TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

Macroarea di Ingegneria
Dipartimento di Ingegneria Civile e Ingegneria Informatica

Introduction to Big Data

Corso di Sistemi e Architetture per Big Data

A.A. 2024/25

Valeria Cardellini

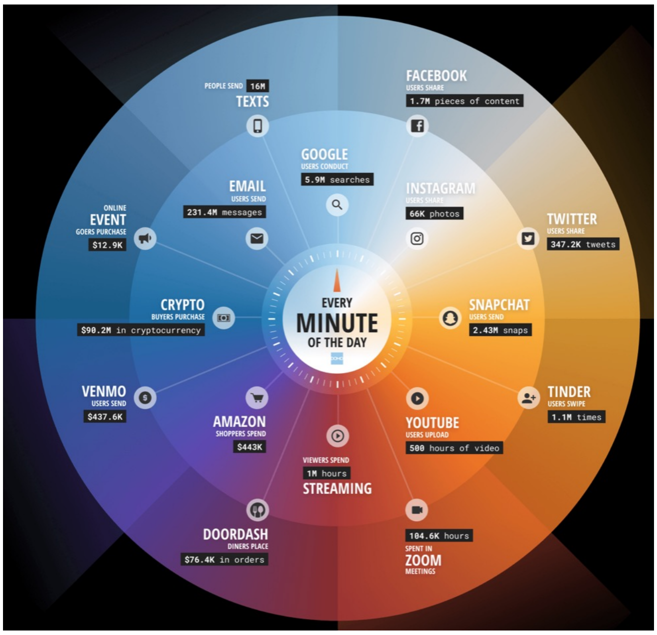
Laurea Magistrale in Ingegneria Informatica

Why Big Data?

How much data is created every single minute of the day?

Global Internet population in Jan. 2025: 5.56 billion (67.9% of world population)

1 billion in 2005



Source: https://www.domo.com/data-never-sleeps

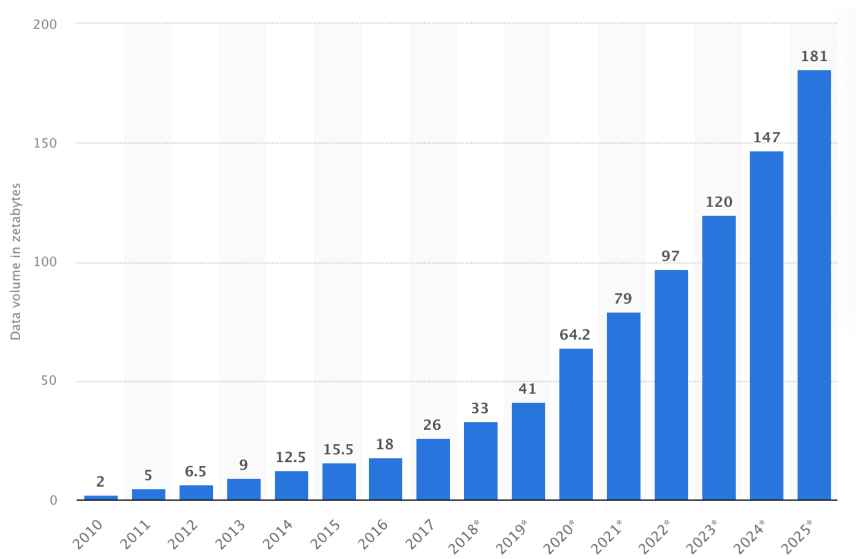
How much data?

- Big data volume: from Terabytes to Zettabytes
  - How big is a Zettabyte?
  - 1 ZB = 2^{70} B = 2^{40} GB ≈ 10^{21} B
    - Remember that 2^{10} = 1024 ≈ 10^3
- 120 Zettabytes of data generated by 2023

  120 Zettabytes ($120 \times 2^{70} \approx 120 \times 10^{21}$) …
  ≈ 120,000 Exabytes ($120{,}000 \times 10^{18}$) ...
  ≈ 120,000,000 Petabytes ($120{,}000{,}000 \times 10^{15}$) ...
  ≈ 120,000,000,000 Terabytes ($120{,}000{,}000{,}000 \times 10^{12}$) …
  ≈ 120,000,000,000,000 Gigabytes ($120{,}000{,}000{,}000{,}000 \times 10^{9}$) …
  ≈ 120,000,000,000,000,000,000,000 bytes!
- Bigger than Zettabytes? Yottabytes!
  - 1 YB = 2^{80} B ≈ 10^{24} B

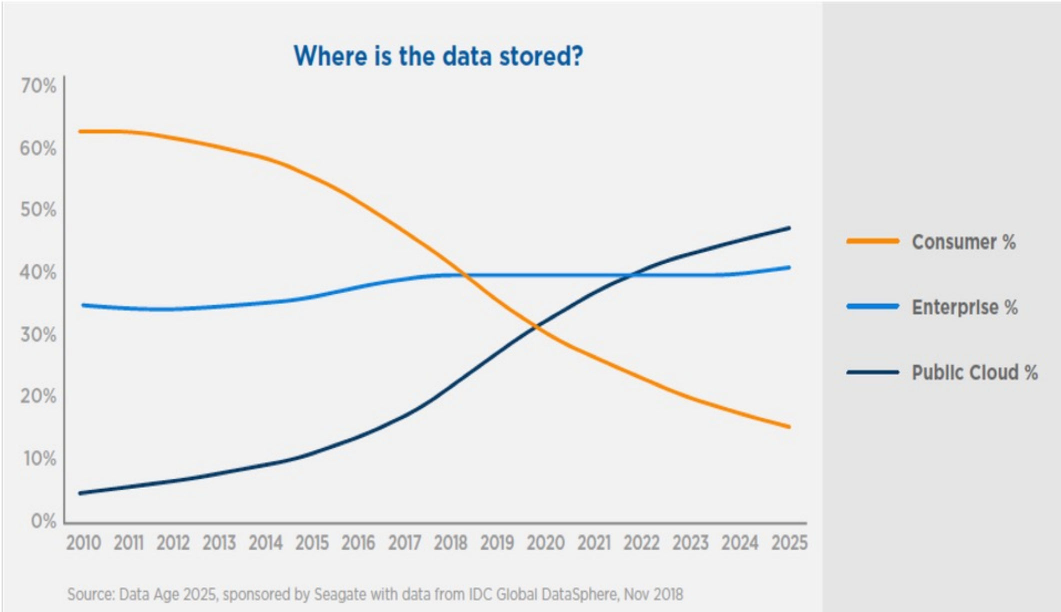
How much data?

- Recent explosion in data volume
  - In 2013: 90% of all the data in the world was generated over the last two years
  - 90x growth from 2010 to 2025



Where is data stored?

