Aalto University
School of Science

Selected Trends/Issues for Big Data Platforms

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

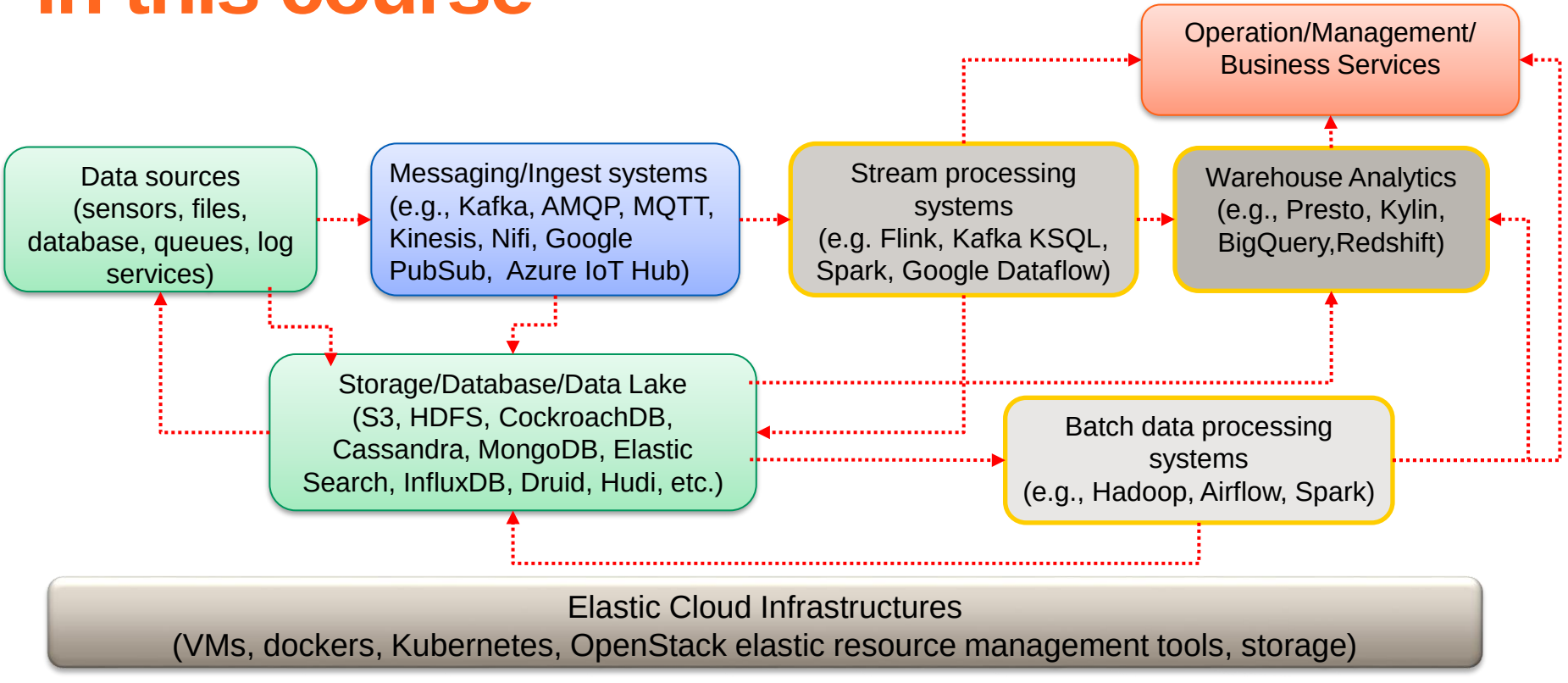
Learning objectives

- **Recap some key points and link them to selected trends/issues that we should address in big data platforms**
- **Understand these challenges in real-world big data design and engineering**

Up to now

- **Big data platforms are complex**
- **We have studied existing works, platforms, and use cases/solutions**
- **Basic, foundational techniques and models**

Big data at large-scale: the big picture in this course



So far

- **Are you happy by just staying with these frameworks?**

- **What would be emerging challenges, relevant/ important aspects that we must investigate further?**
  - think about what could be a good master thesis topic
- **As a researcher or a lead engineer for big data platforms, what could be some aspect that you must study more?**

Extreme big data is still very relevant and hard subject!

Why would EU put for this 30M EUR call? deadline: 05.04.2022

“extreme data mining”

“volume, speed, variety, complexity/diversity/multilinguality of data”

“sparse/missing/insufficient data/extreme variations”

“big data, AI, IoT, HPC, edge/fog/cloud computing”

Topic description

ExpectedOutcome:

Proposal results are expected to contribute to the following expected outcomes:

- provide better technologies, tools and solutions for data mining (searching and processing) of large, constantly growing amounts and varieties of data, and/or extremely sparse/dispersed/heterogeneous/multilingual data (stored centrally or in distributed/decentralized systems), in particular IoT, industrial, business, administrative, environmental, scientific or societal data.

Scope:

The actions under this topic are expected to provide ground-breaking advances in the performance, speed and/or accuracy as well as usefulness of data discovery, collection, mining, filtering and processing in view of coping with “extreme data”: (defined as data that exhibits one or more of the following characteristics, to an extent that makes current technologies fail: increasing volume, speed, variety; complexity/diversity/multilinguality of data; the dispersed data sources; sparse/missing /insufficient data/extreme variations in values). The technologies and solutions are expected to discover and distil meaningful, reliable and useful data from heterogeneous and dispersed/scarce sources and deliver it to the requesting application/user with minimal delay and in the appropriate format. In particular, the advances should enable the development of trustworthy, accurate, green and fair AI systems where quality of data is as important as quantity and/or support industrial distributed decision-making tasks at appropriate level in the computing continuum (edge/fog/cloud). Insofar the results are intended for human use, the design of these tools should take into account the relevant human aspects and interactions with users.

