Welcome

Big Data

Stanford ICME Summer Workshops

Your instructors

Anna-Julia Storch

aka “AJ”

ASK ME ABOUT

- Book recommendations for personal development
- Silicon Valley Startup World
- Skiing, Classic cars & vanlife

FIND ME

/ajstorch

ajstorch@stanford.edu

EDUCATION & EXPERIENCE

- MS Education Data Science @Stanford
- Head Teaching Assistant for Stanford’s premier entrepreneurship classes & Threshold Venture Fellow
- Former business line head for a multinational HR company
- Product & data roles in big tech, consulting & VC-backed startups

amazon

Stanford University

kyte.

Berkeley UNIVERSITY OF CALIFORNIA

McKinsey & Company

Your instructors

Axel Peytavin

aka “Axel” (yep)

ASK ME ABOUT

- Building random software fast
- Climate, sustainability and social impact entrepreneurship
- Climbing & playing guitar

FIND ME

/axel-peytavin

peytavin@stanford.edu

EDUCATION & EXPERIENCE

- M.S. ICME @Stanford x @Centrale Paris
- Head TA for ICME’s core Software Eng class & TA Principled Entrepreneurship
- Computational Modeler of Ocean Plastic Pollution @The Ocean Cleanup
- Threshold Venture Fellow, Founder of getalong.io

THE OCEAN CLEANUP

getalong

Stanford University

CentraleSupélec

BAM

Goals

What we wish we had known before diving into the world of (big) data

Buzzwords

Big Data
Data Lake
Cloud Computing
Data warehouse
Horizontal scalability

Tools

SQL
Apache Spark
Hadoop
AWS
Google colab
Alteryx

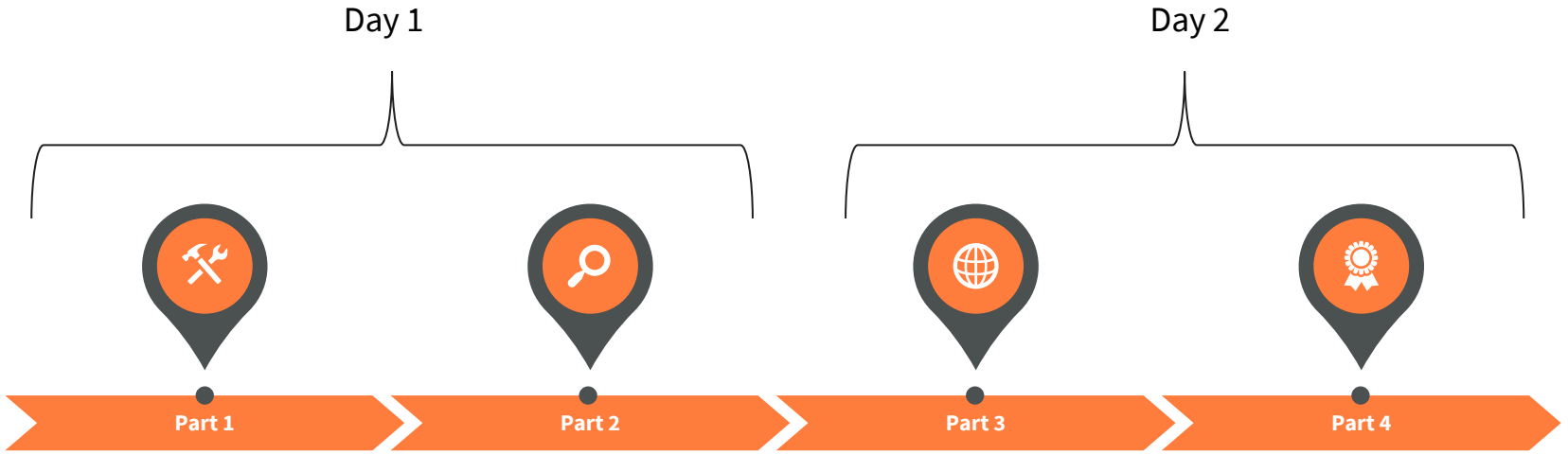
Applications

Education
Social Data
Medicine
Climate
Fashion

Future directions

Synthetic Data
Chat GPT & Generative AI

Goals


Day 1
Day 2
Part 1
Part 2
Part 3
Part 4
WHAT IS BIG DATA?
WHAT CAN WE DO WITH IT?
HOW CAN WE BEST STORE & RETRIEVE BIG DATA?
HOW DO WE PROCESS DISTRIBUTED DATA IN THE CLOUD?
HOW DOES THE FUTURE OF BIG DATA LOOK LIKE?
Quizzes, coding exercises, lectures, tools, readings, courses.. for further use

Get-To-Know Exercise

Share your interest in big data &
tell “The story of
Your name”

Rules of the road

ENGAGEMENT IS FUN

- This is an **interactive** class, you are not here to just sit back and listen, but to **interact.**
  - Quizzes
  - Individual & Team exercises
- Video on
- Regular bathroom/coffee breaks

NO QUESTION IS STUPID. WE MEAN IT

- Raise your hand on Zoom
- Put questions in the chat (we prefer you raise your hand)
- Piazza for longer questions (or before/after workshop)

FEEDBACK RULES

- At the end of each day we collect feedback from you & can adapt even for the next day!
- If you have feedback in between, send us a private message on Zoom!

Let’s get started…

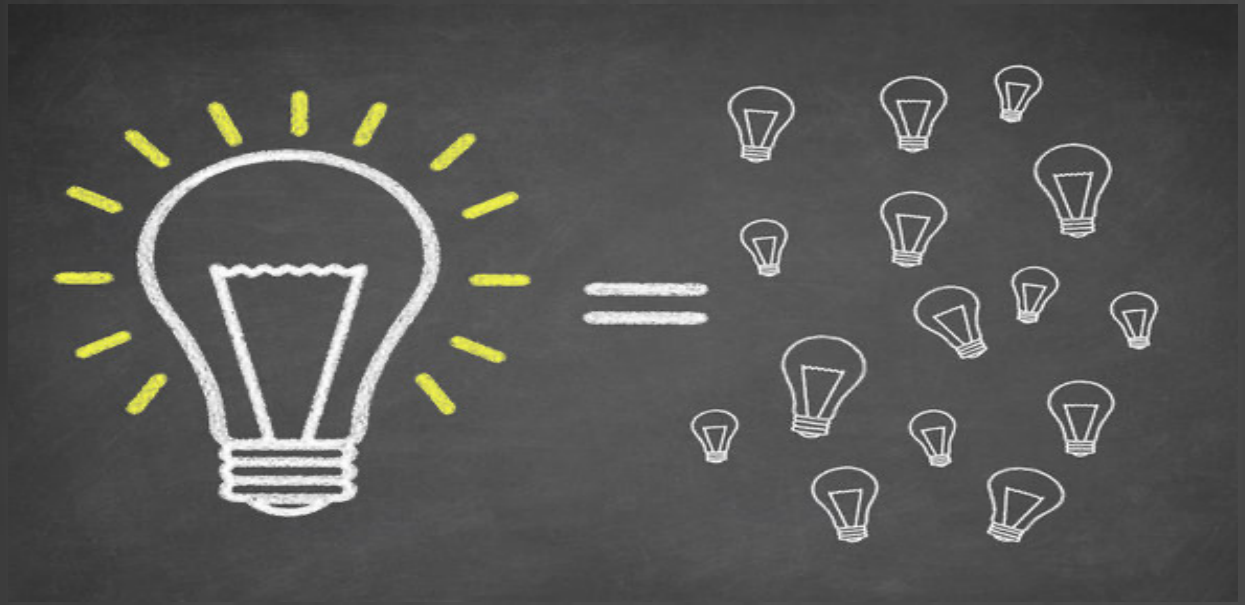
Part 1

What is BIG DATA?

And what can we do with it?


BUT WHAT IS
BIG DATA??
PalZoo.net
memegenerator.net

Big Data = Lots of data?



Quiz!

What is big data?

→ Keep that site open as we are using it throughout the course

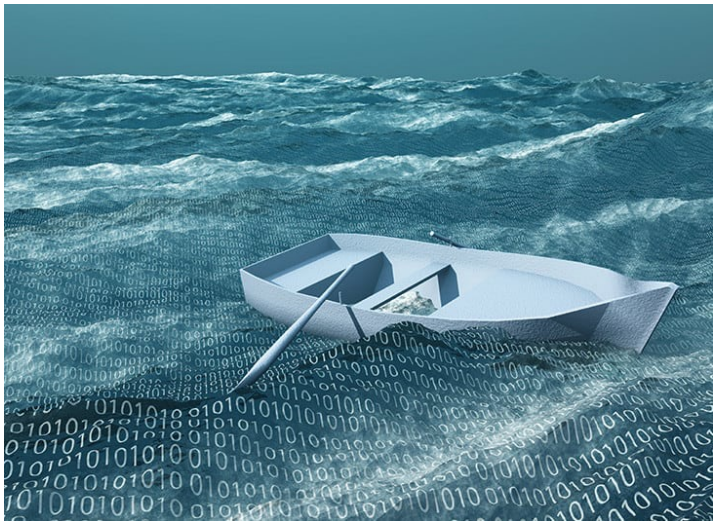
Join at

slido.com

#1813 017

What launched the big data era?

