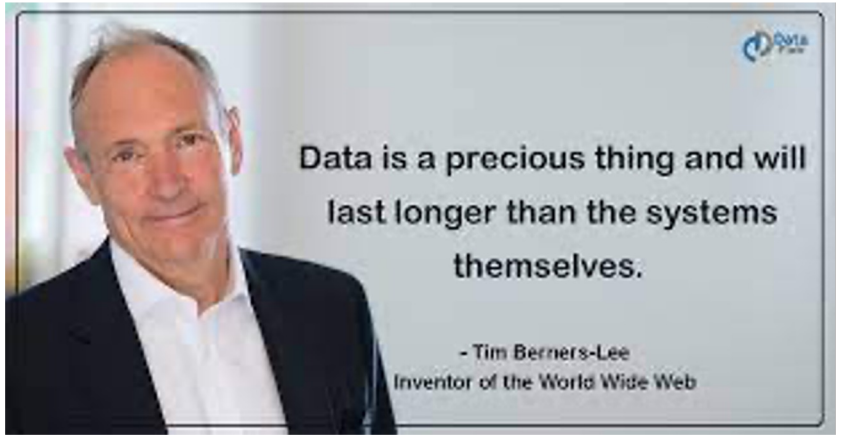
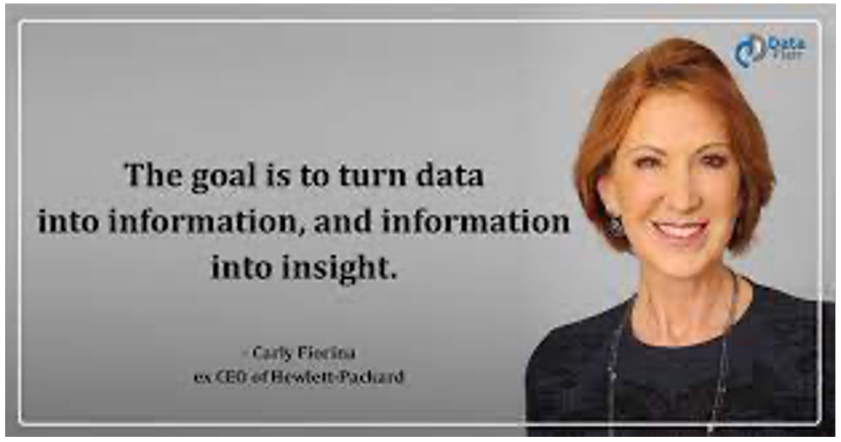
CS760: Systems for Data Management and Data Science

Prof. Angelos Anadiotis
Prof. Anne-Marie Kermarrec

Lecture 1 : Introduction
Feb 20, 2023

Data is one of the most valuable resource





Data science

A data-driven approach to problem solving by analyzing and exploring large volumes of possibly multi-modal data.

It involves the process of collecting, preparing, managing, processing, analyzing, and explaining the data and analysis results.

Data science is interdisciplinary (statistics, computer science, information science, mathematics, social science, visualization, etc.).

