STUTERN

Data Engineering

SGA07_DATASCI

5th May 2020

Book Keeping

- Group Tasks (25% of total course score)
  - Due date 24th April
- The last learning modules (Guest lecture | 5th May)
  - ~~Gist: IoT Engineering~~ Data Engineering
- 4 weeks for Final Project (50% of total course score)
  - Due date 29th May

Module Overview

- Who am I ?
- What is Data Engineering
- Real world example
- Data / ML Ops

STUTERN

Maël Razavet

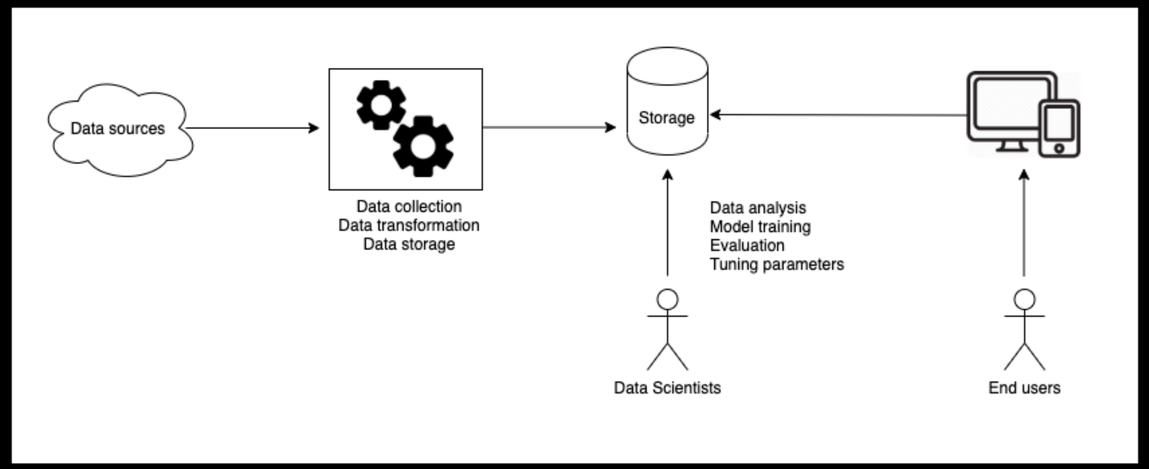
- I'm Data Engineer / AI Engineer
- Graduated from University Of Warwick as a Data scientist
- Started working back in 2014
- Build my startup in 2017 in the HR industry
- I now work as a Freelancer

What is Data Engineering ?

- Design data architectures to host large amount of data
- Build reusable data pipelines for Data Scientists
  - Lambda architecture: Real time ingestion / Batch processing
  - Data exposure
- Industrialise AI models for production
  - From Jupiter notebooks to real world programs
  - Good practices (design patterns, testing, refactoring, …)
  - Packaging
  - Model exposition
  - Deployment

STUTERN

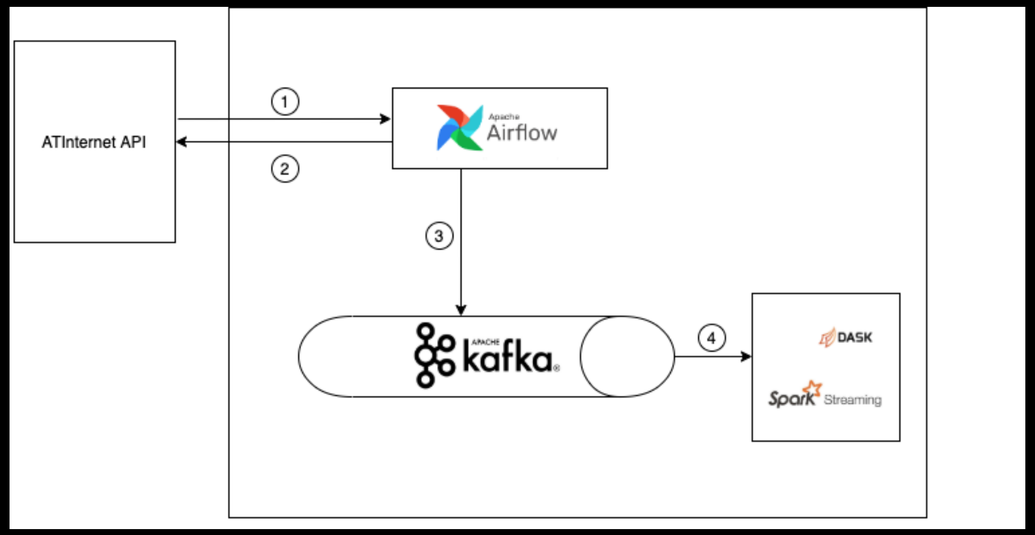
What is Data Engineering ?