1. A flood of data

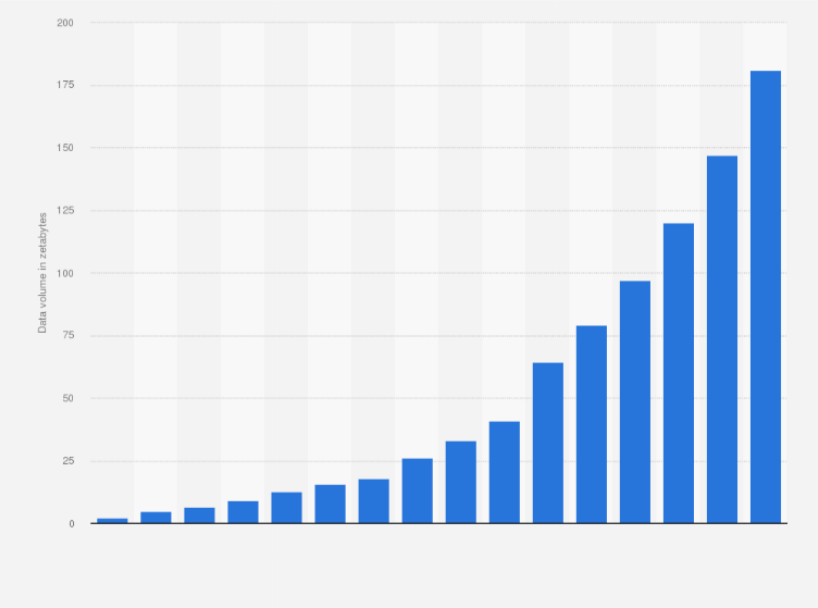


-> We collect more data

-> We store more data

What launched the big data era?

1. A flood of data



Average person creates 1.7 MB / second

90% of the worlds data was generated in the last 2 years

Compared to 2010 we create 60x as much data

What are the largest sources of data?

CONSUMERS

- Social Media (tweets, photos, location, likes)
- Purchasing data (what, where, etc)
- Health data (heart rate, hospital visits)

MACHINES

- Real time sensors in industrial machines
- Environmental sensors
- Health trackers

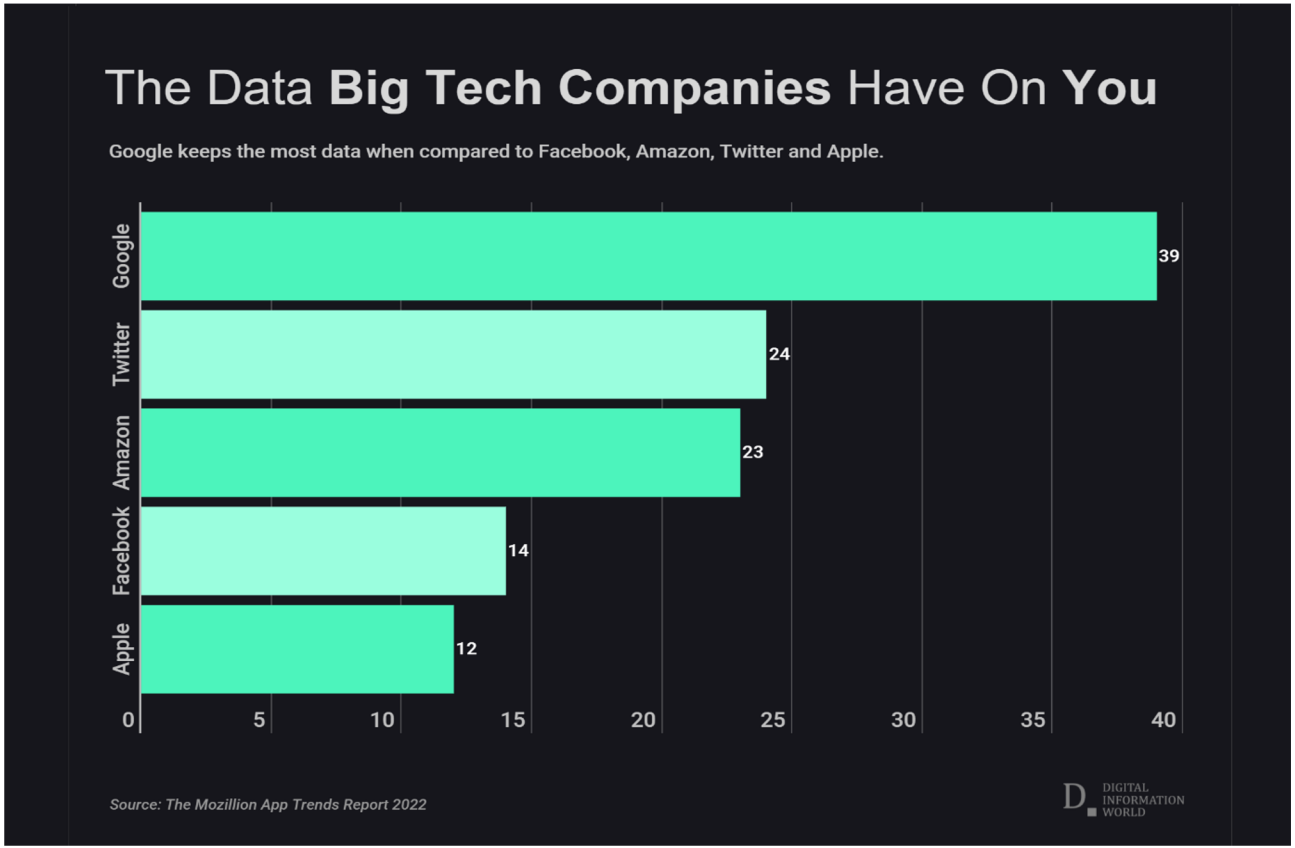
ORGANIZATIONS

- Transaction data
- Employment data

Quiz!

What organizations have the “richest” data in the world?

WHO “owns” the most data in the world?



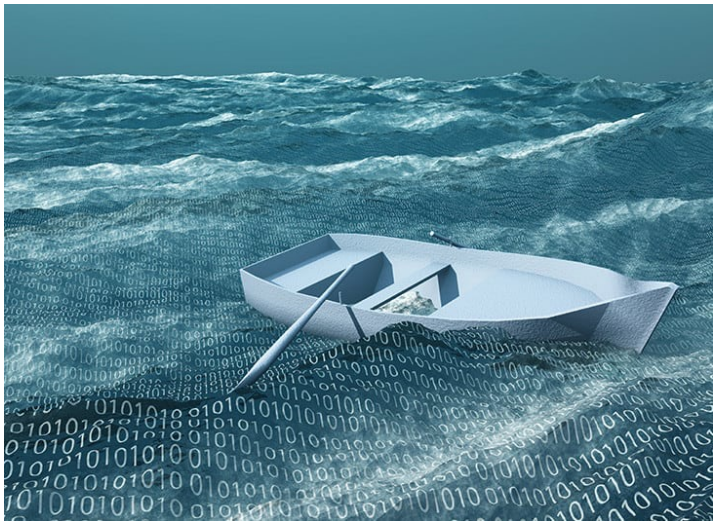
WHO “owns” the most data in the world?

Health Data?

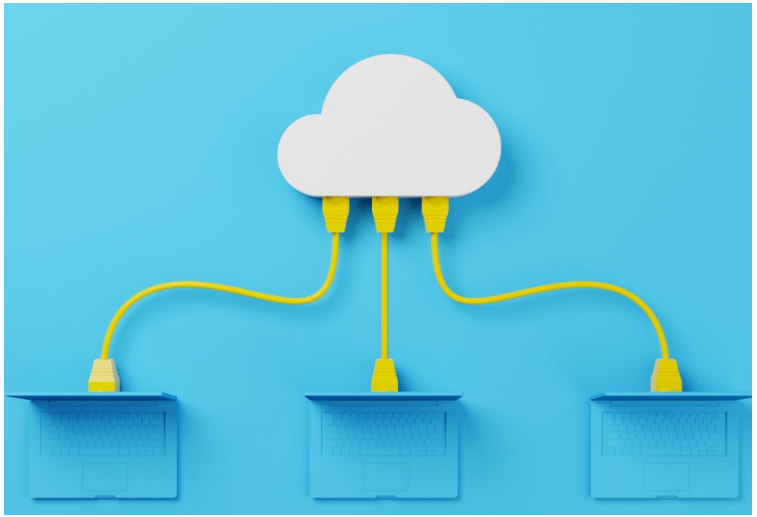
Government?

What launched the big data era?

1. A flood of data

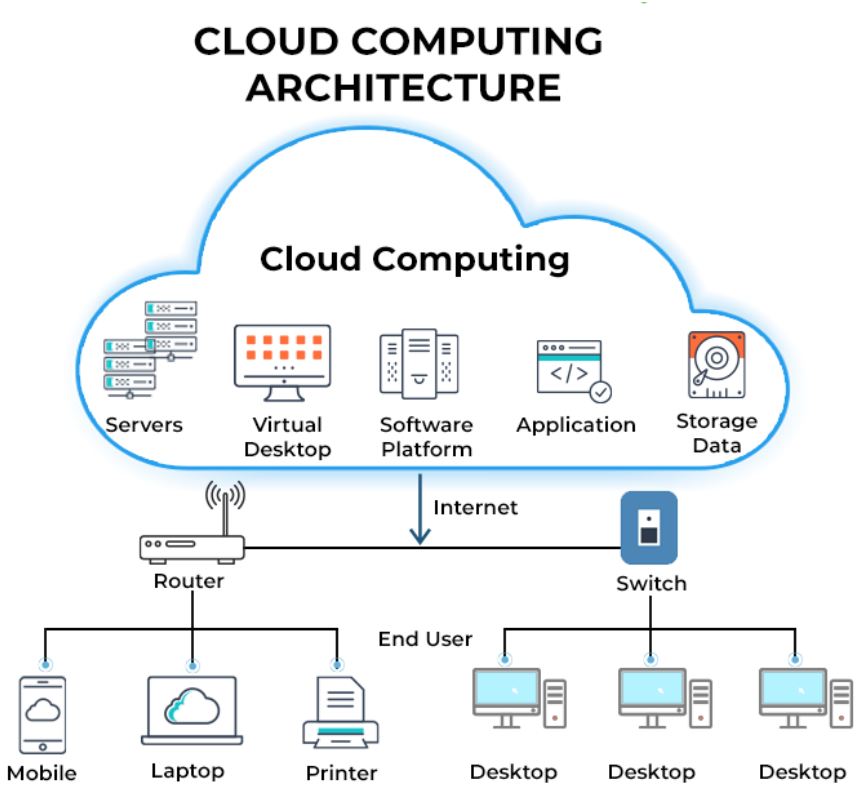


2. Cloud Computing



What launched the big data era?

