What is data engineering?

INTRODUCTION TO DATA ENGINEERING

Vincent Vankrunkelsven
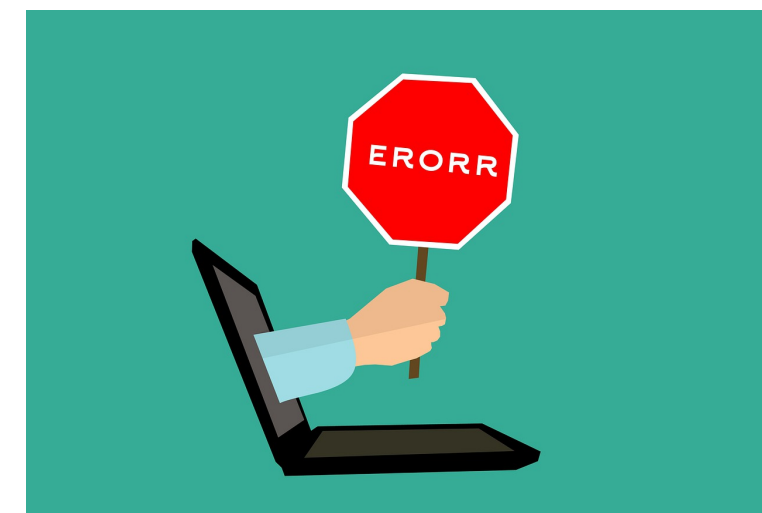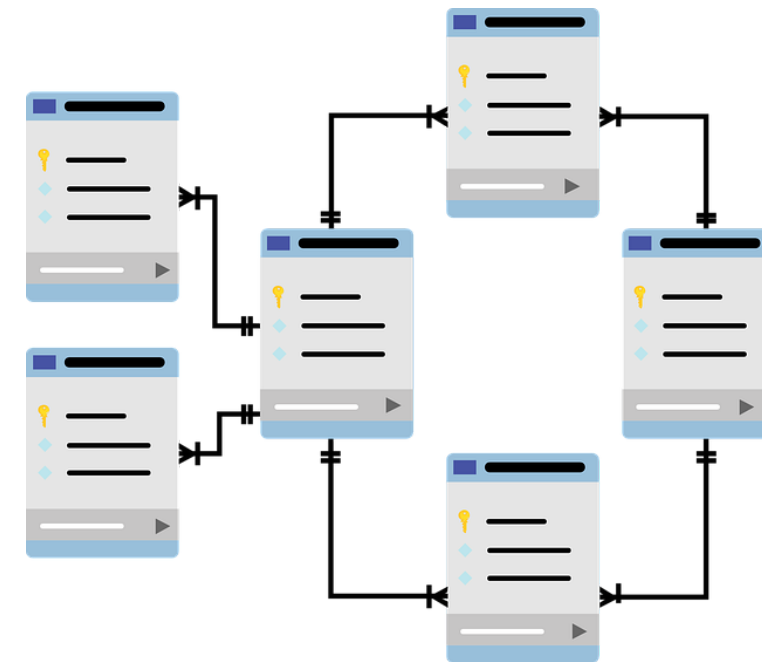Data Engineer @ DataCamp

datacamp

What to expect

- Chapter 1
  - What is data engineering?

- Chapter 2
  - Tools data engineers use

- Chapter 3
  - Extract

  - Transform

  - Load

- Chapter 4
  - Data engineering at DataCamp!

