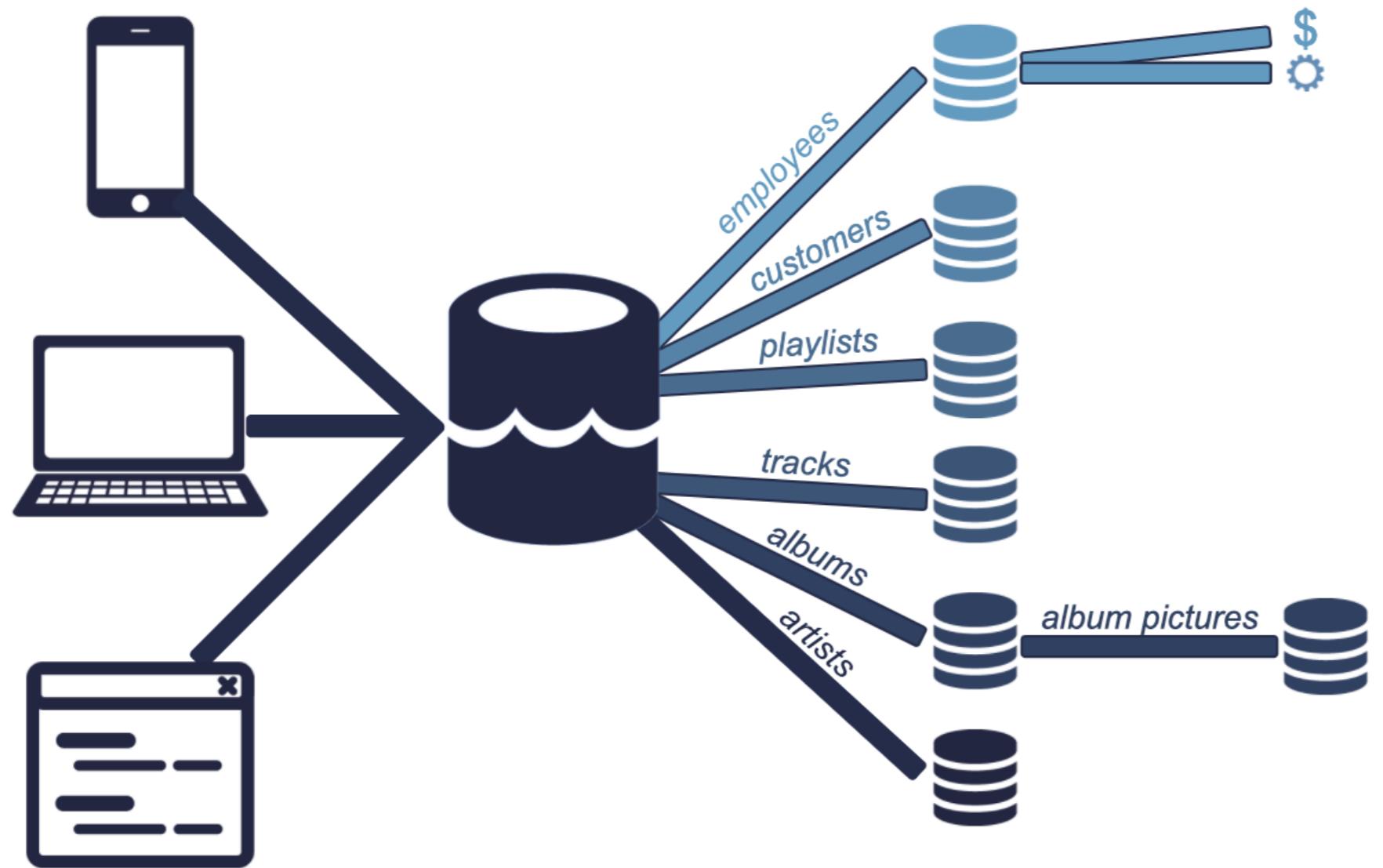


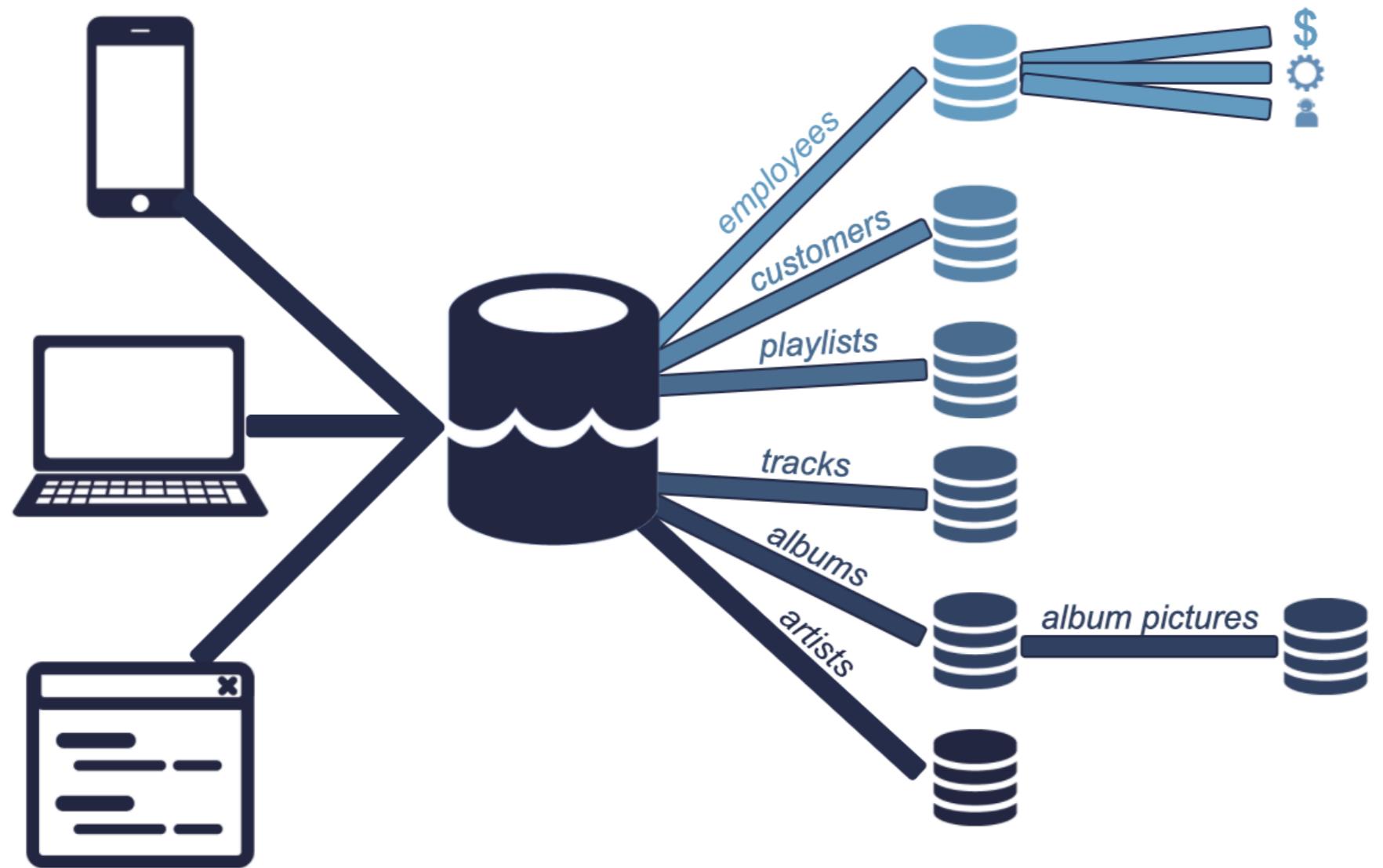