The actions should address the integration of relevant technologies (e.g. big data, AI, IoT, HPC, edge/fog/cloud computing, language technologies, cybersecurity, telecommunications, autonomous systems etc.) as a means towards achieving the goals, and foster links to the respective research, industrial and user/innovator communities (e.g. AI4EU, digital innovation hubs). The use of European data sources (such as Copernicus, Galileo/EGNOS for satellite data) is encouraged in the use cases, where appropriate.

Source: https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/horizon-cl4-2022-data-01-05

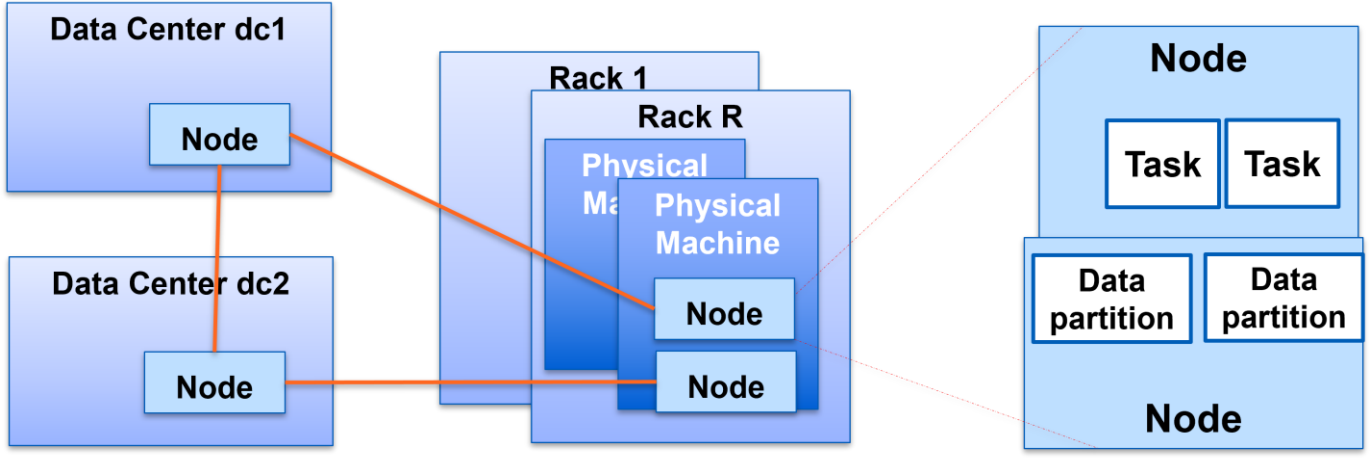
Selected trends/issues for CS-E4640, 2022

- **Alignment of technical design with domain-oriented business and social/organizational structure**
- **Big data in emerging computing continuum**
- **Design for data-first applications (e.g., machine learning)**
- **Sustainability in big data platforms**

Aalto University
School of Science

Design and engineering for complex interactions between data management, analytics, infrastructures and people in a business

Cloud-centric Big Data



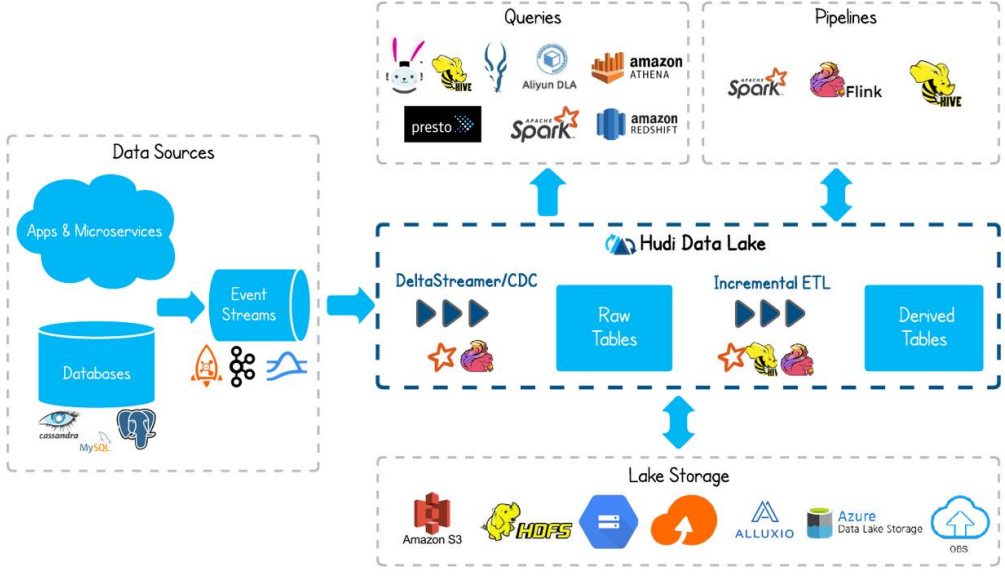
- **What we have learned/observed**
  - different designs for big data
  - complex technologies → complex infrastructures → DevOps, DataOps, service and cloud engineering → change management
  - people working in big data infrastructures may not know data well

Integration of every type of analytical data: data lake

- **All-in-one data integration**
  - massive of datasets, in different collections, in different formats, in different types of data storages
    - *internal/external data, operational/analytical data, raw/clean/training data*
  - for multiple types of analytics/ML
  - example of technologies: Apache Hudi, Delta Lake
- **Integration from technology viewpoint**
  - data from various sources, infrastructure teams, assumption of all kind of possible analytics for different products/departments

Example: Data Lake consolidation and integration

Figure source: https://hudi.apache.org/



different analytics

lake core

different storage

Problems: hard to govern, technology-centric, lack of considerations of data products, cost of effort, etc.