In comes the data engineer

- Data is scattered

- Not optimized for analyses

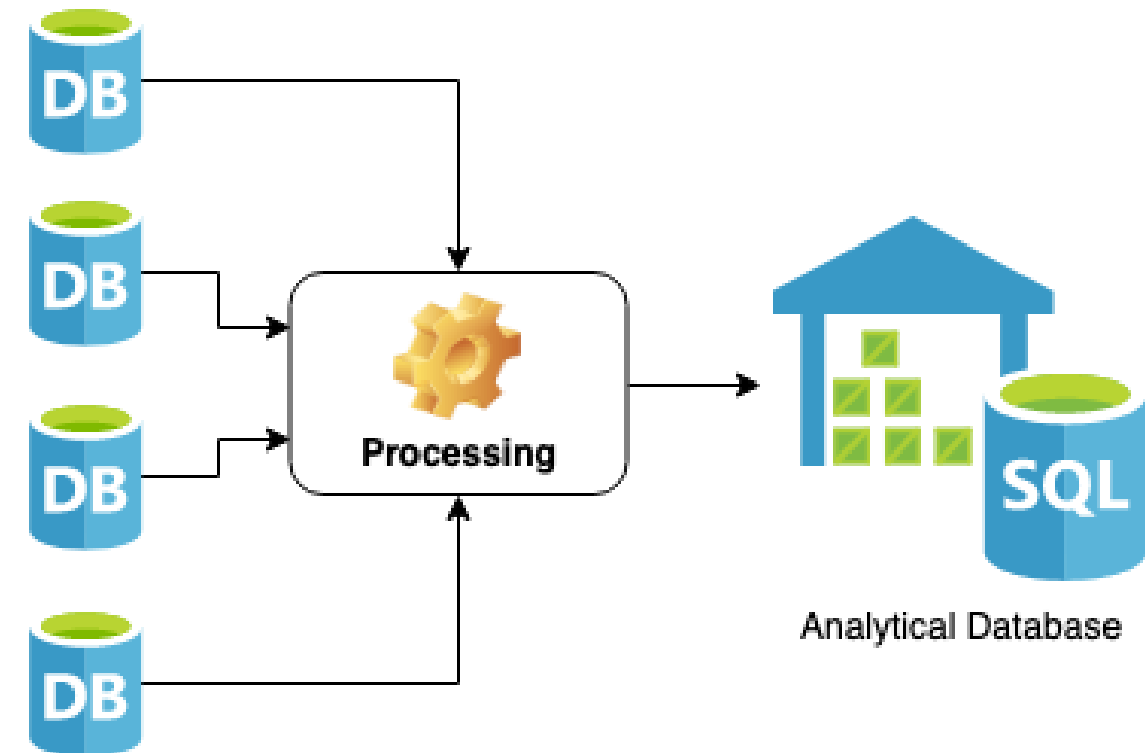- Legacy code is causing corrupt data

Data engineer to the rescue!

Data engineers: making your life easier

- Gather data from different sources

- Optimized database for analyses

- Removed corrupt data

Data scientist's life got way easier!

Definition of the job

An engineer that develops, constructs, tests, and maintains architectures such as databases and large-scale processing systems

- Processing large amounts of data

- Use of clusters of machines

Data Engineer vs Data Scientist

Data Engineer

- Develop scalable data architecture

- Streamline data acquisition

- Set up processes to bring together data

- Clean corrupt data

- Well versed in cloud technology

Data Scientist

- Mining data for patterns

- Statistical modeling

- Predictive models using machine learning

- Monitor business processes

- Clean outliers in data

Let's practice!

INTRODUCTION TO DATA ENGINEERING

Tools of the data engineer

INTRODUCTION TO DATA ENGINEERING

Vincent Vankrunkelsven
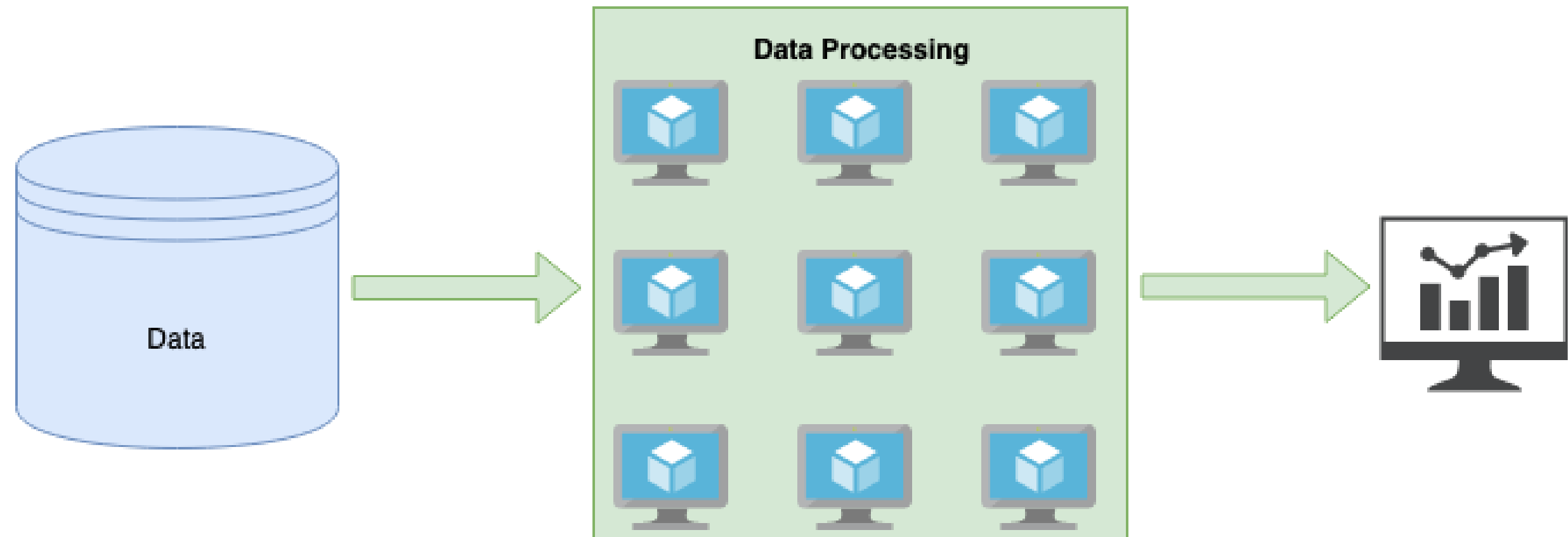Data Engineer @ DataCamp