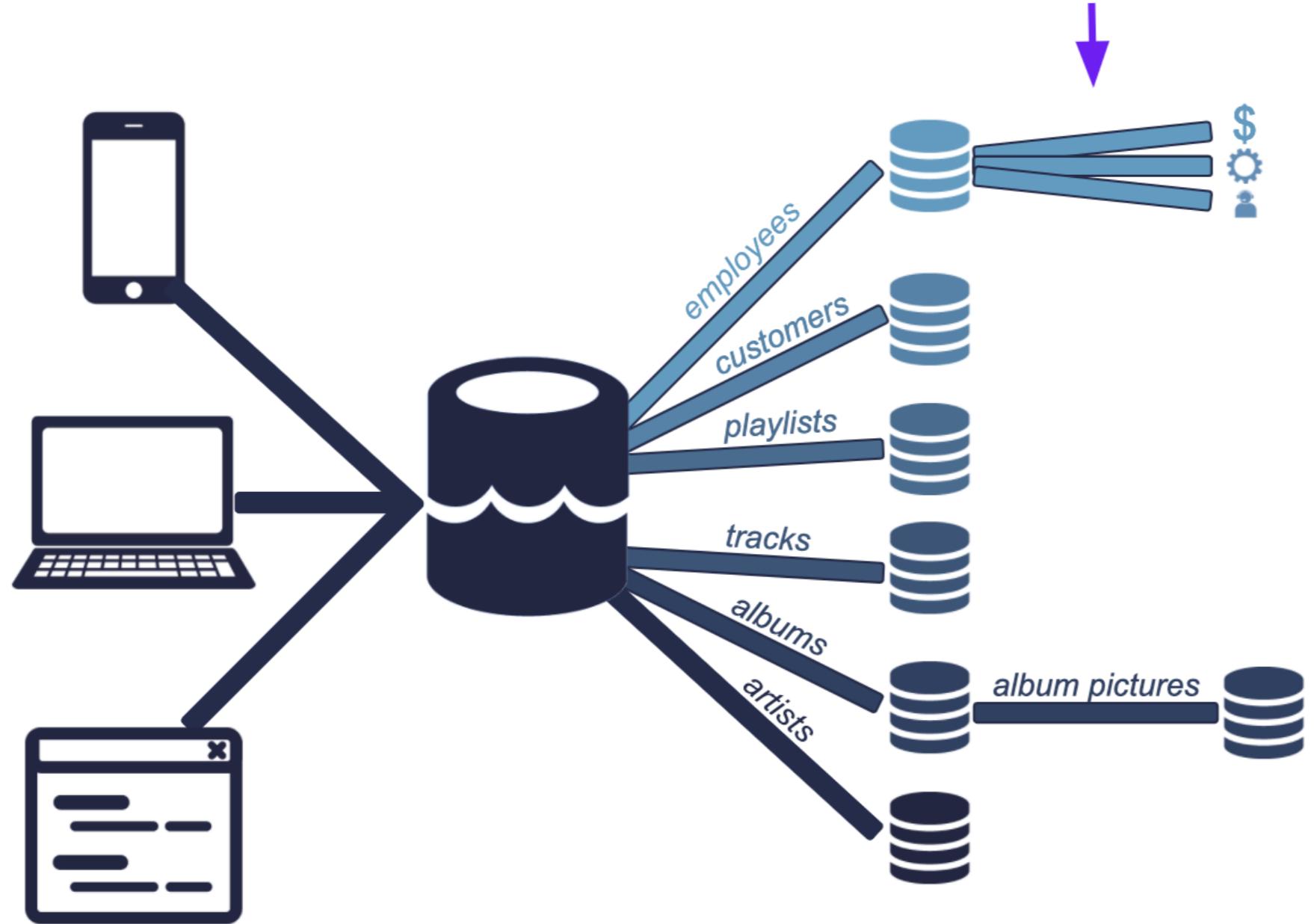


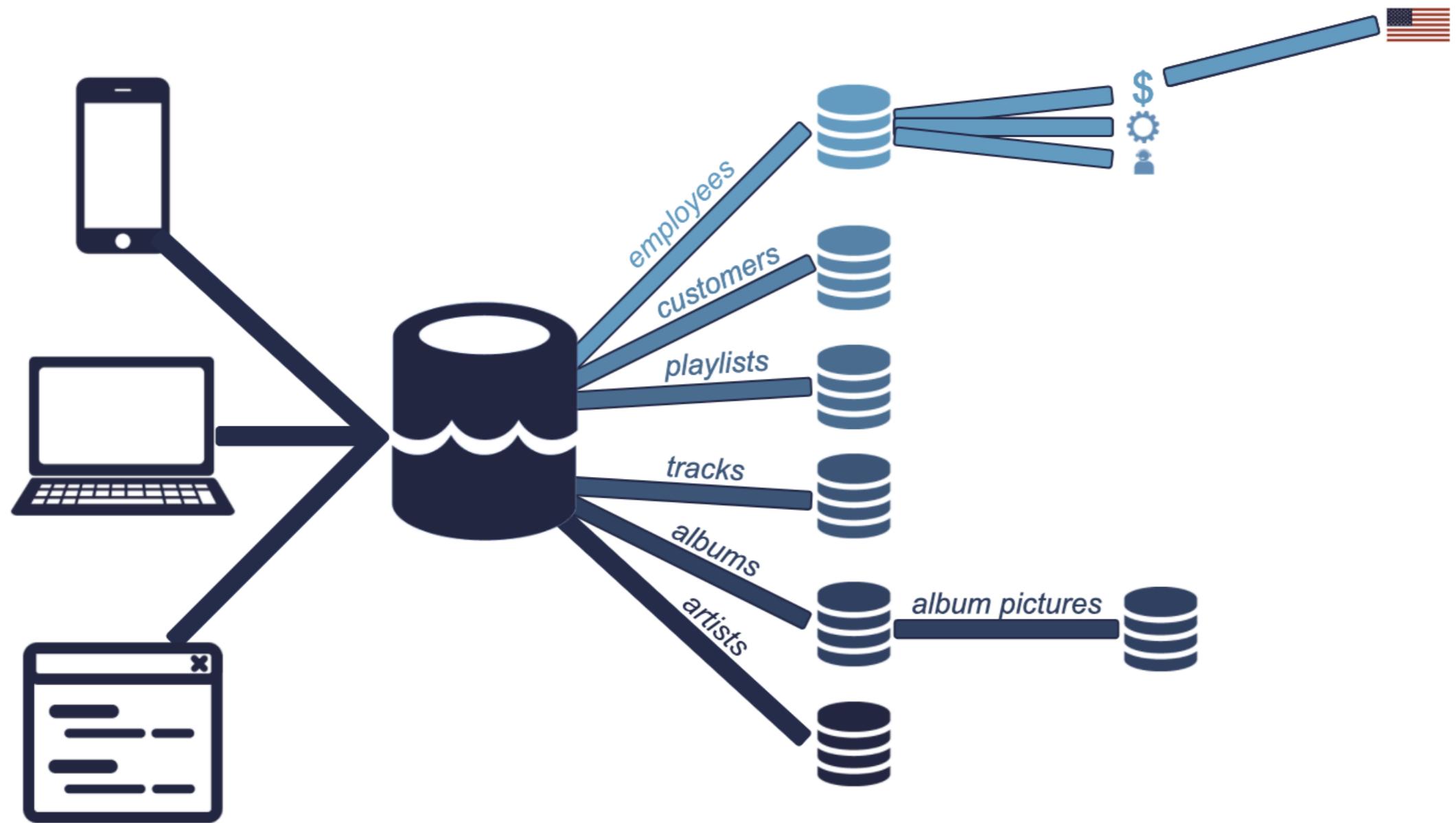


