BIG DATA AND CLOUD PLATFORMS

MODULO 2

whoami

Matteo Francia

- Email: m.francia@unibo.it
- Research fellow @ UniBO

Research topics

- Big data / database
- Geo-spatial analytics

Thesis proposals

- https://big.csr.unibo.it/teaching/

Table of Contents and Exam

Handling data pipelines in the Cloud

- Introduction to data platforms: shifting from databases to well-integrated data ecosystems
- Definition of cloud and taxonomy of cloud services
- Introduction to the most relevant Cloud Platforms
- Introduction to the billing models that lay behind Cloud Computing services
- Cluster migration: on-premises vs on-cloud
- Real case studies

Seminars by companies working with cloud and big data platforms

Connecting the dots

- Information systems, BI, data mining, big data, and machine learning

… all these points will be part of the oral examination! :)

Roadmap

Why going cloud?

From databases to data platforms

Building data platforms

Creating data pipelines in the cloud (in AWS)

Billing and cloud migration

So far

You have acquainted/practiced with **on-premises** solutions

- You were given a working hardware cluster
- ... to deploy software applications on Hadoop-based stack

In the perspective of digital transformation[1], let us guess

- How would you start from scratch?
- How much time would it take?

[1] The process of using digital technologies to create new — or modify existing — business processes, culture, and customer experiences to meet changing business and market requirements

So far

No easy answers

Big-data (distributed) architectures require a lot of skills

- **Configuration**: how do I set up dozens of new machines?
- **Networking**: how do I cable dozens of machines?
- **Management**: how do I replace a broken disk?
- **Upgrade**: how do I extend the cluster with new services/machines?
- (energy and cooling, software licenses, insurance...)

https://aws.amazon.com/compliance/data-center/data-centers/

So far

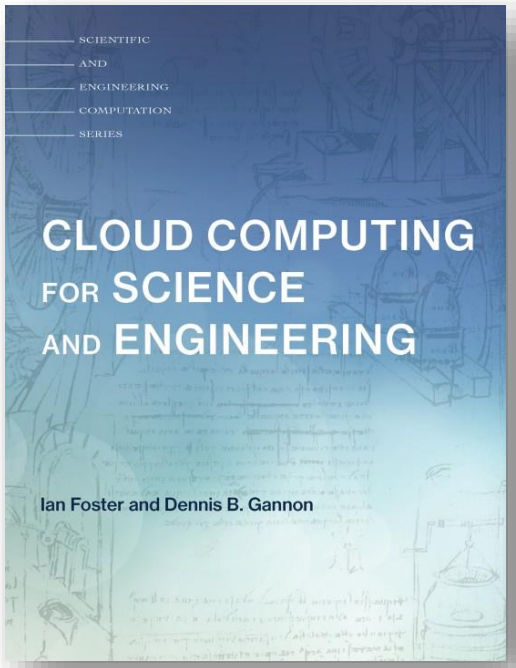
Two sides of the same coin, and your profile is a perfect? fit

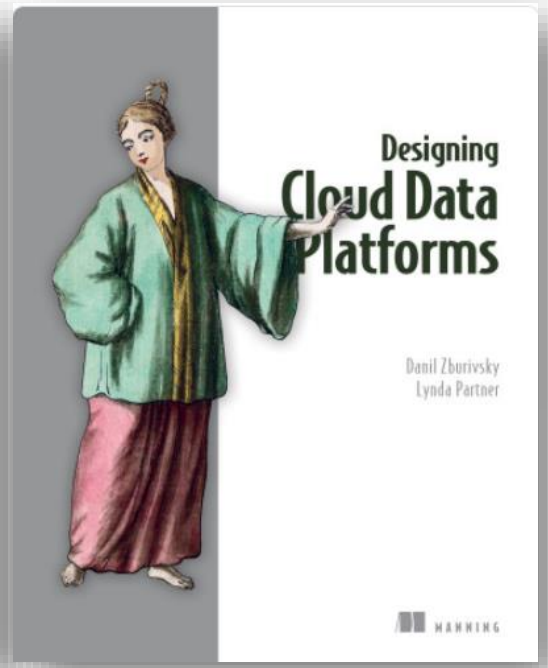
- Technological perspective
  - How do we configure a distributed environment?
  - How do we set up/integrate/control independent services?
  - How do we orchestrate data flows?
- Business perspective
  - Can we afford to spend resources on tasks that are not mission oriented?
  - No free lunch, each choice has cost/benefit
  - How much time does it take to master a technology?
  - How many people do I need?

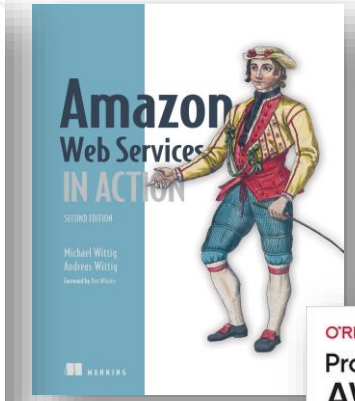
… but first, which are our data needs?

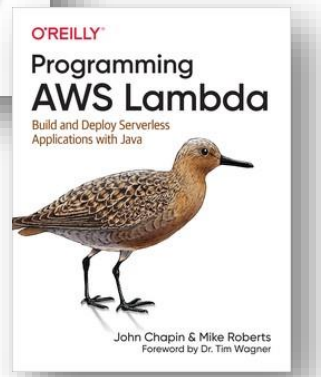
Teaching material

Books



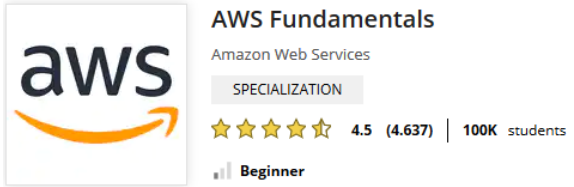


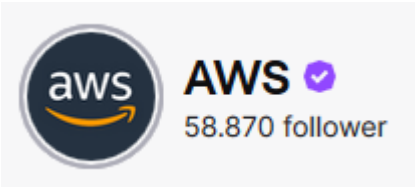




Generic → Specific

Web content





Teaching material

You will find all you need in these slides.

However, keeping up the pace with data platforms and cloud is hard

- There is a rapid development of technologies, and not all of them will survive
- Books are easily outdated with respect to cutting-edge services and technologies
- Research papers (often) describe solutions that are not commercial yet
- (IRL) You will need to dial with a lot of (bad) documentation, online articles, etc.

Rule of thumb

- Understand the general concepts
- Do not be afraid of change