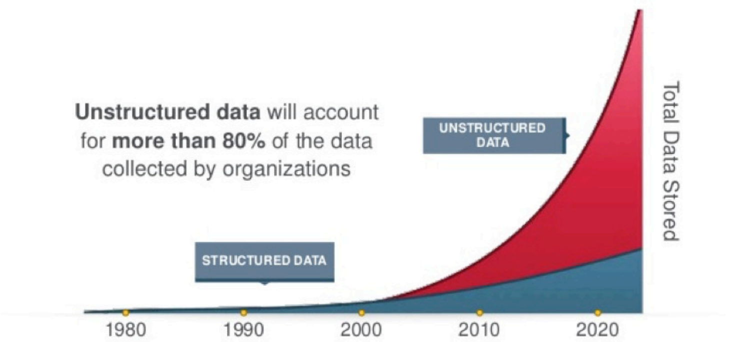
- Data is increasingly stored in public Cloud



Big data driving factors

- Big Data is growing fast
  - Smartphones
  - Social networks
  - Internet of Things (IoT)

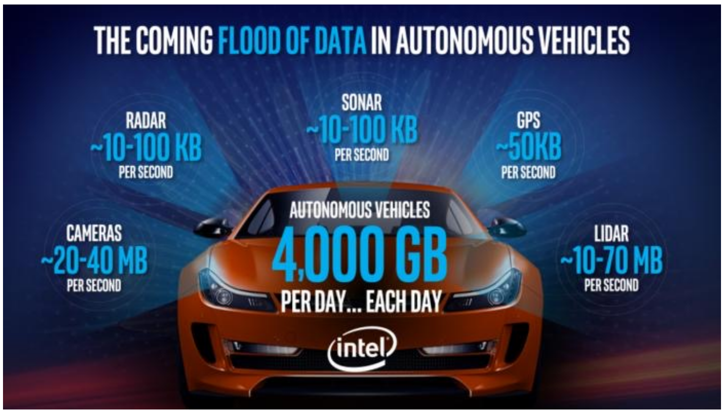
- Unstructured data grows at a fast rate



Source: Human-Computer Interaction & Knowledge Discovery in Complex Unstructured, Big Data

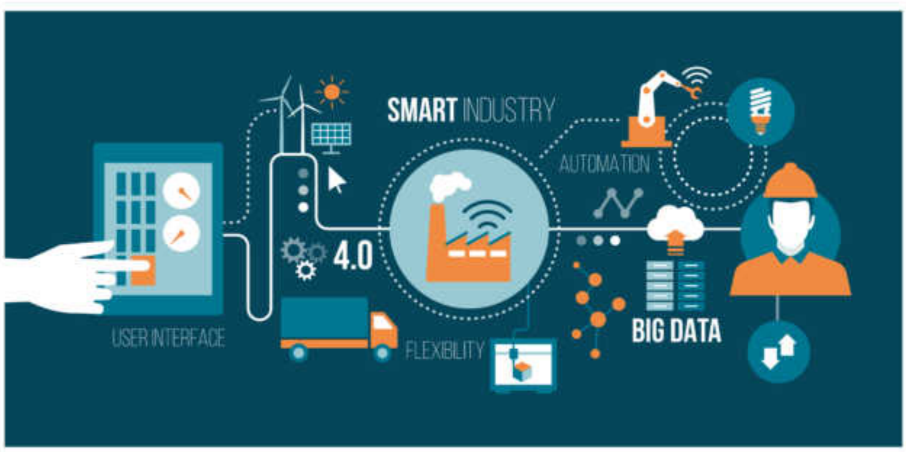
How Big? IoT impact

- IoT largely contributes to increase Big Data challenges
  - By 2023 over 15 billion IoT devices installed worldwide, over 29 billion estimated in 2030
- Example: self-driving cars
  - Just one autonomous car will use 4 TB of data/day



IoT impact: Industrial IoT

- Industrial Internet of Things (IIoT): network of physical objects, systems, platforms and applications that contain embedded technology to communicate and share intelligence with each other, with external environment and with people



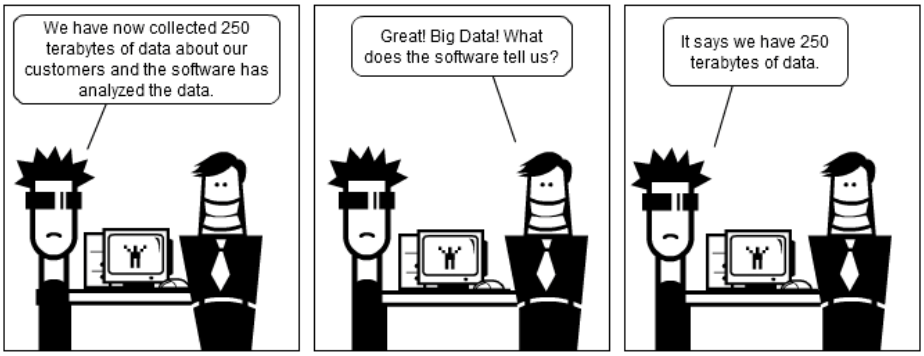
Big Data definitions

Different definitions

- "Big data refers to data sets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze." *The McKinsey Global Institute, 2012*
- "Big data primarily refers to data sets that are too large or complex to be dealt with by traditional data-processing application software." *Wikipedia, 2024*
- "Big data is mostly about taking numbers and using those numbers to make predictions about the future. The bigger the data set you have, the more accurate the predictions about the future will be." *Anthony Goldbloom, Kaggle's founder*

… so, what is Big Data?

- "Big Data" is similar to "small data", but bigger
- But bigger data requires different approaches: scale changes everything!
  - New methodologies, tools, architectures
- …with an aim to solve new problems or old problems in a better way



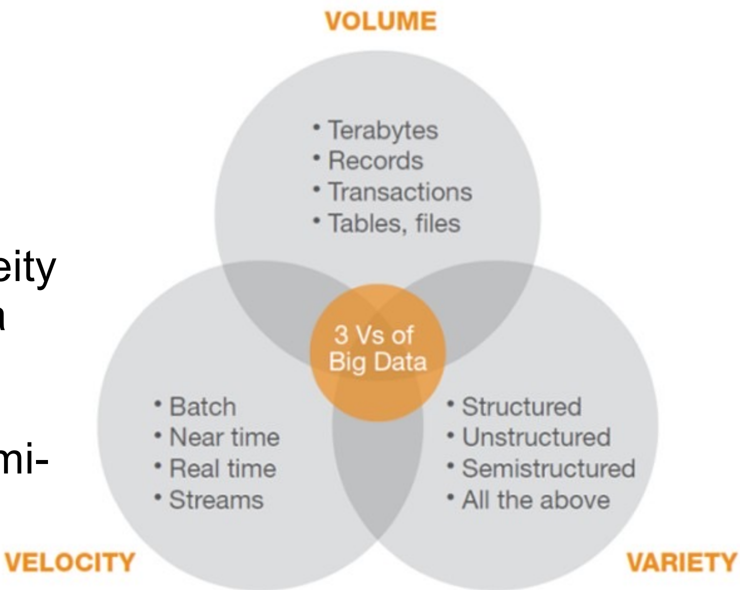
Gartner's Big data definition

- The most-frequently used and perhaps, somewhat abused definition (revised version by Gartner, 2012)

  Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization.

