Aalto University
School of Science

CS-E4640 Big Data Plaforms

How to succeed in the course

Zixuan Liu & Guangkai Jiang
Teaching Assistant
Aalto University

Assignment Structure

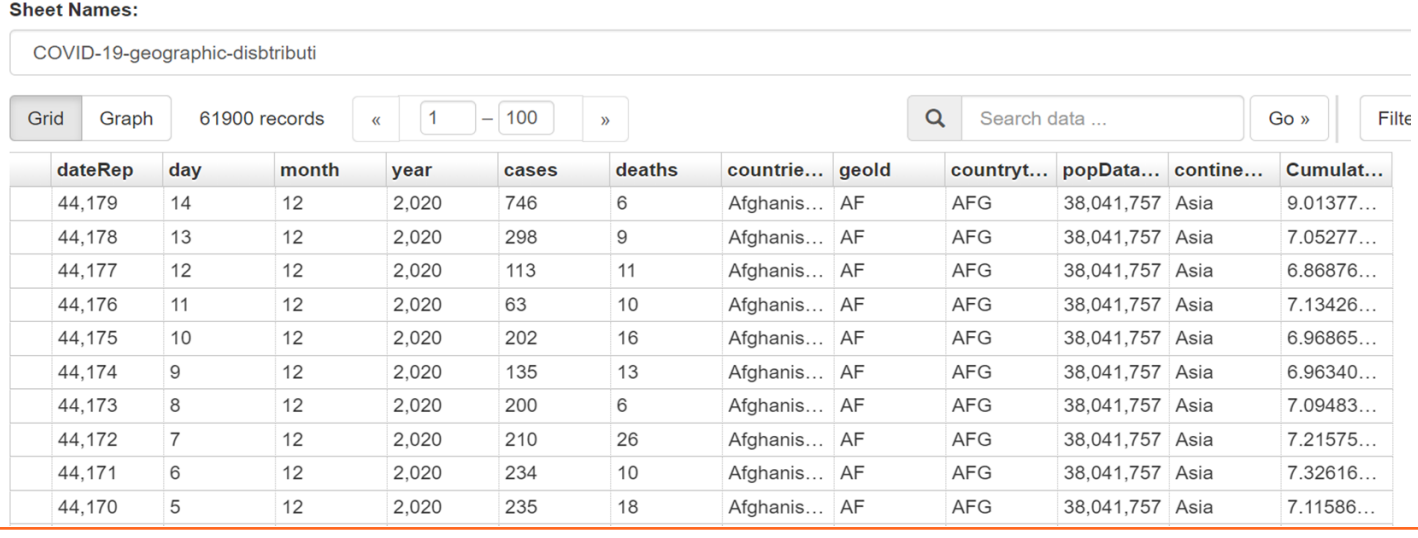
- **Introduction – Goal**
- **For example: (The first assignment)**
  - Design a simple big data platform. The big data platform to be designed will have a set of minimum features built from some key components. You will play two main roles in this assignment:
    - *tenant developer/users: designs tenant data structures and performs simple data ingestion/tests.*
    - *platform designer/provider: provides key big data services for tenants*

Assignment Structure

- **Look at the provided dataset, think of the associate scenario, and choose what you want to work with based on your interest https://version.aalto.fi/gitlab/bigdataplatforms/cs-e4640/-/tree/master/data**

Assignment Structure

- **For example, COVID-19 dataset from EU https://data.europa.eu/data/datasets/covid-19-coronavirus-data?locale=en**



Sheet Names:

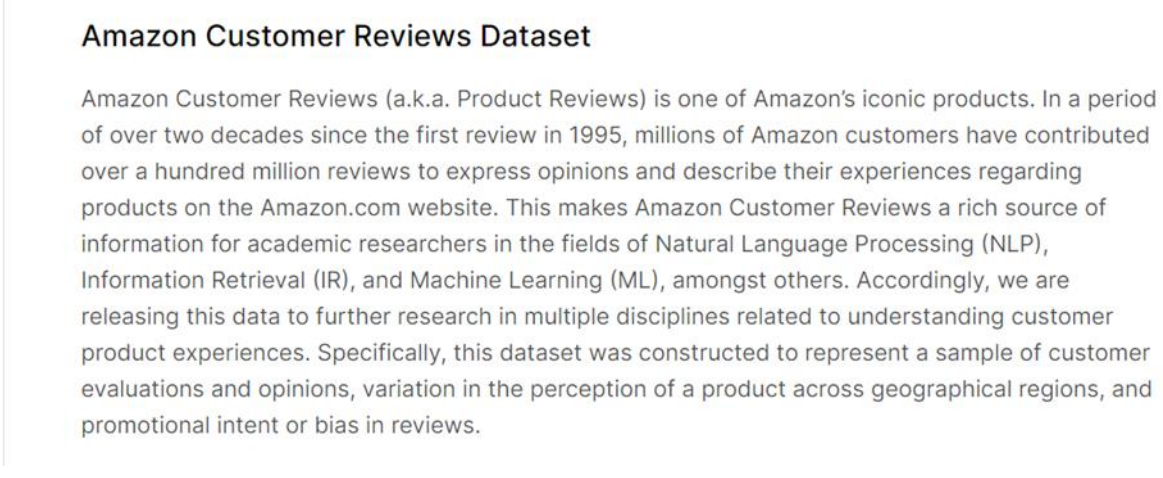
COVID-19-geographic-disbtributi

Grid | Graph | 61900 records | « | 1 – 100 | » | Search data ... | Go » | Filte

dateRep	day	month	year	cases	deaths	countrie...	geold	countryt...	popData...	contine...	Cumulat...
44,179	14	12	2,020	746	6	Afghanis...	AF	AFG	38,041,757	Asia	9.01377...
44,178	13	12	2,020	298	9	Afghanis...	AF	AFG	38,041,757	Asia	7.05277...
44,177	12	12	2,020	113	11	Afghanis...	AF	AFG	38,041,757	Asia	6.86876...
44,176	11	12	2,020	63	10	Afghanis...	AF	AFG	38,041,757	Asia	7.13426...
44,175	10	12	2,020	202	16	Afghanis...	AF	AFG	38,041,757	Asia	6.96865...
44,174	9	12	2,020	135	13	Afghanis...	AF	AFG	38,041,757	Asia	6.96340...
44,173	8	12	2,020	200	6	Afghanis...	AF	AFG	38,041,757	Asia	7.09483...
44,172	7	12	2,020	210	26	Afghanis...	AF	AFG	38,041,757	Asia	7.21575...
44,171	6	12	2,020	234	10	Afghanis...	AF	AFG	38,041,757	Asia	7.32616...
44,170	5	12	2,020	235	18	Afghanis...	AF	AFG	38,041,757	Asia	7.11586...

Assignment Structure

- **Amazon US Customer Reviews Dataset https://www.kaggle.com/datasets/cynthiarempel/amazon-us-customer-reviews-dataset**



Amazon Customer Reviews Dataset

Amazon Customer Reviews (a.k.a. Product Reviews) is one of Amazon's iconic products. In a period of over two decades since the first review in 1995, millions of Amazon customers have contributed over a hundred million reviews to express opinions and describe their experiences regarding products on the Amazon.com website. This makes Amazon Customer Reviews a rich source of information for academic researchers in the fields of Natural Language Processing (NLP), Information Retrieval (IR), and Machine Learning (ML), amongst others. Accordingly, we are releasing this data to further research in multiple disciplines related to understanding customer product experiences. Specifically, this dataset was constructed to represent a sample of customer evaluations and opinions, variation in the perception of a product across geographical regions, and promotional intent or bias in reviews.