Debunking some myths

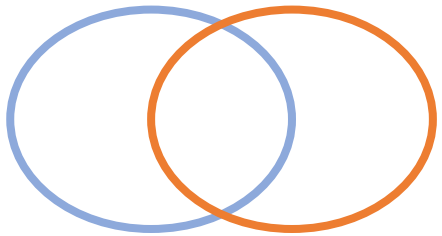
- Myth 1

  Data Science = Big Data

  - Big data = raw material
  - Applications are crucial

- Myth 2

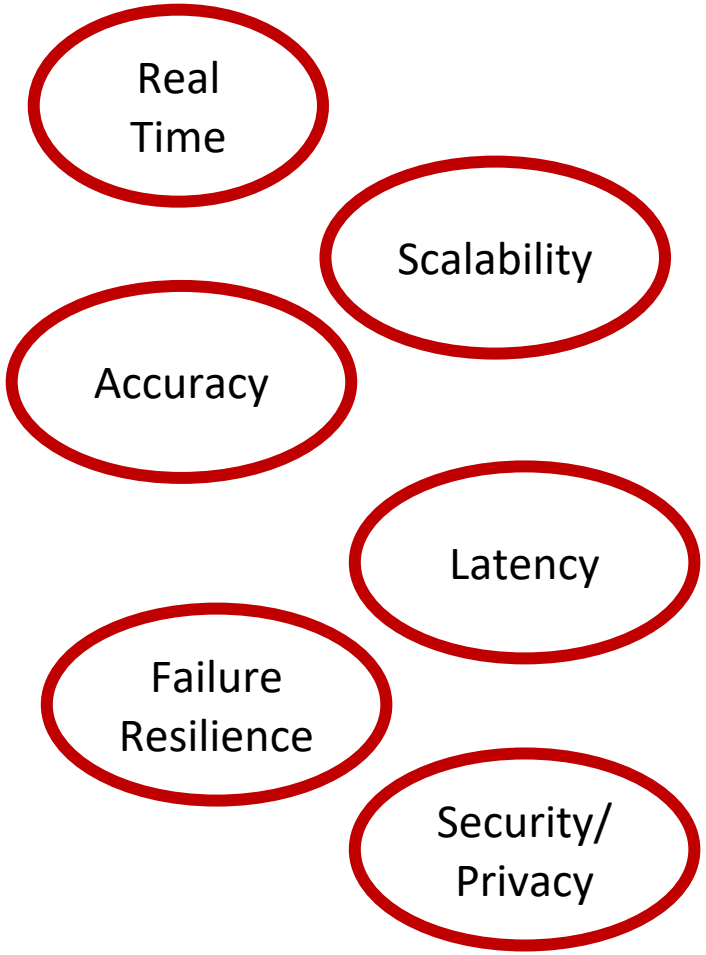
  Data science = Machine Learning



Related but not the same

No Data science without Data science applications

- Fraud and Risk detection
- Gaming
- Recommendation Systems
- Healthcare
- Digital advertisement
- Smart cities
- Finance
- Sustainability
- Image and speech recognition
- Augmented reality



The many faces of data science

Data engineering

Big data management
Data preparation
Large-scale deployment

Data analytics

Data exploration (mining)
Models and algorithms (ML)
Visualizations

Data Security and Privacy

Data integrity
Differential privacy
Cryptography

Data Ethics

Biases (data and algorithms)
Impact on society
Regulations

The many faces of data science

Data engineering

(Data preparation)
Big data management
Large-scale deployment

Why are you here?

Data Science / ML Engineering internships for university students and graduate

Vous avez trouvé un poste qui correspond à vos compétences et vous êtes motivé à relever ce nouveau défi? Postulez et faites-vous connaître de nos spécialistes en recrutement.

VOS TÂCHES

- Join a team of highly skilled engineers and data scientists who are delivering innovative data-driven solution into production
- Work with cutting-edge cloud and on-prem technology (i.e., python, Spark on Databricks, Azure Machine Learning, ...) and rich datasets
- Develop value-generating data solutions that will drive the digital transformation of the company
- Have the independence and time to prototype your ideas and develop your skills
- Apply your skills and knowledge to help drive a company and industry-wide digital transformation

- Experience working with classifiers that scale well with a large number of samples (e.g. approximate kNN).

- Knowledge of clustering algorithms that scale well with large numbers of samples (e.g. minibatch K-means, OPTICS, BIRCH).
- Knowledge of recommender systems (e.g. Matrix factorization).
- Expertise in the processing of large datasets (e.g. via Spark, Azure Databricks) using large-data storage (e.g. Parquet on Azure blob storage, Databricks Delta).