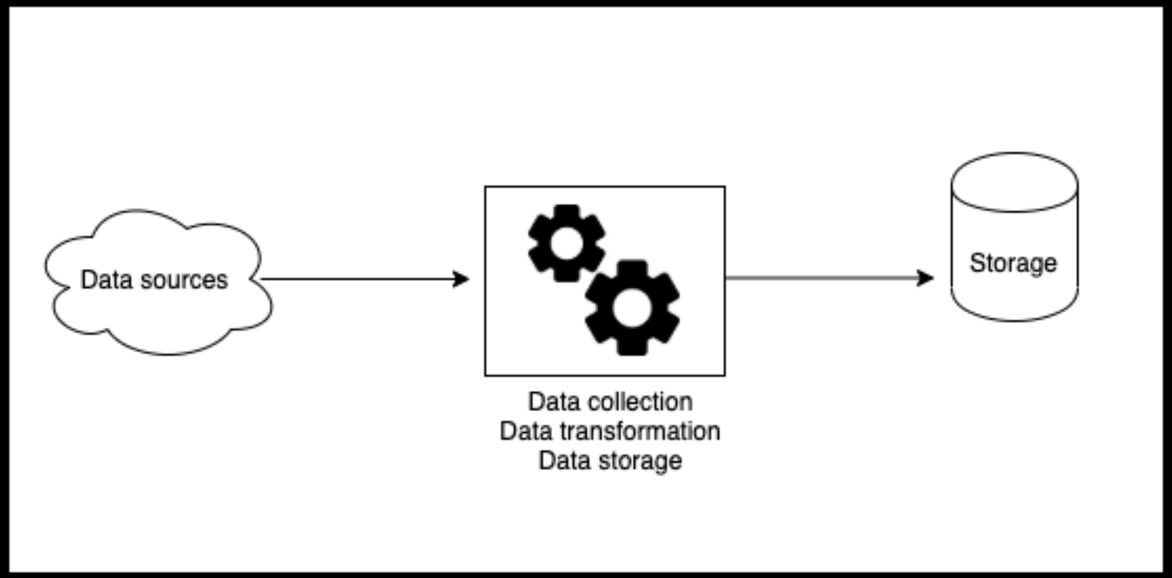
- Technologies matter !!!

Real world example

- Use case needs:
  - Data fetching in real time
  - Error handling
  - Distributed data processing
- Solution:
  - Orchestrator
  - Event bus
  - Hadoop data platform for processing