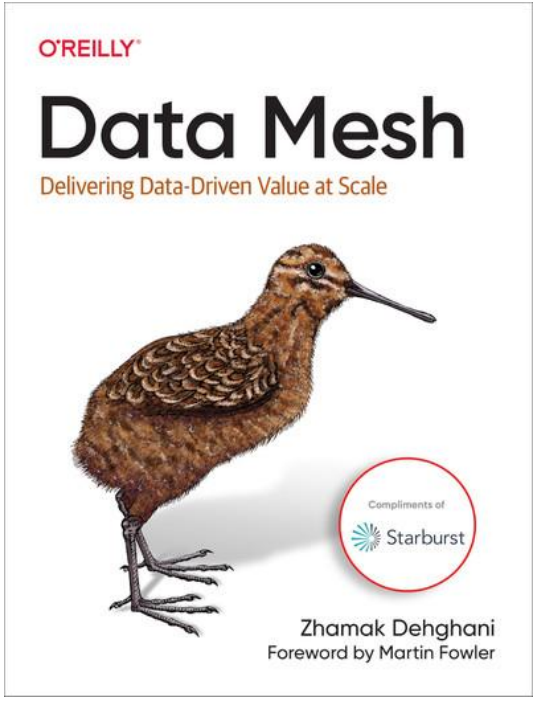
Paradigm change: Data mesh

seems very trendy in industry. What is it?

"data mesh as a *sociotechnical paradigm*: an approach that recognizes the interactions between people and the technical architecture and solutions in complex organizations."

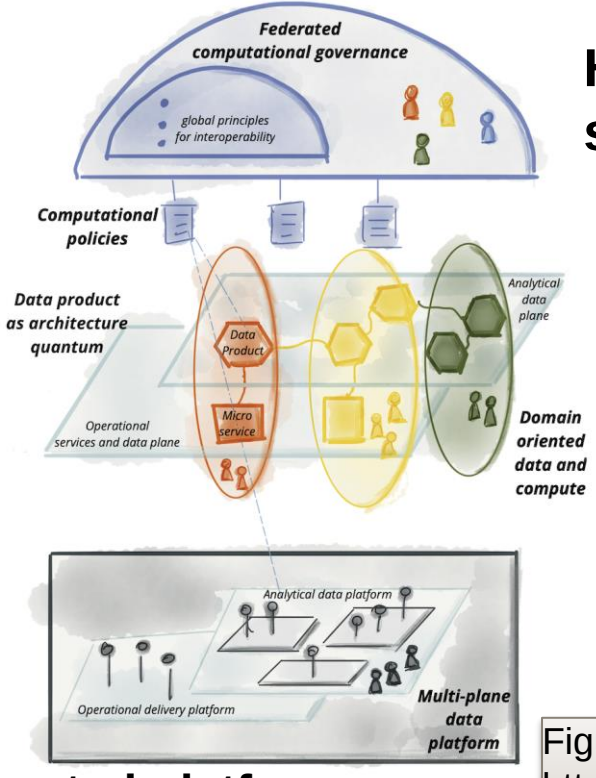
Source: Data Mesh, Zhamak Dehghani

(PS: not like service mesh)



Why is it interesting and relevant?

Treat data as product requires us to have different customer relationships, quality, deliveries, etc. not "data is just part of something or a fragment"



Holistic data governance: see the big picture of data

Domain- oriented ownership: decentralized architecture, data, people (engineer, scientist), rules, etc.

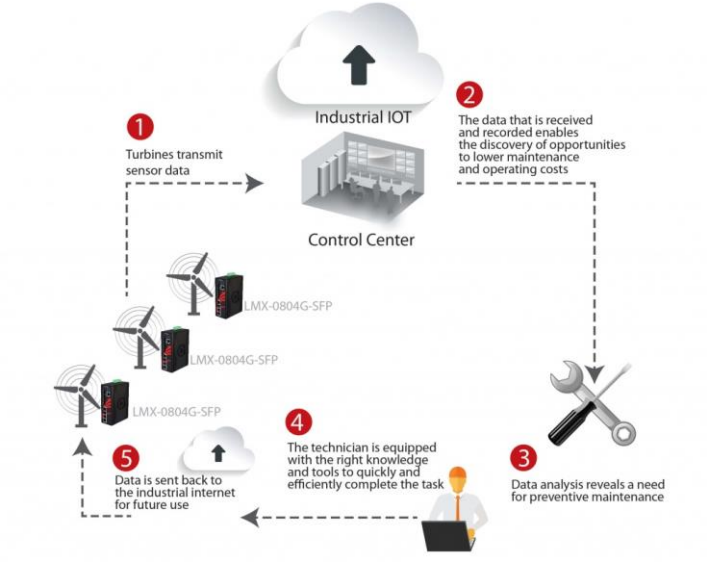
Domain agnostics, integrated platforms, code/data/policy packages for data products

Figure source: https://martinfowler.com/articles/data-mesh-principles.html

Aalto University
School of Science

Design for the entire computing continuum

Data is everywhere in the entire computing continuum: IoT-edge-cloud-HPC continuum



Figures source:
http://www.windpowerengineering.com/design/electrical/controls/wind-farm-networks/talking-turbines-internet-things/

Source: Erik Christensen, http://www.sensorfish.eu/

Data is everywhere in the entire computing continuum

- As a platform design, we must look at the data source infrastructures (e.g., telescopes, cameras, IoT sensors, drones, …)
- Data and tenant's needs spread across edge, cloud and HPC systems