3V model for Big Data

1. **Volume**: data size challenging to store and process (how to index, retrieve)
2. **Variety**: data heterogeneity because of different data types (text, audio, video, record) and degree of structure (structured, semi-structured, unstructured data)
3. **Velocity**: data generation rate and analysis rate



- Defined in 2001 by D. Laney

The extended (3+n)V model

4. **Value**: Big data can generate huge competitive advantages
   – "Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis." (IDC, 2011)
   – "The bigger the data set you have, the more accurate the predictions about the future will be" (A. Goldbloom)
5. **Veracity**: regards quality or insightfulness of data, issues related to uncertainty of accuracy and authenticity of data
6. **Variability**: data flows can be highly inconsistent with peaks
7. **Visualization**

Big Data visualization

- Presentation of data in a pictorial and graphical format
- Why? Our brain processes images 60,000x faster than text
- A first example:

Big Data visualization

- Some example
  - Flight patterns in US https://www.aaronkoblin.com/work/flightpatterns
  - Pollution map https://waqi.info/
  - Ocean surface currents https://catalog.data.gov/dataset/ocean-surface-current-analyses-real-time-oscar-surface-currents-near-real-time-0-25-degree

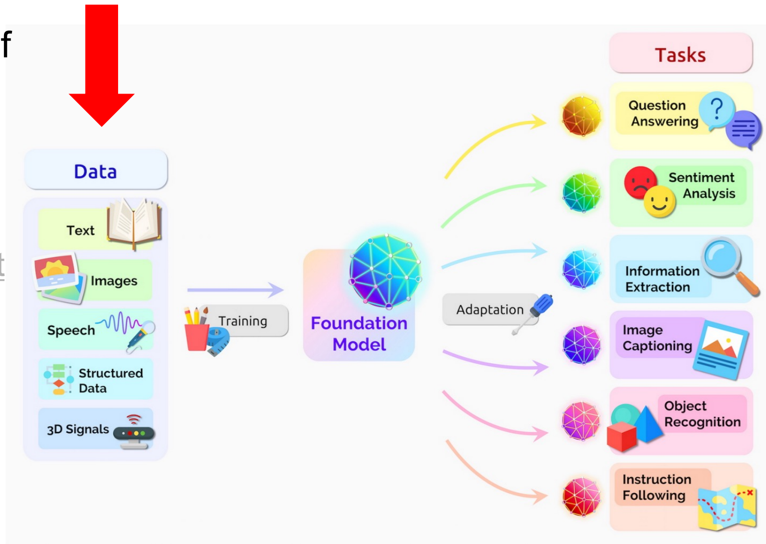
Why now?

- Because we have data
  - Data is already in digital form
  - 22% of data growth from 2023 to 2024
- Because we can store and process data
  - Storage resources: high-capacity disks (e.g., 400€ per 20 TB)
  - Computing resources: more than 40 years of Moore's law and more recently multicore architectures
- But new approaches that require more and more data are on the scene
  - E.g., transformer models
    - OpenAI's GPT-4: trained on hundreds of billions of parameters and datasets containing vast amounts of text, resulting in PBs of data

Transformer model

- Neural network that learns context and thus meaning by tracking relationships in sequential data like the words in this sentence

Aka foundation model:
"The sheer scale and scope of foundation models over the last few years have stretched our imagination of what is possible"
https://crfm.stanford.edu/report.html



The downside of Big Data

- Every day more data to process and store
- How much energy is consumed?
  – In 2022, data centers hosting popular services (e.g., AWS and Google's search engine), used about 1% to 1.3% of world's current electricity use
  – Cryptocurrency mining used another 0.4%
  – By 2027 AI servers could use between 85 to 134 TW hours annually
    - That's similar to what Argentina, the Netherlands and Sweden each use in a year, and about 0.5% of world electricity
- Is it sustainable?

A.I. Could Soon Need as Much Electricity as an Entire Country, NYT, 2023
https://www.nytimes.com/2023/10/10/climate/ai-could-soon-need-as-much-electricity-as-an-entire-country.html

Examples of Big Data applications in very diverse sectors

- Customer analytics in retail industry
  - E.g., to increase customer retention and loyalty
- Predictive maintenance for Industry 4.0
  - E.g., detecting anomalous machine states
- Crime prevention
  - To analyze crime patterns and trends
- Health care
  - E.g., to diagnose and treat genetic diseases
- Finance
  - To anticipate customer behaviors and create strategies
- Education
  - E.g., to improve the learning process, to design a new course
- Space science
  - E.g., astronomical discoveries

Batch vs. real-time analytics

- ***Batch analytics***: analysis of set of data collected over a period of time and that has already been stored
  - We will study **batch processing engines**

- ***Real-time analytics***: analysis of high-velocity, continuous data streams as soon as they are ingested without (or before) storing them
  - Goal: get insights immediately (or very rapidly after) data enters the system
  - We will study **stream processing engines**

Examples of real-time analytics

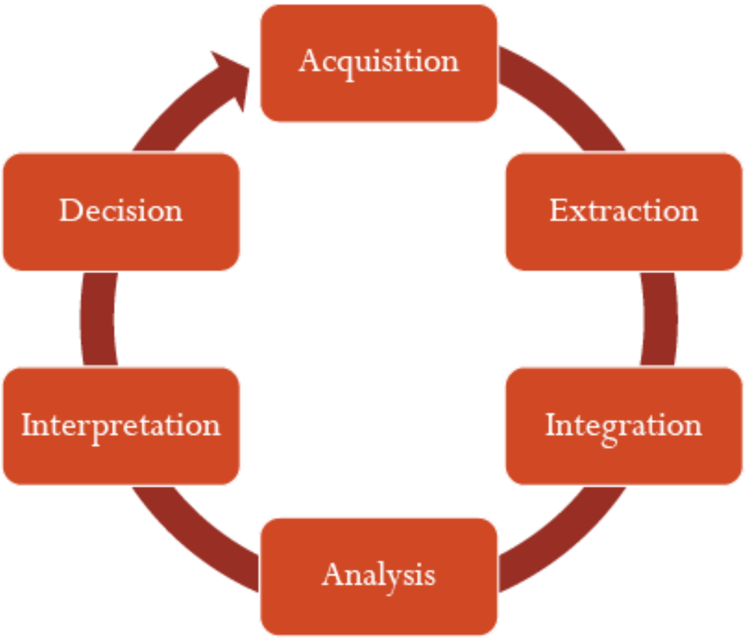
- Grand Challenge at DEBS conference https://debs.org/grand-challenges/
  - Over high volume sensor data: analysis of energy consumption measurements (DEBS 2014)
  - Over high volume geospatial data streams: analysis of taxi trips based on a stream of trip reports from New York City (DEBS 2015)
  - Over social network: to identify posts that trigger the most activity and large communities that are involved in a topic (DEBS 2016)
  - Over maritime transportation data: to predict destinations and arrival times of ships (DEBS 2018)
  - Technical analysis of market data: to compute specific trend indicators and detect patterns resembling those used by traders to decide on buying or selling (DEBS 2022)

… other example of real-time analytics in very diverse sectors

- Medicine
  - To track epidemic diseases, to prevent diseases through wearable health care technologies
- Security
  - To detect frauds or DDOS attacks, to recognize behavioral patterns
- Urban traffic management
  - To address traffic congestion and lack of parking, to plan public transportation

