IYKRA
Data Fellowship Program

(Deep Dive into Modern Data Engineering Landscape)

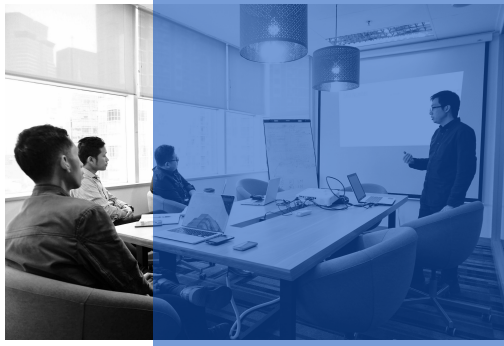
by (Bangun Sasongko)

(date of delivery)

(Section 3)

Trainer Profile

Bangun W. Sasongko
Data Engineer at BRI
Ex. Devops Engineer
Bachelor of Physics, Electrical and Instrumentations
@sasongkobgn
sasongkobgn@gmail.com



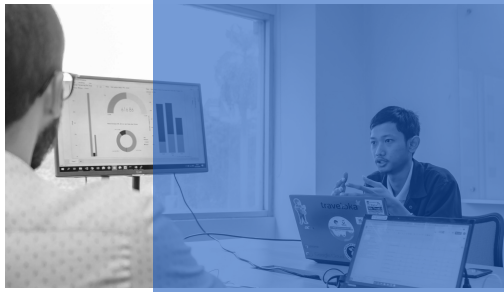
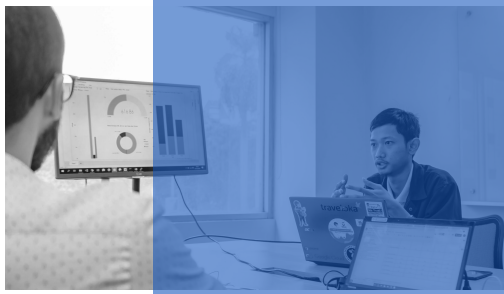


Table of Content

About this Course

Data engineers make life easy for data scientists by preparing raw data for analysis using different processing techniques at different steps. These steps need to be combined to create pipelines, which is when automation comes into play. Finally, data engineers use parallel and cloud computing to keep pipelines flowing smoothly.
Processing data
Scheduling data
Parallel computing
Cloud computing
Finisher wkwkwk



The Objectives

By the end of this course, you will be able to:

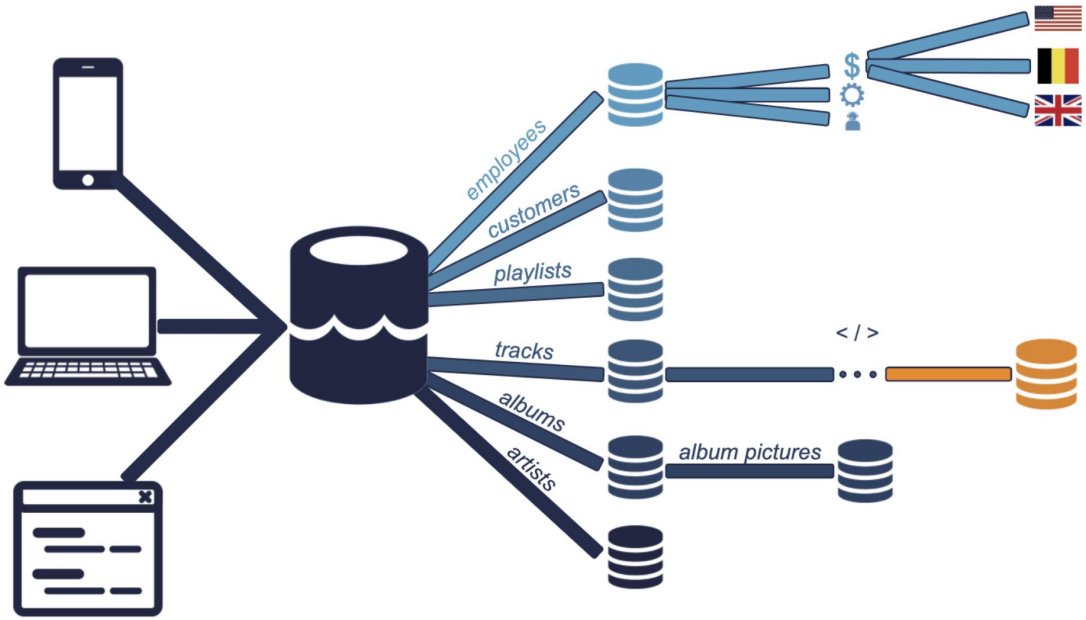
1. Simplify the data analysis process for data scientists by employing various processing techniques at distinct stages to prepare raw data for analysis.

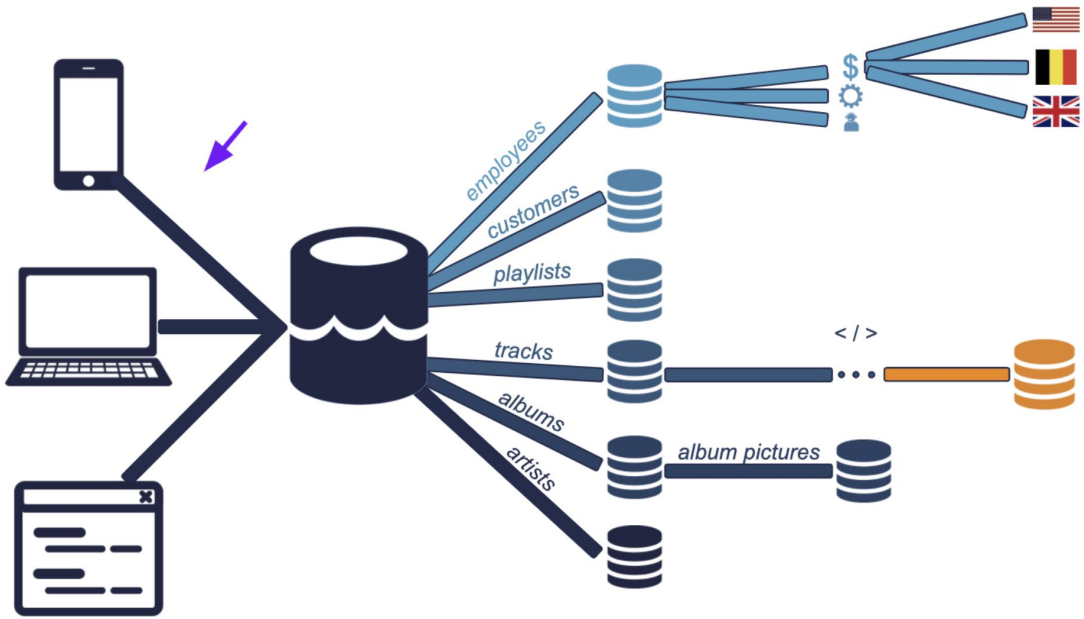
1. Implement automation to seamlessly combine the necessary steps into pipelines, streamlining the overall data preparation process.

