Data analysis pipelines
Big data and data representation

Digital Technology
B. Pernici

 1

1

Quiz

Link: https://forms.office.com/e/jTNpH3gVNH

Topics covered: Python, SQL and Data Quality until Lecture 9.

The quiz can be taken online by connecting to the MS Forms at the given time, 10 minutes to complete it (self timing, the form must be submitted before the 10 minutes expire). Polimi credentials must be used to access the form (try connecting with the link before the quiz starts to verify access). The quiz is closed books.

 2

2

Material

- Chapter 4 of Handbook
- Chapter 5 of Handbook
- Colab

https://colab.research.google.com/drive/1ppx5TOF0wUhyKlq5SUtauFolC25ARTc7?usp=sharing

Datasets

https://www.dropbox.com/sh/ww92mqe6fp0vrfo/AABVDAOIHugWHMse-8GmPq6Na?dl=0

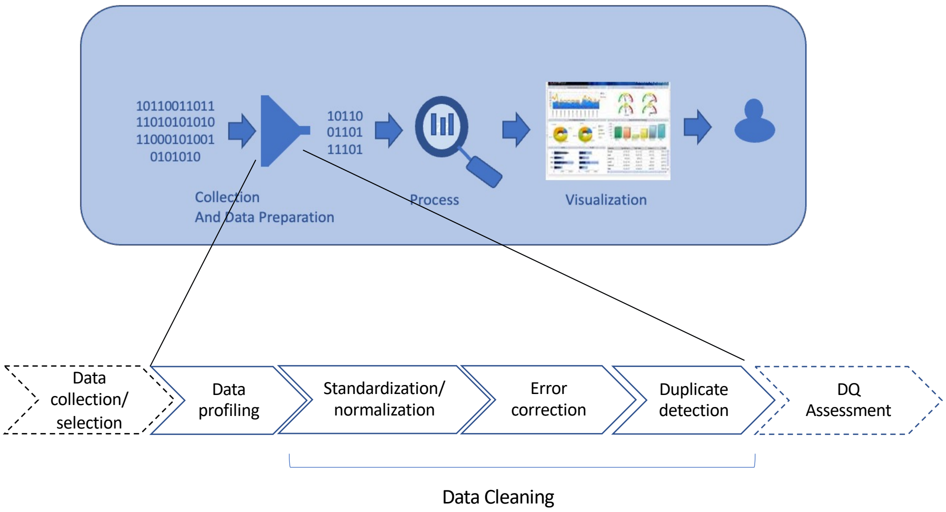
Goals and technological aspects

- Data analysis pipelines
- Data sources and Big Data
- Data warehouses

Pipelines

5

Data-driven Decision Making DDDM



6

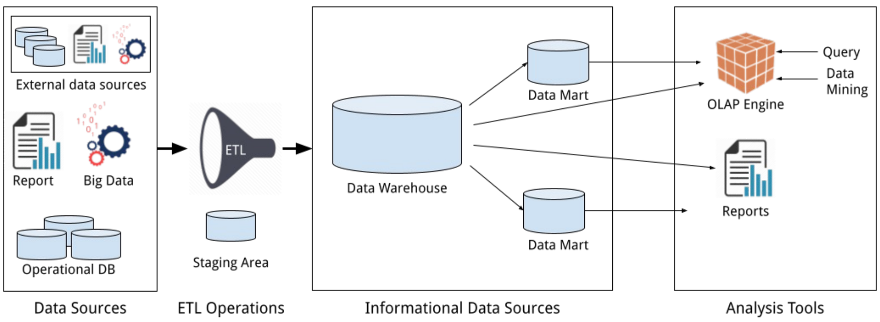
Pipelines – where?