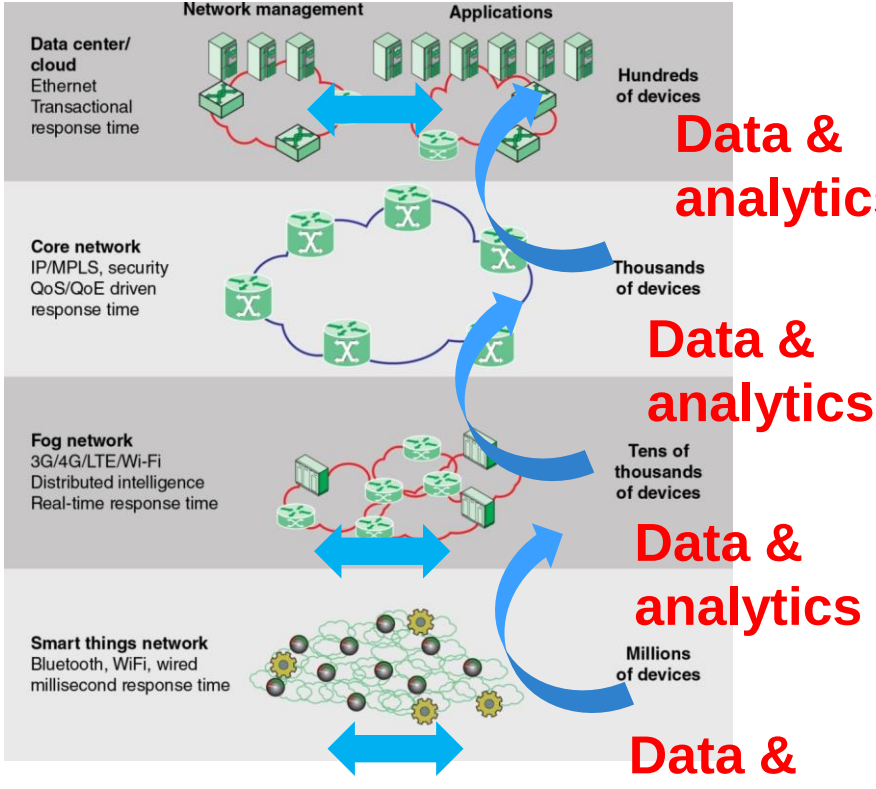
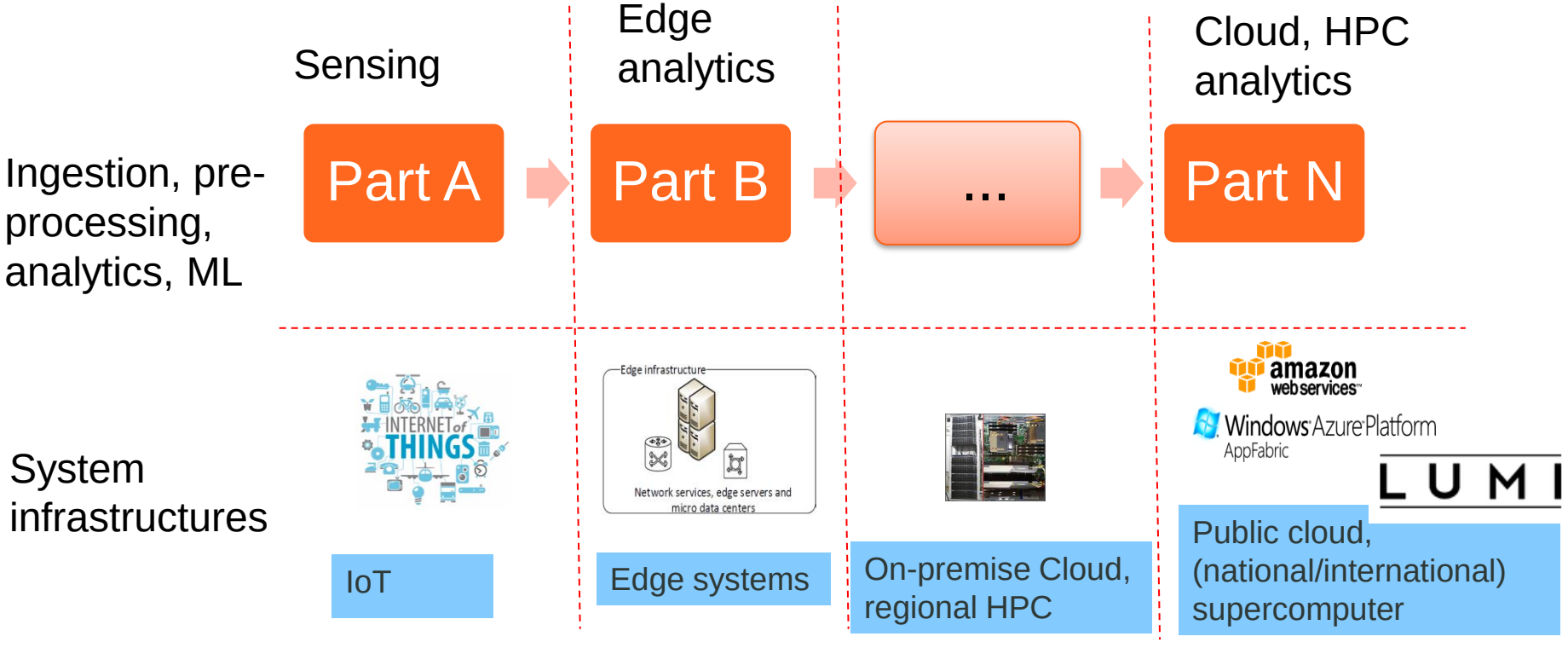


Figure source: From *Foundations of Modern Networking: SDN, NFV, QoE, IoT, and Cloud* by William Stallings (0134175395), Copyright © 2016 Pearson Education, Inc. All rights reserved.

