Data Pipeline in Practice

Marco Morales
marco.morales@columbia.edu

Nana Yaw Essuman
ne2388@columbia.edu

GR5069: Applied Data Science
for Social Scientists

Spring 2025
Columbia University

recap: workflow principles in the Data Science Shop

a) **reproducibility**
   - anyone should be able to arrive to your **same results**

b) **portability**
   - anyone should be able to **pick up where you left off** on any machine

crucial tenets for **collaborative work**

a) **scalability**
   - your project should also work for **larger data sets** and/or be on the path of **automation**

recap: structuring your workspace

```
workspace
|
| -- /src
|      |-- /data              <- code to read/munge raw data
|      |-- /features          <- code to transform/append data
|      |-- /models            <- code to analyze data
|      |-- /visualizations    <- code to create visualizations
|      |-- /functions         <- scripts to centralize functions
|      |-- /config            <- configuration files
|
| -- README.md                <- high-level project description
```

Pro tips

- data is NEVER pushed to GitHub!!!!!!
- {secret keys} are NEVER pushed to GitHub!!!!!!
- references are transferred to GitHub wiki
- TODO is transferred to GitHub projects

question:
what is data?

definition

for our purposes, **data** is

information (as opposed to pure instructions)

encoded in a digital (binary) format

recorded and stored in an **electronic** form

question:

why is data collection important?

- understand your **products and systems** better
- provide means for organizations to make **better data-informed decisions**
- help identify **opportunities or gaps** in a product or system
- measure how your **consumers interact** with your products or system
- understand your **potential market**

In God we trust, all others bring data.

–William E. Deming

but…not all **data** is created equal!

- **Unstructured data**
  - does not have a predefined data model or is not organized in a pre-defined manner
  - examples of unstructured data include audio, video files or No-SQL databases.

- **Structured data**
  - pre-defined data model and ready to analyze
  - examples of structured data are Excel files or SQL databases
  - most traditional form of data storage

but…not all **data** is created equal!

structured

Family Name	Given Name	VIAF ID
Ackersdijck	Willem Cornelis	17959345
Adelung	Friedrich von	22963658
Afzelius	Arvid August	49972119
Amerling	Karel	13331054

```
Family Name,Given Name,VIAF ID
Ackersdijck,Willem Cornelis,17959345
Adelung,Friedrich von,22963658
Afzelius,Arvid August,49972119
Amerling,Karel,13331054
```

Parquet

- fixed structure (schema)

semi-structured

```
{
  "customerID": "103248",
  "name":
  {
    "first": "AAA",
    "last": "BBB"
  },
  "address":
  {
    "street": "Main Street",
    "number": "101",
    "city": "Acity",
    "state": "NY"
  },
  "ccOnFile": "yes",
  "firstOrder": "02/28/2003"
}
```

```
<?xml version="1.0" encoding="UTF-8"?>
<events>
    <event>
        <week>-mtwtf-</week>
        <starttime>06:00</starttime>
        <endtime>11:59</endtime>
        <playlist>morning.m3u8</playlist>
    </event>
    <event>
        <week>SMTWTFS</week>
        <starttime>12:00</starttime>
        <endtime>20:59</endtime>
        <playlist>afternoon.m3u8</playlist>
    </event>
</events>
```

{JSON}

<xml />

- no fixed structure (schema)
- some metadata (*e.g.* tags)

