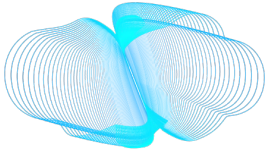
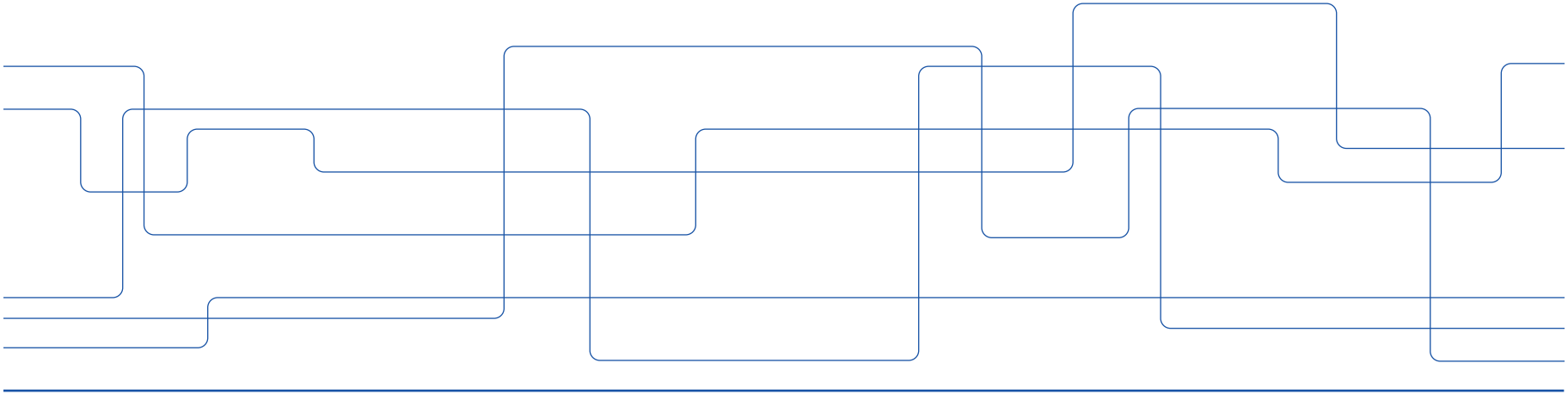
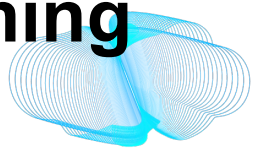


DataCloud project

Tutorial for students of ID2209

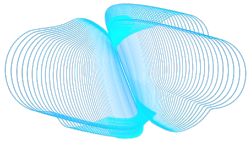


The data analytics process is becoming complex due to



- the characteristics of Big Data,
- the sophisticated tools and technologies involved,
- different interests among stakeholders,
- often changing business needs,
- the lack of a standardized process for the lifecycle