Advanced cross systems big data analytics



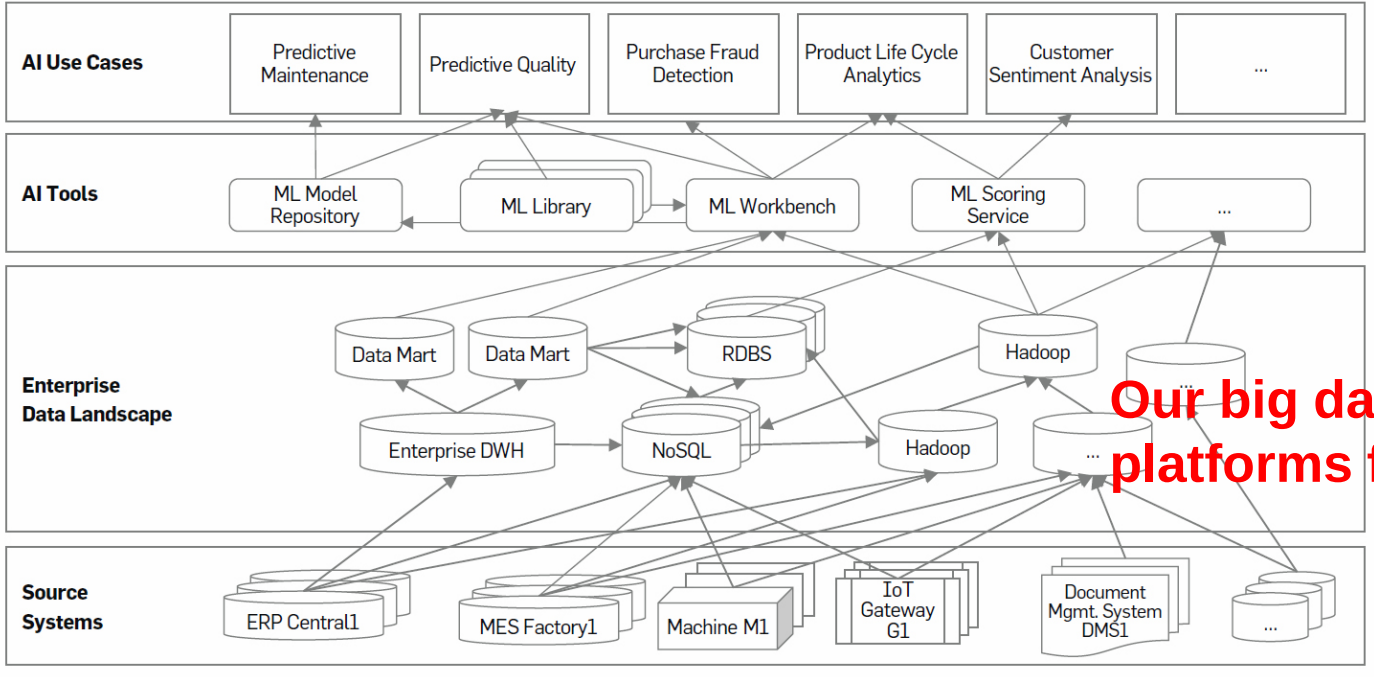
Cross edge-cloud big data analytics

- **Highly distributed edge-cloud data environments**
  - highly parallelized data processing models often seen in large-scale cloud-based and HPC data analytics do not work
- **Workflows across edge and cloud and possible HPC**
- **Quality of analytics is hard to manage**
  - Not just about data but also services
- **Federated data services for federated ML**

Aalto University
School of Science

Data-first applications (ML)

Everyone knows “No AI Without Data”