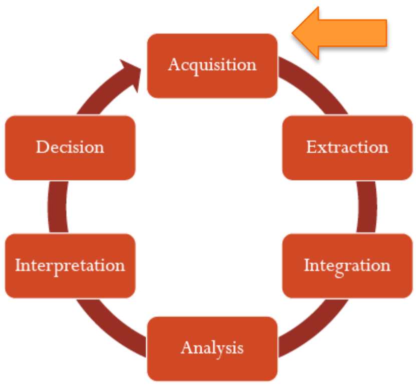
The Big Data process

- 6 stages of the Big data analytics lifecycle



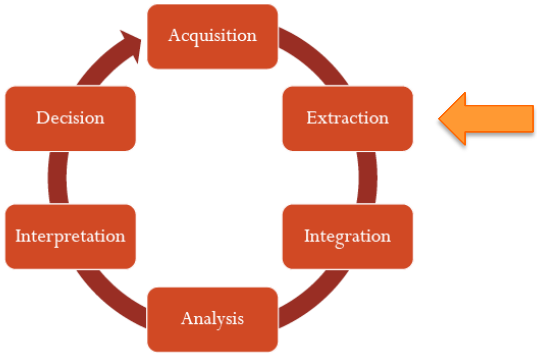
The Big Data process

- Acquisition
  - Requires:
    - Selecting data
    - Filtering data
    - Generating metadata
    - Managing data provenance
      - E.g., GDPR compliance



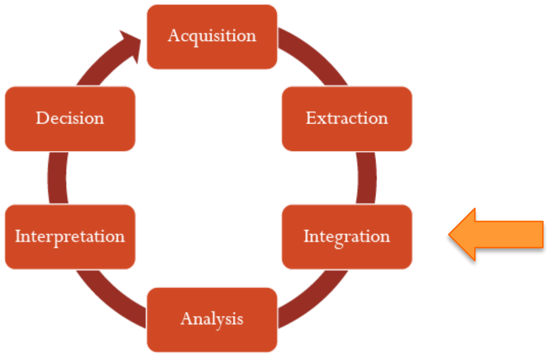
The Big Data process

- Extraction
  - To transform data into a format that can be used by Big data processing frameworks
  - Requires:
    - Data transformation
    - Data normalization
      - E.g., avoid duplication
    - Data cleaning
      - Remove corrupted or inaccurate data (e.g., outliers)
      - Impute missing data using some data imputation technique
    - Data aggregation
      - E.g., from multiple sources (e.g., because of multiple data providers)



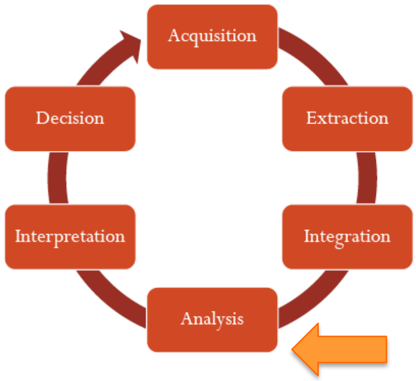
The Big Data process

- Integration
  - Requires:
    - Standardization
    - Conflict management
    - Reconciliation
    - Mapping definition



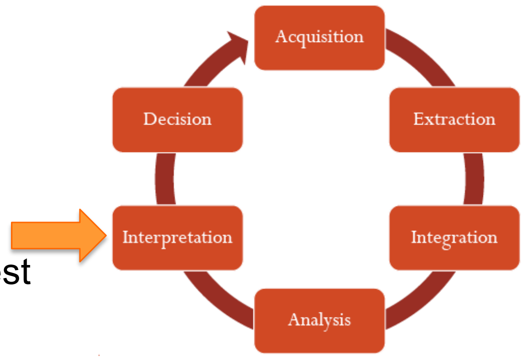
The Big Data process

- Analysis
  - Requires:
    - Data analytics techniques
      - Statistics
      - Data mining
      - Machine learning
      - Visualization



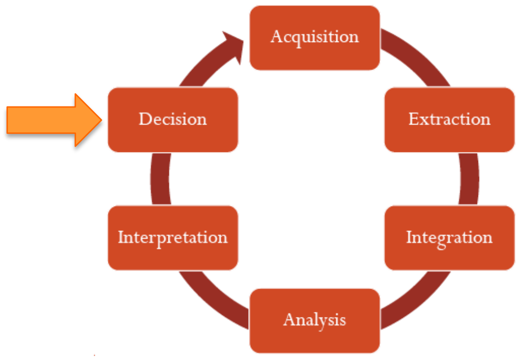
The Big Data process

- Interpretation
  - Requires:
    - Knowledge of domain
    - Knowledge of data provenance
    - Identification of patterns of interest
    - Flexibility of the process



The Big Data process

- Decision
  - Requires:
    - Managerial skills
    - Continuous improvement of the process (loop)



Some techniques for Big Data analytics

BUSINESS ANALYTICS

	Descriptive	Diagnostic	Predictive	Prescriptive
Questions	What happened ? What is happening ?	Why it happened ? Why it is happening ?	What will happen ? Why will it happen ?	What should do ? Why should do ?
Enablers	- Dashboard - Scorecards - Business reporting - Data warehouse	- Business reporting - Dashboard - Data warehouse	- Data mining - Forecasting - Text mining - Web/media mining	- Simulation - Optimization - Decision modeling - Expert system
Outcomes	Minutely defined problems and opportunities	Ability to drill down to the root cause.	Accurate projection of conditions and states	Best possible business prospect

@ Innolever Solutions Pvt Ltd

Added value, complexity →

Some techniques for Big Data analytics

- *Data mining*: anomaly detection, association rule mining, classification, clustering, regression, summarization
- *Machine learning*: supervised learning, unsupervised learning, reinforcement learning
- *Crowdsourcing*
  - Outsourcing human-intelligence tasks to a large group of unspecified people via Internet

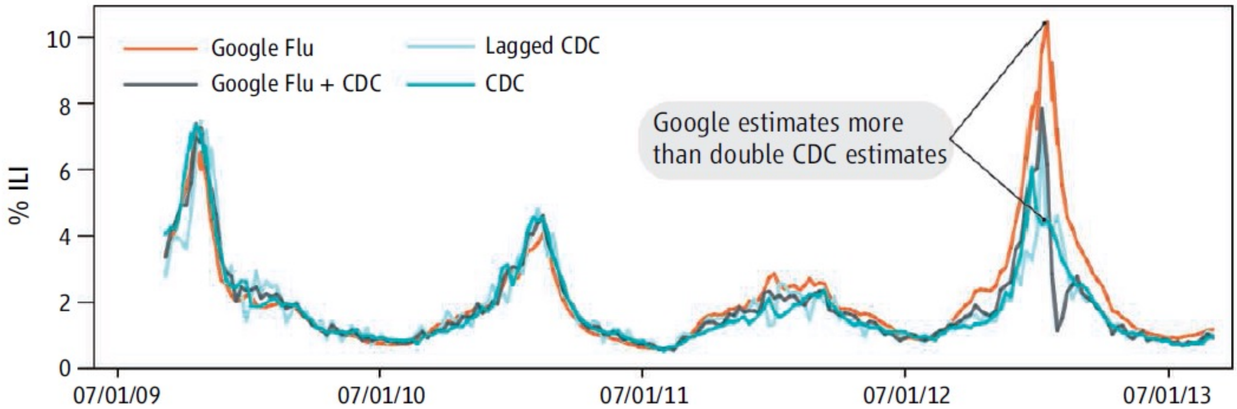
In this course we focus on systems and architectures for Big Data, not on data analysis techniques