2. Cloud Computing



What launched the big data era?

2. Cloud Computing

Largest Data Center in the world: China Telecom Data Center

Size: more than 150 football fields!

Server (as part of a Data center)

A computer program or device that provides a service to another computer program & it's users

≠

Supercomputer (also part of a data center, though not all)

Usually used in academic settings for data-intensive & computation heavy research & engineering e.g. weather forecasting, flight simulations

What launched the big data era?

2. Cloud Computing

OVH Data Center (France)

What

1. A

So finally, what on earth is big data then?

Data sets that are too large or complex to be dealt with by traditional data-processing application software

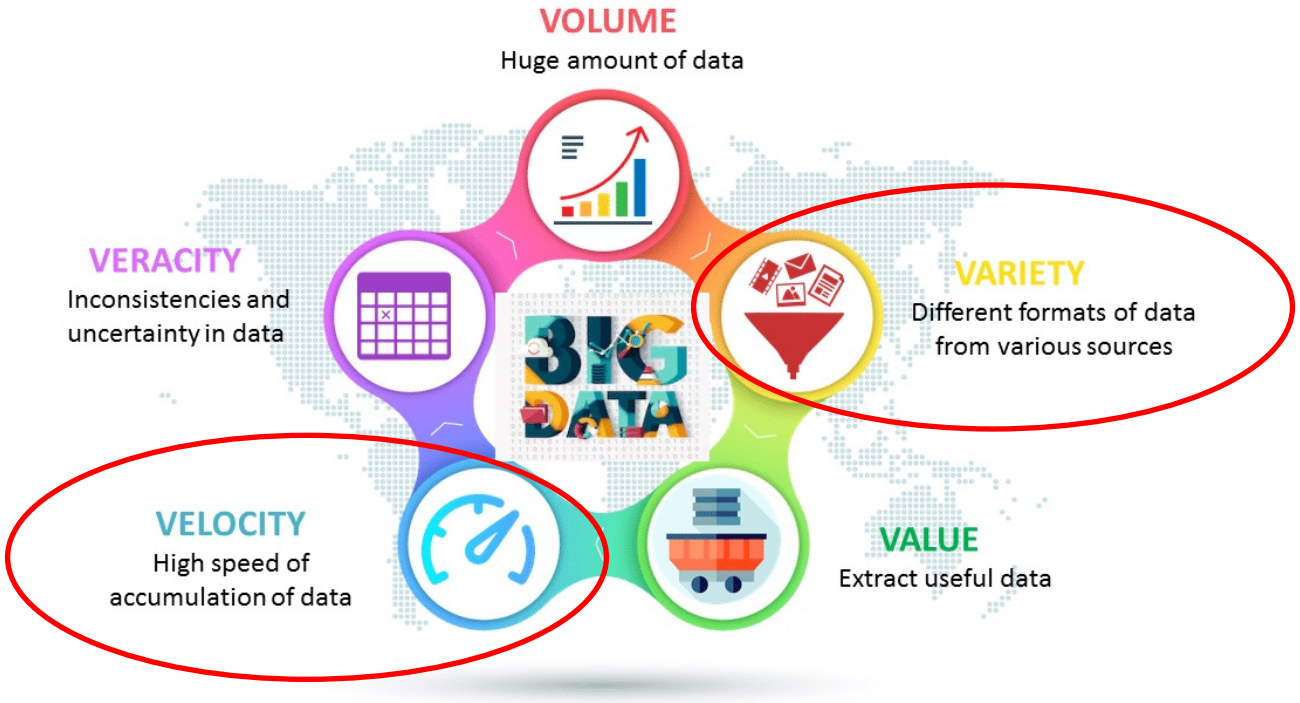
In other words…

- No Excel (sorry..) - ca 1 million rows
- No traditional data processing software
  R: 1 billion rows (with additional efforts); > 1 billion rows →map reduce algorithms processed with Hadoop etc.
- Usually not run on your local computer, but on the cloud

→ Depends!

The V's of big data

Big data is characterized using a number of V's…



The V's of big data

Variety: Different formats of data from various sources

Pictures

Tweets (text)

Video

Graphs

Temperature (Number)

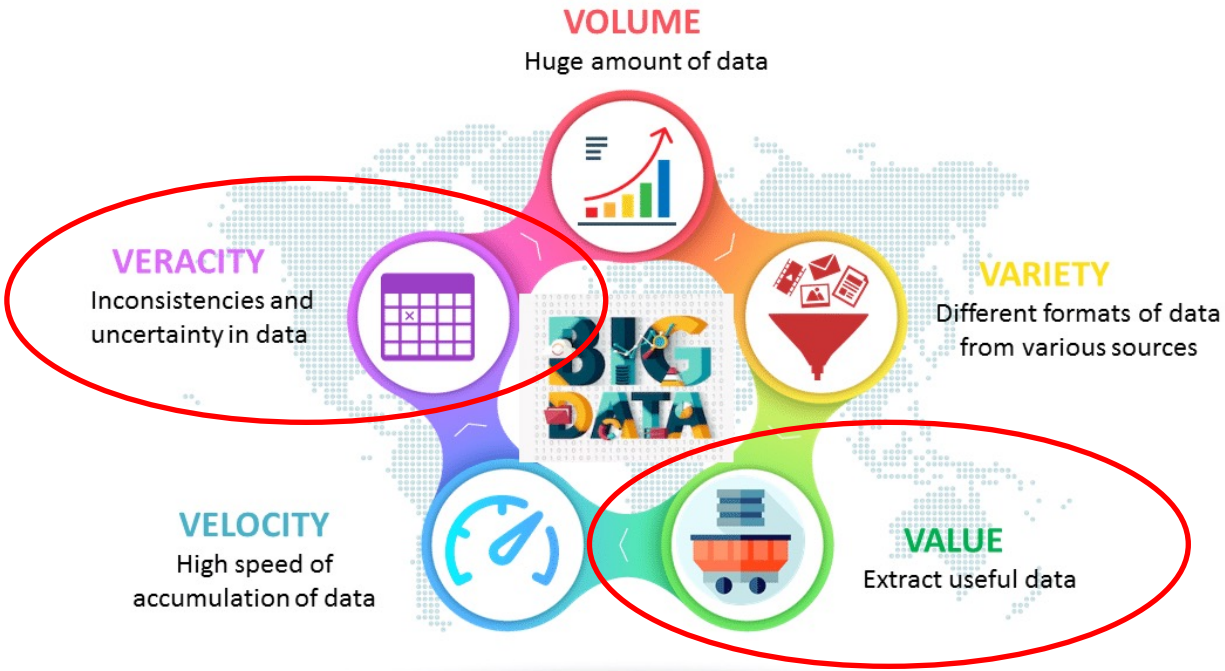
Voice

MANY more! (Brain waves, reaction times, …)

Velocity: High speed of data generation

The V's of big data

Big data is characterized using a number of V's…



The V's of big data: Veracity & Value

Veracity: Data is inaccurate or inconsistent

Table 1

	V1	V2	V3
1	1	NA	10
2	2	6	NA
3	NA	7	NA
4	4	NA	13
5	NA	9	14

NA's…

Dirty

	Ship Mode	First Class			Same Day
	Segment	Consumer	Corporate	Home Office	Consumer
Order ID	Order Date				
CA-2011-100293	14-Mar-13				
CA-2011-100706	16-Dec-13				
CA-2011-100895	02-Jun-13				
CA-2011-100916	21-Oct-13				
CA-2011-101266	27-Aug-13				
CA-2011-101560	28-Nov-13				
CA-2011-101770	31-Mar-13				
CA-2011-102274	21-Nov-13				
CA-2011-102673	01-Nov-13				
CA-2011-102988	05-Apr-13				
CA-2011-103317	05-Jul-13		242.546		
CA-2011-103366	15-Jan-13	149.95			
CA-2011-103807	02-Dec-13				
CA-2011-103989	19-Mar-13		590.762		
CA-2011-104283	27-Jun-13				
CA-2011-106054	06-Jan-13		12.78		
CA-2011-106810	14-May-13				
CA-2011-107573	12-Dec-13				
CA-2011-107811	29-Apr-13				

Clean

Ship Mode	Segment	Order ID	Order Date	Sales
First Class	Consumer	CA-2011-103366	15-Jan-13	149.95
First Class	Consumer	CA-2011-109043	15-Aug-13	243.6
First Class	Consumer	CA-2011-113166	24-Dec-13	9.568
First Class	Consumer	CA-2011-124023	07-Apr-13	8.96
First Class	Consumer	CA-2011-130155	19-May-13	34.2
First Class	Consumer	CA-2011-136861	05-Sep-13	31.984
First Class	Consumer	CA-2011-153927	12-Aug-13	286.65
First Class	Consumer	CA-2011-157784	05-Jul-13	514.03
First Class	Consumer	CA-2011-160094	30-Apr-13	1000.95
First Class	Consumer	CA-2011-164749	23-Mar-13	9.912
First Class	Consumer	CA-2011-166730	30-Dec-13	39.128
First Class	Consumer	CA-2012-102722	18-Apr-14	106.5
First Class	Consumer	CA-2012-102778	21-Nov-14	18.176
First Class	Consumer	CA-2012-117828	23-Dec-14	194.32
First Class	Consumer	CA-2012-130218	23-Mar-14	59.48
First Class	Consumer	CA-2012-132318	30-Oct-14	182.91
First Class	Consumer	CA-2012-137974	16-Apr-14	2298.9
First Class	Consumer	CA-2012-138625	02-Nov-14	197.72
First Class	Consumer	CA-2012-141327	30-Nov-14	440.144
First Class	Consumer	CA-2012-149300	22-Nov-14	32.985
First Class	Consumer	CA-2012-150560	11-Dec-14	196.62

Formatting…

But often it's worse…

Value: Goal is to extract meaningful value (turns out to be hard sometimes!)

