PROBLEM STATEMENT:

A senior researcher approaches you about the FIT study. For this project, participants have been recruited to wear wrist-worn fitness trackers. The fitness trackers record heart rate and steps, among other metrics. The fitness trackers are a commercial product, and the data is stored on the supplier's own cloud systems, for which they offer an API. The data is obtained via a third-party integration that makes use of the API and has been specially developed for Tilburg University. The extraction pipeline tool runs on Tilburg's own servers and is hosted in Kubernetes and Docker, based on minIO. The data is extracted via a web interface, where users manually enter the time range and study participants they want to download data for. Because this requires quite a bit of manual work, it is typically done by students.

The researcher points out that she has very limited technical understanding of how the data collection, the hosting of the pipeline, and the downloads actually work. So, when she found out that the downloaded data is in a non-human-readable format, she did not know what to do with it and decided to contact you.

QUESTIONS:

Q1. What is your approach in the first meeting with the researcher? What would you prepare in advance and what questions would you ask her? What solution(s) would you pitch to her and how would you explain them?

> **Preparation:** For preparation, I would start by revisiting the specifics of hosting the pipeline myself. These concepts must be fresh in my knowledge for me to be useful to the researcher. If I am not familiar with the researcher and their past work, I would also make sure to look them up on Google Scholar to get a better idea of their domain, and the methods they commonly use in their research.
>
> **Questions:** In terms of questions, I would first like to get an overview of what the research goal is for the study. Specifically the variables of interest in the collected data. This way, while preparing the data I understand what information is necessary, and what information can be considered redundant (redundancy will be confirmed with the researcher before I take any actions). Additionally, I would like to get a better idea of the skillset of the researcher- whether or not they have any programming experience, to tailor my technical explanations to the required level.
>
> **Solutions:** Parallel to my preparation, I would prepare a brief and high-level summary of the data hosting pipeline (see Figure 1). It is not yet extremely important to present fine details of the process such as how the API is used, and the hosting details of Kubernetes/Docker. What's more important at the moment, is a general idea of how the cloud computing technology is being used, and the usage of the web interface.
>
> A critical aspect that I would highlight to the researcher would be the frequency of data extraction from the supplier's cloud to Tilburg's cloud storage. A quick overview of the data structure of the 'FIT_study.zip'- (see Figure 2) shows that there is one CSV file for every data point (which corresponds to one date). I would explain the implications of this first, i.e., this makes navigating through the folders and accessing data tedious even with a programming language, and virtually impossible if one wants to do it manually.

Next, I would recommend looking into this, and checking if it is possible to have one CSV per month instead of one day. The data corresponding to the date could just be added as another column in the CSV file. If the researcher explains that they are unsure how this can be done, I would offer to look into this problem by scheduling another session with them and reviewing it in more detail. Although this may be a minute and rather technical aspect, I believe it could be used as a starting point for the researcher to get more familiar with the data extraction/hosting pipeline.

Lastly, depending on their experience with a programming language (R/ MATLAB/ Python), I would recommend learning the basics usage of some important libraries such as **numpy, pandas and matplotlib**. These libraries can be immensely important when it comes to data handling. As someone in the Research Support Team, my job should not just be providing direct solutions to the researcher's problems, but also trying to subtly introduce them to important technical skills that make them more independent and capable. Learning the usage of specific libraries can be a very good starting point and may make them more capable of tackling issues by themselves.