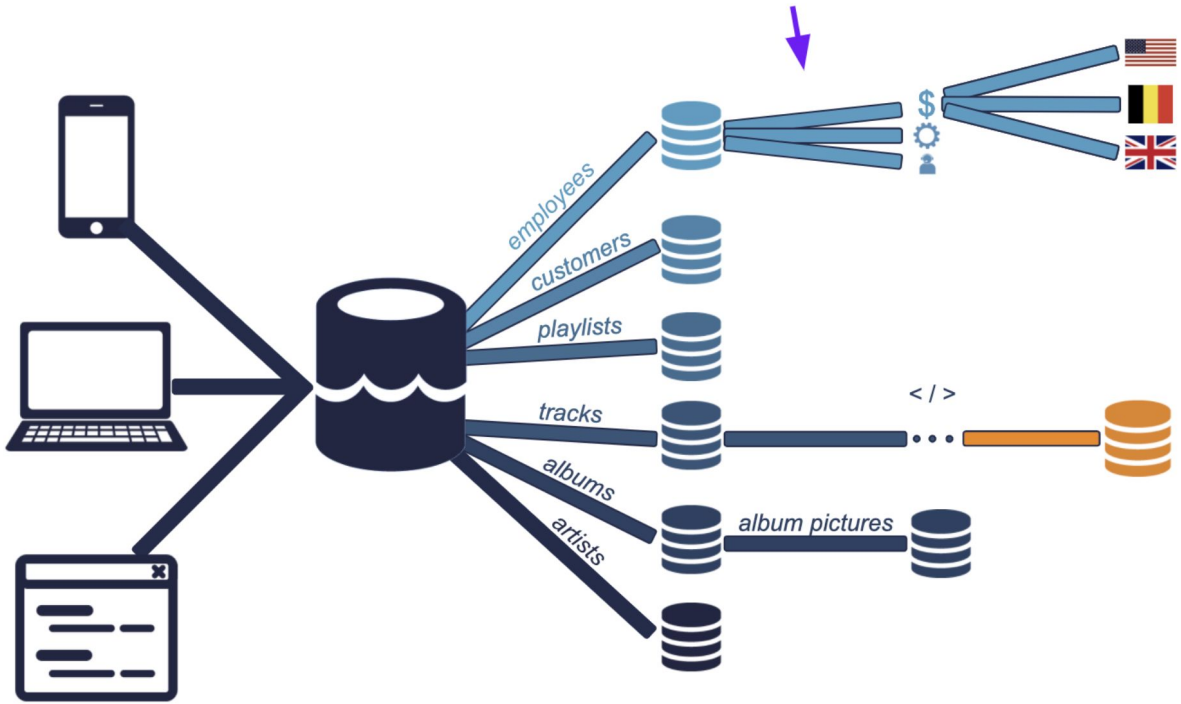
1. Ensure the continuous and efficient flow of pipelines by leveraging parallel and cloud computing technologies in the realm of data engineering.