Quiz!

What is big data?
What are the V’s of big data?

Join at

slido.com

#1813 017

ICME

Questions?

Part 1

What is BIG DATA?

And what can we do with it?

Why is Big Data important?

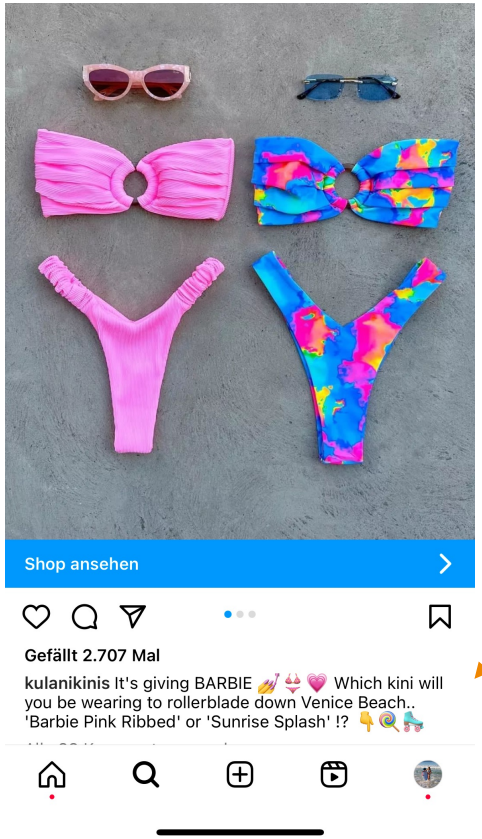
Big Data → Better Models → Higher precision

More insights

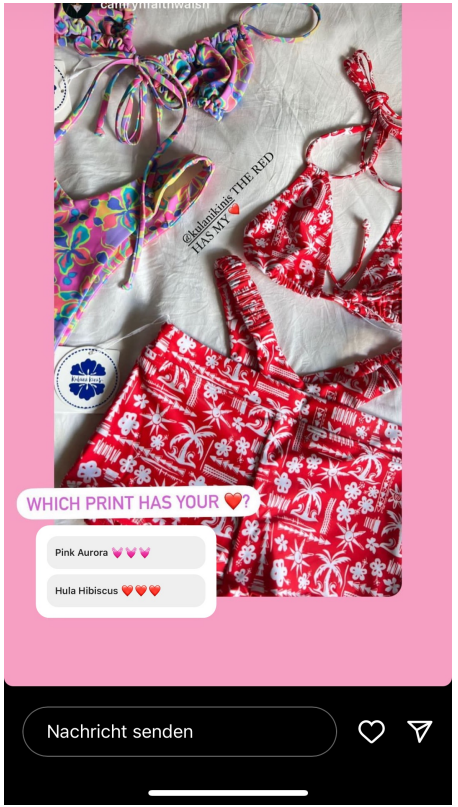
Analytics → Prediction → Generation

Real world big data problems

Real world big data problems



Fashion companies predict sales (& thus production needs) through social media insights



Real world big data problems

target

Real world big data problems

There are over **three billion base pairs (sites) on a human genome**: sequencing a whole genome generates more than 100 gigabytes of data

Genome Sequencing

Understand genetic variations, mutations, and their associations with diseases and traits.

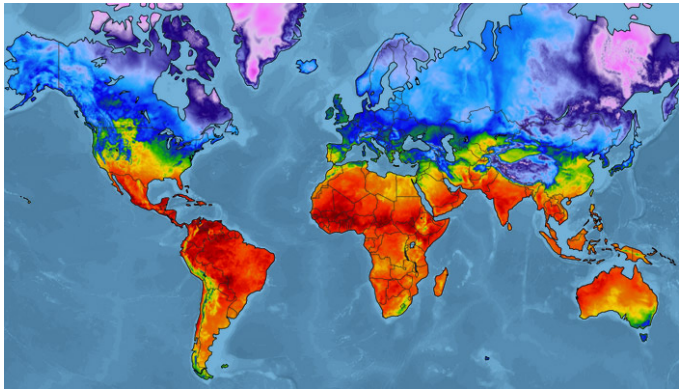
Personalized Medicine

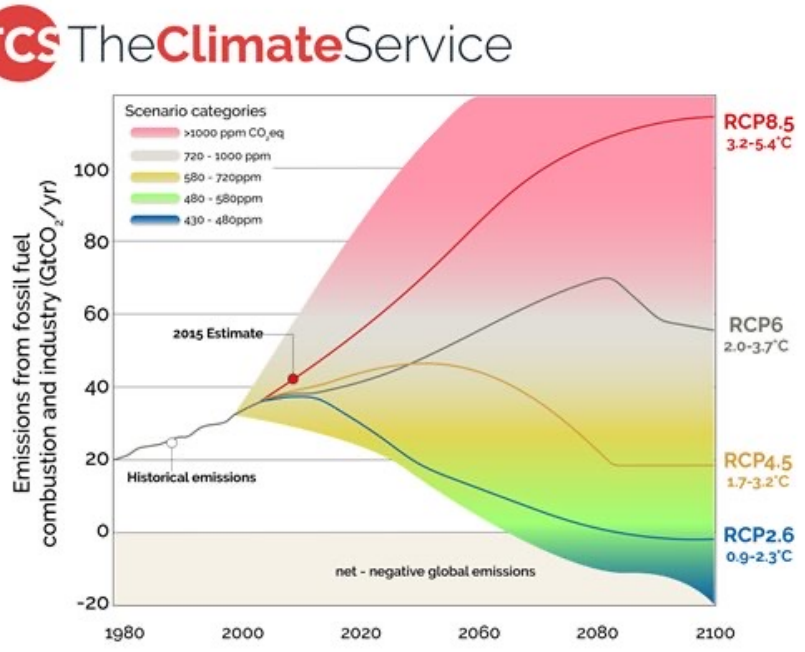
Tailor treatment plans to a patient's unique genetic makeup, increasing treatment efficacy and reducing adverse effects

Real world big data problems

Climate risk modeling

- Climate Scientists
- Insurance companies
- Governments





Group Exercise!

What questions can you answer using big data within your organization/life?

Break

Take a 5 minute break

Part 2

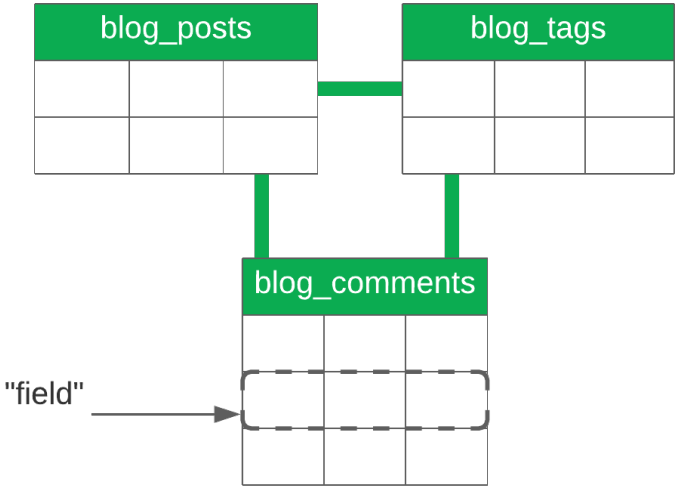
How can we best store and retrieve big data?

2 primary types of data storage

Relational Databases

The „original“ data store

Data organized in tables, keys (usually “ID) uniquely identify a row



Non-Relational (NoSQL) Databases

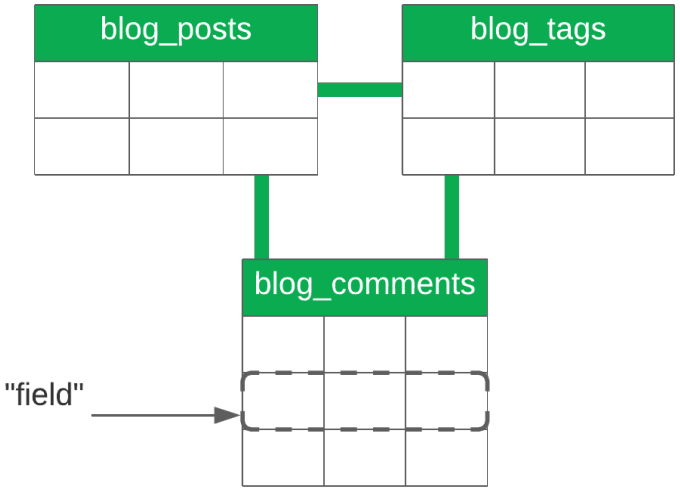
Doesn‘t enforce any „schemas“ – more loosely conntected

2 primary types of data storage

Relational Databases

The „original“ data store

Data organized in tables, keys (usually “ID) uniquely identify a row



Notion

Demonstration

2 primary types of data storage

Relational Databases

Use it when

1.Structured Data:

- If data can be organized into tables with well-defined relationships between them

2. **Data Integrity, constraints, reliability**

- Can enforce integrity rules & constraints for accuracy & consistency
- Provides strong transaction support → crucial for applications that require data integrity and reliability, such as financial systems

3. Need for Mature Ecosystem and Support:

- robust tooling, libraries, and support → easier to find developers with experience + wealth of resources for troubleshooting and optimization

4. Scalability for Moderate Data Sizes

5. Reporting and Business Intelligence:

- Usually when complex data analysis and aggregations are required

Do NOT use it when

Handling Unstructured or Semi-Structured Data:

- E.g. JSON, XML, or text documents

analytics or
aming platforms

- massive amounts of data and requires horizontal scalability across distributed systems, NoSQL databases (like column-family stores or distributed key-value stores might be a better fit)

2 primary types of data storage

- **Relational Databases**

Most common Relational Database Manageagement Systems (RDBMS)

Language to query (retrieve) data from a relational database

Structured Query Language → **You will learn how to use it NOW**

Coding Exercise!