- Work in cross functional teams to drive the adoption of data privacy solutions (e.g. data anonymization) and technologies (e.g. Confidential computing, federated learning etc)

Systems for data science: Learning objectives

- Solid foundations for understanding large-scale distributed systems used for storing and processing massive data
- Cover advanced topics in data-intensive computing platforms
- Solve a real problem on multiple machines

1. Explaining data-intensive platforms
2. Storing and retrieving data in distributed stores
3. Processing data
4. Building advanced applications on a cluster of computers

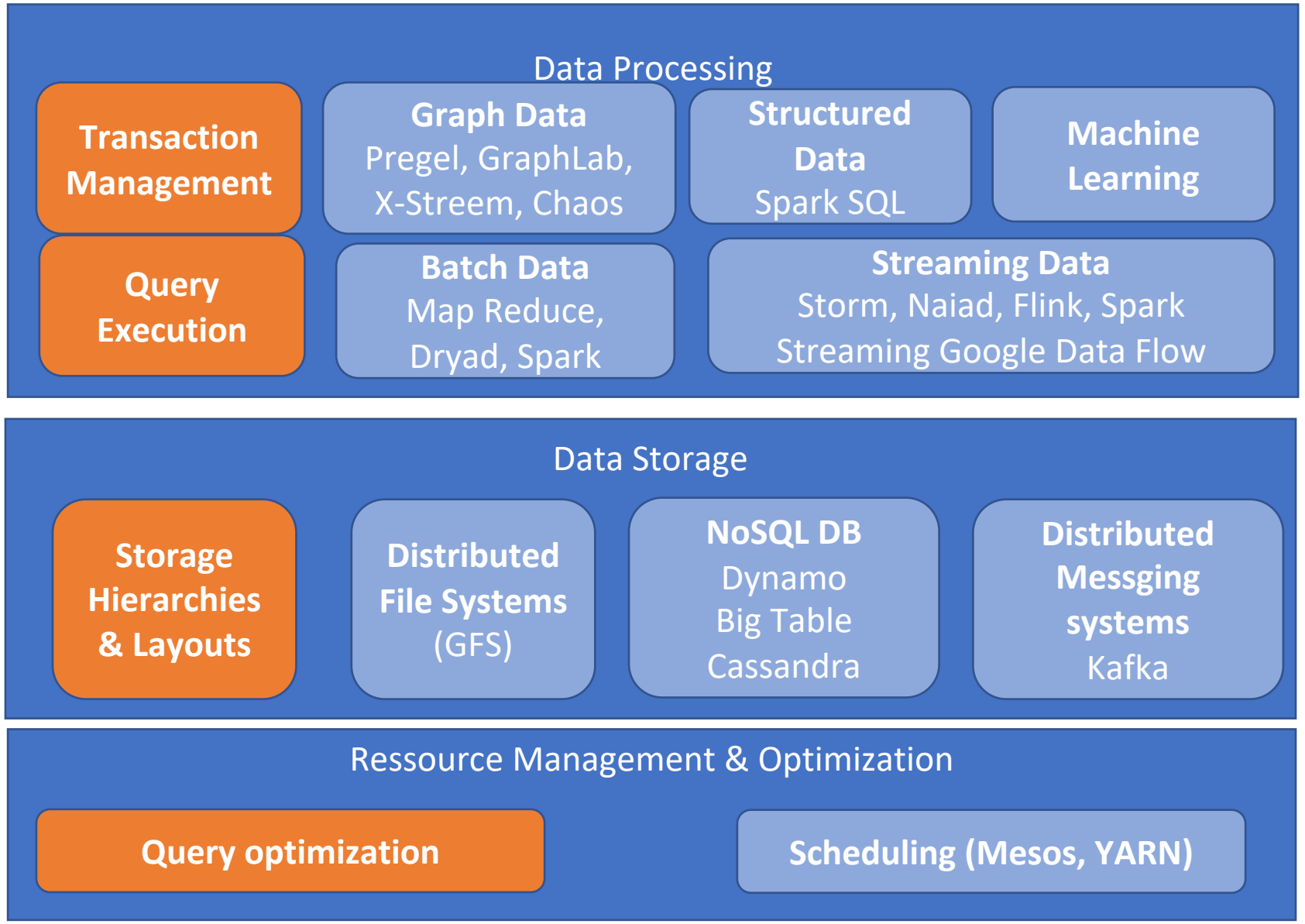
What we will cover in this class

- Architecture
- Storage systems
- Data processing
- Concurrency and fault tolerance

What we will cover

- Gossip Protocols
- Consistency protocols
  CAP Theorem
- Distributed/decentralized systems