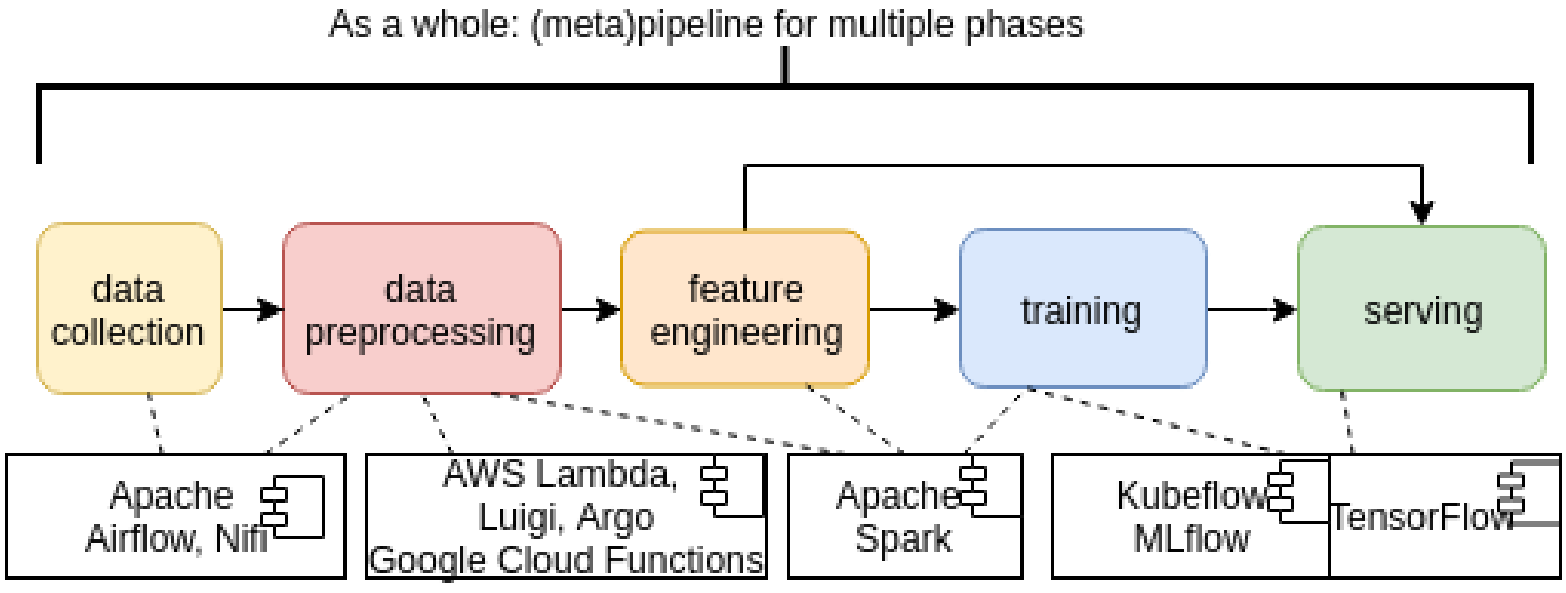


DWH: Data Warehouse, ERP: Enterprise Ressource Planning, MES: Manufacturing Execution System, ML: Machine Learning, RDBS: Relational Database System

Our big data platforms focus

Figure source: There Is No AI Without Data, by Christoph Gröger, Communications of the ACM, November 2021, Vol. 64 No. 11, Pages 98-108, 0.1145/3448247

Big data and ML pipelines



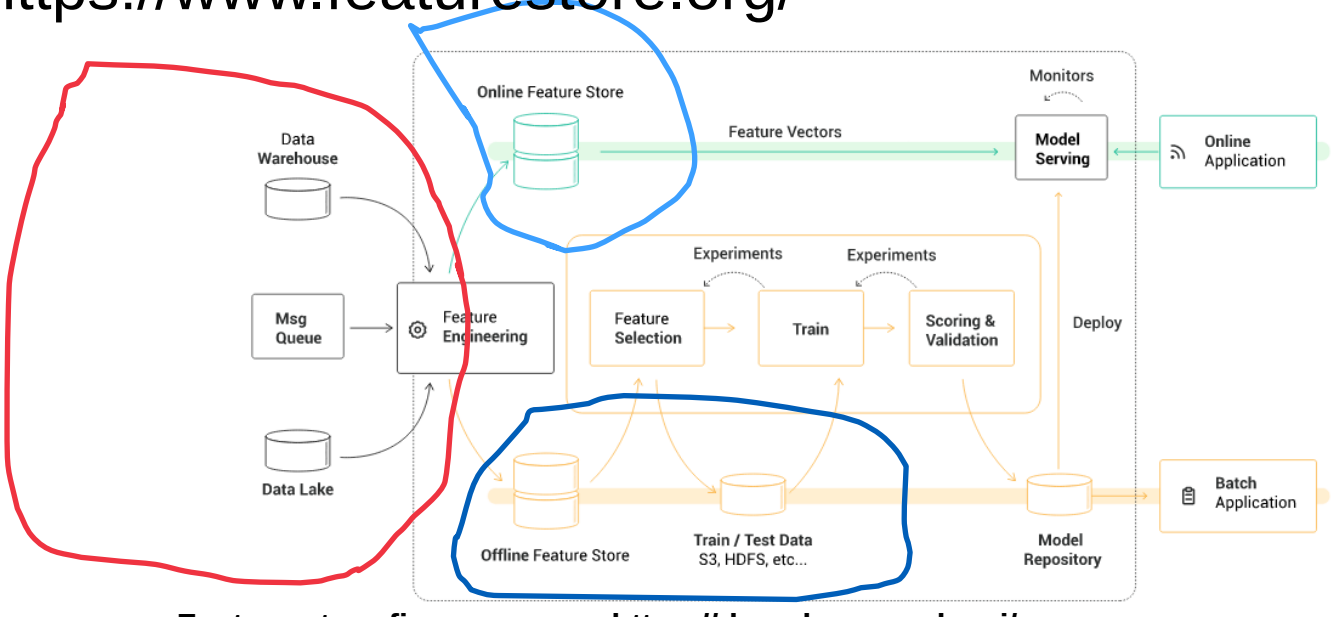
Examples of subsystems: different components and internal workflows

What would be the impact of AI/ML on big data platforms design

- **What if ML is a key part that big data platforms must support (actually, it happens now)?**
  - We must look at the detail of ML requirements/processes
- **Data management for ML**
  - enable sharing, retrieval and managing features for ML training and models
- **Integration of big data processing with ML**
  - parallel/distributed data analysis within ML
- **Data governance**