- Datawarehouses
- Data science
- Machine learning

 7

7

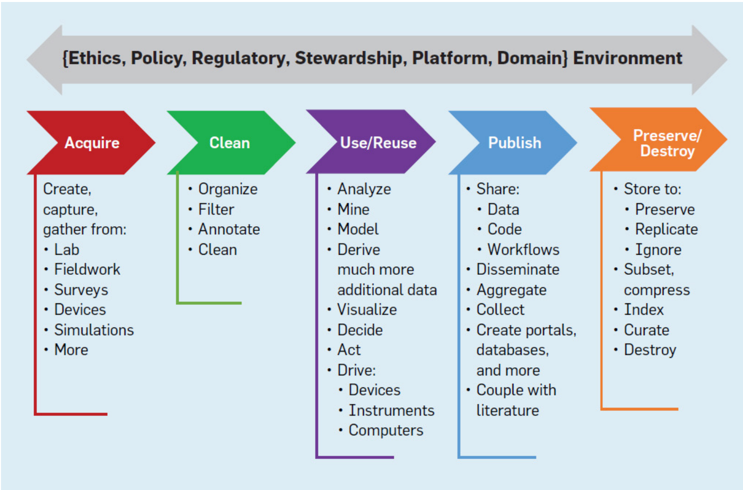
Data Warehouses



E Extraction
T Transformation
L Loading

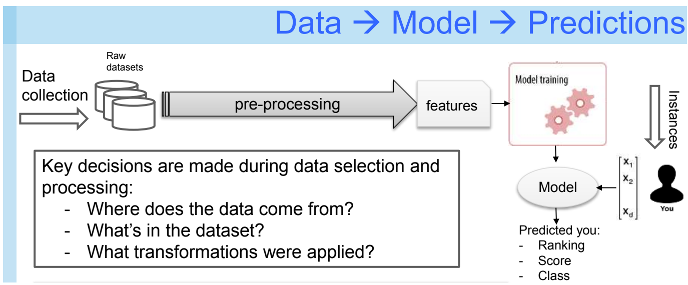
8

Data science pipelines



9

ML pipelines - Supervised learning



Missier 2021

10

Sources of data

- Traditional data sources
- Big data
- Source selection

Sources of data

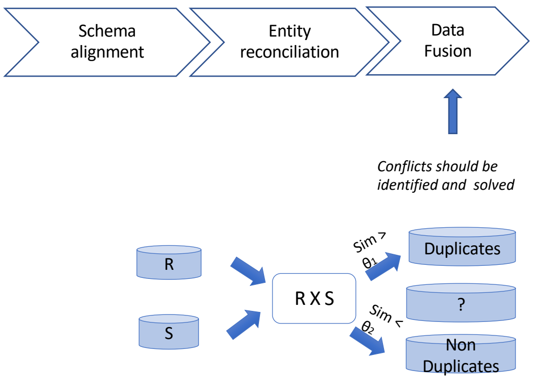
- Internal sources

From OLTP

Variable structures
Quantity
Texts

- External sources

In case of multiple sources, data integration is also needed



Note: structured data – from relational DBs

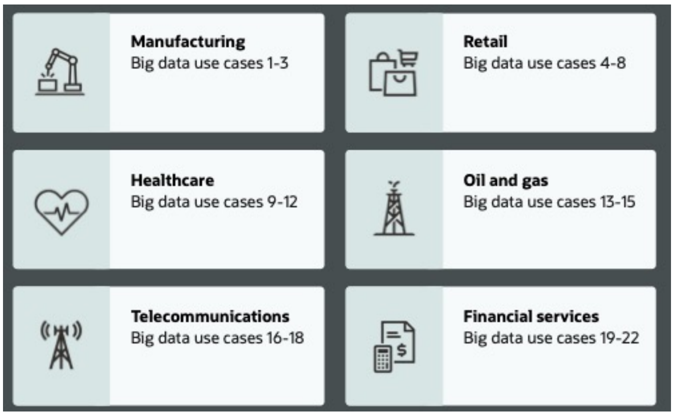
13

Big Data

 14

14

Use cases



https://www.oracle.com/a/ocom/docs/top-22-use-cases-for-big-data.pdf

Manufacturing

- Predictive maintenance
- Operational efficiency
- Production optimization

Challenges

- Companies must integrate data coming from **different formats** and identify the signals that will lead to optimizing maintenance.
- balance the data volume with the growing **number of sources**, users, and applications
- analyze their production equipment data, material use, and other factors. Combining the **different kinds of data** can pose a challenge.

https://www.oracle.com/a/ocom/docs/top-22-use-cases-for-big-data.pdf

Big Data

Volume

Scale of Data

...40 zettabytes of data will be created by 2020, an increase of 300 times from 2005...

