

# Data Pipeline in Practice

Marco Morales

marco.morales@columbia.edu

Nana Yaw Essuman

nanayawce@gmail.com

GR5069

Topics in Applied Data Science  
for Social Scientists

Spring 2023  
Columbia University

# recap: workflow principles in a 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**

## c) **scalability**

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

# recap: structuring your workspace

some basic principles...

1. use **scripts for everything** you do
  - ▶ **NEVER** do things **manually**
2. organize your scripts in a sequence
  - ▶ **separate activities** in sections
  - ▶ keep an early section for **definitions**
  - ▶ call **other scripts** when necessary
3. write **efficient** (aka lazy) code
  - ▶ turn code used multiple times into **functions**
  - ▶ **re-use functions**: make them generic enough
4. rely on **version control** (git)

# recap: structuring your workspace

some basic principles...

```
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
```

## ***ProTips:***

- ▶ data is **NEVER** pushed to GitHub!!!!!!
- ▶ {secret keys} are **NEVER** pushed to GitHub!!!!!!
- ▶ reports could live in GitHub (depends)
- ▶ references are transferred to GitHub **wiki**
- ▶ TODO is transferred to GitHub **projects**

# data collection

# why is data collection important?

- ▶ understand your products and systems better
- ▶ provides means for organizations to make better data-informed decisions
- ▶ helps identify opportunities or gaps in a product or system
- ▶ measures how your consumers interact with your products or system
- ▶ understanding your potential market

**In God we  
trust, all  
others bring  
data.**

–William E. Deming



# types of data

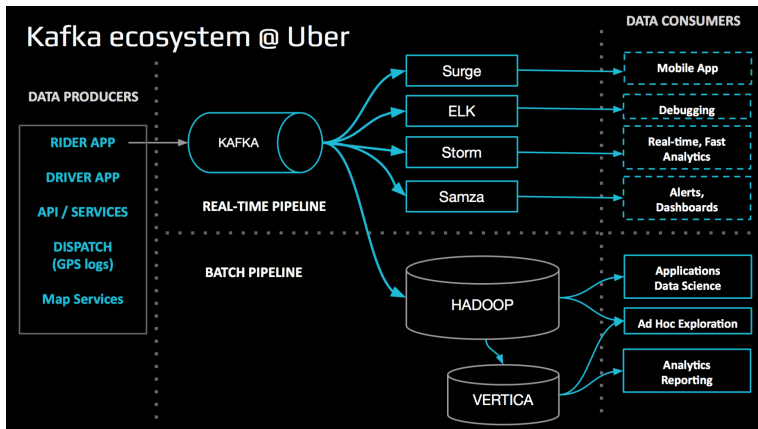
- ▶ unstructured data
  - ▶ does not have a predefined data model or is not organized in a pre-defined manner
  - ▶ examples of unstructured data include audio, video files or No-SQL databases.
- ▶ structured data
  - ▶ pre-defined data model and ready to analyze
  - ▶ examples of structured data are Excel files or SQL databases
  - ▶ most **traditional** form of data storage



# levels of datasets

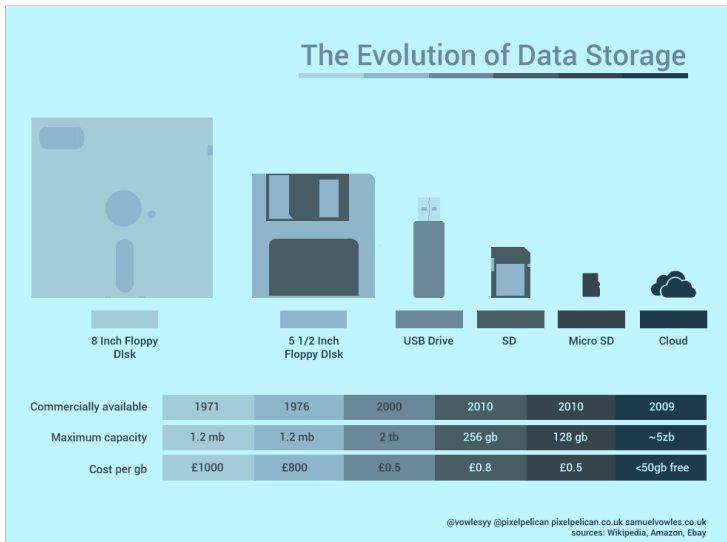
- ▶ first party datasets
  - ▶ data generated by your own product or systems
  - ▶ the most **useful** and **valuable** data you can collect about your consumers
- ▶ second party datasets
  - ▶ someone else's first-party data but useful to your organization
  - ▶ arrangement with trusted partners who 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)

# Data Ingestion Pipeline



# data storage

# evolution of data storage



# 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 system
  - ▶ a file system with data stored on a server. The 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
- ▶ 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

# ways of storing data - cont'd

- ▶ 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

nanayawce@gmail.com

GR5069

Topics in Applied Data Science  
for Social Scientists

Spring 2023

Columbia University