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

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, <u>skills</u>, 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/big-data/

Amazon Web Services

- **Database services**
  - https://aws.amazon.com/products/databases/
- **Analytics services**
  - https://aws.amazon.com/big-data/datalakes-and-analytics/

Apache *

- https://hadoop.apache.org/
- https://spark.apache.org/
- https://cassandra.apache.org/
- https://avro.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

- **MongoDB**
  - https://www.mongodb.com/
- **Neo4J**
  - https://neo4j.com/
- **SAP HANA**
  - https://www.sap.com/products/hana.html
- **Etc.**

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
- Price, privacy, security, programming support, etc.

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

Aalto University
School of Science

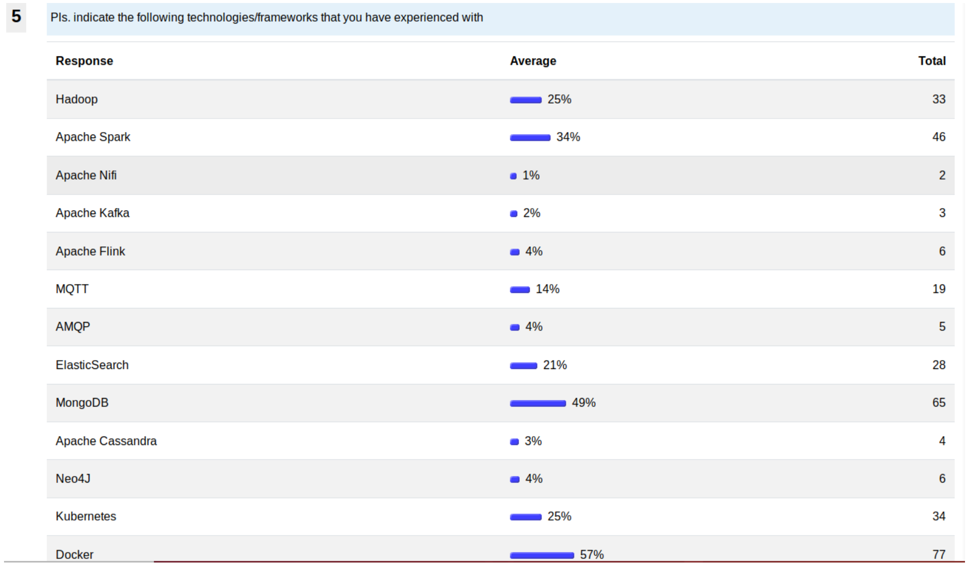
15 minutes breaking sessions for group and self activities:

let us explore/discuss the technologies you know

Tech Radar

Are you happy with your tech radar?

2019 CS-E4640 student survey



5 Pls. indicate the following technologies/frameworks that you have experienced with

Response	Average	Total
Hadoop	25%	33
Apache Spark	34%	46
Apache Nifi	1%	2
Apache Kafka	2%	3
Apache Flink	4%	6
MQTT	14%	19
AMQP	4%	5
ElasticSearch	21%	28
MongoDB	49%	65
Apache Cassandra	3%	4
Neo4J	4%	6
Kubernetes	25%	34
Docker	57%	77

LA TUA OPINIONE E' IMPORTANTE PER UN SERVIZIO MIGLIORE.
Come valuti la tua esperienza di oggi ai controlli di sicurezza?
YOUR FEEDBACK IS IMPORTANT FOR A BETTER SERVICE.
How was your experience with security control today?

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

Final remark

- **Can you build your tech radar and share/discuss it?**
  - 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?

Thanks!

Hong-Linh Truong
Department of Computer Science

rdsea.github.io