Data science software stack



Week	Date	Topic	Prof
1	20/02	Introduction- Course Logictics – Gossip Protocols	AMK
2	27/02	DHT – Consistency models	AMK
3	06/03	Transactions	AA
4	13/03	Key-value Store – CAP Theorem	AMK
5	20/03	Storage Hierarchy	AA
6	27/03	Batch Processing - Map-Reduce	AA
7	03/04	Mid-term	
8	17/04	Scheduling	AMK
9	24/04	Query Execution & optimization	AA
10	01/05	Stream Processing	AMK
11	08/05	Distributed Transactions	AA
12	15/05	Distributed Learning Systems	AMK
13	22/05	Guest Lecturer	Guest Lecturer
14	29/05	Day off	

Expected outcome

Course

Learn the internals of a (distributed) platform for data science

Breadth coverage

Exercises

Put the course in practice

Programming skills

Exam preparation

Background for the project

Project

Acquaintance with a real platform

Going in depth

Intended as a practical work

Will not be related to every part of the course

Course logistics

- CS460 Moodle: all the material needed, updated every week

- Schedule
  - Lecture (Monday 1:15-3 pm) – CM2- Recorded, not streamed.
  - Exercises (Monday 3:15-5 pm) – CM2
  - Individual Project (Wed 2-4 pm : time slot to work on the project, no need to be available or present during that time)

- Grading scheme
  - Project (40%)
  - Midterm (30%)
  - Exam (30%)

Important dates

- Week 1: Presentation of the project (exercise slot)
- Week 2: Presentation of Scala (exercise slot)
- Project
  - No intermediate milestone
  - Mostly Automatic Grading
  - Deliverables: report, source code
  - Deadline: 19/05/2023
- Programming exercises
- Mid-term: 03/04/2023
- Exam: Exam session

CS-460 Project

- Data analysis and movie recommendations with Apache Spark

- Three milestones
  - Milestone 1: Analyzing data with Spark
  - Milestone 2: Data processing pipelines
  - Milestone 3: Movie Recommendation Serving

- Individual project
- Deadline: 19th May 2023
- Auto-graded on GitLab

Lecturers

Dr Angelos Anadiotis

Prof. Anne-Marie Kermarrec

TA Team

Rafael (Postdoc)

Ioan (AE)

Giacomo (AE)

Hamish (TA)

Aunn (TA)

Rishi (TA)

Akash (TA)

Who does what?

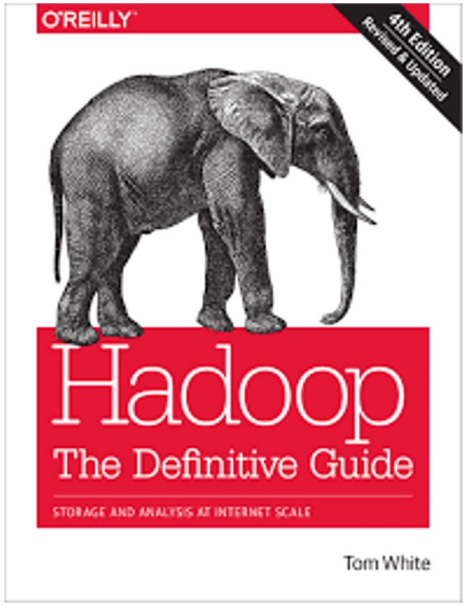
Professors: Angelos.anadiotis@epfl.ch, Anne-Marie.Kermarrec@epfl.ch