Velocity

Analysis of streaming data

...e.g., sensors collect numerous data that have to be analyzed...

Variety

Different forms of data

...sensors, videos, databases, posts, tweets...

Veracity

Uncertainty of data

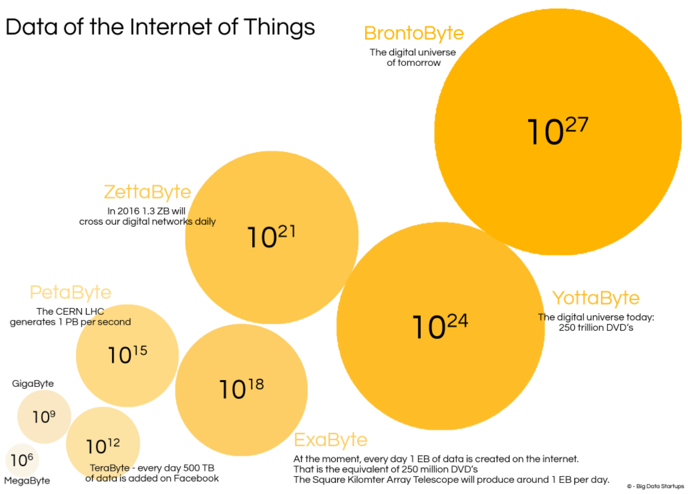
..1 in 3 business leaders don't trust the information they use to make decisions...

http://www-01.ibm.com/software/data/bigdata/

 17

17

Volume

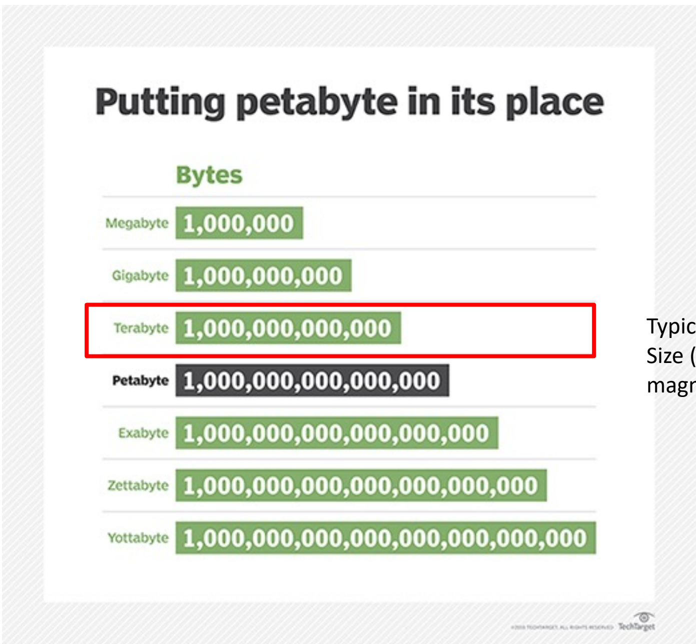


 18

18

The Challenge of Big Data

- Big data
  - Massive sets of unstructured/semi-structured data from web traffic, social media, sensors, and so on
- **Volumes** too great for typical DBMS
  - Terabytes 10^{12}, Petabytes 10^{15}, Exabytes of data 10^{18}, Zettabyte 10^{21}
- Can reveal more patterns, relationships and anomalies
- Requires new tools and technologies to manage and analyze



Typical database
Size (order of
magnitude)

Big data: many sources
Volume

- **Conversation text data**: e.g. Twitter, Facebook
- **Photo and video Image data**: e.g. Youtube
- **Audio files**: e.g. call centers
- **Sensor data**: e.g. smart phones, geo seismic data
- **The Internet of Things data**: e.g. smart devices
- **Web customer data**: e.g. Web logs
- **Traditional customer data**: e.g. receipts, loyalty programs, traffic data of telephone/Internet operators