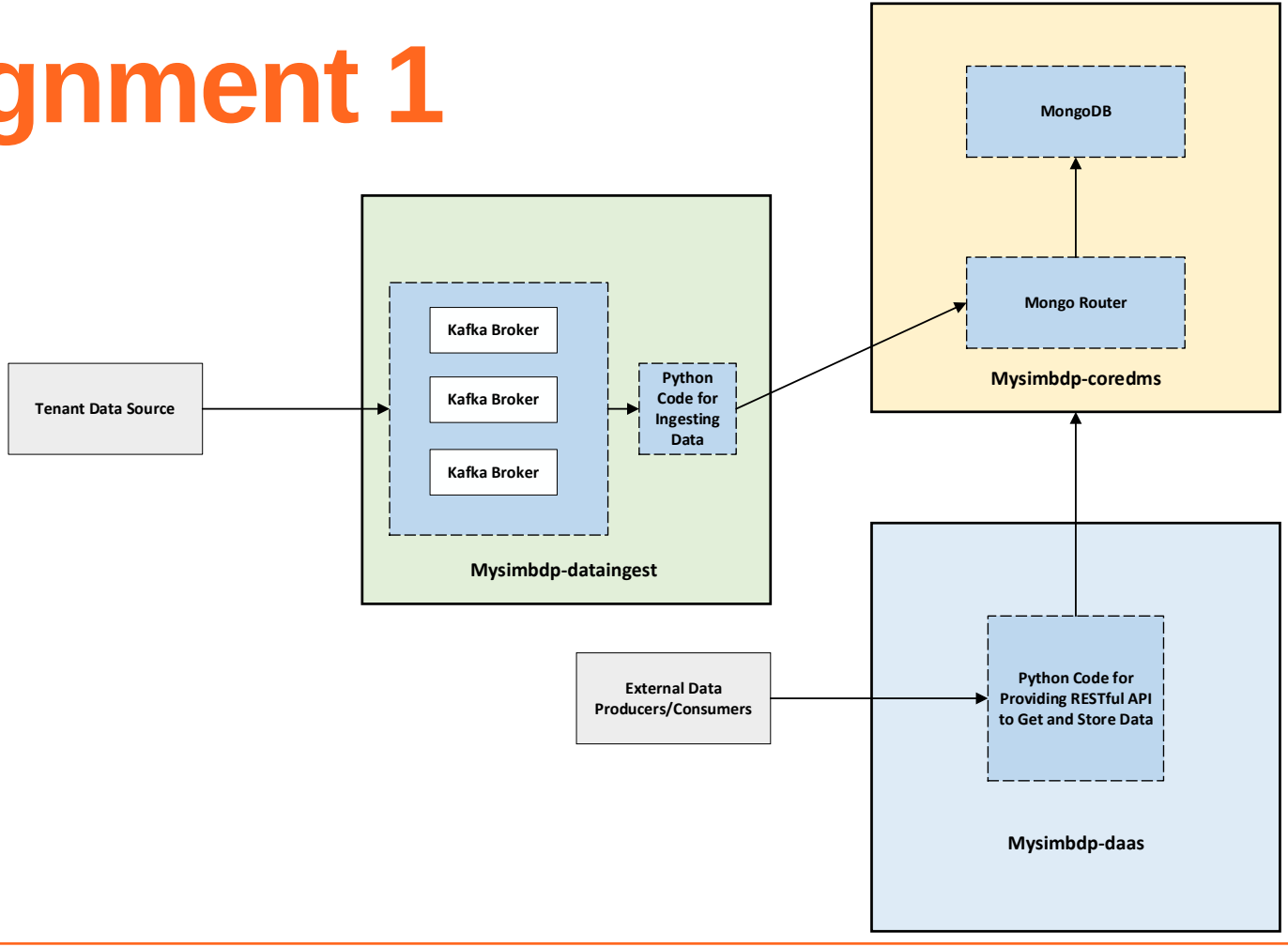
Assignment Structure

- **Constraints and inputs for the assignment**
  - key components: for example, **data ingestion components**, to read data from data sources
  - assumptions: Multi-tenant, use the same datasets/models
  - frameworks/technologies can be used in this assignment (message broker, databases, datasets, **programming languages**, stream processing)

