Data Engineering for Everyone

Wednesday, 18 November 2020 10:27 PM

What is data engineering?

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 engineers are responsible for the first step of the process: ingesting collected data and storing it.

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

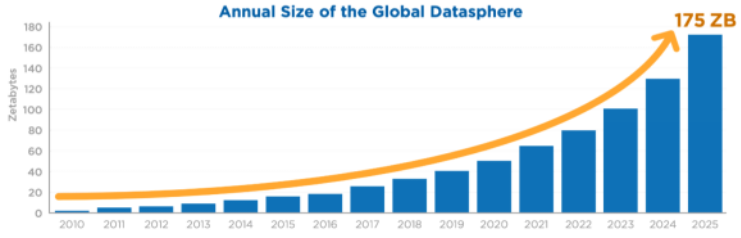
- Ingest data from different sources
- Optimize databases for analysis
- Remove corrupted data
- Develop, construct, test and maintain data architectures

Data engineers and big data

- Big data becomes the norm => data engineers 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)



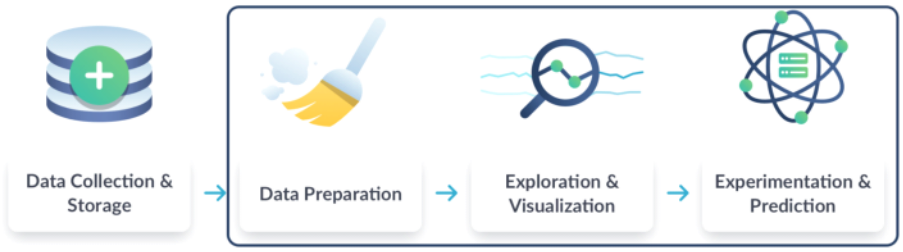
The five Vs

- 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

Data scientists



Data scientist intervene on the rest of the workflow: they prepare the data according to their analysis needs, explore it, build insightful visualizations, and then run experiments or build predictive models.

Data engineers lay the groundwork that makes data science activity possible.

Data engineers enable data scientists

Data engineer

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

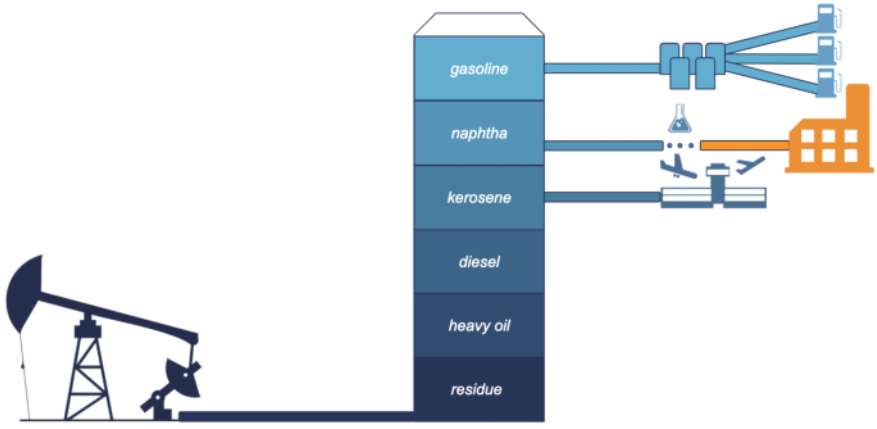
Data scientist

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

Summary

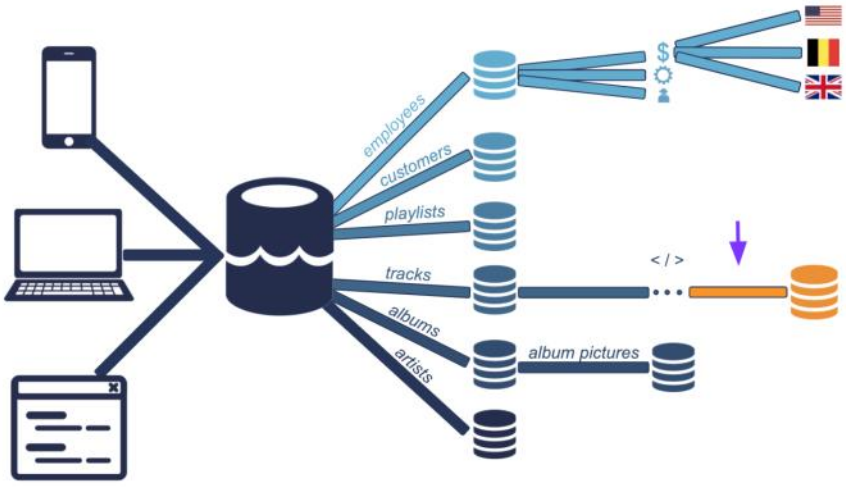
- At which stages data engineers and data scientists intervene
- How data engineers enable data scientists

If data is the new oil...



Back to data engineering

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



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

Storing data

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

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

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)
);
```

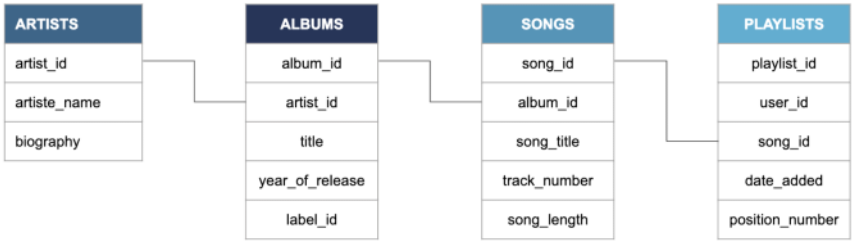
SQL for data scientists

- 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



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

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

Moving and processing data

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

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

Summary

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

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
- 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

Summary

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

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

Benefits and risks of parallel computing

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

Summary

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

Cloud computing for data processing

Servers on premises

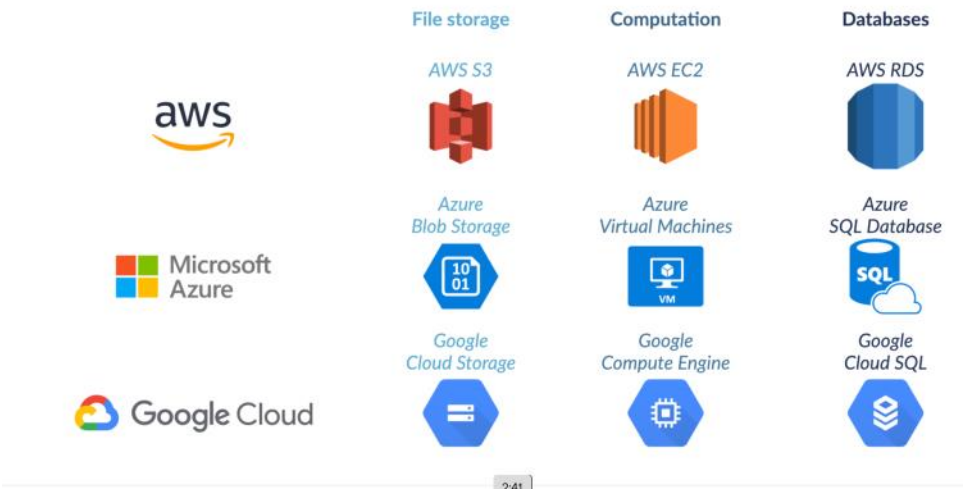
- Bought
- Need space
- Electrical and maintenance cost
- Enough power for peak moments
- Processing power unused at quieter times

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



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