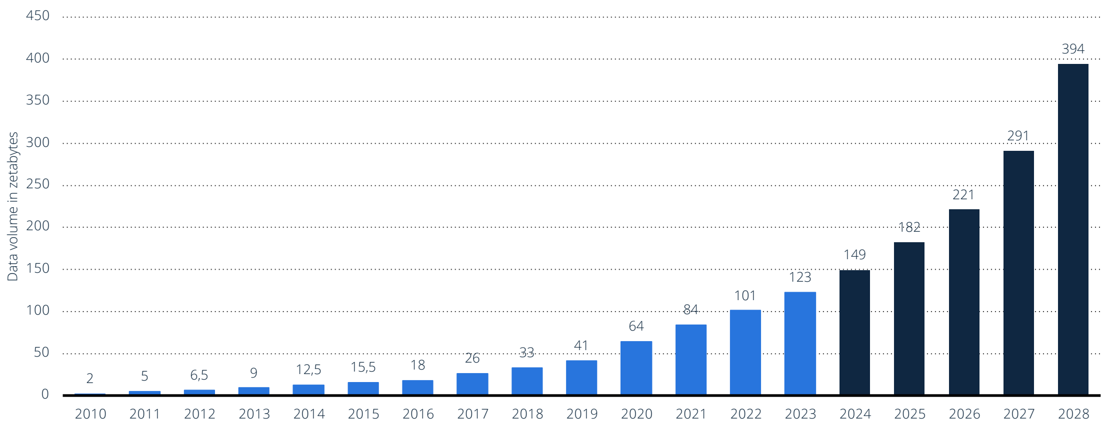
Typical database data in IS

21

21

Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2023, with forecasts from 2024 to 2028 (in zettabytes)

Amount of data created, consumed, and stored 2010-2023, with forecasts to 2028

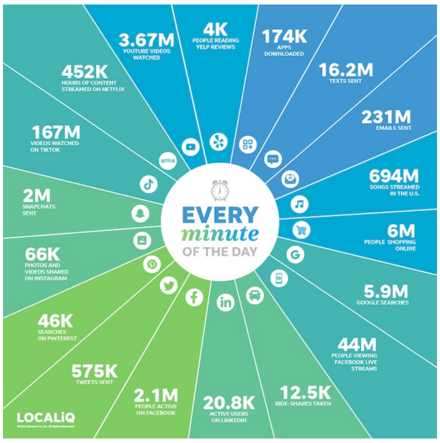


Note(s): Worldwide; May 2024; All figures have been gathered by Statista, see the "Details" tab for more information
Further information regarding this statistic can be found on page 8.
Source(s): IDC; Statista estimates; ID 871513

2

statista

23



Source: LocaliQ 2022

All that content and constant bombardment with messages presents an equally huge challenge for marketers who need to get their message heard through the storm.

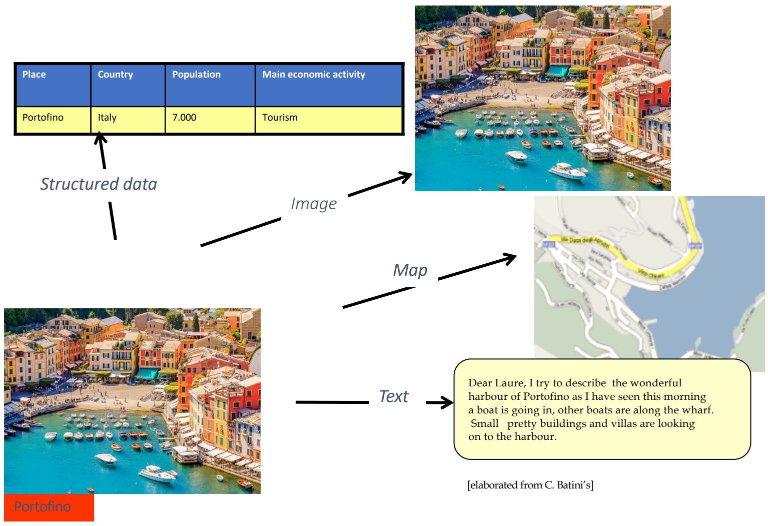
There's even a term for it - Content shock - coined by Mark Schaefer

Big data: many sources -> Variety

Conversation text data	• e.g. Twitter, Facebook
Photo and video Image data	• e.g. Youtube
Audio files	• e.g. call centers
Sensor data	• e.g. smart phones, geo seismic data
The Internet of Things data	• e.g. smart devices
Web customer data	• e.g. Web logs
Traditional customer data	• e.g. receipts, loyalty programs, traffic data of telephone/Internet operators

Structure of data
Different in different sources
Not only in relational form

Variety



[elaborated from C. Batini's]

Big data: Velocity

From data to be queried on demand to
dynamic data to be analyzed in real time
(Data streaming)

"Sometimes 2 minutes is too late. For time-sensitive processes such as catching fraud, big data must be used as it streams into your enterprise in order to maximize its value.