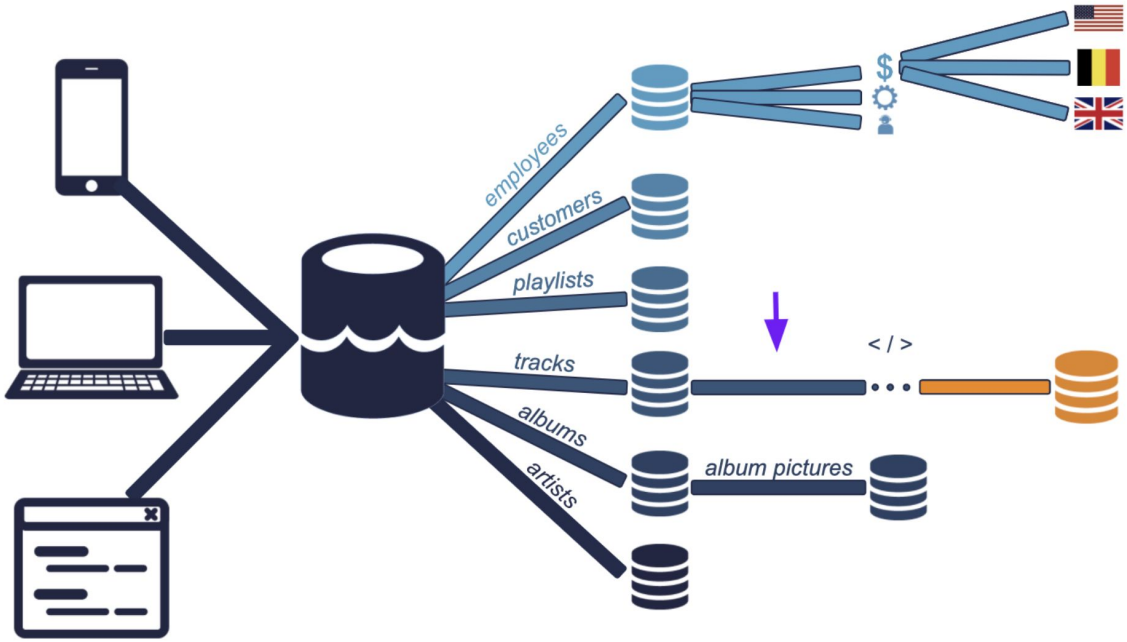
Data Fellowship

Processing data









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 Spotify

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

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

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

How data engineers process data

How data engineers process data

Data Structure: Summary

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

Data Fellowship

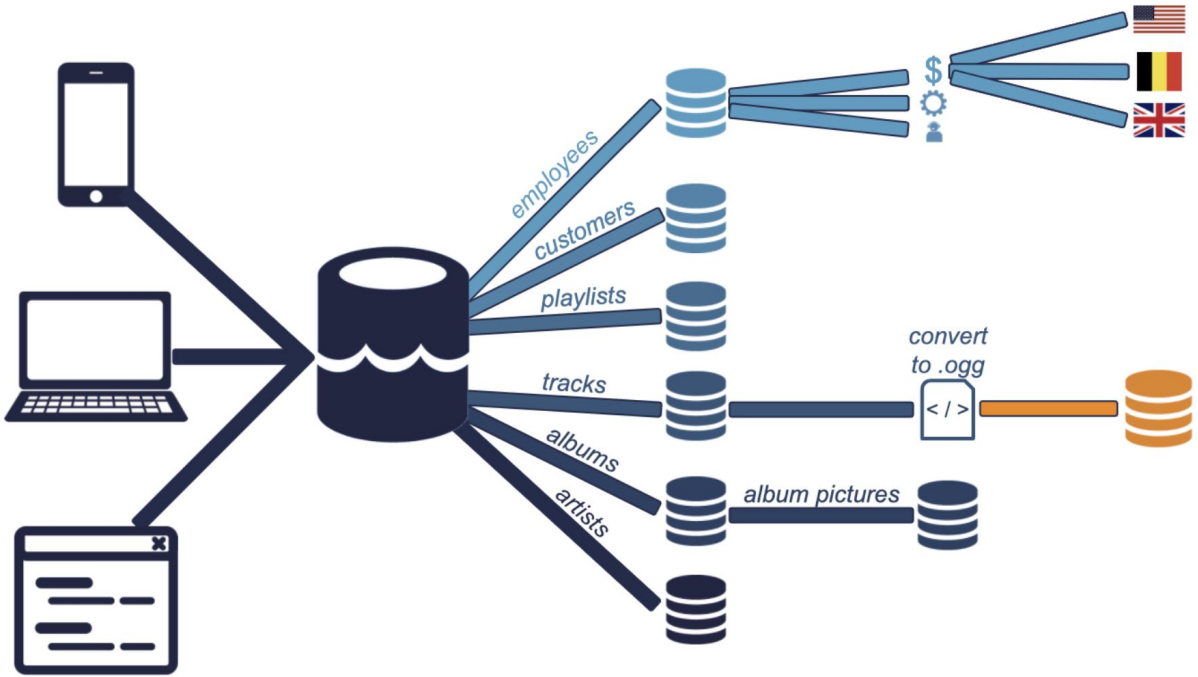
Scheduling data

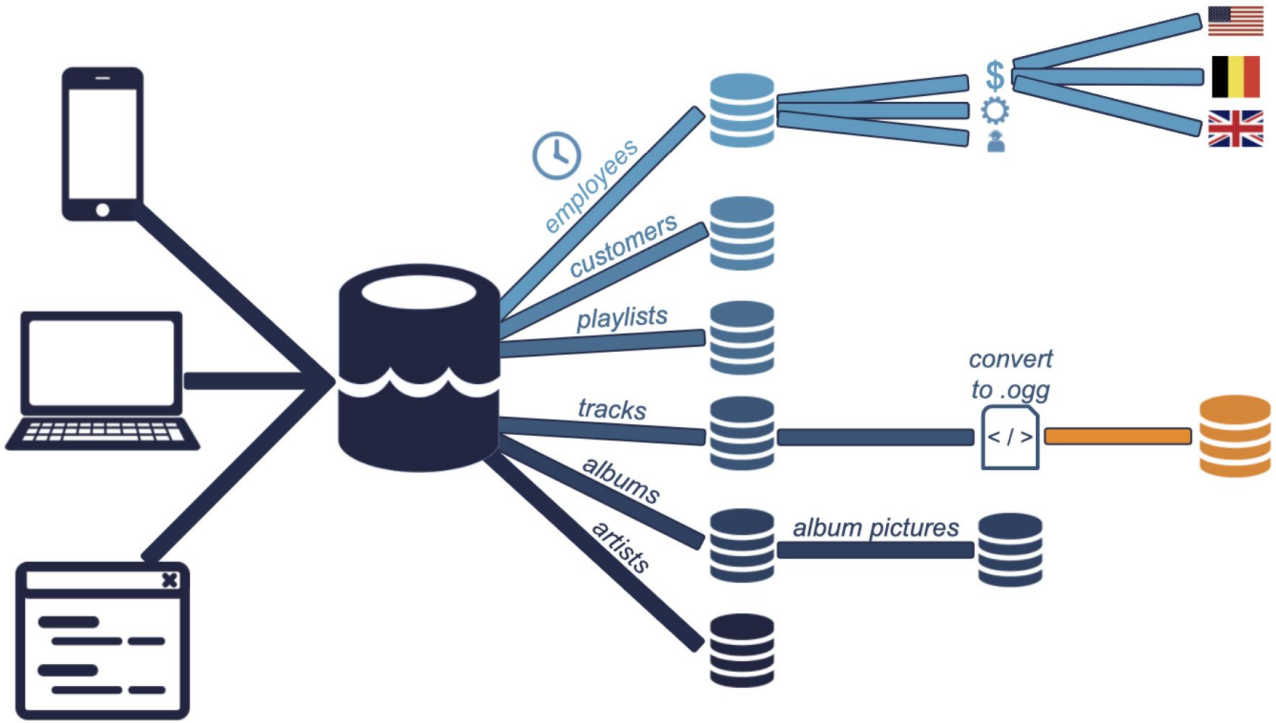
Scheduling

- 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

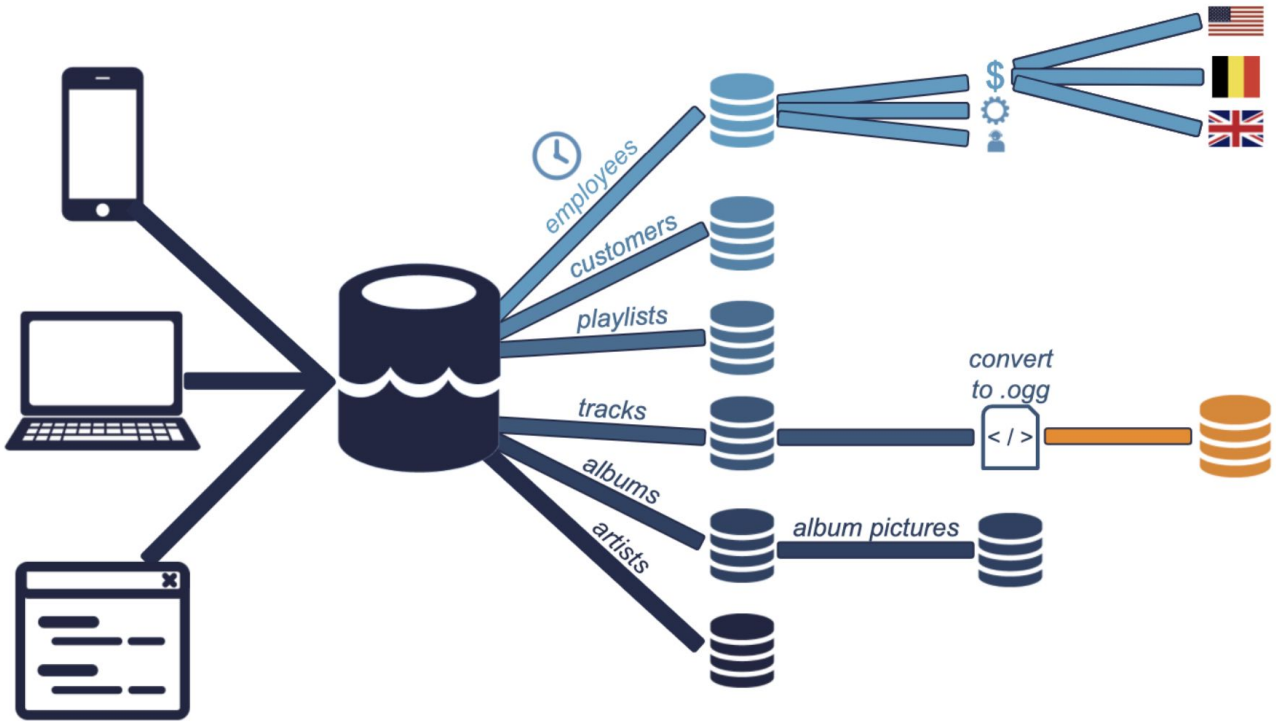
- Manually
- Manually update the employee table

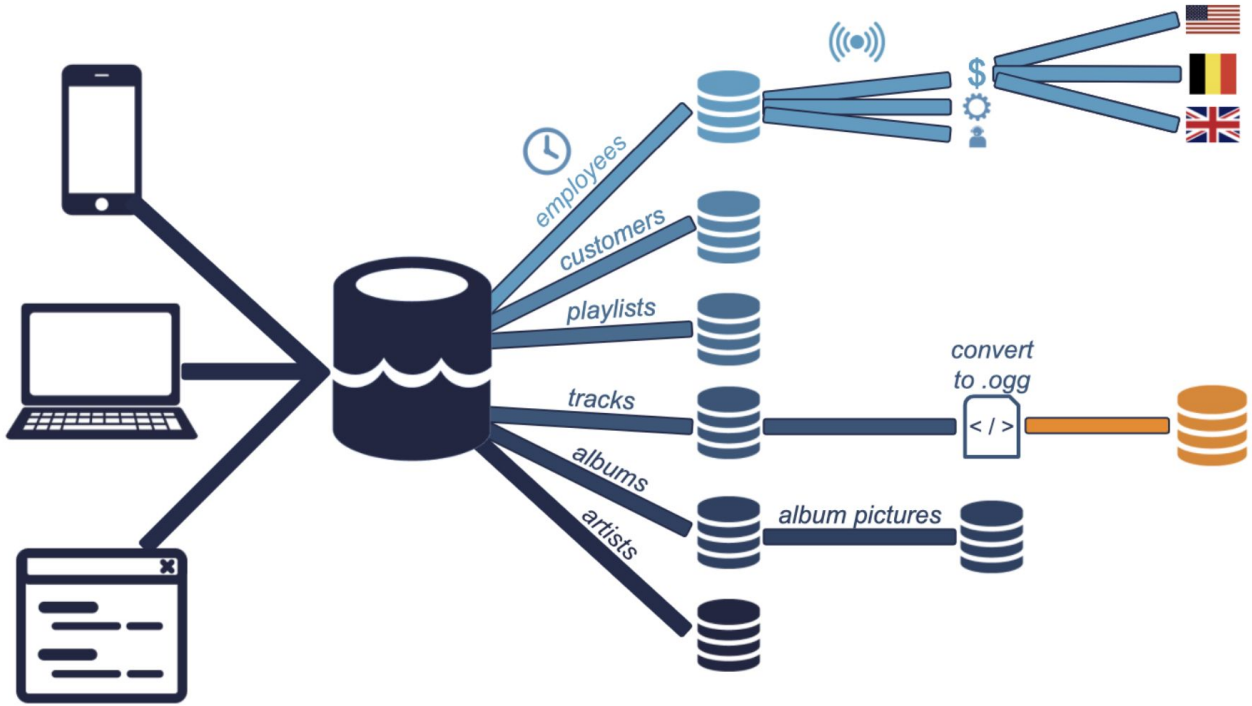




Manual, time and sensor scheduling

Manually	Manually update the employee table
Automatically run at a specific time	Update the employee table at 6 AM
Automatically run if a specific condition is met ● Sensor scheduling	





Batches and streams

Batches ● Group records at intervals ● Often cheaper	● Songs uploaded by artists ● Employee table ● Revenue table
Streams ● Send individual records right away	● New users signing in ● Another example: online vs. offline listening

Scheduling tools

Scheduling Data: Summary

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

Data Fellowship

Parallel computing

Parallel computing

- Basis of modern data processing tools
- Necessary:
  - Mainly because of memory
  - Also for processing power