unstructured


text


image


audio


video

logs

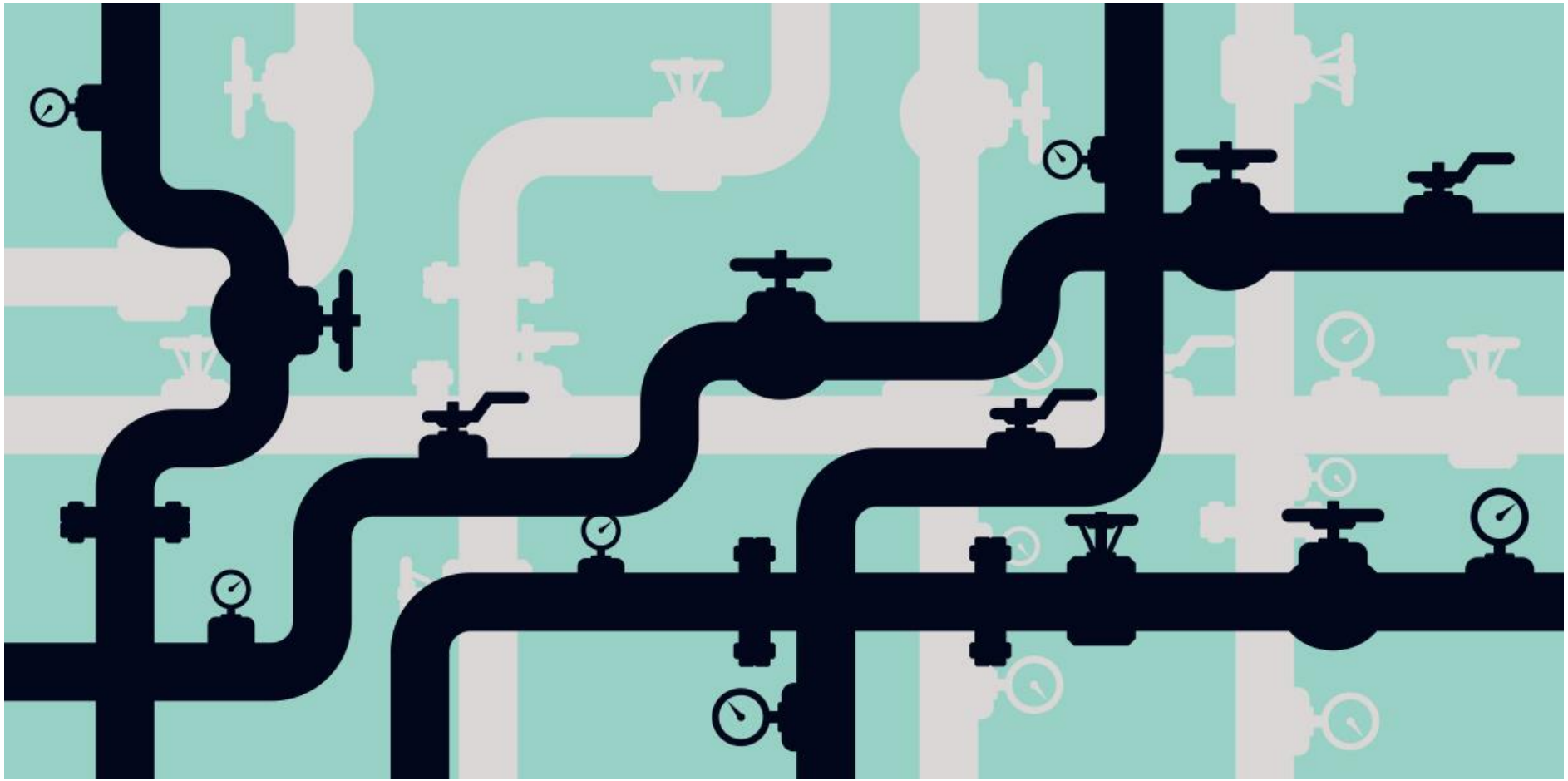
- no pre-defined structure (schema)

levels of datasets

- **First party datasets**
  - data generated by your **own product or systems**
  - the most useful and valuable data you can collect about **your consumers**
- **Second party datasets**
  - **someone else's first-party data** but useful to your organization
  - arrangement with trusted partners who are willing to share their customer data with you (and vice versa)
- **Third party datasets**
  - data that is **widely accessible to competitors**, so you aren't gaining unique advantage
  - great for demographic, behavioral, and contextual targeting
  - data that you buy from outside sources that are not the original collectors of that data (**data aggregators**)

question:
how do we get data?

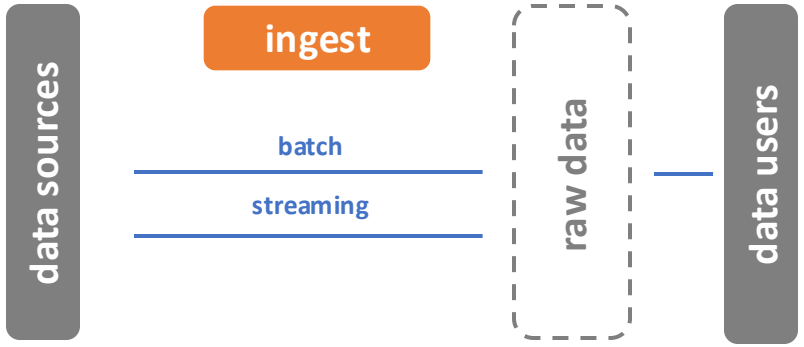
let’s introduce a concept: **data pipelines**