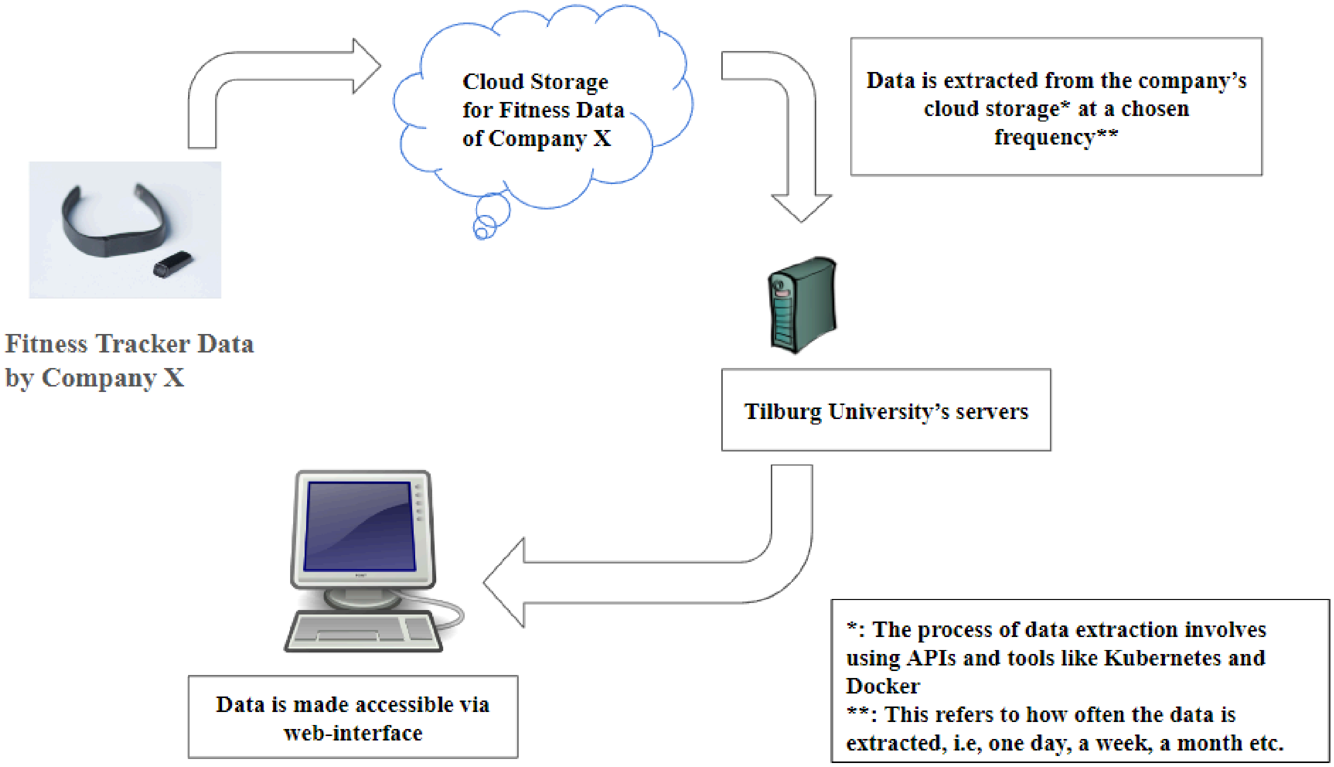
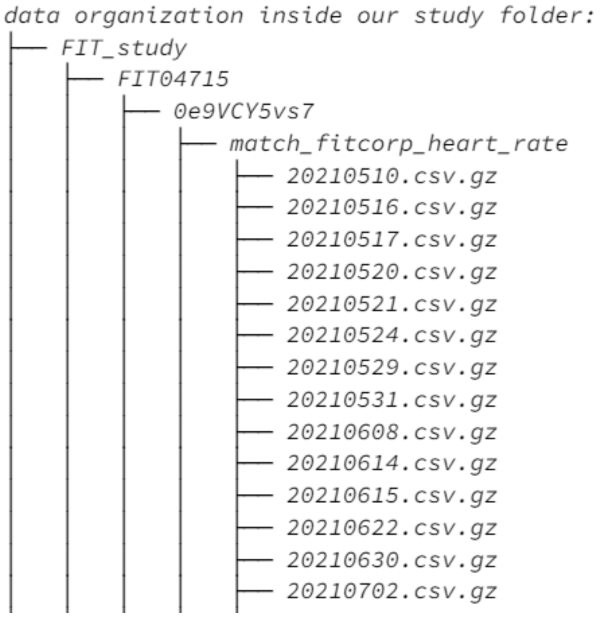

Cloud Storage for Fitness Data of Company X
Data is extracted from the company's cloud storage* at a chosen frequency**
Fitness Tracker Data by Company X
Tilburg University's servers
Data is made accessible via web-interface
*: The process of data extraction involves using APIs and tools like Kubernetes and Docker
**: This refers to how often the data is extracted, i.e, one day, a week, a month etc.

Figure 1: A high-level overview of the data hosting pipeline.



```
data organization inside our study folder:
├── FIT_study
│   ├── FIT04715
│   │   ├── 0e9VCY5vs7
│   │   │   ├── match_fitcorp_heart_rate
│   │   │   │   ├── 20210510.csv.gz
│   │   │   │   ├── 20210516.csv.gz
│   │   │   │   ├── 20210517.csv.gz
│   │   │   │   ├── 20210520.csv.gz
│   │   │   │   ├── 20210521.csv.gz
│   │   │   │   ├── 20210524.csv.gz
│   │   │   │   ├── 20210529.csv.gz
│   │   │   │   ├── 20210531.csv.gz
│   │   │   │   ├── 20210608.csv.gz
│   │   │   │   ├── 20210614.csv.gz
│   │   │   │   ├── 20210615.csv.gz
│   │   │   │   ├── 20210622.csv.gz
│   │   │   │   ├── 20210630.csv.gz
│   │   │   │   ├── 20210702.csv.gz
│   │   │   │   │
```

Figure 2: A glimpse into the data organization inside the FIT_study folder.

Q2: Some time later, a follow-up study uses the same method for data collection but involves more participants and a higher sampling frequency of the fitness data. Researchers run into the problem that the cleaning takes much too long. What steps would you explore and suggest to resolve this? What could it depend on, and what would be a viable solution? Explore a couple of alternatives if you can think of any.

If the cleaning is taking much longer and the data volume has also increased significantly, three aspects come to mind: the methods used for cleaning the data, batch processing (using smaller chunks or volumes of data to be processed at once) and computing resources (using the high performance computing (HPC) cluster).

My first go-to would be to examine scripts and methods they use to clean data. When it comes to processing large volumes of data, smaller changes to the methods used may have a significant impact (using vectorization instead of looping through the rows of a dataset would be more efficient, for instance).

Next, I would explore ways of batch processing with the HPC. Given the information at hand, I would try to schedule jobs to the HPC such that periodically a fixed volume of data is sent to the cluster, avoiding any manual work and speeding up the process.

Overall, involving the HPC could play a monumental role in approaching this issue– with larger volumes of data and optimized methods to clean them, parallel processing resources offered could significantly faster data cleaning and preparation.