•Scrutinize 5 million trade events created each day to identify potential fraud

•Analyze 500 million daily call detail records in real-time to predict customer churn faster "

[https://www-01.ibm.com/software/in/data/bigdata/]

Big data Vs

- Volume
- Variety
- Velocity

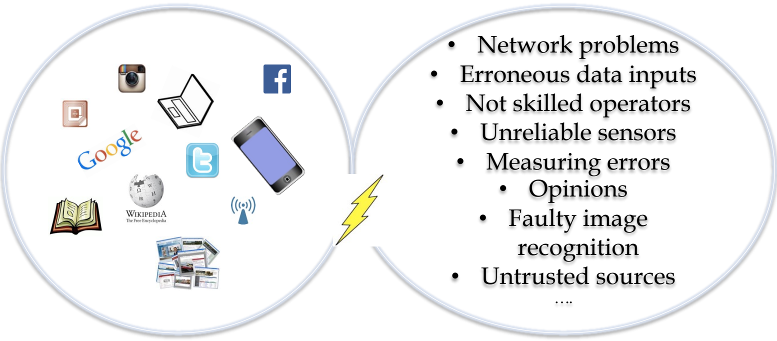
Many other Vs

- Veracity

 28

28

Veracity



DATA UNCERTAINTY

29

29

Big data and data quality

Big Data analysis allows to understand customer needs, improve service quality, and predict and prevent risks.

High quality data are the precondition for guaranteeing the quality of the results of Big Data analysis.

Big Data tried to overcome **Data Quality issues with Data Quantity. But quality is still an issue.**

Cai, Li, and Yangyong Zhu. "The challenges of data quality and data quality assessment in the big data era." *Data Science Journal* 14 (2015).

30

Big data challenges

Diversity of data sources (Variety)

- Abundant data types - internal + external data sources
- Complex data structures - structured, semi-structured, IoT
- Difficult data integration - ETL and traditional approaches useless due to data volume and velocity

Tremendous data volume (Volume)

- Data quality profiling and assessment (collection, cleaning, and integration) is difficult to execute in a reasonable amount of time.

Timeliness of data is very short (Velocity)

- Data is updated continuously. If data is not collected and analysed in real time, information becomes outdated and invalid.

Missing standard for Data Quality (Veracity)

- Standards have been proposed for DQ of traditional data sources but not for big data.

Cai, Li, and Yangyong Zhu. "The challenges of data quality and data quality assessment in the big data era." *Data Science Journal* 14 (2015).

31

Source selection

Selection criteria

- Accessibility (internal, external), licensing
- Quality of data
  - Accuracy
  - Completeness
  - Timeliness (real-time updates)
- Availability of metadata, simplicity of integration
- Reputation of source
  - Reuse
  - Data ecosystems
- Cost

DATA WITH VARIABLE STRUCTURES
Storage and interchange

Digital Technology
B. Pernici

Goals and technological aspects

- Semi-structured information, irregular structures
- Data formats
- Geographical data

Structured data and unstructured data

Structured data
- Regular data structure common to all available information of the same type
- Composed of structured elements
- Tabular format, csv files

Unstructured data
- Free text
- Irregular structures
- Other types of data (images)

Semi-structured data

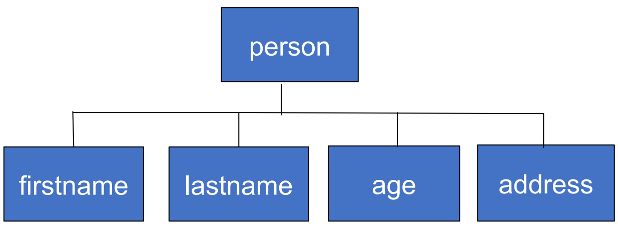
Problems to be addressed

- Multiple values for an attribute
- Hierarchical structures
- Variable structures
- “labels”
- Interoperability
- Data and metadata

 36

36

From structured data ....