Big data management and ML

A hot topic: Feature Store and Data Engineering
https://www.featurestore.org/



Feature store figure source: https://docs.hopsworks.ai/

Big data platforms and ML

Example: Michelangelo ML from Uber

It is easy to see the role of big data components we study in this figure

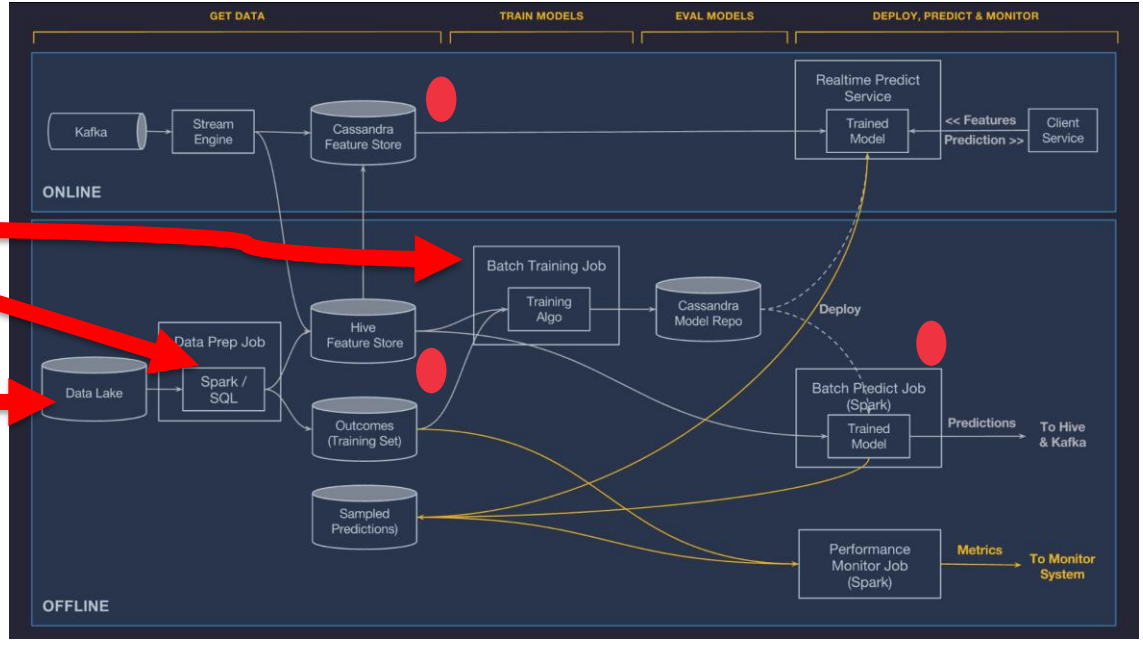
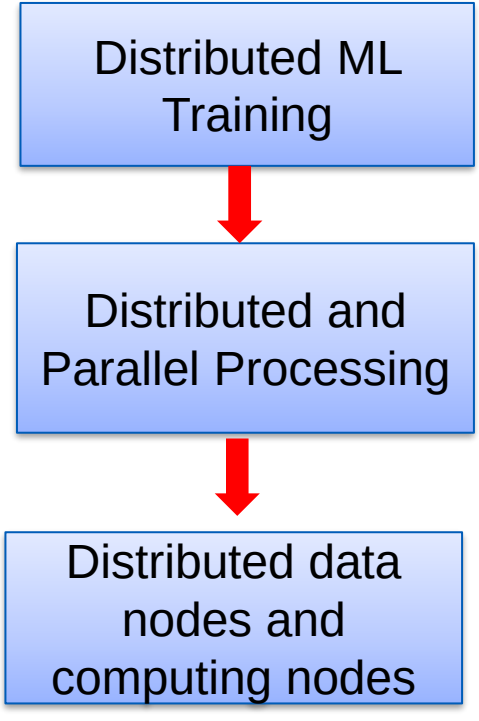


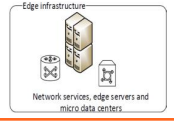
Figure source: https://eng.uber.com/michelangelo-machine-learning-platform/

Distributed training

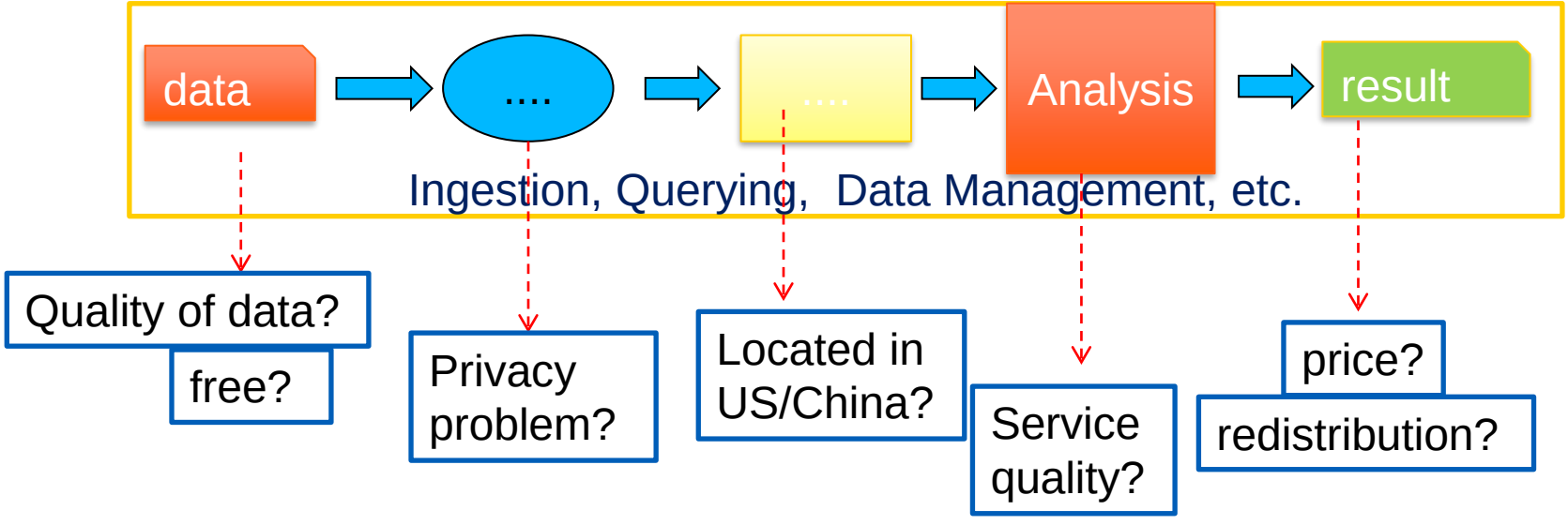
Now if we look at from the perspective of big data platforms: how would you design your big data platforms so that they can be easily integrated with distributed training in "data hungry" ML?