Risks and challenges of Big Data

- Effectiveness of data analysis
- Performance
  - Efficiency
  - Scalability and elasticity
    - Scale linearly as workloads and data volumes grow
  - Fault tolerance
  - Sustainability
    - Data grows faster than energy on chip
- Heterogeneity
  - Data, processing environment, network latencies, …
- Flexibility
- Privacy and security
- Costs

Effectiveness of Big data analysis

- A famous example of inaccurate analysis
- Google Flu Trends' predictions
  - Sometimes very inaccurate: over the interval 2011-2013, when it consistently overestimated flu prevalence and over one interval in the 2012-2013 flu season predicted twice as many doctors' visits as those recorded



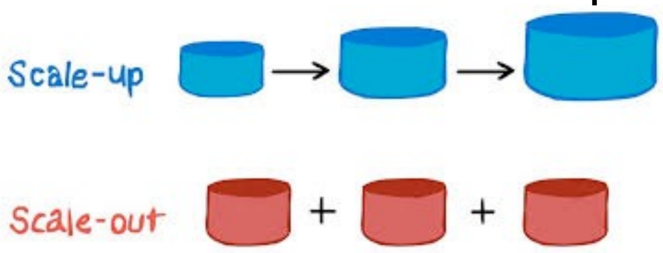
Lazer et al., "The Parable of Google Flu: Traps in Big Data Analysis". *Science,* 2014 https://gking.harvard.edu/files/gking/files/0314policyforumff.pdf

Taming performance: distribution and replication

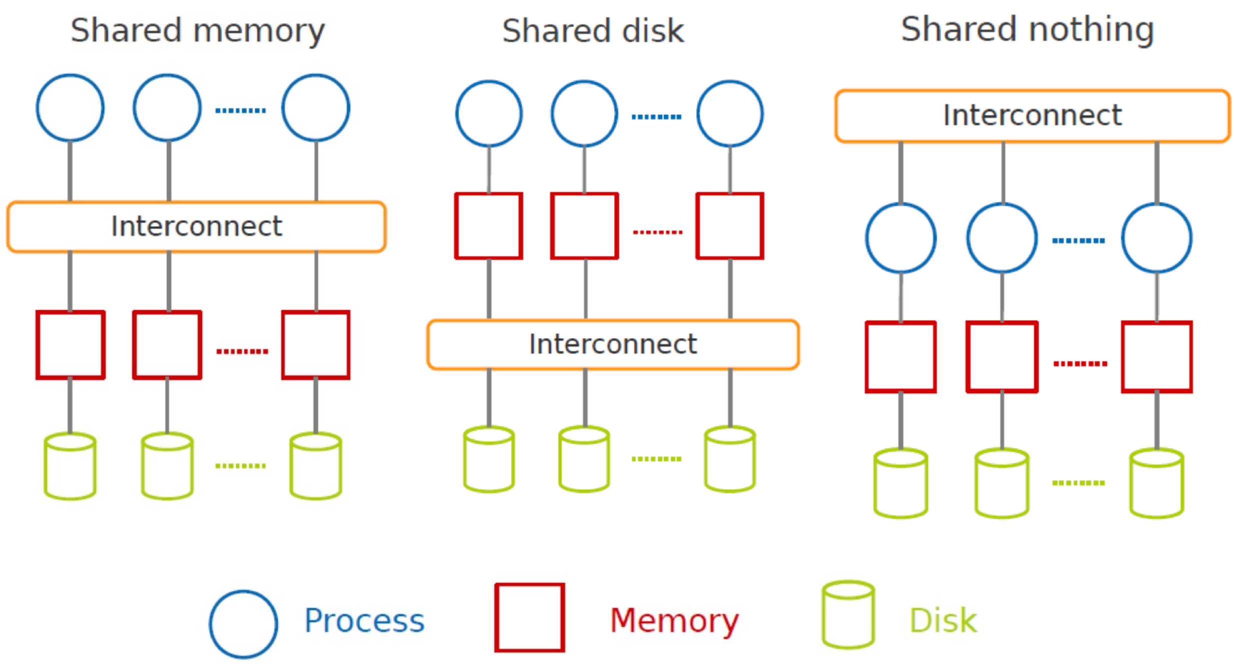
- Distributed architecture
  - Common architectural solution for Big Data processing: cluster of commodity hardware resources, also in Cloud
  - Scale out (horizontally), not up (vertically)!
  - Challenges: *elasticity* and data processing at the *network edges*
- Distributed processing
  - **Shared-nothing** model
  - New programming paradigms, e.g., functional programming
- Resource replication
  - The well-known solution to achieve fault tolerance
  - Eventual consistency (CAP theorem!)

Scaling out vs. scaling up

- Two different ways of addressing the need for more processor capacity, memory and other resources
- **Scaling up** (or vertical scalability) refers to purchasing and installing a more powerful server
  - E.g., with more processing capacity and RAM
- **Scaling out** (or horizontal scalability) means adding other lower-performance servers to collectively do the work of a much more powerful one

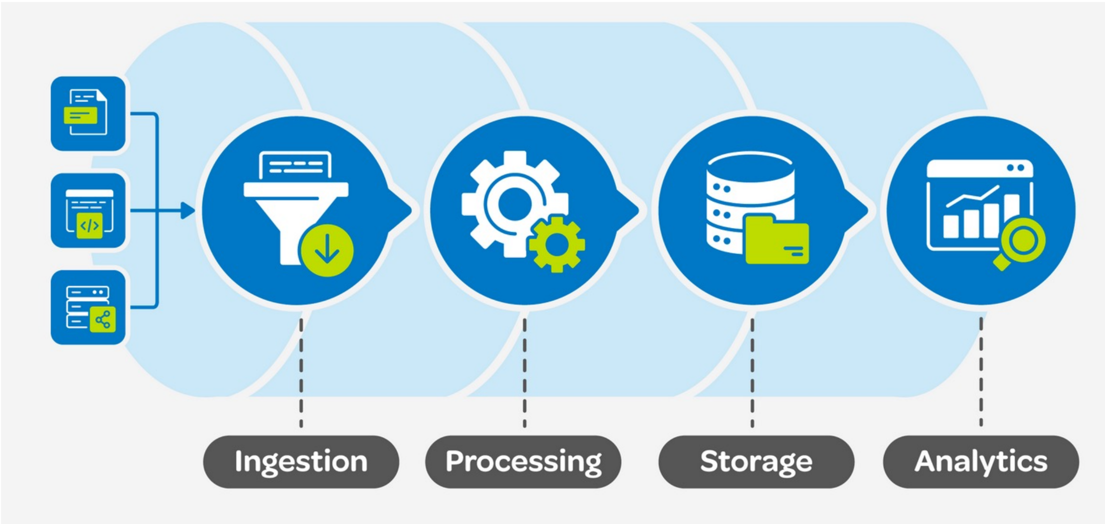


Shared nothing vs. other parallel architectures



DeWitt and Gray, "Parallel database systems: the future of high performance database systems", *Comm. ACM*, 1992
https://dl.acm.org/doi/pdf/10.1145/129888.129894