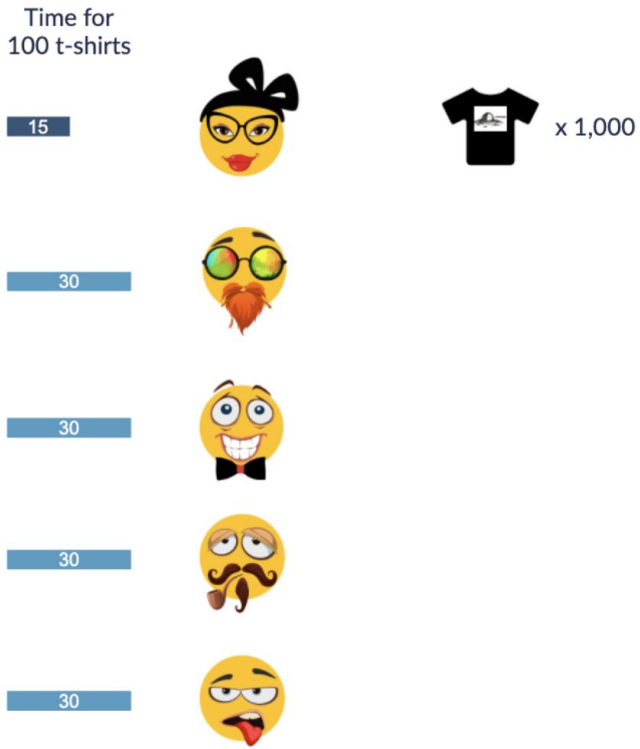
- How it works:
  - Split tasks up into several smaller subtasks
  - Distribute these subtasks over several computers