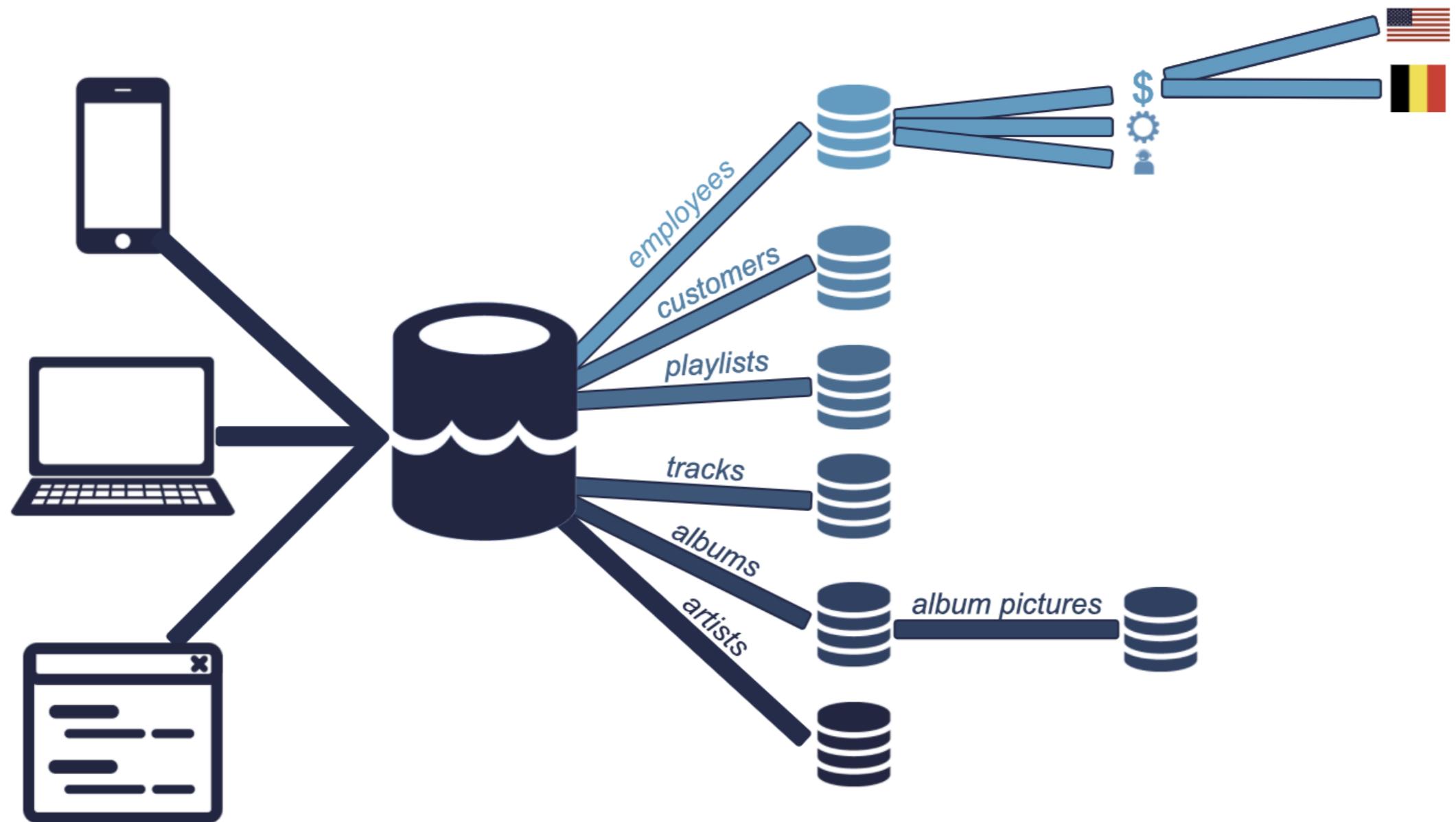