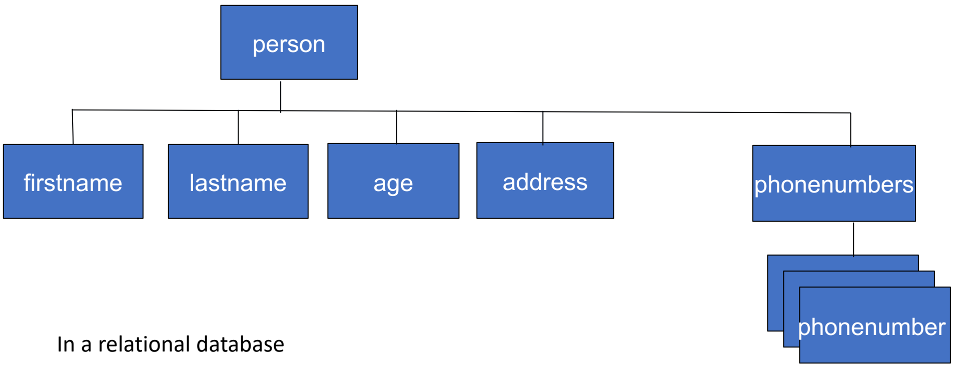
Relation schema

PERSON(firstname, lastname, age, address)

 37

37

Multiple values



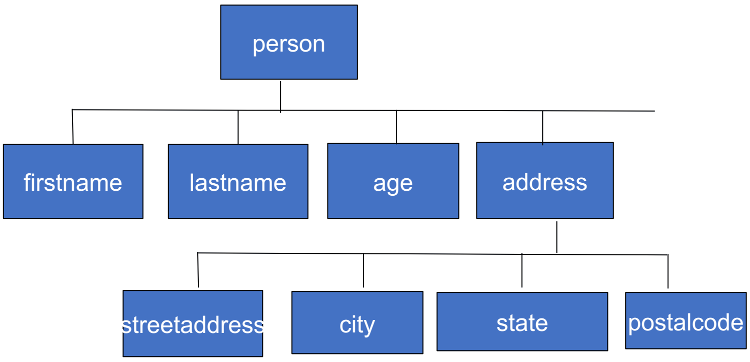
In a relational database

PERSON(<u>firstname</u>, <u>lastname</u>, age, address)
PHONENUMBER(<u>firstname</u>, <u>lastname</u>, <u>phonenumber</u>)

 38

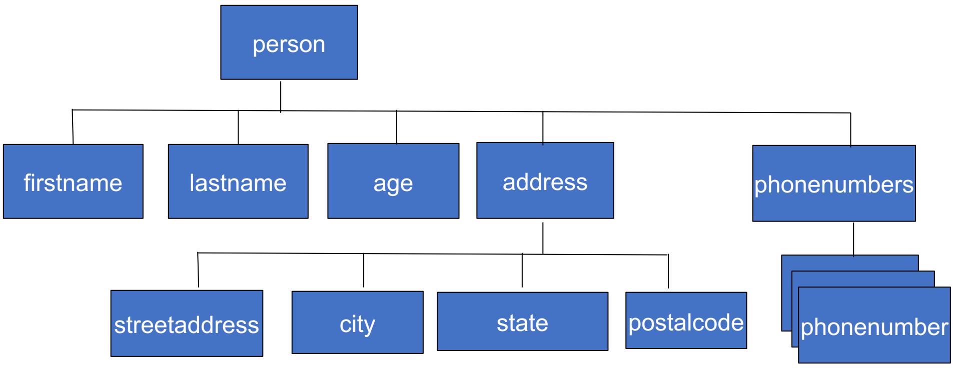
38

Hierarchical structures



 39

39



 40

40

Document formats – infrastructure level

- In all cases:
  - Schema
  - Data structured according to schema
- Uses
  - In noSQL databases
  - Interchange format