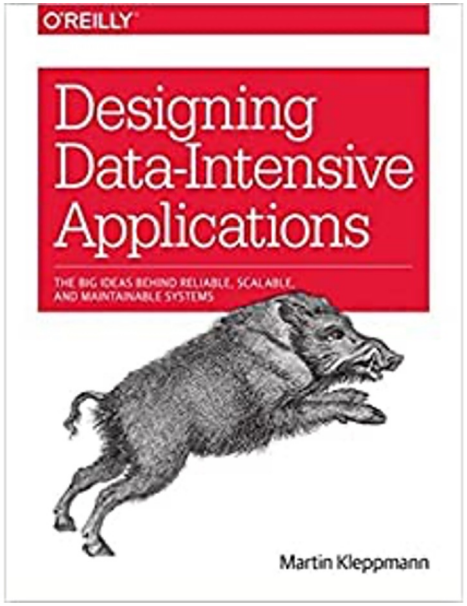
Project/exercices: Rafael.Pires@epfl.ch

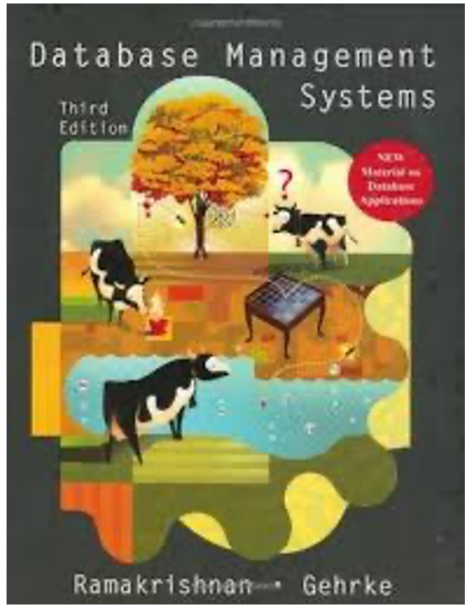
rishi.sharma@epfl.ch (TA), akash.dhasade@epfl.ch (TA), aunn.raza@epfl.ch (TA), hamish.nicholson@epfl.ch (TA), giacomo.orsi@epfl.ch (AE), ioan.nitu@epfl.ch (AE)

Course material

Will be available on Moodle before the session (watch the recommended reading as well)







Scalable Computing Systems Laboratory

Computing systems that make human sense of big data are now ubiquitous. Equipped with powerful AI algorithms, they are now present in all aspects of our life: they drive cars, do surgery, control the lighting in your home, recommend movies and books, and are even about to replace banks.

Data-intensive applications

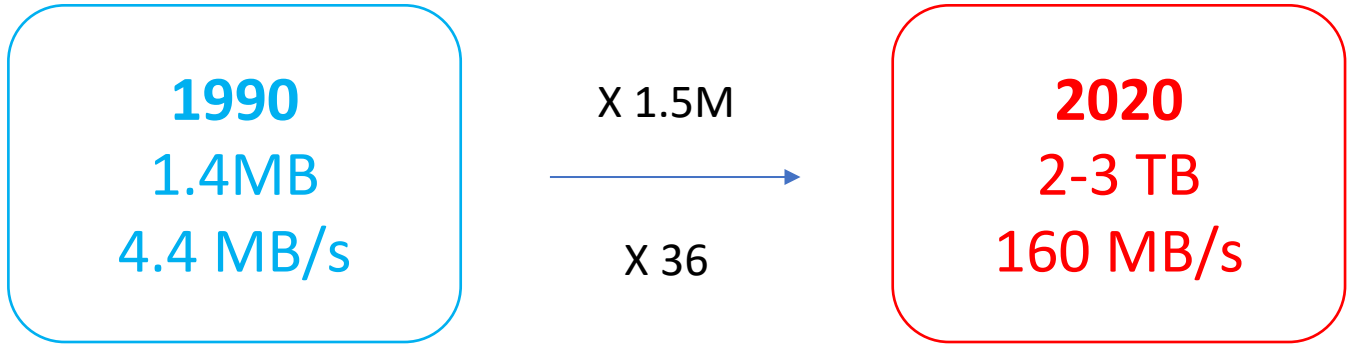
A daily 2.5 quintillion

2.5 quintillion bytes of new data are produced by humans every day

- 2.5 quintillion pennies would cover the earth 5 times
- ¼ of all individual insects alive at any time
- 250 million human brains in neurons

Data storage and analysis

- Storage capacities of hard drive have increased massively
- But not the access speed



Data underlying Data Science: Big Data

Whatever data we cannot handle on a single computer and that does not fit in RAM

And data is not only big but it is also fast/dynamic

Data-intensive applications

- Data-intensive application versus compute-intensive
  - Volume, complexity, speed of changing