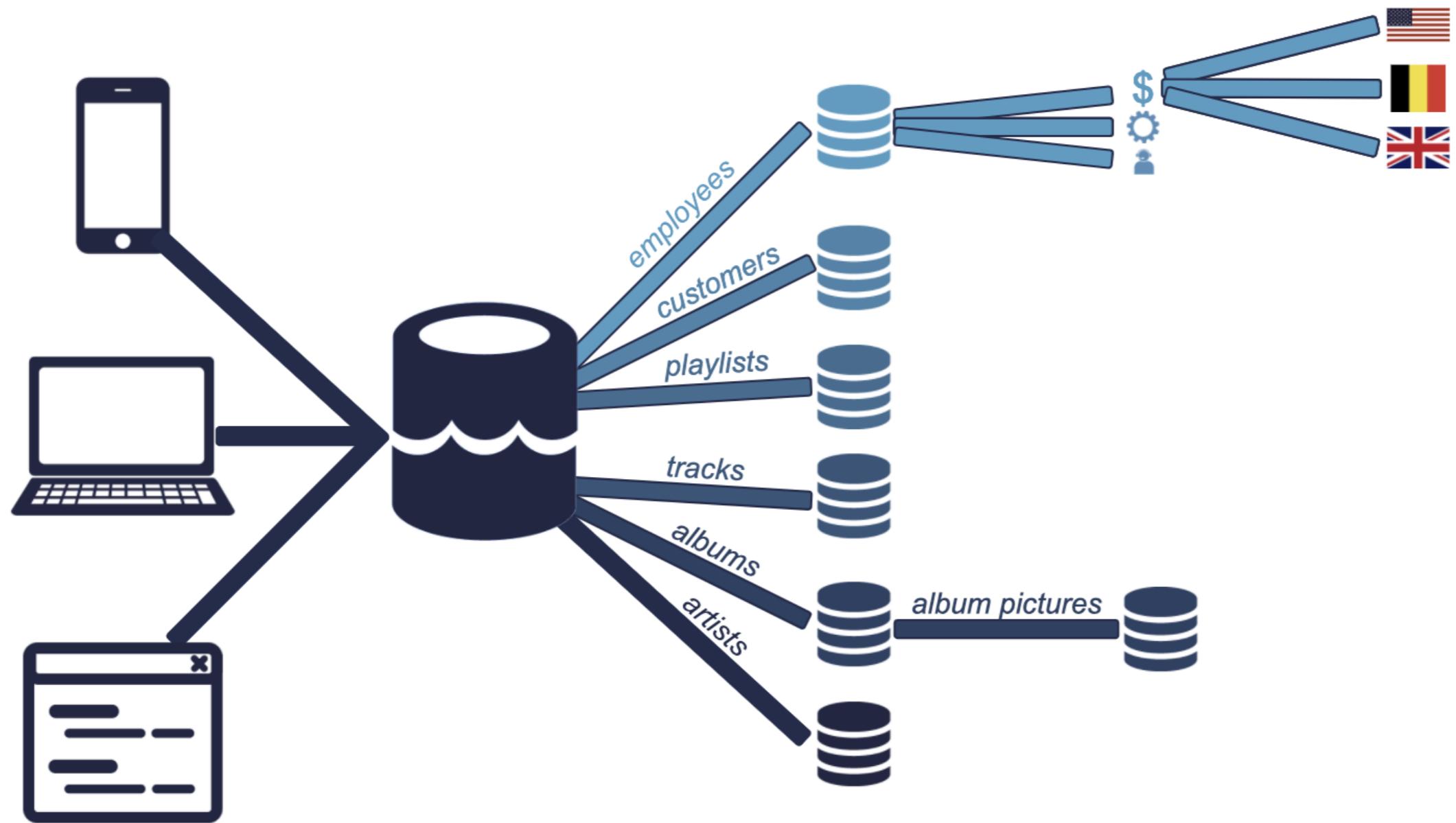


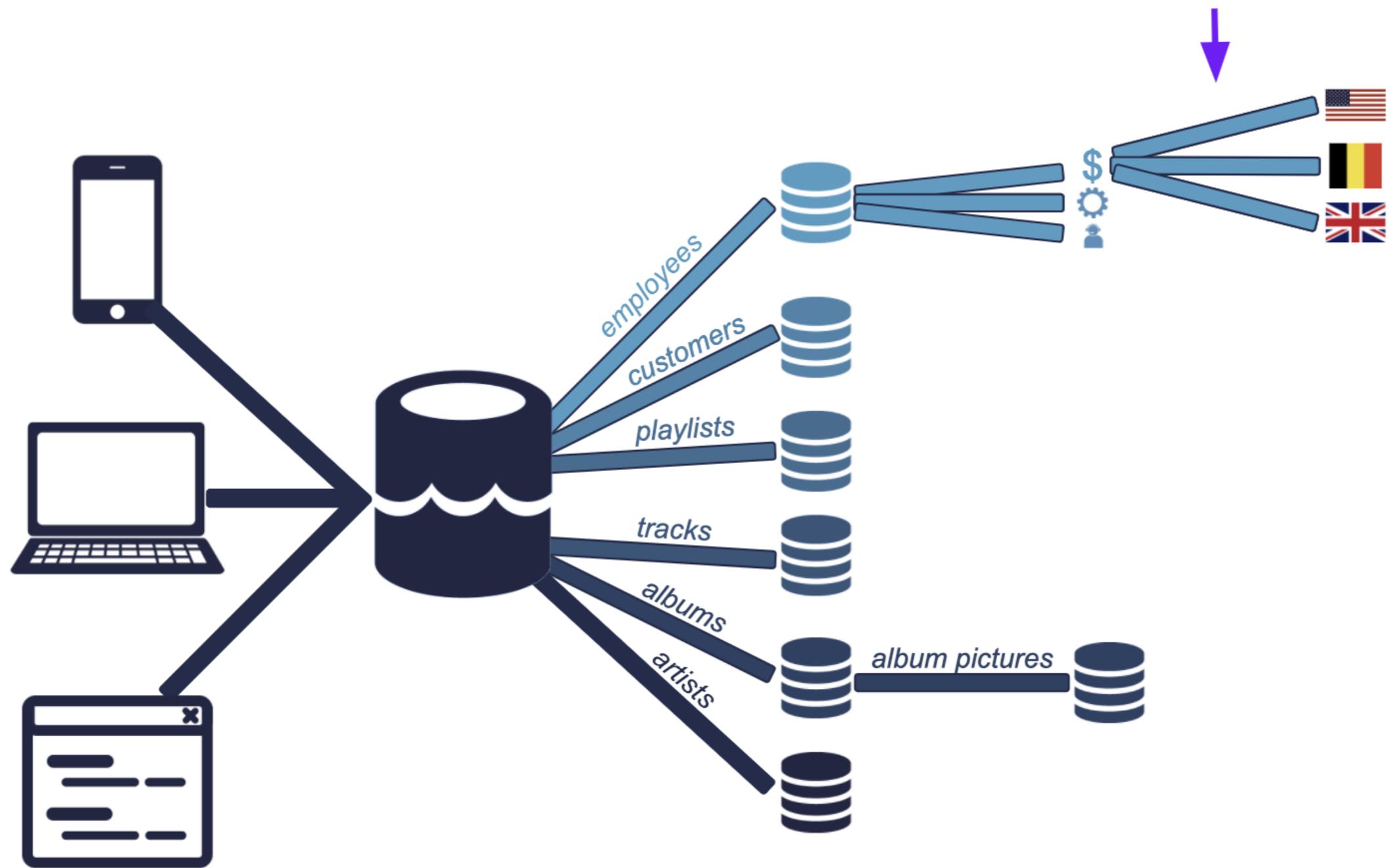