Databases

- Hold large amounts of data

- Support application

- Other databases are used for analyses

Product
name
price
stock_amount

Processing

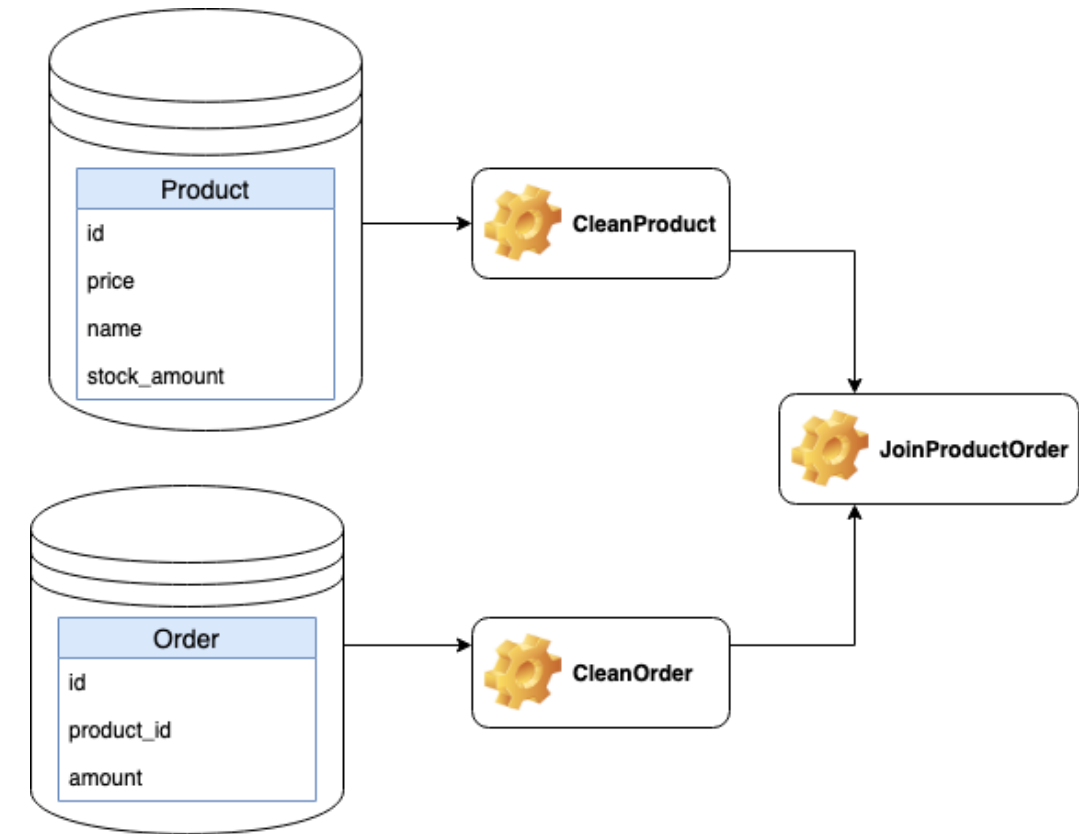- Clean data

- Aggregate data

- Join data

Processing: an example

```python
df = spark.read.parquet("users.parquet")


outliers = df.filter(df["age"] > 100)


print(outliers.count())
```

Data engineer understands the abstractions.