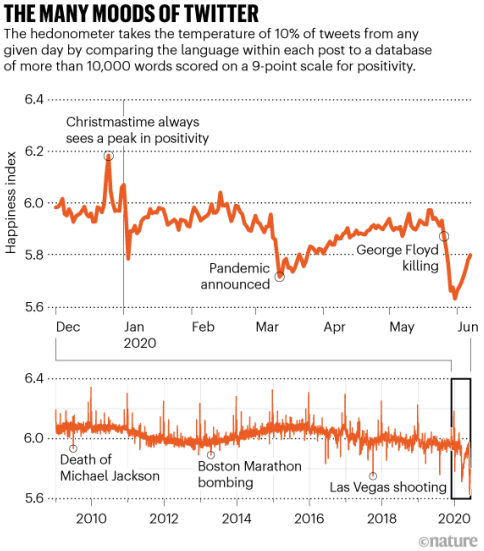
Use SQL to retrieve data from a relational database

Case Theme: Twitter data!

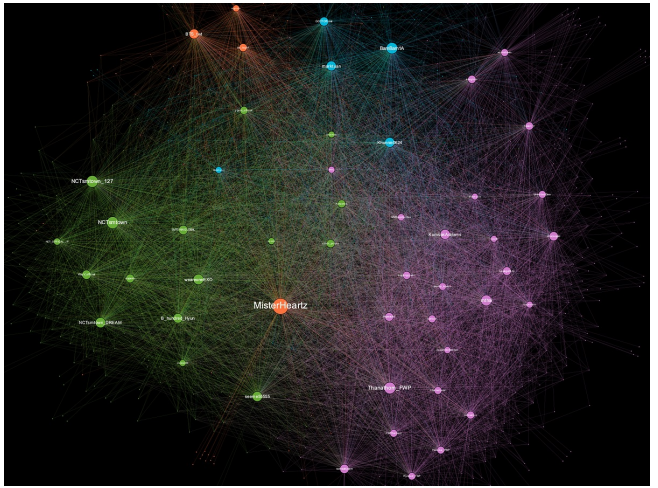
Or should we say…

What can we do with Twitter data ?

Observe trends in society



Study the spread of Fake News



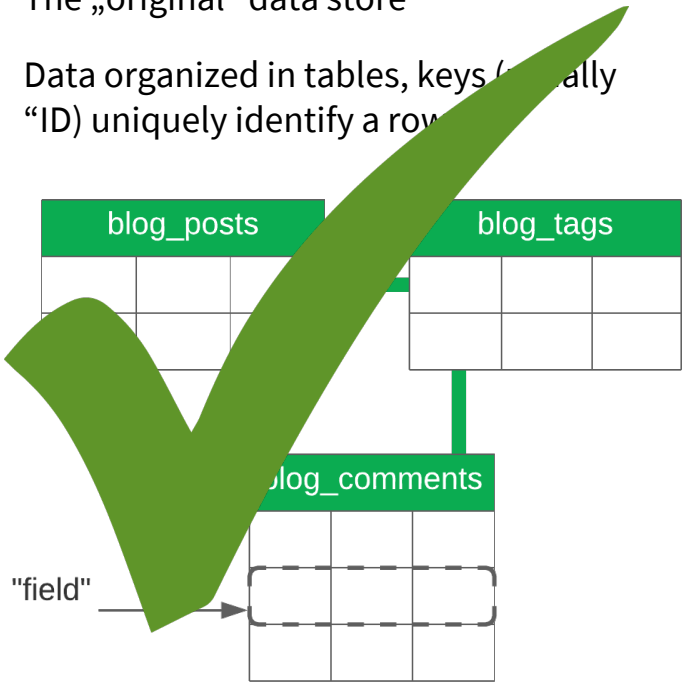
This is more complex data…

2 primary types of data storage

Relational Databases

The „original“ data store

Data organized in tables, keys (
“ID) uniquely identify a row



Non-Relational (NoSQL) Databases

Doesn‘t enforce any „schemas“ –
more loosely conntected

2 primary types of data storage

Non-Relational (NoSQL) Databases

Doesn‘t enforce any „schemas“ –
more loosely conntected

Notion

Demonstration

Questions?

2 primary types of data storage

Non-Relational (NoSQL) Databases

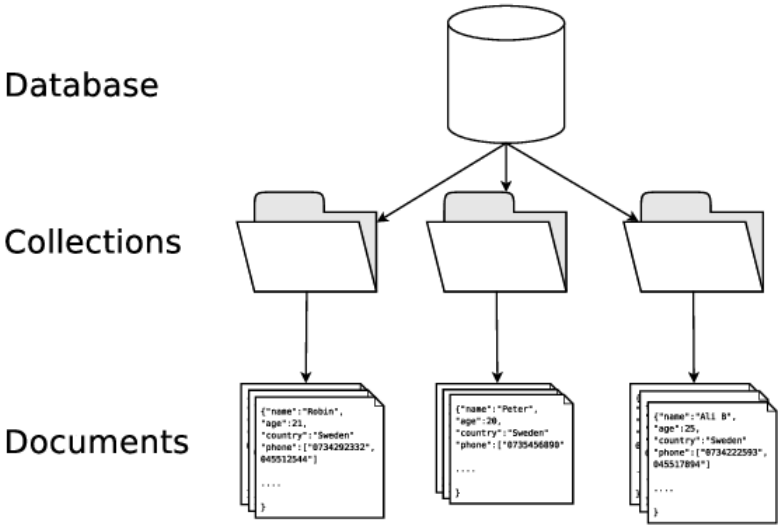
Doesn‘t enforce any „schemas“ –
more loosely conntected

Many different types!

- Document-based databases
- Key-Value Stores
- Wide-Column Stores
- Graph Databases
- Time Series Databases (TSDB)
- Search Engine Databases

Non-Relational Databases

Document-based databases

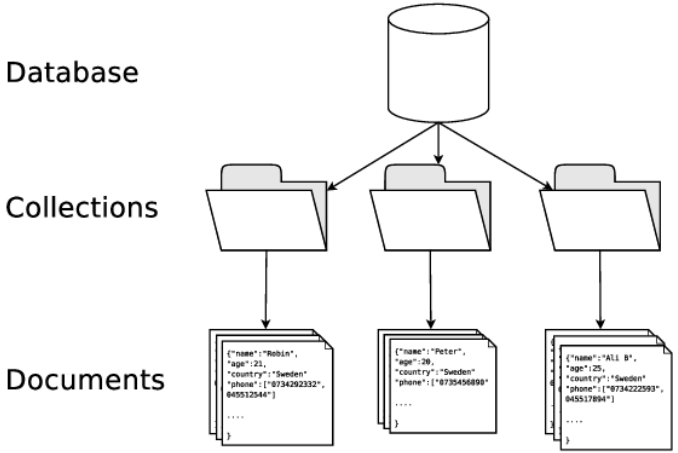



```
{
  "_id": "5cf0029caff5056591b0ce7d",
  "firstname": "Jane",
  "lastname": "Wu",
  "address": {
    "street": "1 Circle Rd",
    "city": "Los Angeles",
    "state": "CA",
    "zip": "90404"
  }
  "hobbies": ["surfing", "coding"]
}
```

Horizontally Scalable! – Good for running big data applications

Non-Relational Databases

Document-based databases



What to use it for?

- Content Management Systems (CMS)
- Product Catalogs
- User Profiles and Session Data
- Internet of Things (IoT) Applications
- Collaborative Platforms

Company that uses it

Notion

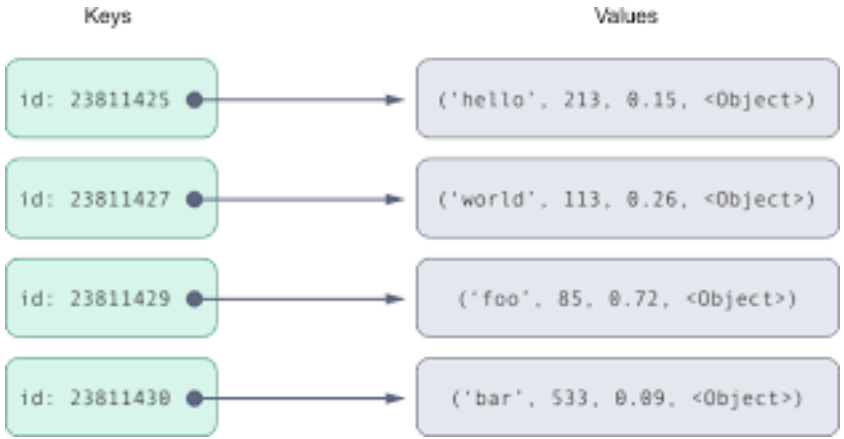
Popular Software

mongoDB.