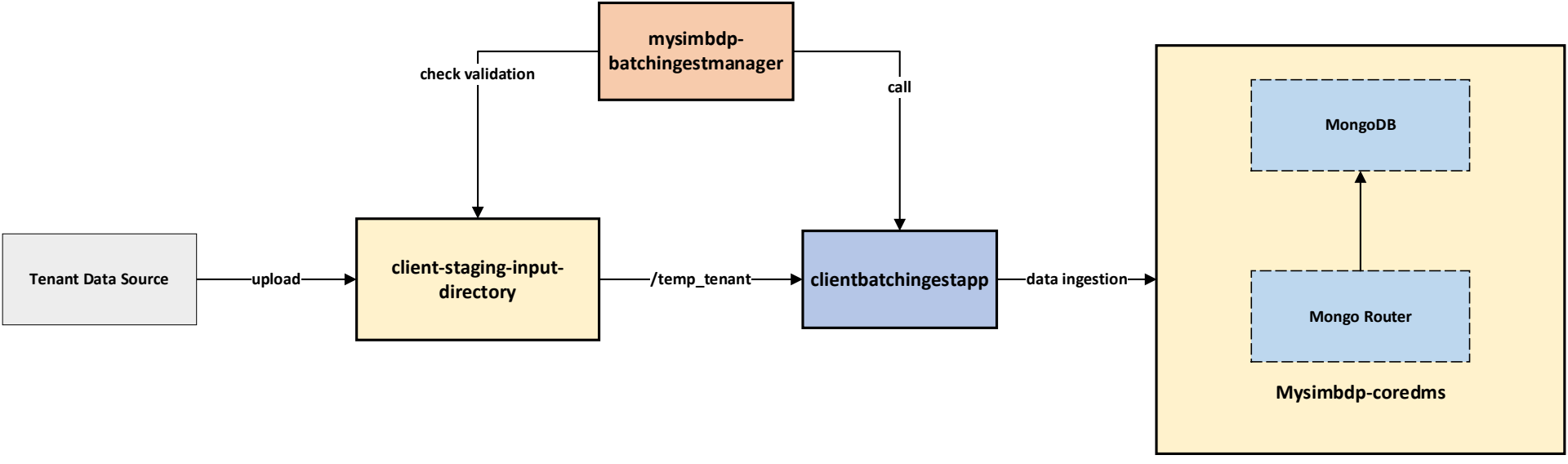
Some reasons of my choice

- **Language**
  - Python
- **Databases**
  - MongoDB: Easy operation in python. Good documentation. Support Semi-Structured data storage
- **Message Broker**
  - Kafka-python: Good documentation
- **Stream Processing**
  - Apache Spark: Support Python well