Scheduling

- Plan jobs with specific intervals

- Resolve dependency requirements of jobs

`JoinProductOrder` needs to run after `CleanProduct` and `CleanOrder`
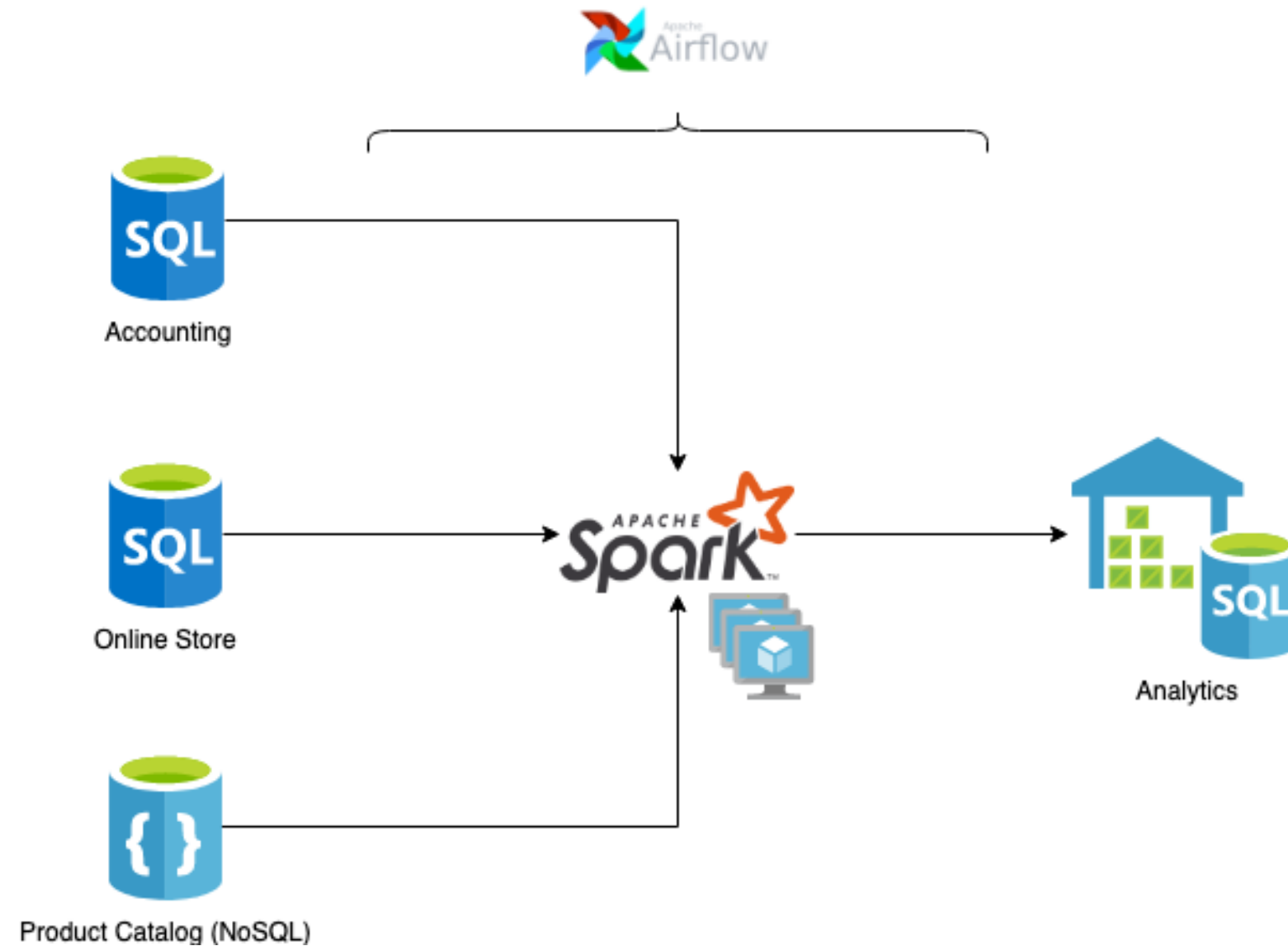
Existing tools

Databases

Scheduling
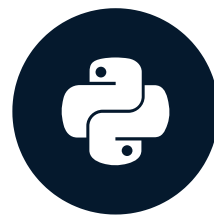
A data pipeline

Let's practice!

INTRODUCTION TO DATA ENGINEERING

Cloud providers

INTRODUCTION TO DATA ENGINEERING

Vincent Vankrunkelsven
Data Engineer @ DataCamp

Data processing in the cloud

Clusters of machines required

Problem: self-host data-center

- Cover electrical and maintenance costs

- Peaks vs. quiet moments: hard to optimize

Solution: use the cloud

Data storage in the cloud

Reliability is required

Problem: self-host data-center

- Disaster will strike

- Need different geographical locations

Solution: use the cloud

The big three: AWS, Azure and Google

32% market share in 2018

17% market share in 2018

10% market share in 2018

- Storage

- Computation

- Databases.

Storage

Upload files, e.g. storing product images

Services

- AWS S3

- Azure Blob Storage

- Google Cloud Storage

Computation

Perform calculations, e.g. hosting a web server

Services

- AWS EC2

- Azure Virtual Machines

- Google Compute Engine

Databases

Hold structured information

Services

- AWS RDS

- Azure SQL Database

- Google Cloud SQL

Let's practice!

INTRODUCTION TO DATA ENGINEERING