a **process** to **move data** across systems

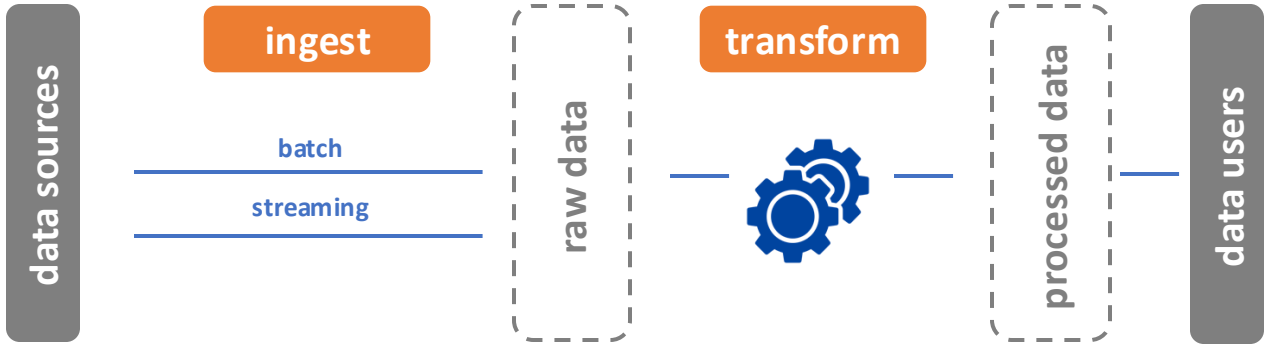
a data pipeline

- “**pull**” data in
- “**push**” data to new location

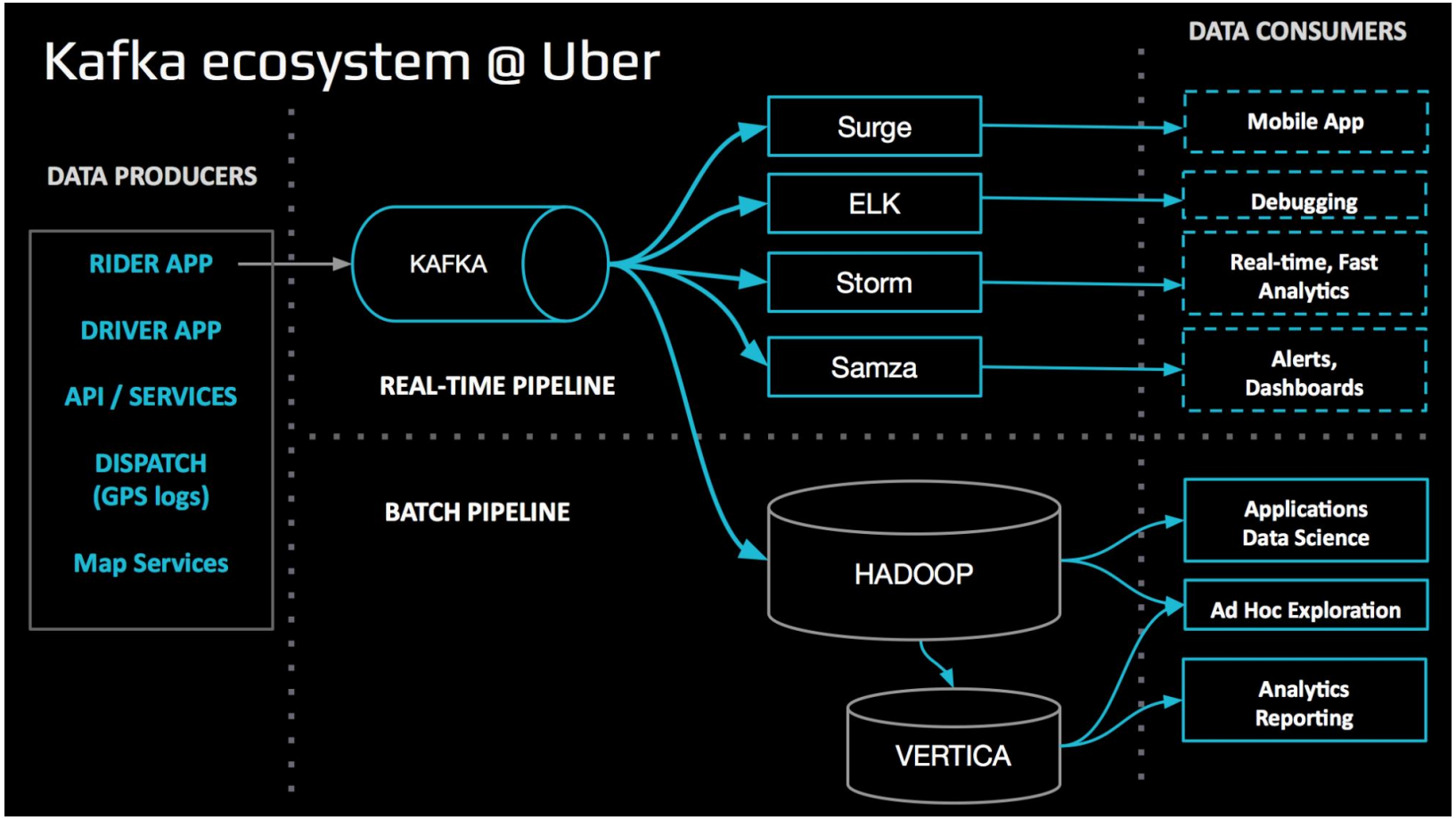


also a data pipeline

- “**pull**” data in
- **transform** data
- “**push**” data to new location



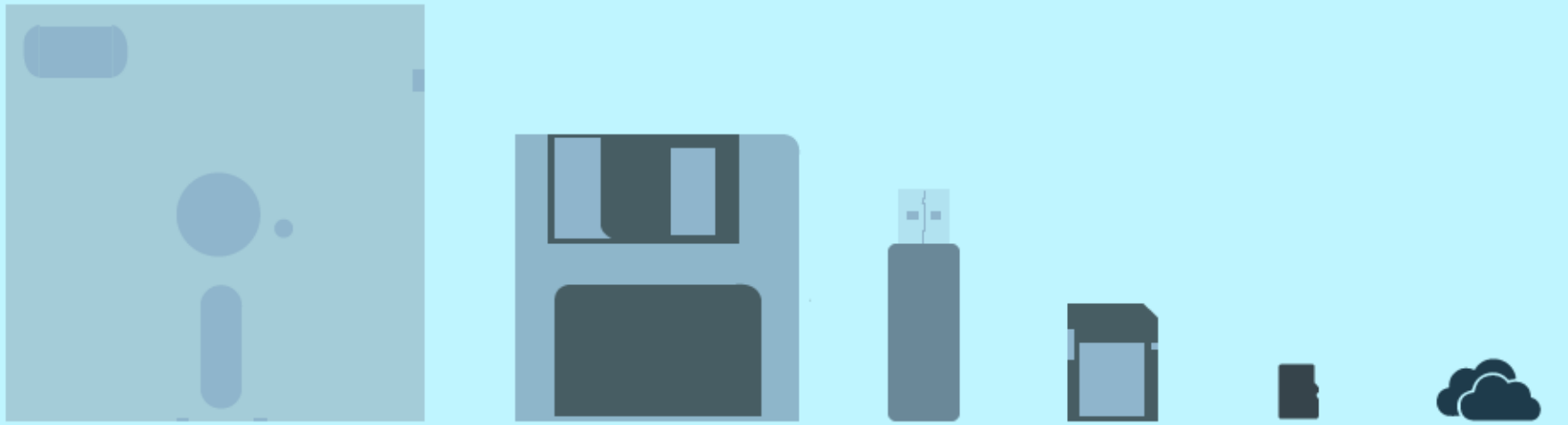
a **data pipeline** matches its **use case**



question:

how is data stored?

The Evolution of Data Storage



	8 Inch Floppy DIsk	5 1/2 Inch Floppy DIsk	USB Drive	SD	Micro SD	Cloud
Commercially available	1971	1976	2000	2010	2010	2009
Maximum capacity	1.2 mb	1.2 mb	2 tb	256 gb	128 gb	~5zb
Cost per gb	£1000	£800	£0.5	£0.8	£0.5	<50gb free

@vowlesyy @pixelpelican pixelpelican.co.uk samuelvowles.co.uk
sources: Wikipedia, Amazon, Ebay

Ways of storing data

- **Object storage**
  - is a way of structuring stored data so that it’s characterized as objects that can be manipulated in different ways by hardware and network storage systems
  - the objects are not in a file-folder hierarchy
  - object stores are scalable, fast data retrieval and cost effective

- **Distributed file systems**
  - a file system with data stored on a server.
  - data is accessed and processed as if it was stored on the local client machine
  - convenient to share information and files among users on a network in a controlled and authorized way

Ways of storing data

- **Relational Databases**
  - uses a structure that allows us to identify and access data in relation to another piece of data in the database
  - data in a relational database is organized into tables

- **NoSQL Databases**
  - a non-relational way of storing data
  - mostly used to store documents, key-value pair data
  - storing a large volume of data, and you don’t want to lock yourself into a schema

Hands on workshop

Data Pipeline in Practice

Marco Morales
marco.morales@columbia.edu

Nana Yaw Essuman
ne2388@columbia.edu

GR5069: Applied Data Science
for Social Scientists

Spring 2025
Columbia University