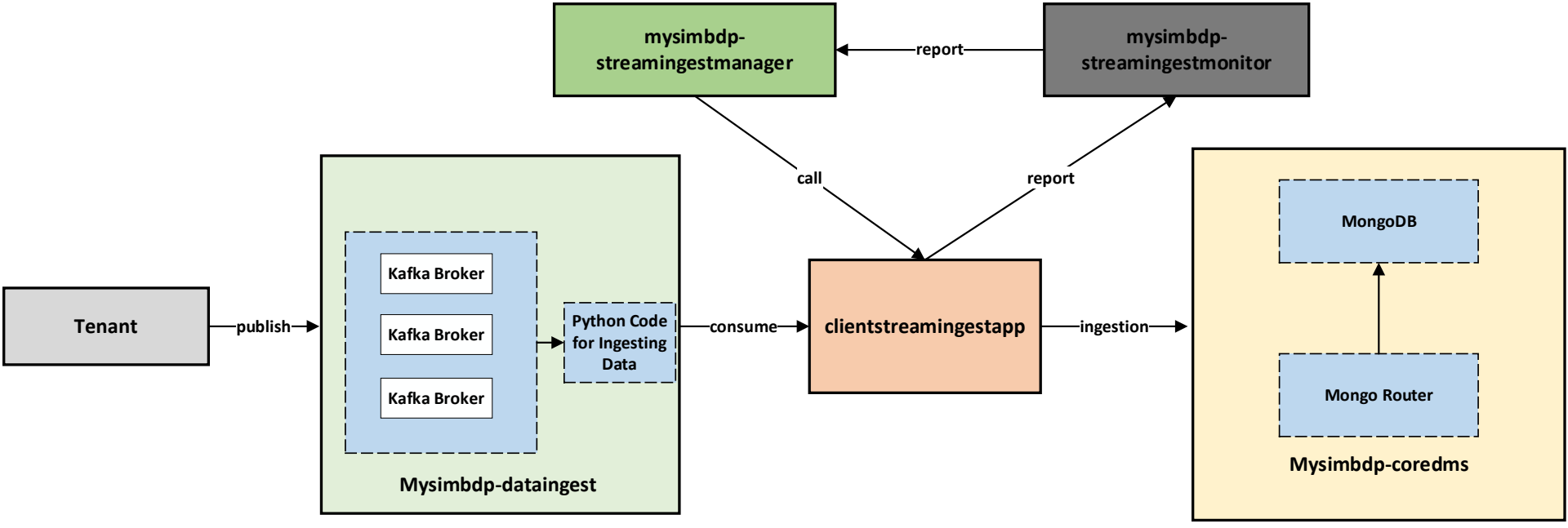
Assignment 1



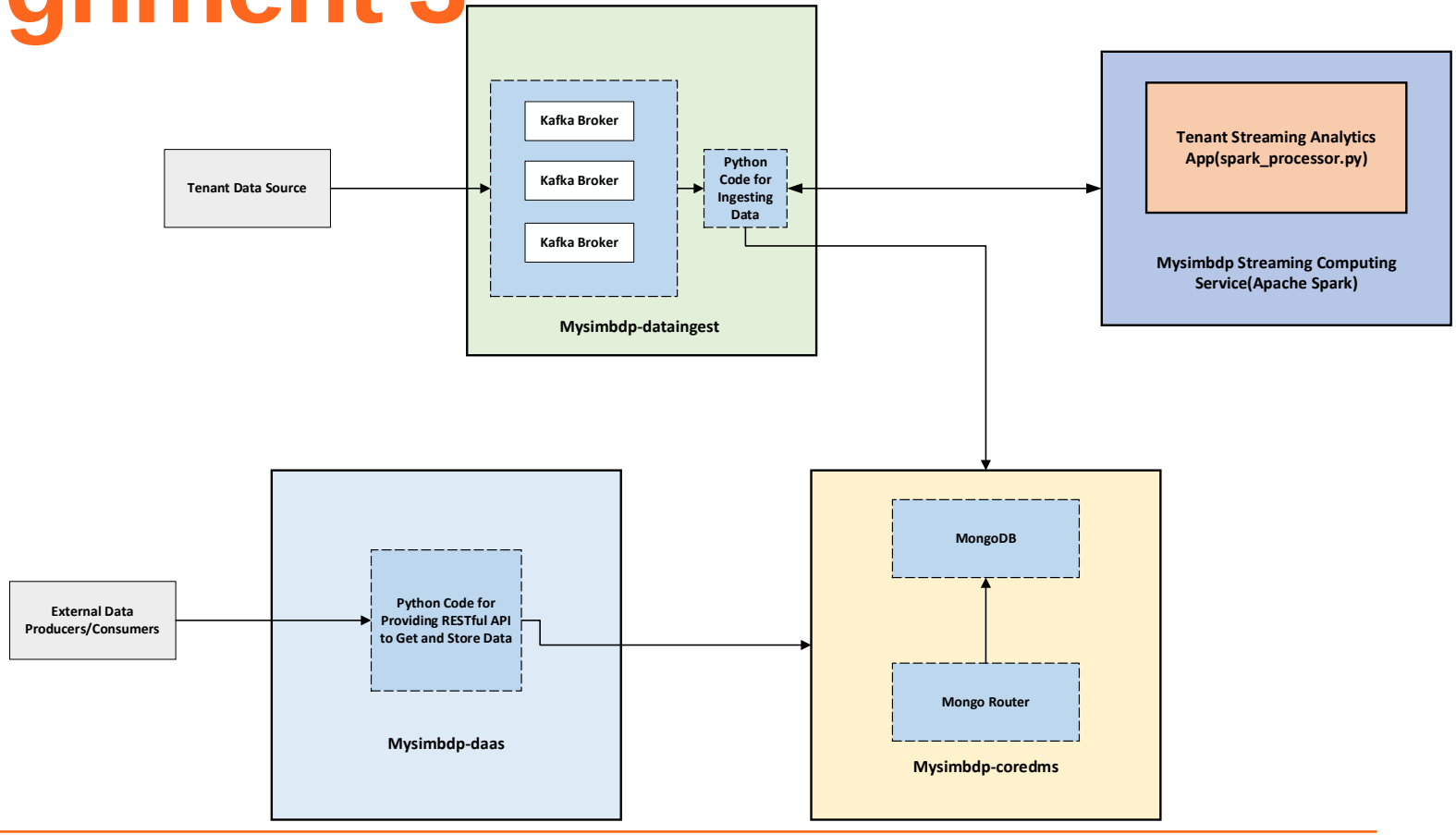
Assignment 2



Assignment 2



Assignment 3



Assignment Structure

- **Requirement and Delivery**
  - **Design**: Think about why you're using a certain technology before using it in your design.
  - **Implementation**: We don't need to build a complete product, but a proof of concept. We are not looking for perfect performance
  - **Extension**: Add additional features (logging), protect data, external service integration

Report nicely

- **Installation and deployment (operating system, dependencies)**

Deploy mysimbdp-daas

1. Open a terminal and `cmd` to `assignment-1-1011315-master/code/mysimbdp-daas` folder.
2. Build from dockerfile: `docker build . --tag mysimbdp-daas`
3. Run our container: `docker run --net=host mysimbdp-daas`

- **Use graph to show your design and evaluation**
- **Honesty => if you reuse exist component/code, specify clearly**

Report nicely

- **Not too long but explains your design choices, dataflow, and results. One good review is to see if others can understand your code just by reading your report**
- **Report problem even if you don't know why (e.g. performance went down dramatically when clients number increase)**

Submission

- **Pack your assignment into one zip file**

- **Do not put external libraries in your repository! Should automatically download somehow. (e.g. pip3 install -r requirements.txt)**

Submission

- **Packaging using docker is not mandatory, but a good practice (docker build .)**

```
FROM python:3-alpine as BASE

RUN apk update && apk add py3-pip && apk add bind-tools

COPY requirements.txt requirements.txt

RUN pip3 install -r requirements.txt

WORKDIR /daas

COPY MysimbdpDaas.py /daas/MysimbdpDaas.py

CMD ["python3","MysimbdpDaas.py"]
```

Don’t panic

- **It’s not as scary**
  - Tough but doable