- Basic functionalities
  - Store data
  - Speedup reads
  - Filter data (search indexes)
  - Asynchronous communication (stream processing)
  - Periodic computation on data (batch processing)

Reliability

- Application functions as expected
- Fault tolerance (HW, SW, human)
- Performance remains good enough
- System prevents unauthorized access and abuse

Scalability

- What if your system grows from 50,000 concurrent users to 10M
- Scalability: ability to cope with increasing load
- Load: number of requests/second, ratio of reads/writes in a database, number of simultaneously active users…

- Measure performance
  - Latency/Response time: duration for a request to be handled
  - Average versus percentiles
    - The 95th: response time at which 95% of requests are faster than that threshold
  - Tail latency: refers to high latencies that clients see fairly infrequently

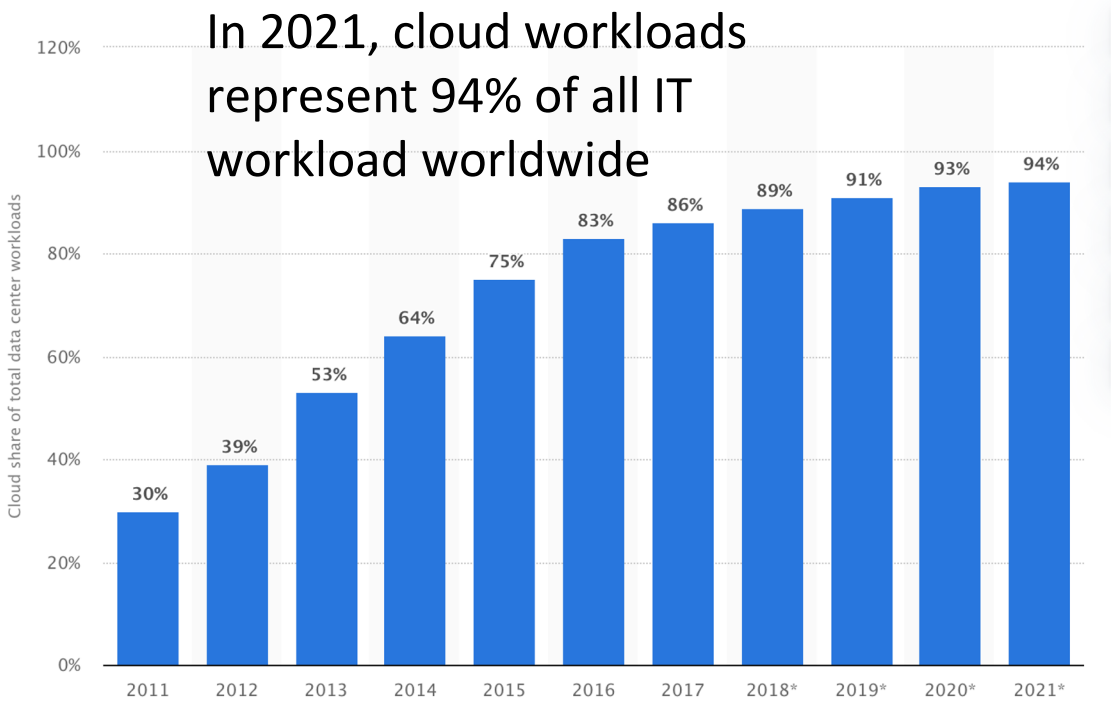
Prophecy

In 1965, MIT's Fernando Corbató and the other designers of the Multics operating system envisioned a computer facility operating “like a power company or water company”.