Non-Relational Databases

Key Value Stores

- Simplest type of NoSQL database.
- Every single item in the database is stored as an attribute name (or key), together with its value.



What to use it for?

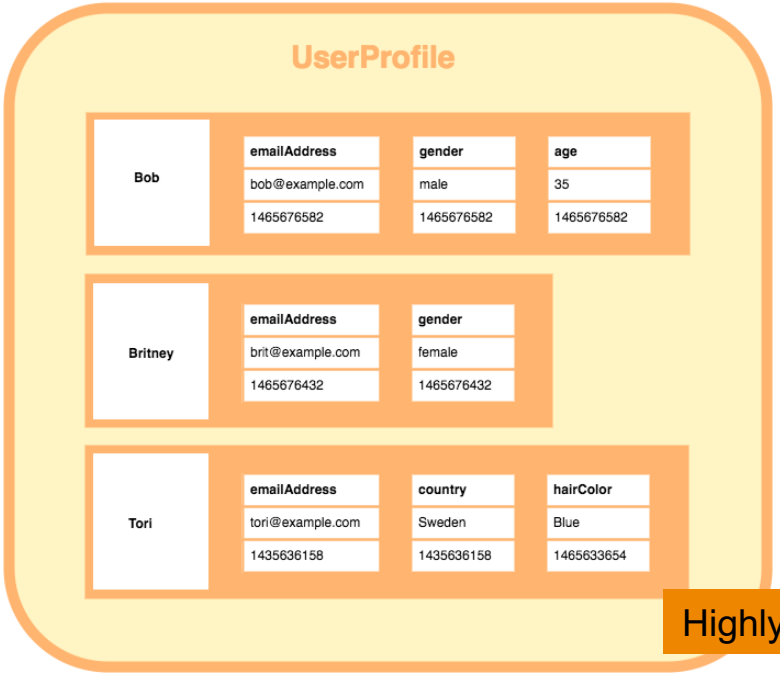
Key-value stores are good for simple applications that need t**o store simple objects temporarily**. An obvious example is a cache (speed up applications by minimizing reads & writes to slower disk-based systems)

- Web Applications: User session details
- Real time recommendations & advertising

redis riak Amazon DynamoDB

Non-Relational Databases

Wide Column Stores



What to use it for?

Generally when columns are not always the same for each row

- OLAP & Business intelligence,
- Real-time analytics, high number of concurrent users
- Big data – efficient storage & retrieval of data

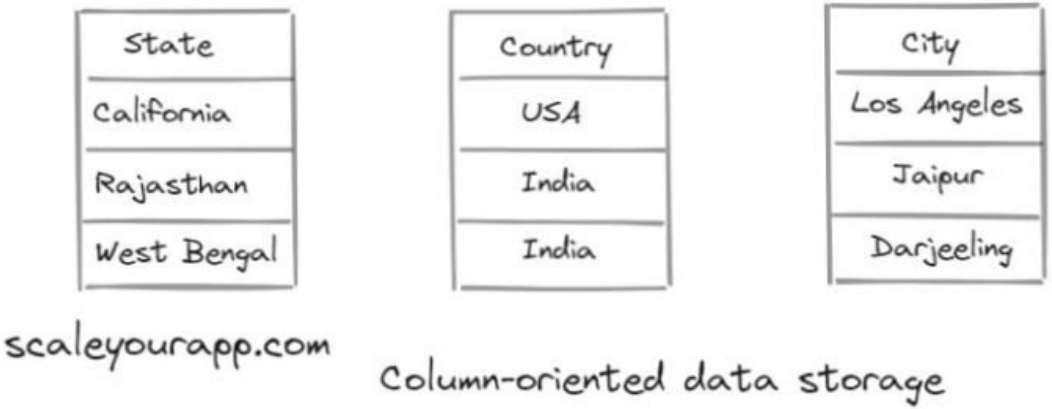
BigQuery

amazon REDSHIFT

Highly scalable!

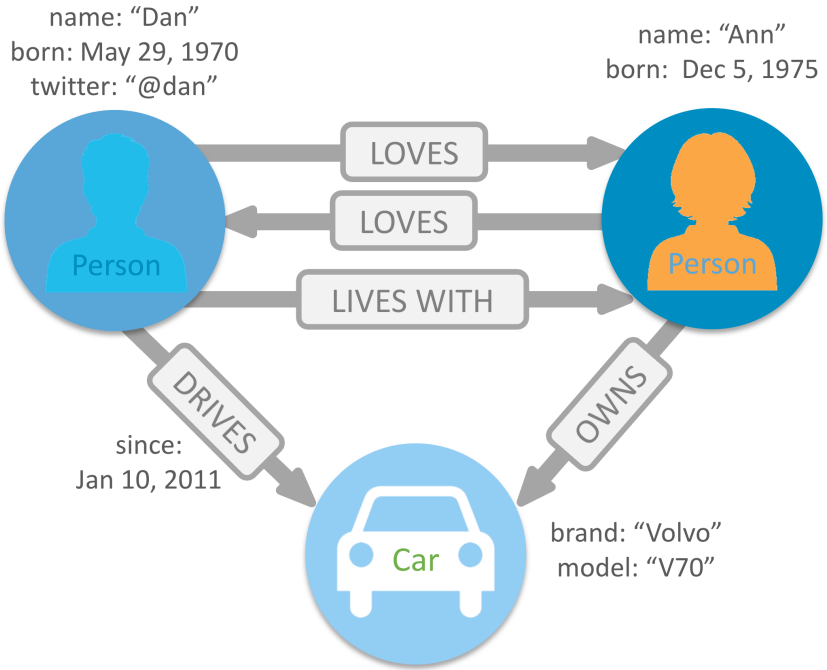
Non-Relational Databases

- **Wide Column Stores – more explanation**



Non-Relational Databases

Graph Databases



Non-Relational Databases

Graph Databases

What to use it for?

Anywhere where relationships are a key element

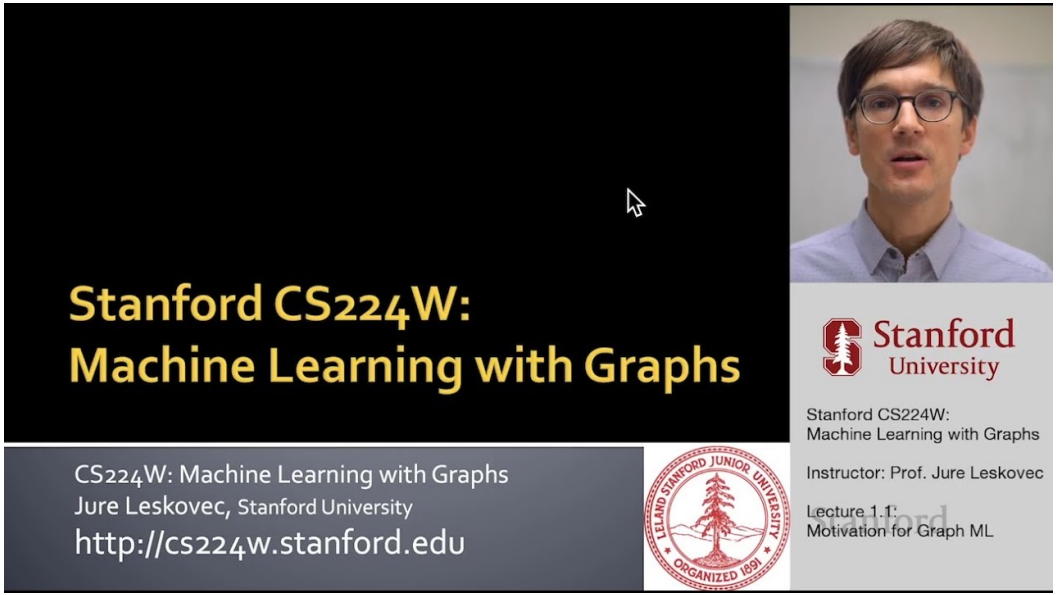
- Contact tracing
- Social Media analytics: “People you could know...”
- Fraud detection (graph of money transactions)
- Knowledge management (e.g. Siemens)

JanusGraph

neo4j

Non-Relational Databases

Graph Databases



Learn more about how to use Graphs & Machine Learning!

For more introductory classes reach out to the teaching team

Non-Relational Databases

Time Series Databases

Measurement Time	Air Quality Index (AQI)	The density of PM2.5
2018/01/01 00:00	156	45
2018/01/01 01:00	101	29
2018/01/01 02:00	97	19
…	…	…
2018/12/31 21:00	133	34
2018/12/31 22:00	135	36
2018/12/31 23:00	141	43

What to use it for?

Store data based on time stamps

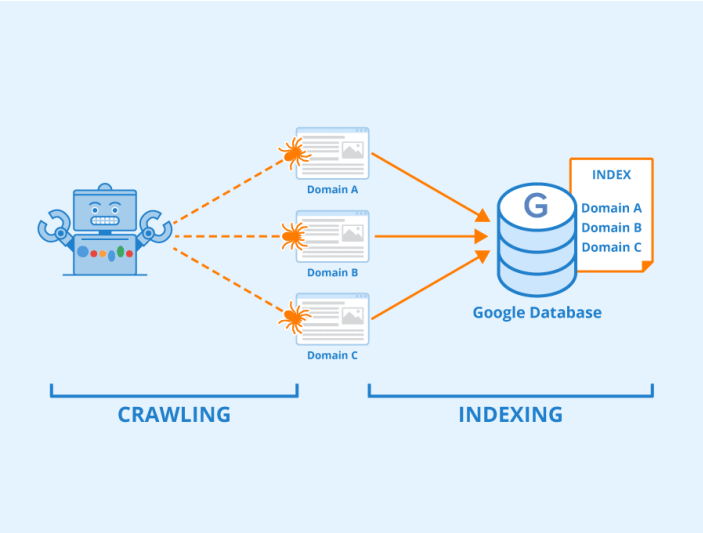
- Store and access IoT data
- Understand financial trends (stock prices, etc.)
- Process self-driving vehicle data
- Sales forecasting

influxdb

OpenTSDB

Non-Relational Databases

Search Engine Databases



Solr

elasticsearch

What to use it for?