 41

41

Document formats – infrastructure level

- Schema
- Data structured according to schema
- JSON JavaScript Object Notation
- geoJSON

 42

42

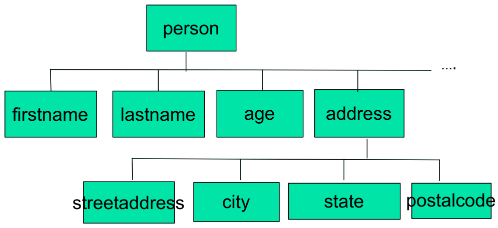
JSON – pairs name: value

- pairs
  “name”: “value”
- Separated by commas
  “name1”: “value1”,
  “name2”: “value2”
- **Nesting** through parentheses (hierarchical structures)
  { }
  objects: collections of name-value pairs
- **Lists** in square brackets (multiple values)
- [{“name1”:”value11”}, {“name1”:”value12”}, {“name1”:”value13”}]

 43

43

JSON – person representation




```
{
 "firstName": "John",
 "lastName": "Smith",
 "age": 25,
 "address": {
  "streetAddress": "21 2nd Street",
  "city": "New York",
  "state": "NY",
  "postalCode": "10021"
  },
 ....
```

Tool:
https://jsonformatter.org/

44

```
"phoneNumber": [
  {
   "type": "home",
   "number": "212 555-1234"
  },
  {
   "type": "fax",
   "number": "646 555-4567"
  }
 ],
}
}
```


45

Import/export json files

import json

json.read()
json.dump()

JSON OBJECT	PYTHON OBJECT
object	dict
array	list
string	str
null	None
number (int)	int
number (real)	float
true	True
false	False

 46

46

Examples in colab

- Using JSON data
  - From JSON strings to dataframes

  - Colab
  https://colab.research.google.com/drive/1ppx5TOF0wUhyKlq5SUtauFolC25ARTc7?usp=sharing
  Datasets
  https://www.dropbox.com/sh/ww92mqe6fp0vrfo/AABVDAOIHugWHMse-8GmPq6Na?dl=0

- Prettyprint JSON code
https://jsonformatter.org/json-pretty-print/

 47

47

Geojson

- GeoJSON is a format for encoding a variety of geographic data structures.

{ "**type": "Feature",**
"**geometry**":
{ "**type**": "Point",
"**coordinates**": [125.6, 10.1] },
"**properties**": { "name": "Dinagat Islands" }
}

Coordinates: lon, lat

48

Geojson

GeoJSON supports the following

Geometry types:

Point, LineString, Polygon, MultiPoint, MultiLineString, and MultiPolygon.

Geometric objects with additional properties are **Feature objects**.

Sets of features are contained by FeatureCollection objects.

49

geoJSON tools

- https://www.openstreetmap.org/#map=5/42.088/12.564
- https://nominatim.openstreetmap.org/ui/search.html
- https://nominatim.openstreetmap.org/search?q=cathedral,Milan&format=geojson
- http://geojson.io/#map=18/45.46390/9.19077

To try how data are collected and displayed From Copernicus Emergency service mapping

- Open https://emergency.copernicus.eu/mapping/list-of-activations-rapid
- Select an event
- Download the data
- In the folder you find json files (geoJSON)
- Open one and copy the content into

https://geojson.io/
(in the right box, selecting JSON as source)

- Explore different views (standard, satellite, OSM)

Non-relational (Not-only relational) Databases

- Not-only relational databases: "NoSQL"
  - More flexible data model
  - Data sets stored across distributed machines
  - Easier to scale
  - Handle large volumes of unstructured and structured data

- Many relational DBMS have NoSQL extensions
  - Eg Postgres

Storing non relational data

- JSON or JSON-like documents DBMS
  - PostgresSQL extensions
  - MongoDB
  - CouchDB, with emphasis on replication

- Time series DBMS
  - PostgresSQL extensions
  - Graphite

    Big companies such as **Booking.com, Reddit and GitHub** use it on a daily basis to be able to easily detect **outage** on their architecture.

Recap

- Representation of data at conceptual, logical, physical level

- Non-relational structures and formats
  - JSON
  - geoJSON

- Database technology
  - Relational
  - NoSQL

