Aalto University
School of Science

Some industrial and open source big data platforms for your tech radar

Hong-Linh Truong
Department of Computer Science
linh.truong@aalto.fi, https://rdsea.github.io

CS-E4640 Big Data Platforms, Spring 2024, Hong-Linh Truong
18/1/2024

Why should we be aware of possibilities?

- **Course lectures give you concepts, techniques, methods and principles**
- **Hands-on allow you to test these techniques and methods with concrete tools and technologies**
  - These tools and technologies are powerful but they may not be 100% of the ones you have to master for your real-world development
- **The technology stack for your real-world applications may change**
  - Being aware of possible technology stack and platforms is IMPORTANT for applying the learned concepts, techniques & methods

Hard decision in practice!

- **Building a big data platform**
  - Complex requirements
  - Complex and diverse available technologies
- **If you are not familiar with existing technologies, where should you start?**
- **If you know some technology stacks: are they suitable for your requirements?**

⇒ Our learning objective is to build a "tech radar" for our "big data platforms" design and development

Hard decision in practice!

- **Many cloud technologies and software stacks**
- **But you/your organization will need to decide**
  - Case 1: use free open sources and build everything
  - Case 2: use free open sources and build platforms but not infrastructures
  - Case 3: use enterprise versions and build everything
  - Case 4: use enterprise versions …
  - Case 5: …

There are many constraints: functionality, budget, data regulation, skills, etc. (for study or for real product)!

In the course, you will have to exercise your decision for your assignments!

The first goal is to be aware of potential solutions!

Let us walk around some stacks/ecosystems

Google for Big Data Platforms

- As a solution catalog
  - https://cloud.google.com/solutions/smart-analytics
- As technologies based on data lifecycle
  - https://cloud.google.com/solutions/data-lifecycle-cloud-platform

Azure for big data platforms

- As service catalog for analytics
  - https://azure.microsoft.com/en-us/services/#analytics
- As solution catalog
  - https://azure.microsoft.com/en-us/solutions/cloud-scale-analytics/

Amazon Web Services

- **Database services**
  - https://aws.amazon.com/products/databases/
- **Analytics services**
  - https://aws.amazon.com/big-data/datalakes-and-analytics/
- **Well-architected Framework - Data Analytics Lens**
  - https://docs.aws.amazon.com/wellarchitected/latest/analytics-lens/analytics-lens.html?did=wp_card&trk=wp_card

Apache *

- https://hadoop.apache.org/
- https://spark.apache.org/
- https://cassandra.apache.org/
- https://hudi.apache.org/
- https://iceberg.apache.org/
- https://hbase.apache.org/
- http://tinkerpop.apache.org/
- https://kafka.apache.org/
- https://pulsar.apache.org/
- https://airflow.apache.org/
- Etc.

Other stacks

- **ELK Stack (ELK, ElasticSearch, Kibana, Logstash)**
  - https://www.elastic.co/elastic-stack
- **The TICK Stack (Telegraf, Infuxdb, Chronograf, Kapacitor)**
  - https://www.influxdata.com/time-series-platform/

Many more software/services: free and commercial

- MongoDB
  - https://www.mongodb.com/
- CockroachDB
  - https://www.cockroachlabs.com/
- Presto vs Trino
  - https://prestodb.io/getting-started/ vs https://trino.io/
- Clickhouse
  - https://clickhouse.com/

Notes on services for big data platforms in existing cloud providers

- Different providers but similar functionality (and built from similar software)
- Coupling with underlying cloud infrastructures
- Coupling among services
- Management features
- Price, privacy, security, programming support, etc.