Data pipeline architecture



Data pipeline architecture: stages

- **Ingest** data
  - Raw data is captured or collected and brought into pipeline from various sources
  - Distributed file system, object store, message queue, message oriented middleware
- **Process** data
- **Store** data
  - Processed data is stored for later retrieval and analysis
  - Databases, data stores, data lakes, data warehouses
- **Analyze** data

Where to process and analyze Big Data

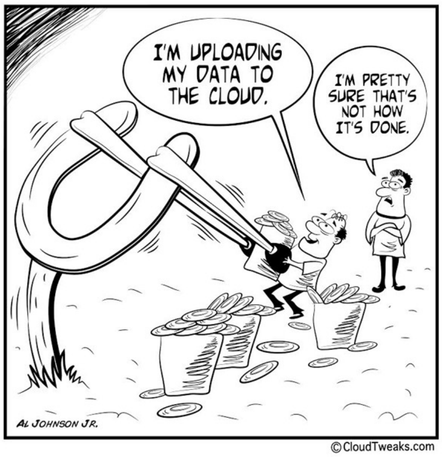
- Traditional way: using a cluster of servers on premises
  - Compute nodes are stored on racks
    - 8-64 compute nodes on a rack
  - There can be many racks of compute nodes
    - Rack nodes are connected by a network, typically Gb Ethernet
    - Racks are connected by another level of network or a switch
    - Bandwidth of intra-rack communication is usually much greater than that of inter-rack communication
- Cons:
  - ✗ Need to manage hardware infrastructure and processing platforms (acquire, install, configure, …)

Where to process and analyze Big Data

- The Cloud way: using Cloud analytics services
- Some example
  - Amazon EMR, Google Dataproc, Azure HDInsight: Hadoop and Spark (plus other frameworks) in the Cloud
- Pros:
  - Gain Cloud scalability and elasticity
  - Do not need to manage and provision the infrastructure and the platform

Where to process and analyze Big Data

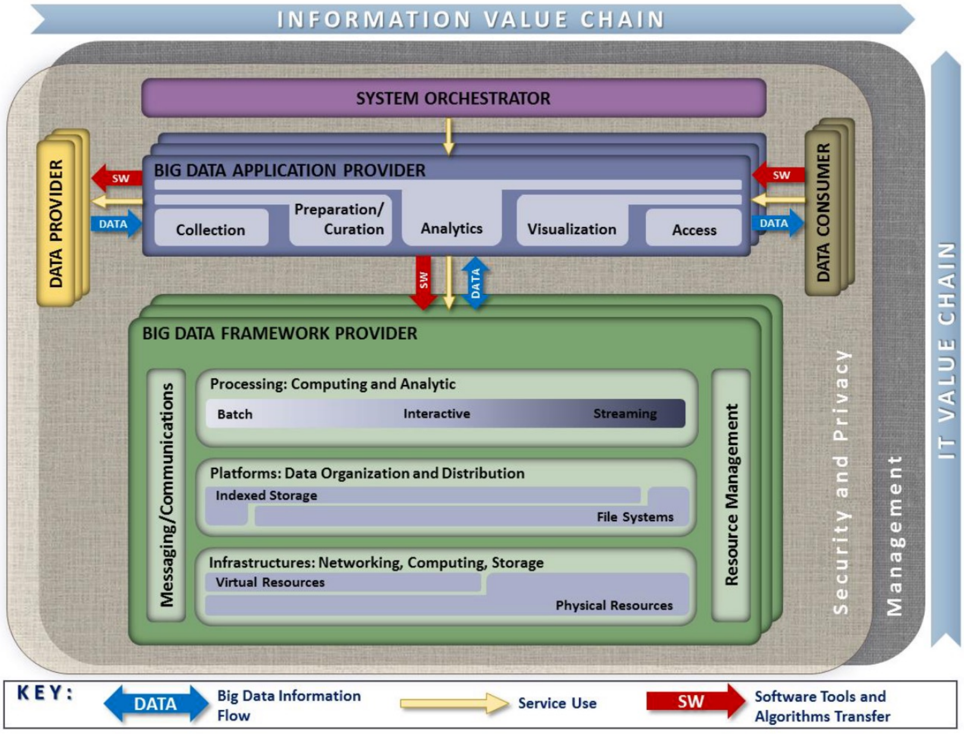
- Cloud data centers are located in the network core
- Data is collected everywhere and can be accessed anywhere
- Challenges:
  – Move data to Cloud
    - Latency is not zero (because of speed of light)!
    - Minor issue: network bandwidth
  – Data security and privacy



Where to process and analyze Big Data

- The new scenario: Compute continuum
  – Progressive convergence between Cloud, Fog and Edge computing, resulting in a continuum
  – Not only Cloud data centers, but also many micro data centers located at network edges
  – Move data processing and analytics close to data producers and consumers