- **Web Search Engines:** powering web search engines like Google, Bing, and Yahoo.
- **Enterprise Search:** enable employees to search across various internal data sources, including documents, emails, databases, and knowledge bases.
- **E-commerce Search:** enable efficient product searches for online shoppers.
- **Content Management Systems (CMS):** enable fast and relevant content searches for website visitors. This includes searching articles, blog posts, images, and multimedia conten

Break 5 min

Coding Exercise!

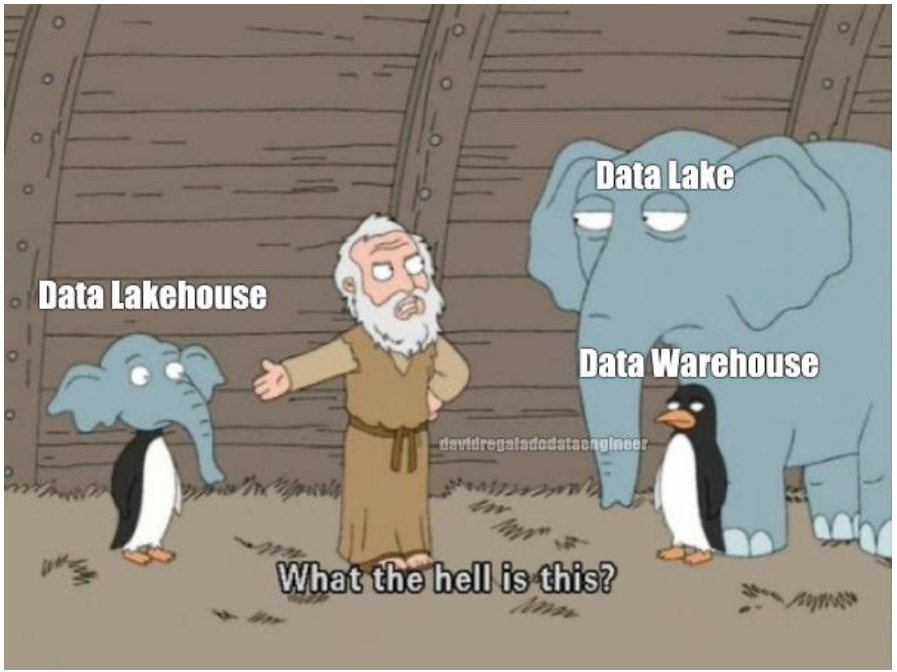
Non-relational Database

Data Warehouse? Data Lake?

Data base

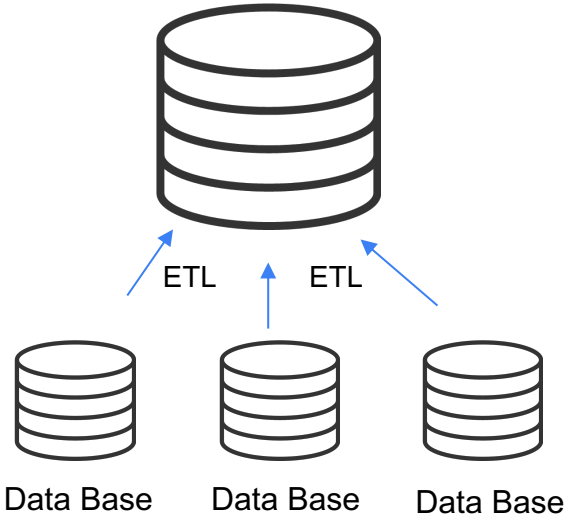
Data Warehouse???

Data Lake???



Data Warehouse

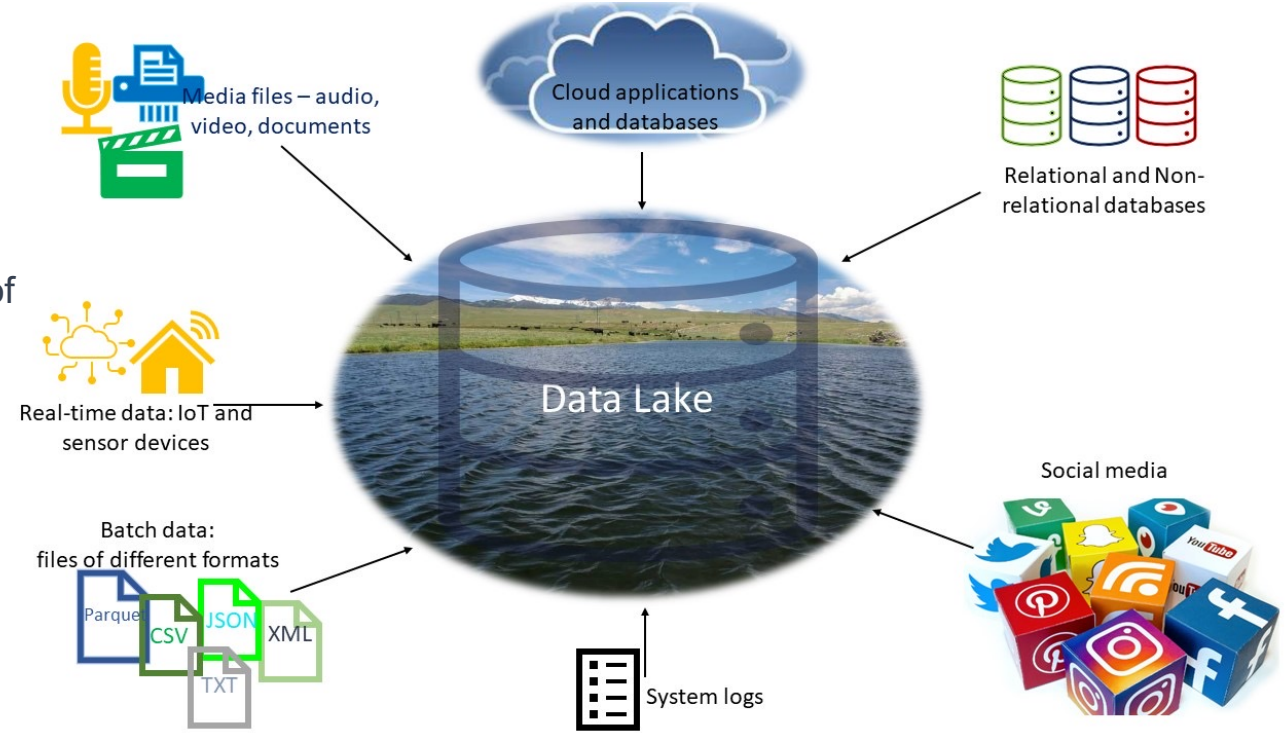
Data Warehouse **--- is a ---** **Data Base**



Data Base	Data Warehouse
Store transactions, etc	For analytics & reporting
Refreshed & detailed	Refreshed periodically & summarized
Work slowly for quering data	Generally faster

Data Lake

- centralized and scalable **storage repository** that
- holds a vast amount of **raw, unstructured, and structured data from multiple sources**,
- allowing for flexible data exploration and analysis.



DATA LAKE vs DATA WAREHOUSE

Data

unstructured

Users

Data Scientists,
Data Analysts

Use cases

Stream Processing,
Machine Learning,
Real time analysis

Raw

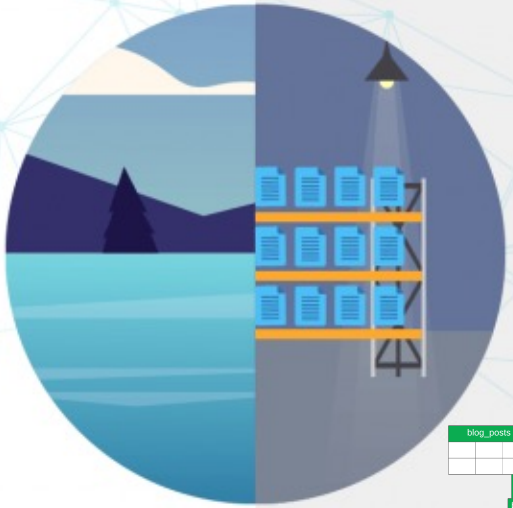
Data Lakes contain unstructured, semi structured and structured data with minimal processing. It can be used to contain unconventional data such as log and sensor data

Large

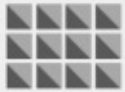
Data Lakes contain vast amounts of data in the order of petabytes. Since the data can be in any form or size, large amounts of unstructured data can be stored indefinitely and can be transformed when in use only

Undefined

Data in data lakes can be used for a wide variety of applications, such as Machine Learning, Streaming analytics, and AI



Data



Structured

Users

Business Analysts

Use cases

Batch Processing,
BI, Reporting

Refined

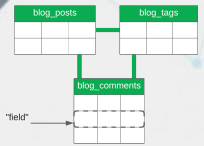
Data Warehouses contain highly structured data that is cleaned, pre-processed and refined. This data is stored for very specific use cases such as BI.

Smaller

Data Warehouses contain less data in the order of terabytes. In order to maintain data cleanliness and health of the warehouse, Data must be processed before ingestion and periodic purging of data is necessary

Relational

Data Warehouses contain historic and relational data, such as transaction systems, operations etc



How do they all work together?

Option 1

Use a data warehouse to do analysis

Image (above): Land data in a data warehouse, analyze the data, then share data to use with other analytics and machine learning services

How do they all work together?