⇒ We can select a subset of services/software for practicing design and concepts in the course

Notes on key requirements for building a big data platform

- **Common questions for your customer could provide a lot of conditions for selecting technologies**
- **Examples:**
  - What would be the big data platform strategy for a long run?
    - *In-house/On-premise vs public cloud vs hybrid ⇒ would remove many frameworks/tools*
  - How would the big data platform be integrated with existing services/software
    - *external cloud based services would influence a lot of your choices*

Tech Radar

Personal Techradar

- **Techradar**
  - https://www.thoughtworks.com/radar
  - Core principles: identify and assess relevant frameworks, services and techniques for your work!
- **Guide and Example**
  - http://nealford.com/memeagora/2013/05/28/build_your_own_technology_radar.html
  - https://medium.com/@ckoster22/whats-on-your-tech-radar-9ad8769c8c1
- **Focus the radar for this course:**
  - only the Big Data Platforms context for your big data platform story
- **Example of a company specific radar:**
  - https://opensource.zalando.com/tech-radar/

Build and share your techradar

- **Select a suitable real-world dataset (for a domain) and imagine that you need to handle such data in your big data platform**
- **Scan software and services for building your big data platform**
  - *Google Cloud Platform*
  - *Microsoft Azure Cloud*
  - *Amazon Web Services*
  - *Apache *, ELK stack, TICK stack, ...*
- **Why do you think that the tools in your radar are suitable for you?**

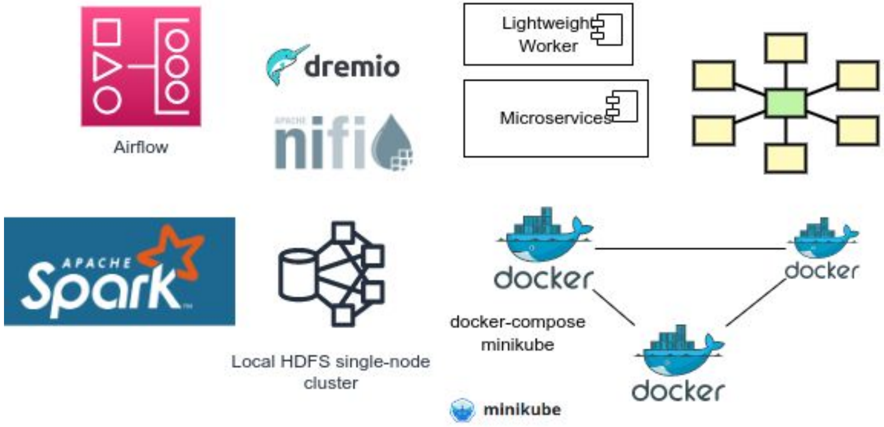
Study testbed - Hybrid Configuration

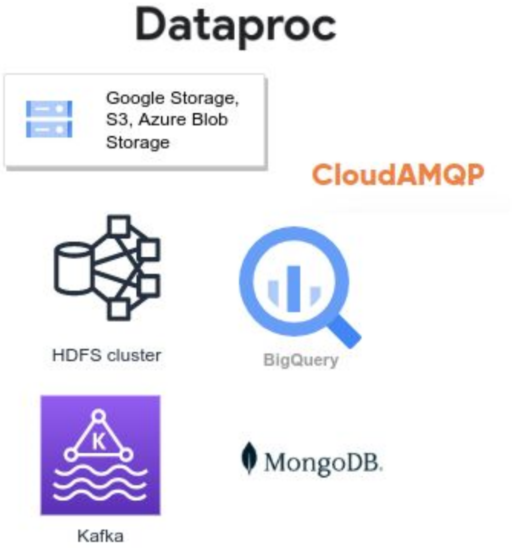
- **Software in your techradar**
  - Some available only in the cloud, some requires a lot of resources
- **Our own computing resources are not enough**
- **Hybrid configurations for study**
  - Computing resources: combine your resources with free/other resources
  - Services: some free, limited services and your own deployment
- **It is important to make right design**
  - e.g., not tightly coupled design because “I have only 4 core CPUs”

Hybrid configurations for studies

Your own resources (laptops, VMs on premise, CSC or cloud)

Cloud services: e.g.,





Key concerns: can you control the configuration for testing?

Thanks!

Hong-Linh Truong
Department of Computer Science

rdsea.github.io