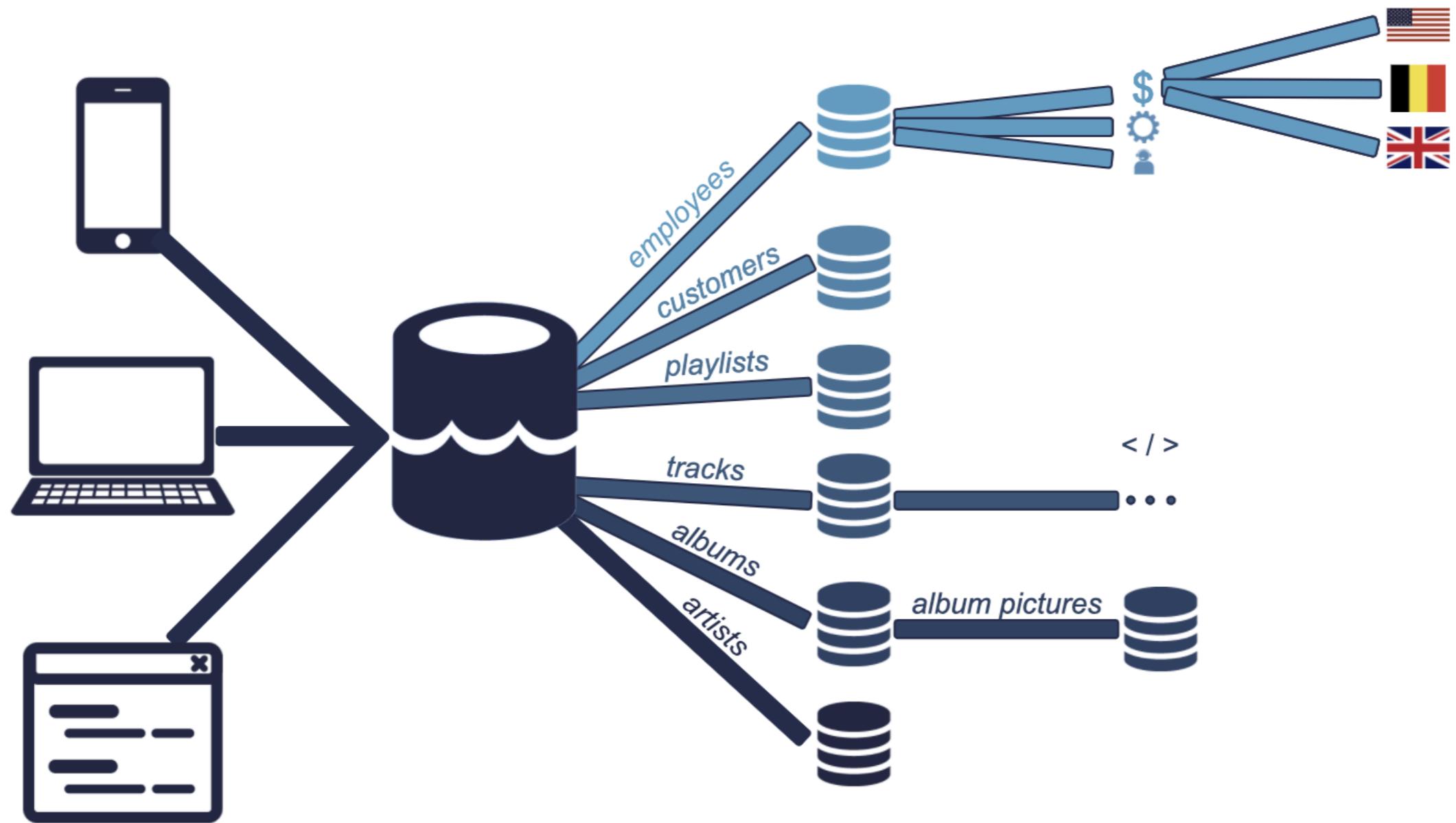


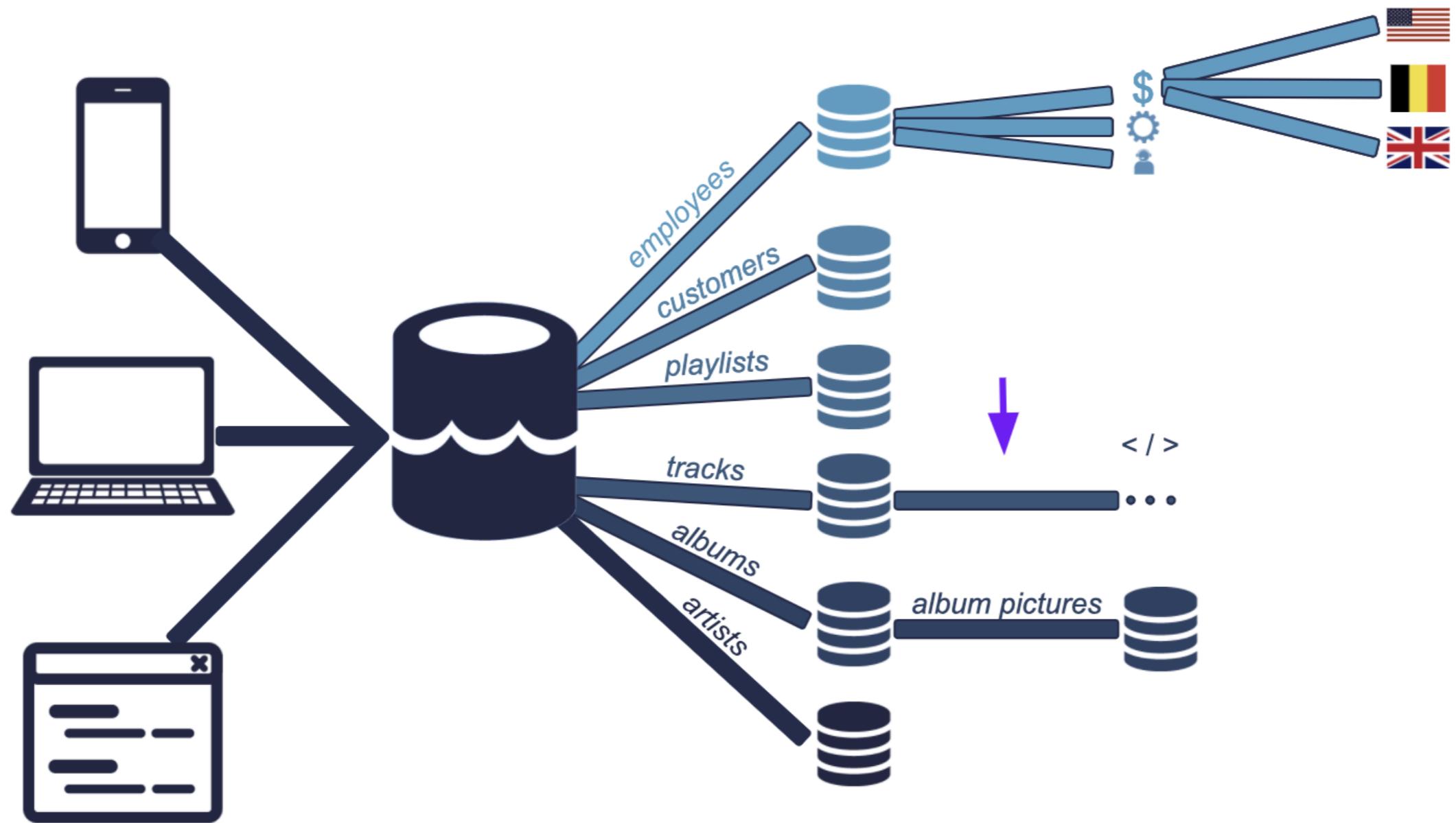