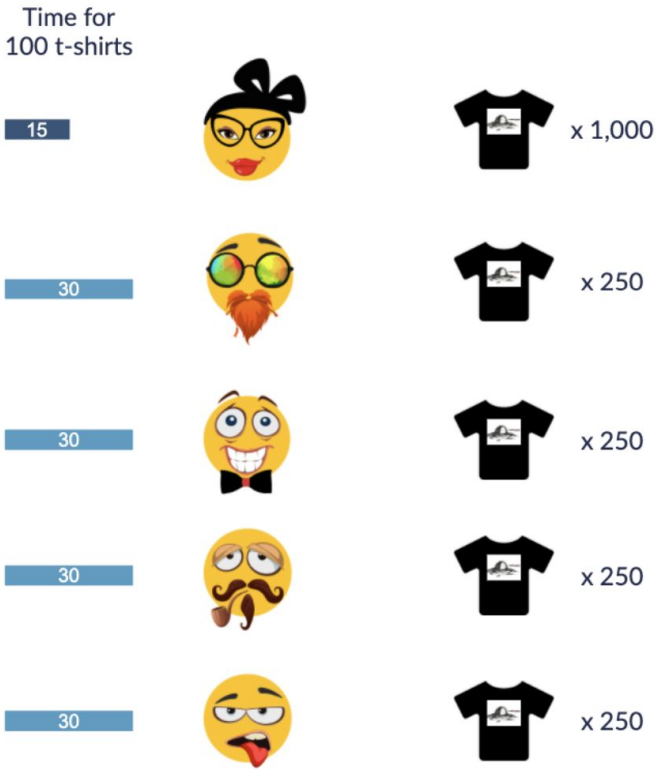
x 1,000

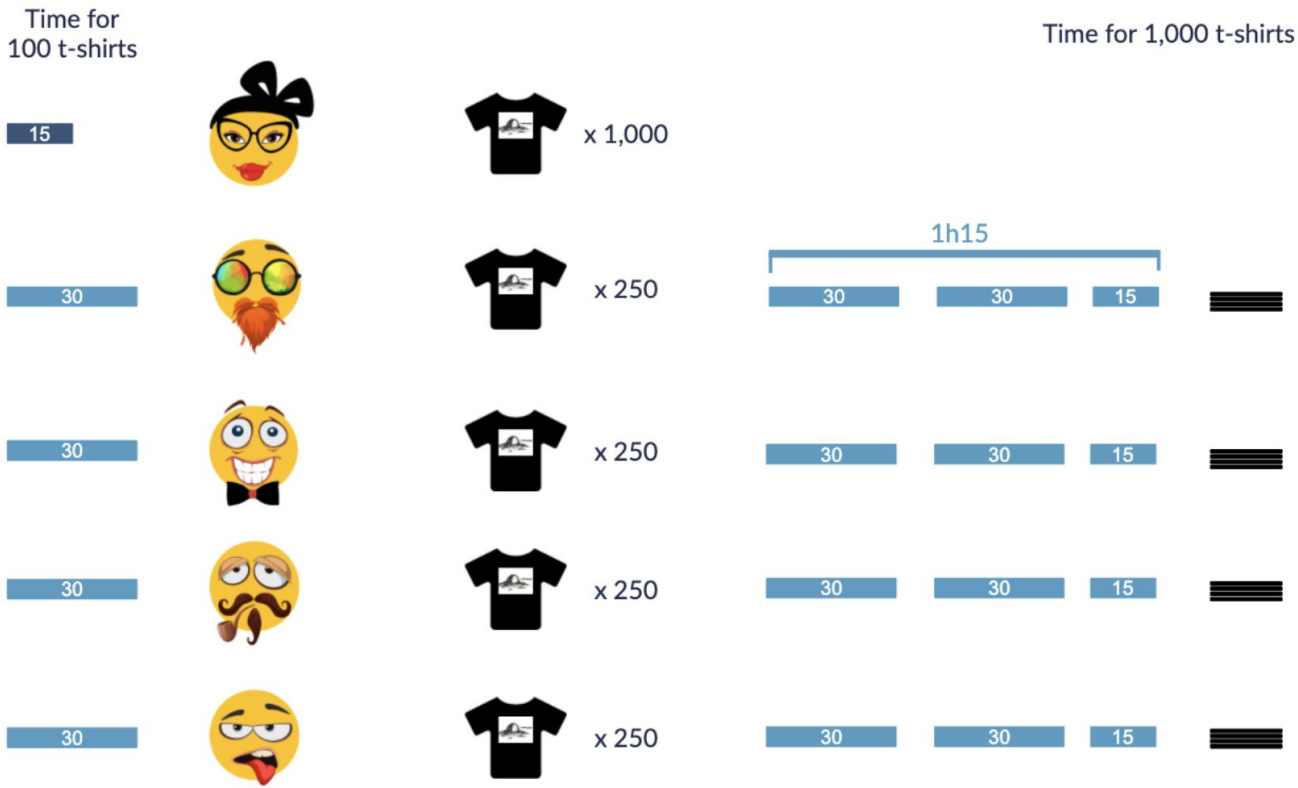
Time for
100 t-shirts

15

x 1,000







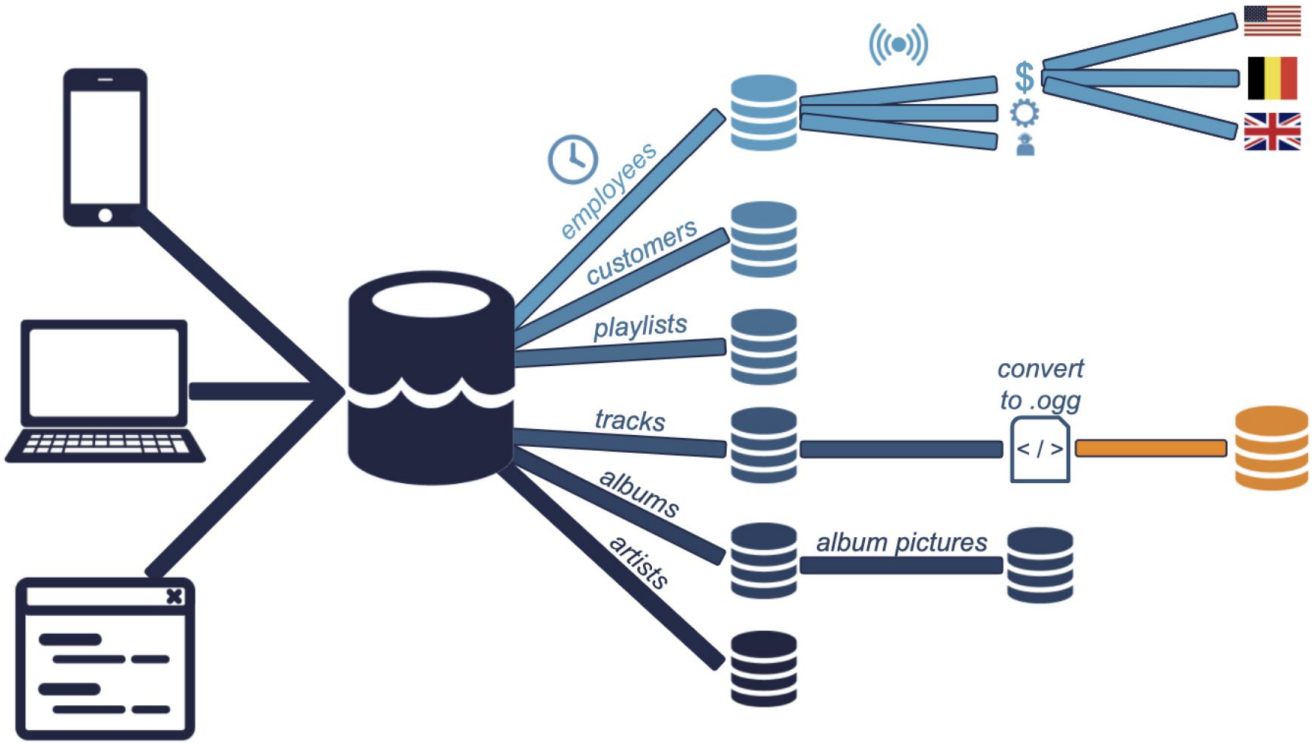
Parallel computing

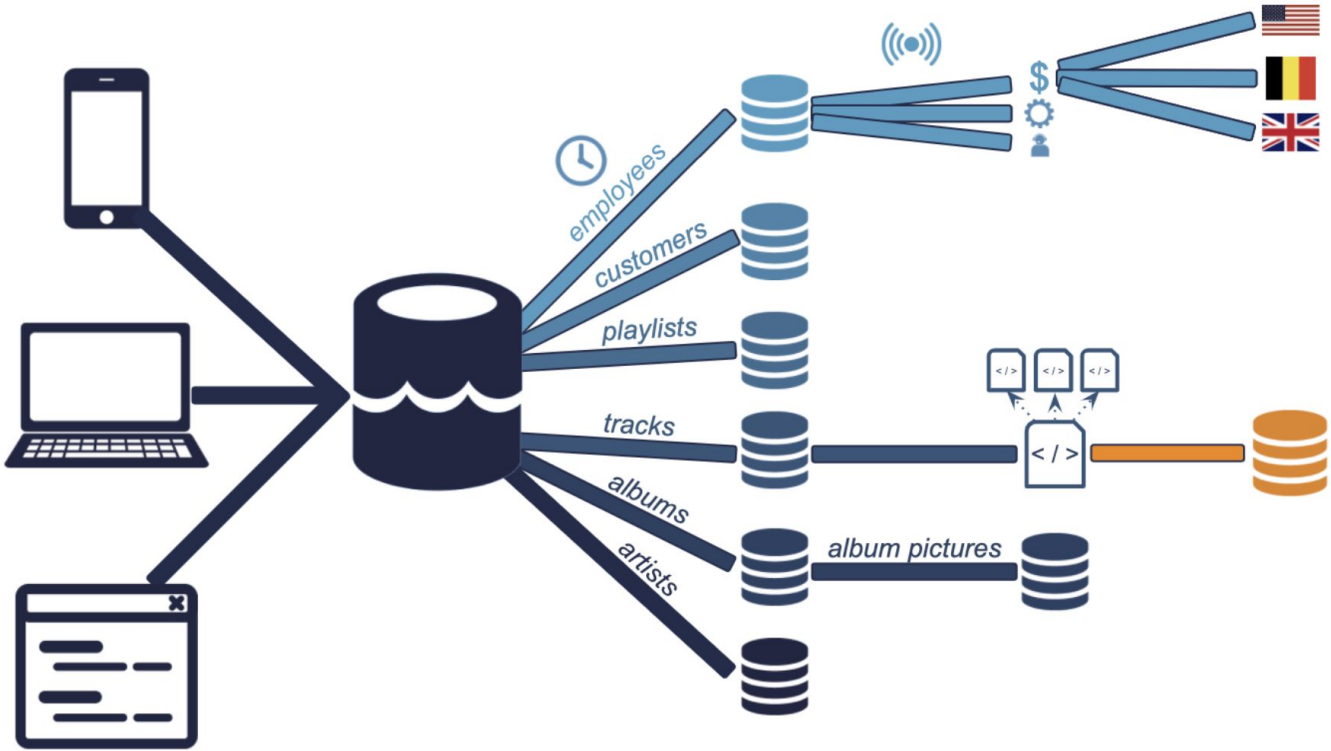
Benefits and risks of parallel computing