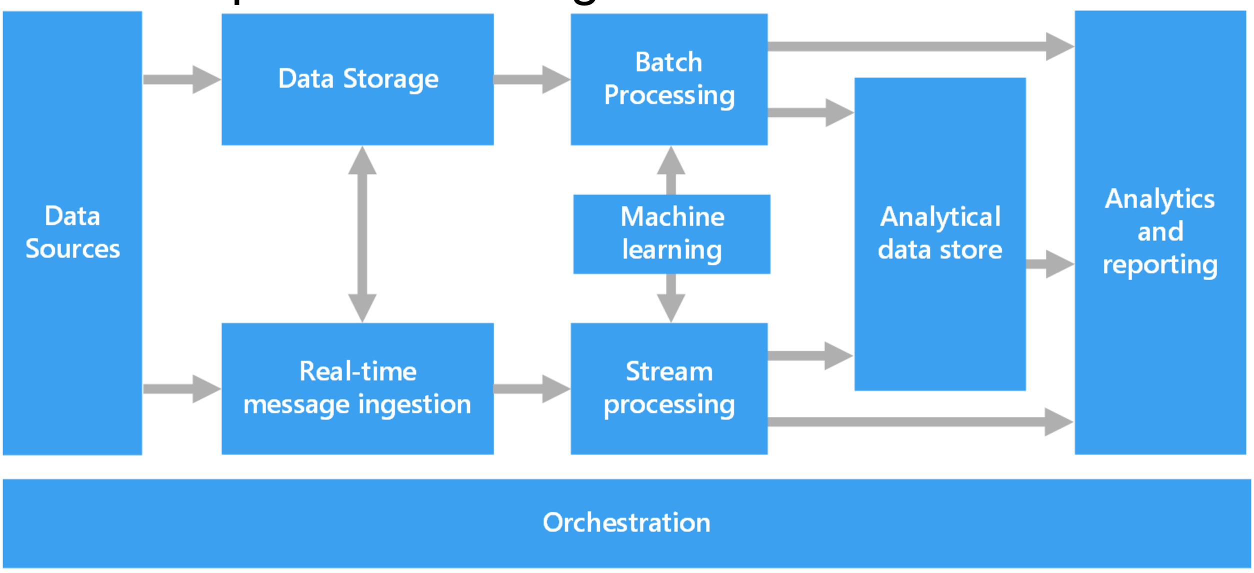
NIST Big Data reference architecture



https://doi.org/10.6028/NIST.SP.1500-6r2

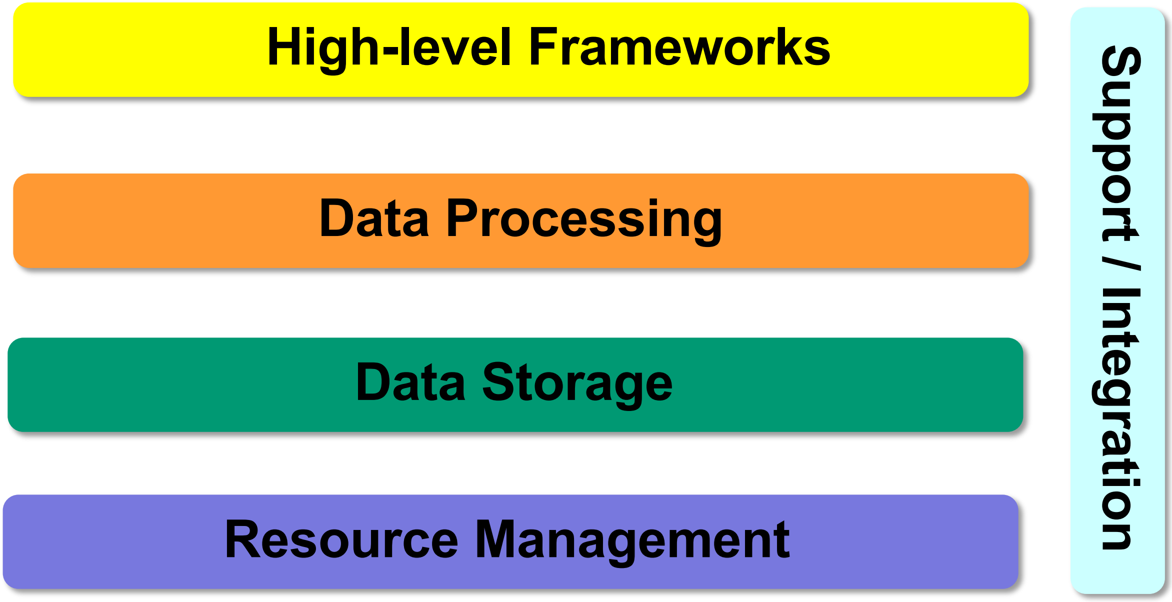
Big data architecture

- Designed to handle ingestion, processing, and analysis of Big data
- Components of a big data architecture



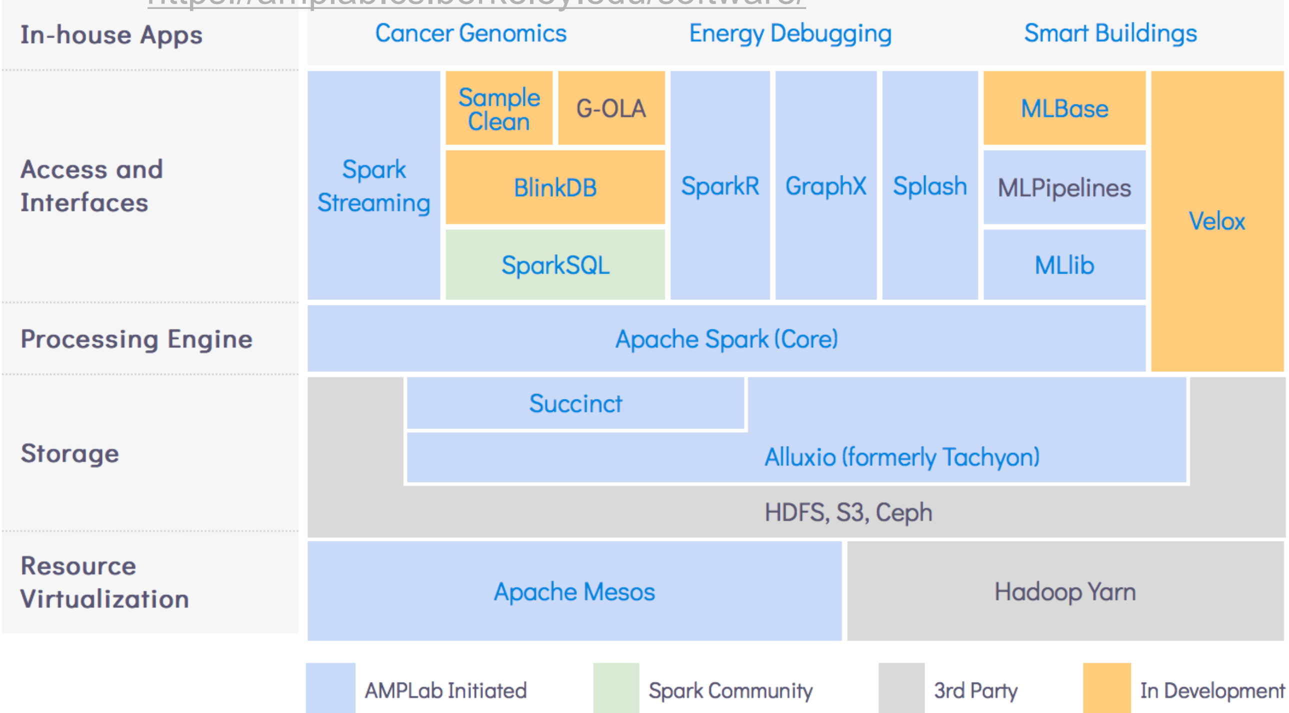
https://docs.microsoft.com/en-us/azure/architecture/data-guide/big-data/