Open MPI



Data governance



- Ethical consequences?
- Regulation-compliant platforms: e.g., GDPR

Data about datasets, including quality

- **Data governance is hard**
  - Must have a data about datasets, e.g., possible datasheets/metadata with 55 questions?
- **How do we automate data governance processes?**
  - Tools (Google Data Catalog, Apache Atlas, Linkedin DataHub) are not enough
  - Federated governance: technologies + people + automation?

Template: https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4t8QB

Aalto University
School of Science

Sustainability

Sustainable Big Data Platforms

- **UN Sustainability Development Goals (SDGs):**
  - https://www.un.org/sustainabledevelopment/sustainable-development-goals/
- **Obviously, big data platforms play a big role in supporting SDGs**
  - Collect, process and manage data for SDGs in many domains
- **But what about "sustainable big data platforms"?**
  - A broad view: critical infrastructures for industry and innovation, responsible production, human effort, energy consumption
  - E.g., in your experiences with assignments, how much does automation help reduce your effort?

Do big data platforms also introduce a lot of waste (time, energy, space, etc.)

Original Research/Scholarship | Open Access | Published: 23 December 2019

Big Data, Big Waste? A Reflection on the Environmental Sustainability of Big Data Initiatives

Federica Lucivero

Science and Engineering Ethics **26**, 1009–1030 (2020) | Cite this article

14k Accesses | **18** Citations | **32** Altmetric | Metrics

See: https://link.springer.com/article/10.1007/s11948-019-00171-7

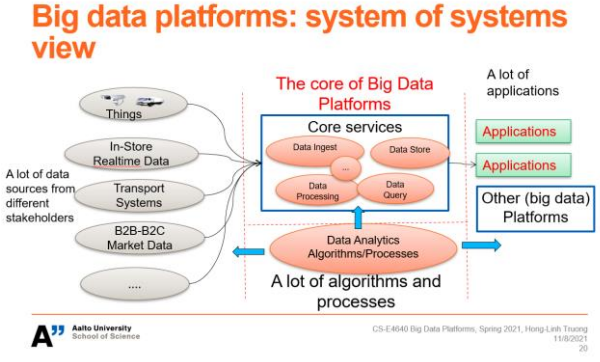
Examples of best practices to reduce energy and carbon footprints

- **From Google (the 4Ms): reduce "energy by 100x and emissions by 1000x".**
  - "efficient ML Model"
  - optimized "Machine"
  - "Mechanization" w.r.t cloud and on-premise
  - "Map Optimization" w.r.t green/location-specific data centers

Source papers:
https://ai.googleblog.com/2022/02/good-news-about-carbon-footprint-of.html
https://www.techrxiv.org/articles/preprint/The_Carbon_Footprint_of_Machine_Learning_Training_Will_Plateau_Then_Shrink/19139645

We probably see a lot of wastes in our own big data platforms due to our design and implementation?

Integrating sustainability into the design requirements for big data platforms



Key technical requirements

- **Reliability**
- **Resilience**
- **Elasticity**

change

Sustainability Integration

- **Reliability**
- **Resilience**
- **Elasticity**
- ***Sustainability***

Efficiency and sustainability

- **Highly efficient might not be good for sustainability**
- **Sustainability within big data platforms**
  - energy consumption, reusability, extensibility

Design and implementation

- **Reliability**
- **Resilience**
- **Elasticity**

How to have it for

- ***Sustainability***

Sus. goal | Platform design | Concrete techniques | Verification

Start with something we are familiar with

- **Less energy consumption in the design**
  - we can consider it is one objective for supporting the SDG *"Take urgent actions to combat climate changes and its impacts"*
- **Reuse/recycle and automation in the design**
  - we consider it as one objective for supporting the SDG *"Ensure sustainable consumption and production patterns"*

Work on what, when/where, and how!

What: Sustainability aspect	When/Where: Data/Services design in Big Data Platforms	How: Concrete Techniques	How: Verification of the Benefit
Less energy consumption (ecological sustainability to reduce CO2)	• In handling data to reduce data transfers • In elastic service provisioning to reduce under utilized services	• Better techniques: e.g., Filter and compress data during the data ingestion pipeline • Using elastic techniques to manage services	It is hard to measure without real energy consumption tools but we can measure the CPU and time of computing resources to verify if the design helps to reduce energy consumption
Reuse/recycle and automation (economic sustainability w.r.t efficiency, social/cultural sustainability w.r.t. human resources)	• In data pipeline, service APIs, interface to external systems, integration to external providers • When automating deployments and configuration	• Use pipeline design composition and microservices with containers in your design • Standard APIs for the services • Build generic connectors that can be reused • Select green services from providers • Implement automation techniques	Measure effort to deploy services and pipelines in different computing environments, manual effort due to the use of automation (also examine how we reduce the human workforces by implementing automation techniques)

Summary

- **Complex design and engineering issues in big data platforms**
  - many big systems, very complex designs and requirements
  - require us to continuously upskill

- **Still, what we learn so far are just "foundation"**
  - social-technical paradigm challenges in big data for data products
  - continuum computing and big data platforms
  - designs for data-first applications
  - sustainability

Thanks!

Hong-Linh Truong
Department of Computer Science

rdsea.github.io