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

Parallel computing

Parallel computing





Parallel computing: Summary

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

Data Fellowship

Cloud computing

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%

17.6%

6%

File storage

32.4%

17.6%

6%

aws

32.4%

Microsoft Azure

17.6%

Google Cloud

6%

File storage

AWS S3

aws

32.4%

Microsoft Azure

17.6%

Google Cloud

6%

File storage

AWS S3

Azure Blob Storage

32.4%

17.6%

6%

File storage

AWS S3

Azure
Blob Storage

Google
Cloud Storage

	File storage	Computation
aws 32.4%	AWS S3 	AWS EC2 
Microsoft Azure 17.6%	Azure Blob Storage 	
Google Cloud 6%	Google Cloud Storage 	

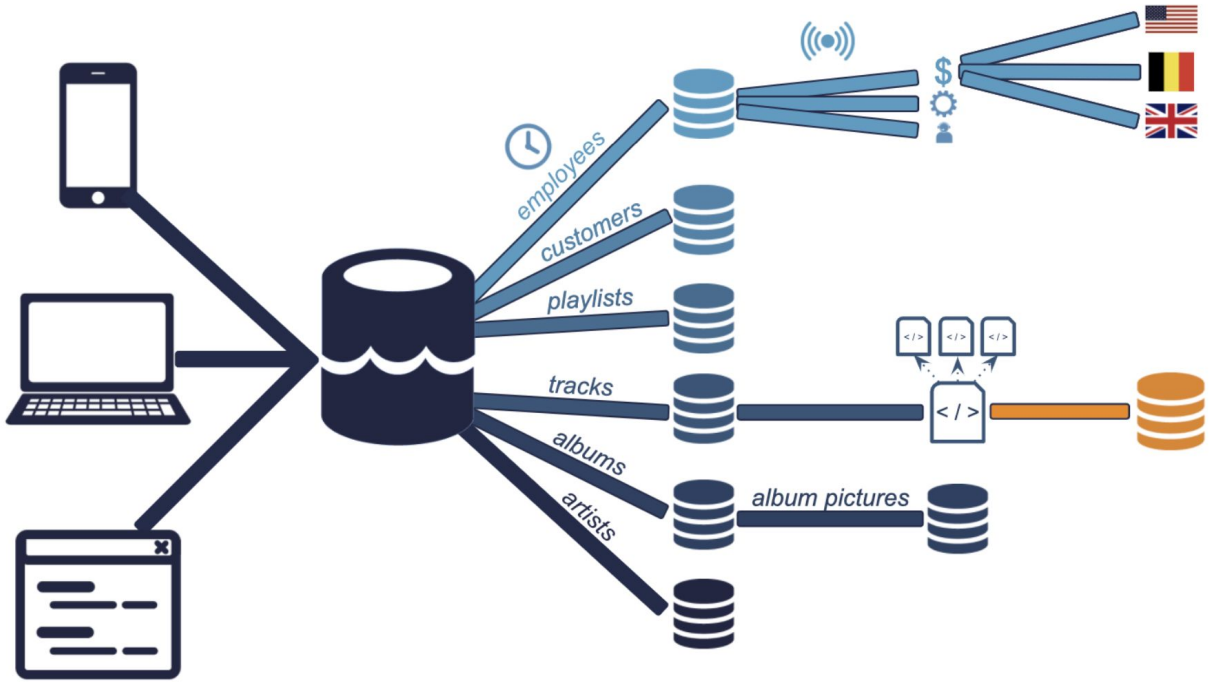
	File storage	Computation
aws 32.4%	AWS S3 	AWS EC2 
Microsoft Azure 17.6%	Azure Blob Storage 	Azure Virtual Machines 
Google Cloud 6%	Google Cloud Storage 	

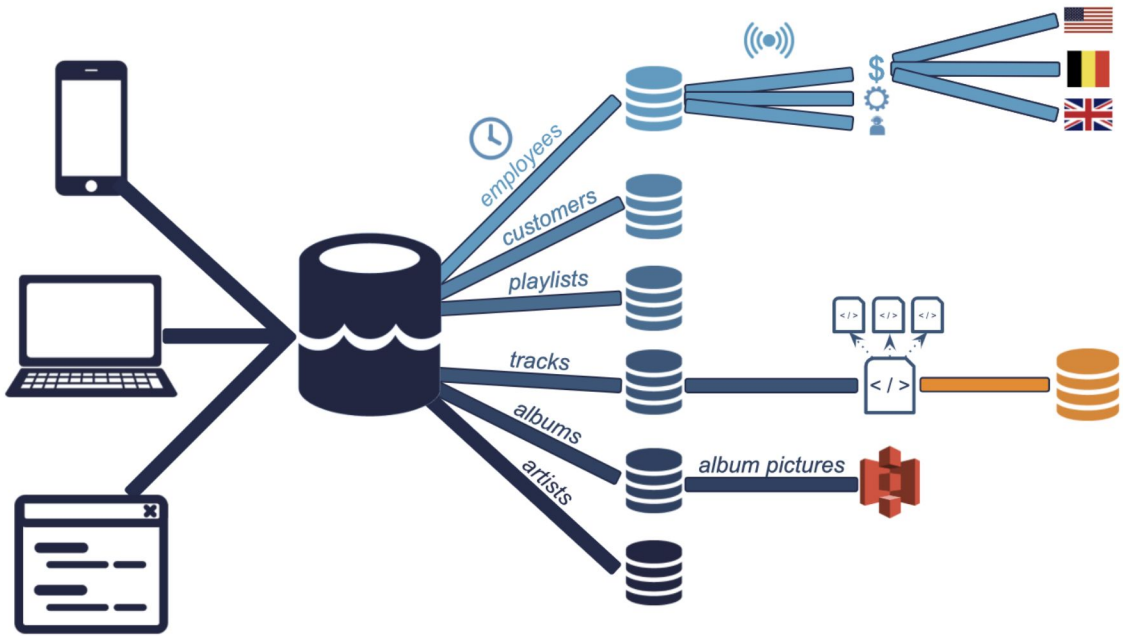
	File storage	Computation
aws 32.4%	AWS S3 	AWS EC2 
Microsoft Azure 17.6%	Azure Blob Storage 	Azure Virtual Machines 
Google Cloud 6%	Google Cloud Storage 	Google Compute Engine 