Our Big Data stack



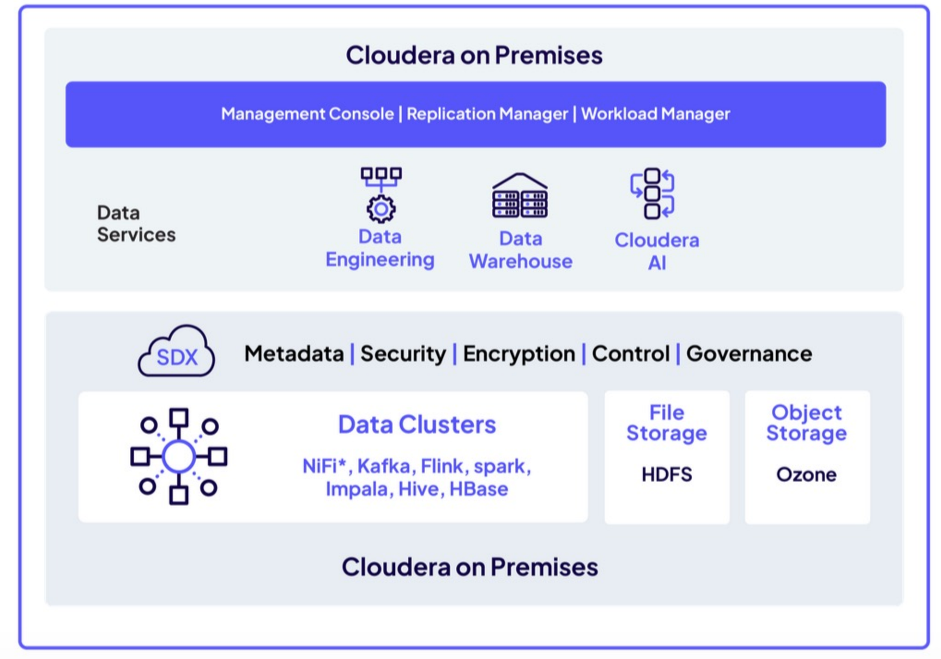
Example of Big Data stack: BDAS

- BDAS: Berkeley Data Analytics Stack
  https://amplab.cs.berkeley.edu/software/



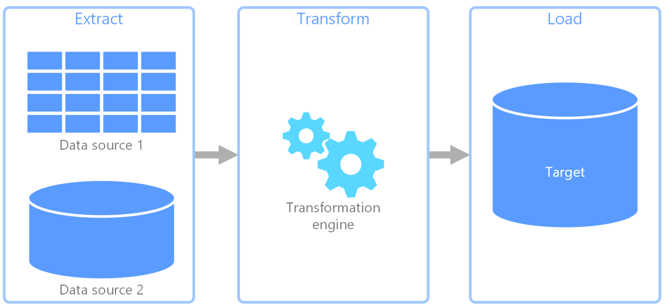
Example of Big Data stack: Cloudera

https://www.cloudera.com/products/cloudera-data-platform/private-cloud.html



Approaches for data ingestion: ETL

- Extract, Transform, and Load (**ETL**)



- ➢ *Extract* data from different sources
- ➢ *Transform* data into a usable (i.e., proper format) and trusted resource for storing
  - ➢ Data is not usually loaded directly into the target system, but instead it is common to have it uploaded into a staging DB
- ➢ *Load* data into the target system, i.e., DB or data warehouse (DWH)

https://databricks.com/it/glossary

Data warehouse

- For many years, relational DBMSs were sufficient for companies' needs
- 1990s: no longer On Line Transaction Processing (OLTP) but also On Line Analytical Processing (OLAP)
- Data warehouses were born to unite companies' structured data under one roof
- Emerged as technology that brings together an organization's collection of RDBMs under a single umbrella, allowing data to be queried and viewed as a whole
- Typically run on expensive, on-premises appliance-based hardware from vendors (e.g., Teradata and Vertica), also available as cloud services (e.g., AWS Redshift)

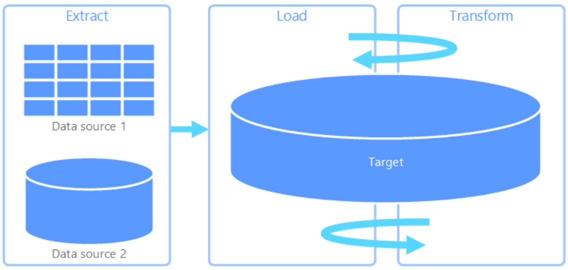
Data warehouse: pros and cons

- Pros
  - ✓ Integration of many data sources
  - ✓ Data optimized for read access
  - ✓ Ability to run quickly ad hoc analytical queries
  - ✓ Data audit, governance and lineage

- Cons
  - ✗ Inability to store unstructured, raw data
  - ✗ Expensive, proprietary hardware and software
  - ✗ Difficult to scale due to tight coupling of storage and compute

Approaches for data ingestion: ETL

- Extract, Load, and Transform (**ELT**)

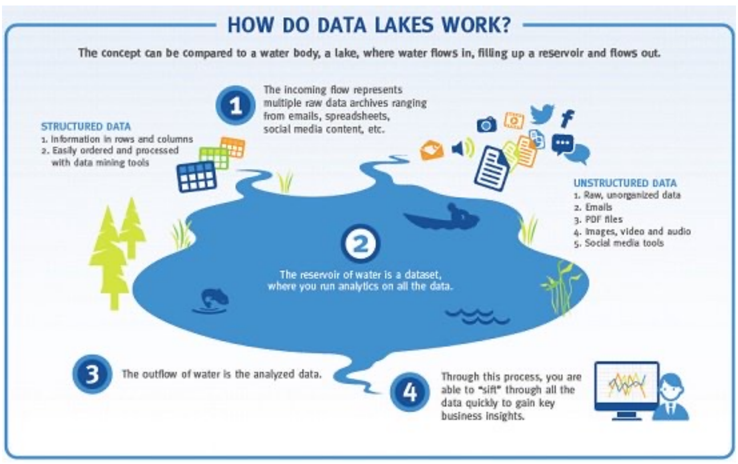


  - *Extract* data from different sources
  - *Load* data into a **data lake**, where data is held in original format
  - *Transform* data using the processing capabilities of target system
- Pros:
  - – No need for separate transformation engine
  - – Data transformation and loading happen in parallel
  - – More effective when speed is critical
  - – Works well when target system is powerful enough to transform data efficiently

Data lake

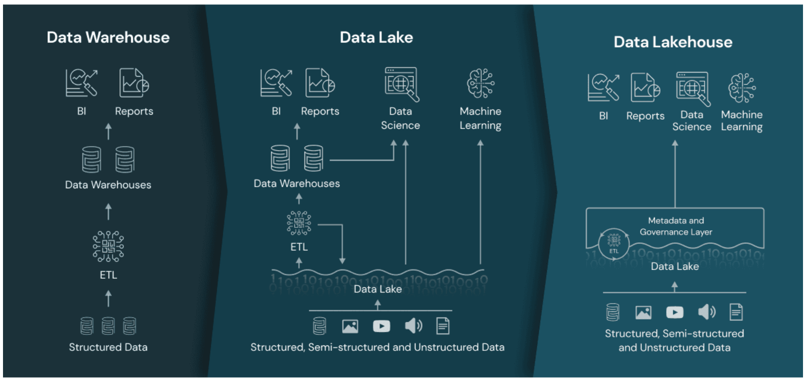
- Method of storing data within a system or repository, in its native, raw format, that facilitates collocation of data in a variety of formats (structured, unstructured, semi-structured), using object blobs or files
- Designed for *quickly changing data*

  https://www.databricks.com/discover/data-lakes/introduction



The new trend: data lakehouse

- Data lakehouse: data lake + data warehouse
- A new architectural unification strategy
  - Integrates data lake and data warehouse to improve performance, scalability, flexibility, and cost-effectiveness and eliminate data silos and ETL processes
  - Unifies all data to simplify data engineering processes and support business intelligence (BI) and ML workloads on all data www.databricks.com/glossary/data-lakehouse



References

- Big Data - ACM Panels in Print, *Comm. ACM*, 2017 https://dl.acm.org/doi/pdf/10.1145/3079064
- Jagadish et al., Big Data and Its Technical Challenges, *Comm. ACM*, 2014 https://dl.acm.org/doi/pdf/10.1145/2611567
- Wing, The Data Life Cycle, *Harvard Data Science Review*, 2019 https://hdsr.mitpress.mit.edu/pub/577rq08d/release/4