Option 2

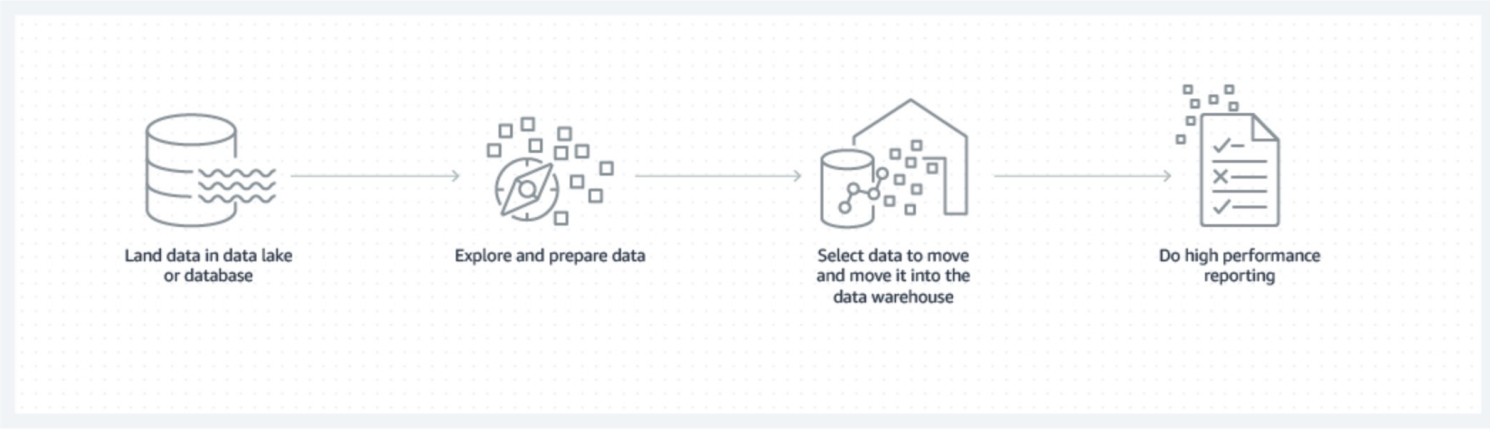


Image (above): Land data in a database or datalake, prepare the data, move selected data into a data warehouse, then perform reporting.

How do they all work together?

Option 3

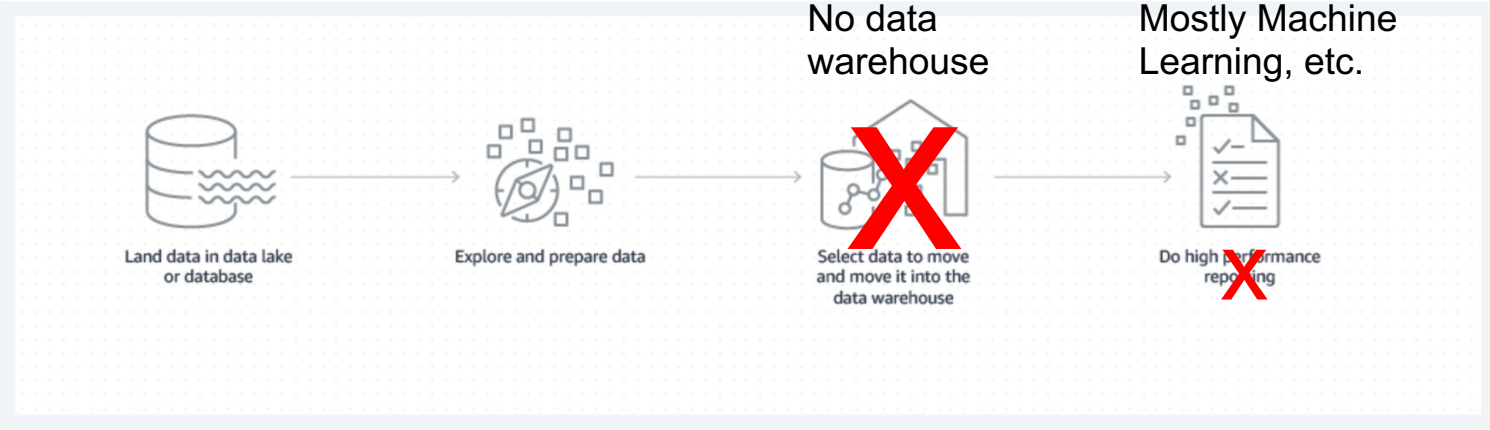


Image (above): Land data in a database or datalake, prepare the data, move selected data into a data warehouse, then perform reporting.

Practical Exercise!

Explore data in an AWS S3 **Data Lake!**

Recap and Q&A

Part 1

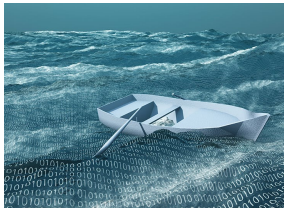
What is BIG DATA?

And what can we do with it?

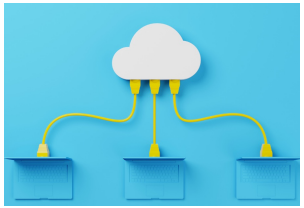
What is big data?

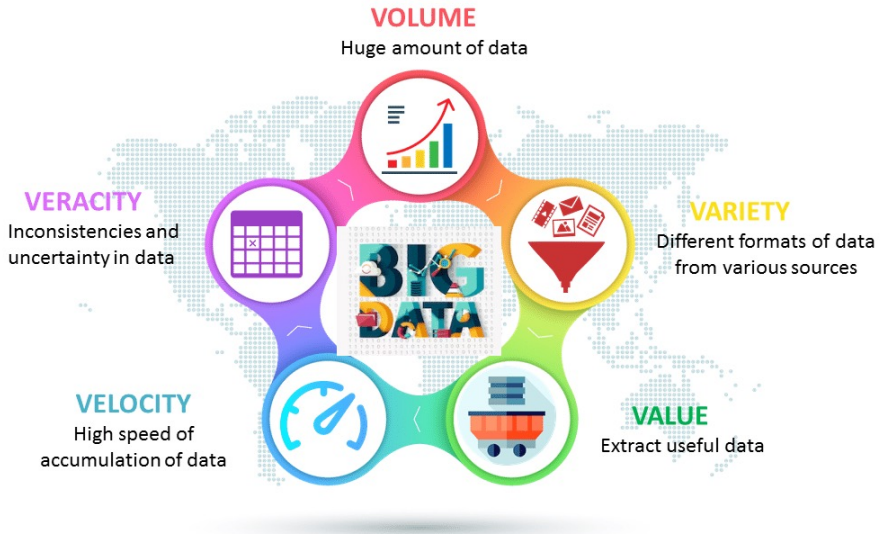
Data sets that are too large or complex to be dealt with by traditional data-processing application software

1. A flood of data



2. Cloud Computing





Part 2

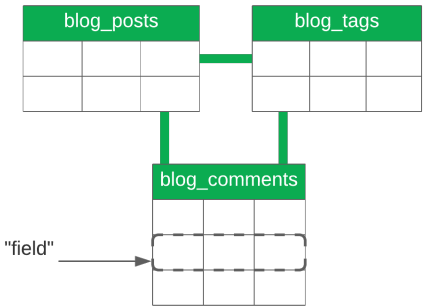
How can we best store and retrieve big data?

2 primary types of data storage

Relational Databases

The „original“ data store

Data organized in tables, keys (usually “ID) uniquely identify a row

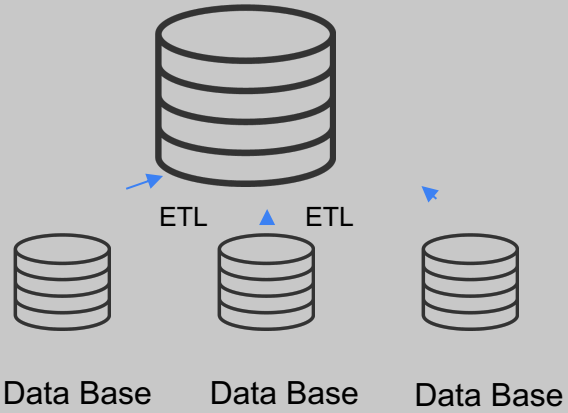


Non-Relational (NoSQL) Databases

Doesn‘t enforce any „schemas“ –
more loosely conntected

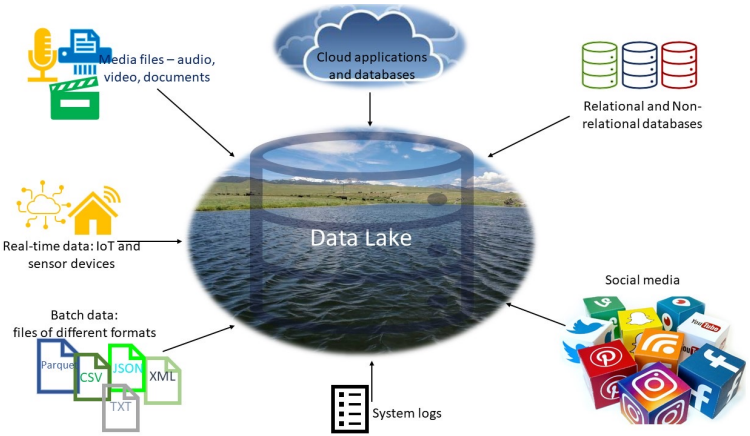
Data Warehouse vs Data Lake

Data Warehouse



Structured, relational data

Data Lake



Unstructured, raw data that is unrelated

Any Questions?

Raise your hand!

Fill out feedback survey!