What we saw

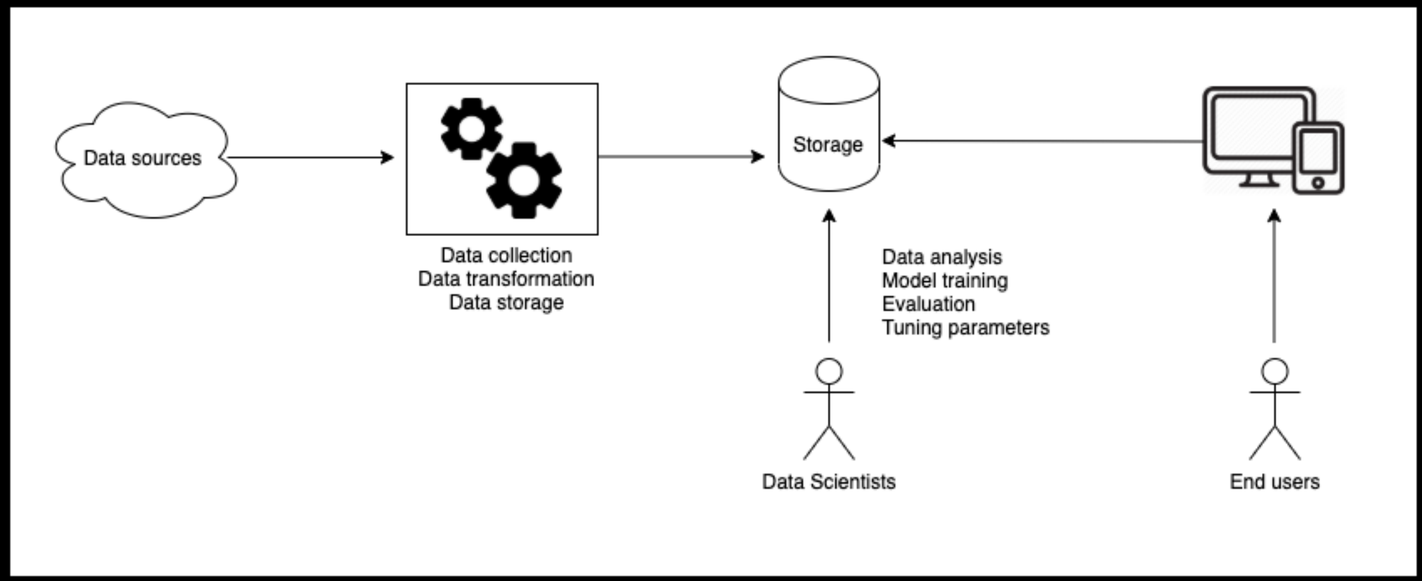

Data sources
Storage
Data collection
Data transformation
Data storage

End-to-end ML lifecycle

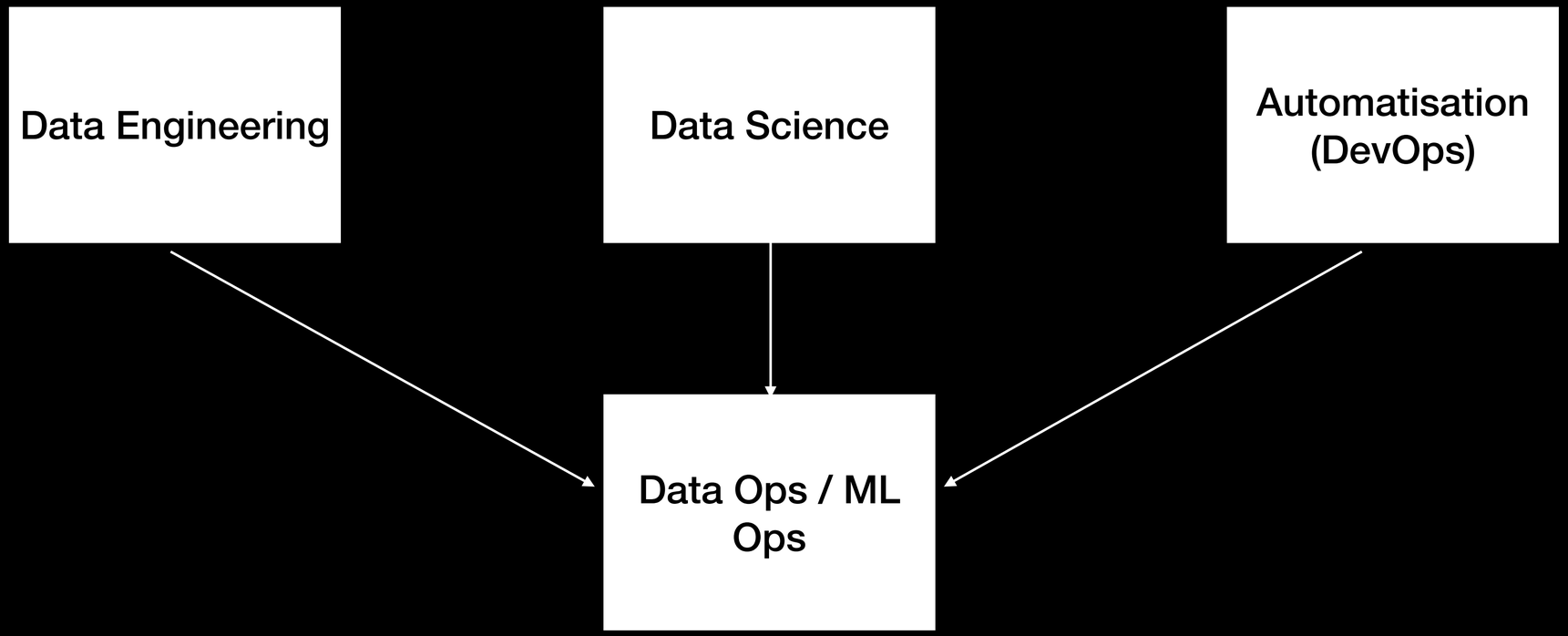
- Develop & train model with reusable pipelines
  - What frequency ?
- Package model
  - Using containers or API endpoints
- Deploy model
  - To the cloud or on premise for use in real-time / streaming / batch processing
- Monitor model in terms of business value
  - When to replace / deprecate a stale model ?

STUTERN

What we saw



This is the new trend



Suggested Material

- https://www.dataquest.io/blog/what-is-a-data-engineer/
- https://medium.com/@rchang/a-beginners-guide-to-data-engineering-part-i-4227c5c457d7
- https://www.kubeflow.org/
- https://www.mlflow.org/
- https://medium.com/data-ops/the-dataops-enterprise-software-industry-2019-a862904857ef