- **You can succeed**
  - Time and effort
- **Reassurance**
  - I am here, aren’t I
- **Focus on learning**
  - The course is rewarding, don’t fixate on grade but actual learning

Start Early Enough

- **Project duration about 2 weeks (Design + Implementation + Reporting)**
  - **Design** - Takes time to understand the requirement and constraints, and research in different technologies and what to choose
  - **Implementation** – Takes time to read documentation and understand how certain libraries works (Main architecture, Testing, Deploying)
  - **Reporting** – Make notes for the report along your design and implementation! You also need to answer some theoretical/extensibility questions.

Don’t procrastinate

- **Start working early**
  - A week – danger zone
  - 3 days – unlikely
  - Over night - impossible
- **Buffer time**
  - Unexpected issue
  - Troubleshoot or workaround
  - Startover

How to success

- **Understand big data**
- **Clear design**
  - Real thing, but demo, not production
- **Implementation choice**
  - Language, Real tech stack, Data domain
- **Work environment**
  - Kubernetes, VM, docker, not bare-metal
- **Learn but not copy**

Don’t reinvent wheels

- **Use available open source library**
- **Code in a modular way from the beginning since you’ll most probably need to reuse for the later assignments**

Be honest

- **With yourself**
  - Do I need help?
- **With TAs**
  - Do I understand this?
- **With design**
  - Know is know, don’t know is don’t know
  - Not perfect but working
  - Your own work

Career opportunity

- **Applications**
  - ML/AI, Data mining, scalable apps
- **Job opportunity**
  - ML engineer, Data analysist, Cloud DevOps

Ask for help

- **Do communicate with TA/prof etc. for unclear points**
- **Contacts:**
  - zixuan.liu@aalto.fi
  - guangkai.jiang@aalto.fi
- **Help answer other people's question**

Thank you!

Any Question?