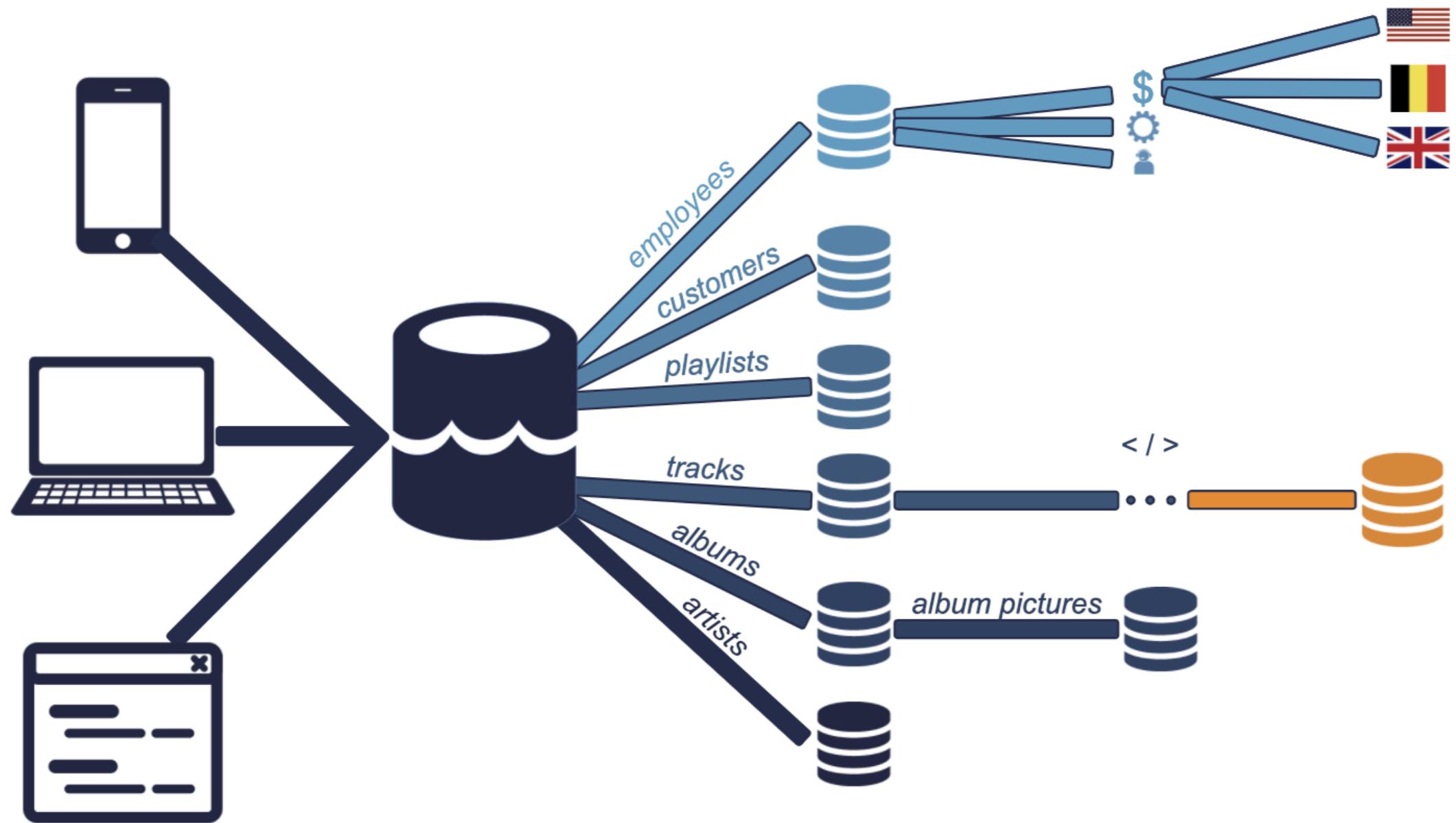


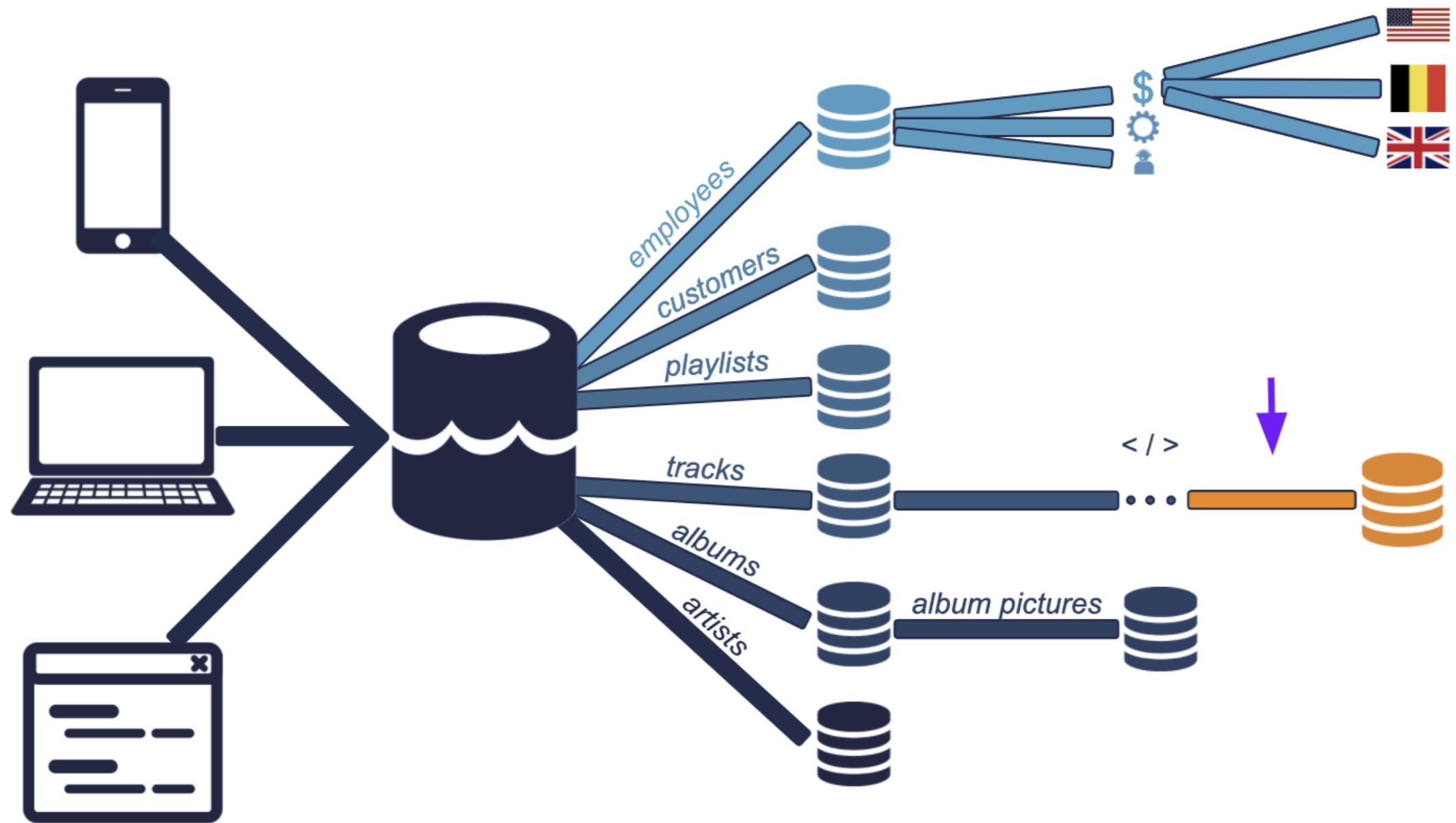
















2004

Everybody gets a pipeline !!!



# 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!**

**UNDERSTANDING DATA ENGINEERING**