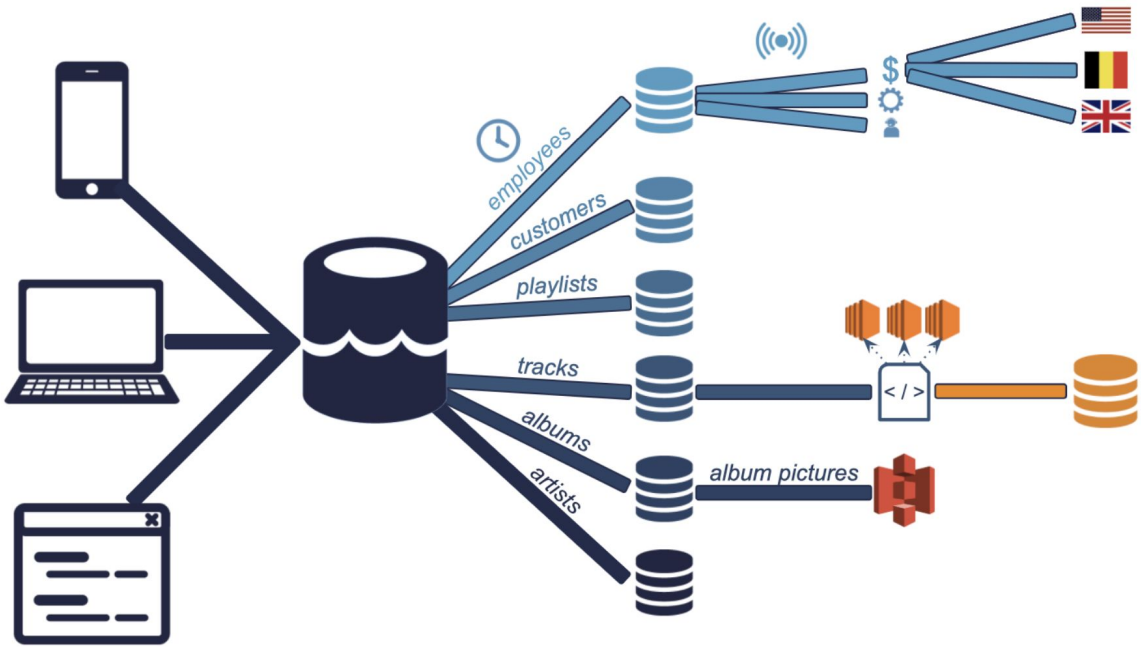
	File storage	Computation	Databases
aws 32.4%	AWS S3	AWS EC2	AWS RDS
Microsoft Azure 17.6%	Azure Blob Storage	Azure Virtual Machines	
Google Cloud 6%	Google Cloud Storage	Google Compute Engine	

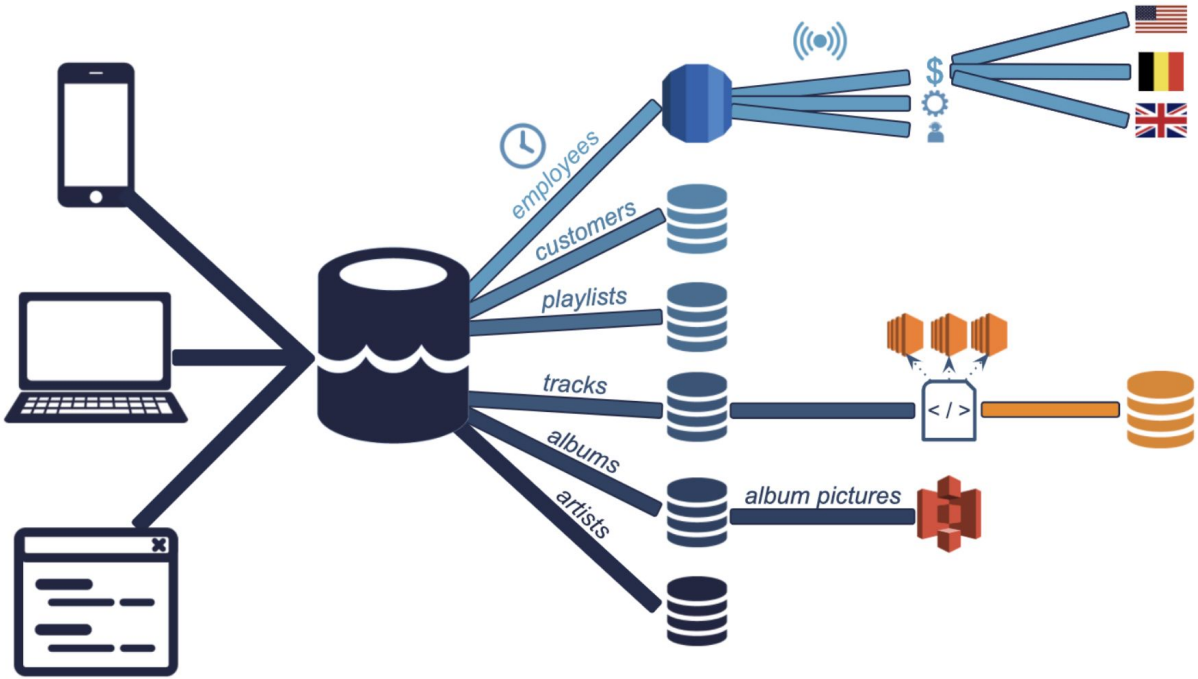
	File storage	Computation	Databases
aws 32.4%	AWS S3 	AWS EC2 	AWS RDS 
Microsoft Azure 17.6%	Azure Blob Storage 	Azure Virtual Machines 	Azure SQL Database 
Google Cloud 6%	Google Cloud Storage 	Google Compute Engine 	

	File storage	Computation	Databases
aws 32.4%	AWS S3 	AWS EC2 	AWS RDS 
Microsoft Azure 17.6%	Azure Blob Storage 	Azure Virtual Machines 	Azure SQL Database 
Google Cloud 6%	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
$
< / >





Multi Cloud

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

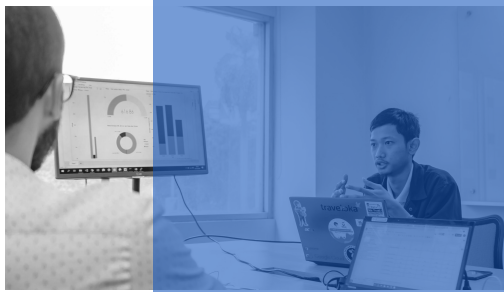
- Cloud providers try to lock in consumers
- Incompatibility
- Security and governance

Cloud computing: Summary

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

References

Datacamp : understanding data engineering



iykra Data Fellowship

Thank you!