This is what Cloud Computing addresses

The rise of the Cloud

Move Big Data storage and processing to the cloud



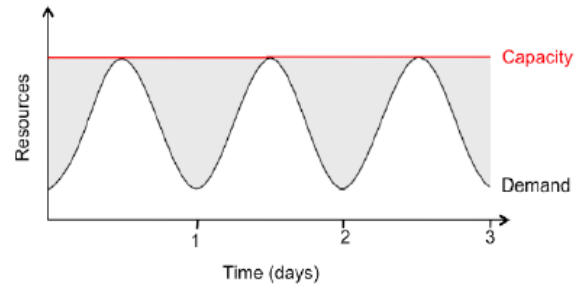
Statista 2021

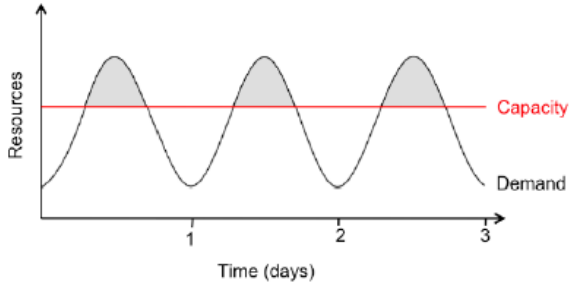
FB spent 16Bn $ on datacenter in 2019
Google spent 13Bn $ datacenter in 2019

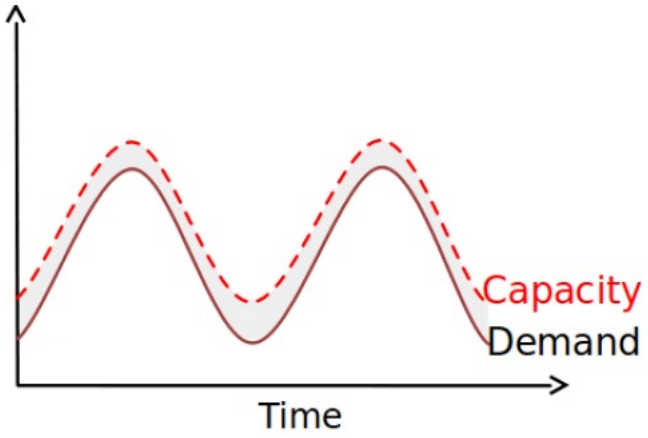
Hyperscale datacenters (fitting more IT in less space, scale hugely and quickly to increasing demand (computing ability, memory, networking infrastructure, storage resources) have been growing at a historic rate over the past 5 years

Cloud Characteristics

- On-demand service
- Ubiquitous network access
- Location transparent resource polling
- Rapid Elasticity
- Measured service with pay per use







Features in Today's Clouds

1. Massive scale.
2. On-demand access: Pay-as-you-go, no upfront commitment, and anyone can access it
   - AWS Elastic Compute Cloud (EC2): a few cents to a few $ per CPU hour
   - AWS Simple Storage Service (S3): a few cents per GB-month
   - HaaS-Hardware as a Service: your own cluster
   - IaaS-Infrastructure as a Service:
     - Access to flexible computing and storage infrastructure.
     - Virtualization: allocate physical resources and enforce isolation (Kubernetes, Dockers, VMs,…).
     - Ex: Amazon Web Services (AWS: EC2 and S3), OpenStack, Eucalyptus, Rightscale, Microsoft Azure, Google Cloud.
   - PaaS- Platform as a Service Access to flexible computing and storage infrastructure, coupled with a software platform (often tightly coupled) - Ex: Google's AppEngine
   - SaaS-Software as a Service: Access to software services, when you need them (Ex: Google docs)

Features in Today's Clouds

1. Massive scale.
2. On-demand access: Pay-as-you-go, no upfront commitment, and anyone can access it
3. Data-intensive Nature: What was MBs has now become TBs, PBs and XBs.
4. New Cloud Programming Paradigms: Easy to write and run highly parallel programs in new cloud programming paradigms:
   - Amazon: Elastic MapReduce service (pay-as-you-go)
   - Google (MapReduce)
     - Indexing: a chain of 24 MapReduce jobs
     - ~200K jobs processing 50PB/month (in 2006)
   - Yahoo! (Hadoop + Pig)
     - WebMap: a chain of several MapReduce jobs
     - 300 TB of data, 10K cores, many tens of hours (~2008)
   - Facebook (Hadoop + Hive)
     - ~300TB total, adding 2TB/day (in 2008)
     - 3K jobs processing 55TB/day
   - NoSQL: MySQL is an industry standard, but Cassandra is 2400 times faster!

An increasingly popular alternative

- Citizen-friendly alternative
- Decentralized infrastructure
- Privacy-aware