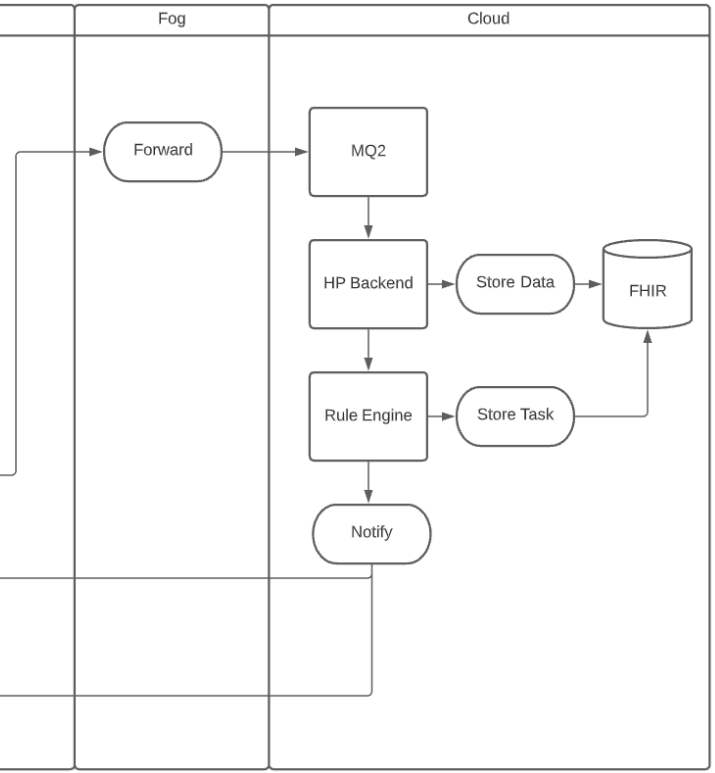
Big Data piipelines



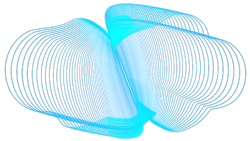
Because of the complexity of Big Data analysis tasks, the software supporting such analysis requires a combination of a broad spectrum of trusted software components.

- Such a combination involves integrating components into pipelines that take care of the pipeline execution and data transfer.
- The design and usage of Big Data pipelines increase the efficiency of data analysis, while at the same time require support for designing and managing the pipelines.
- However there are still critical challenges in their implementation, such as the heterogeneity of involved stakeholders and limited knowledge reuse.

Data Processing in a eHealth application



Big Data pipelines

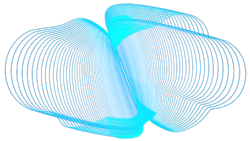


- The entire pipeline lifecycle should be supported by advance tools.

- Most of recent tools have focus mainly on runtime execution of pipelines rather than on pipeline definition.

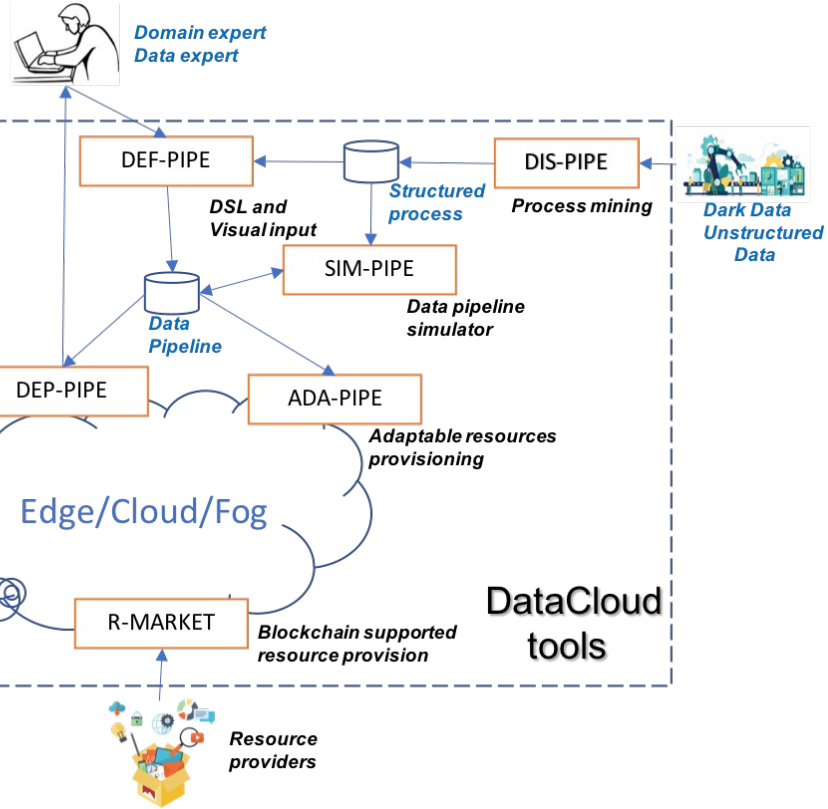
- The tools should allow their usage by domain experts

How to create pipelines?

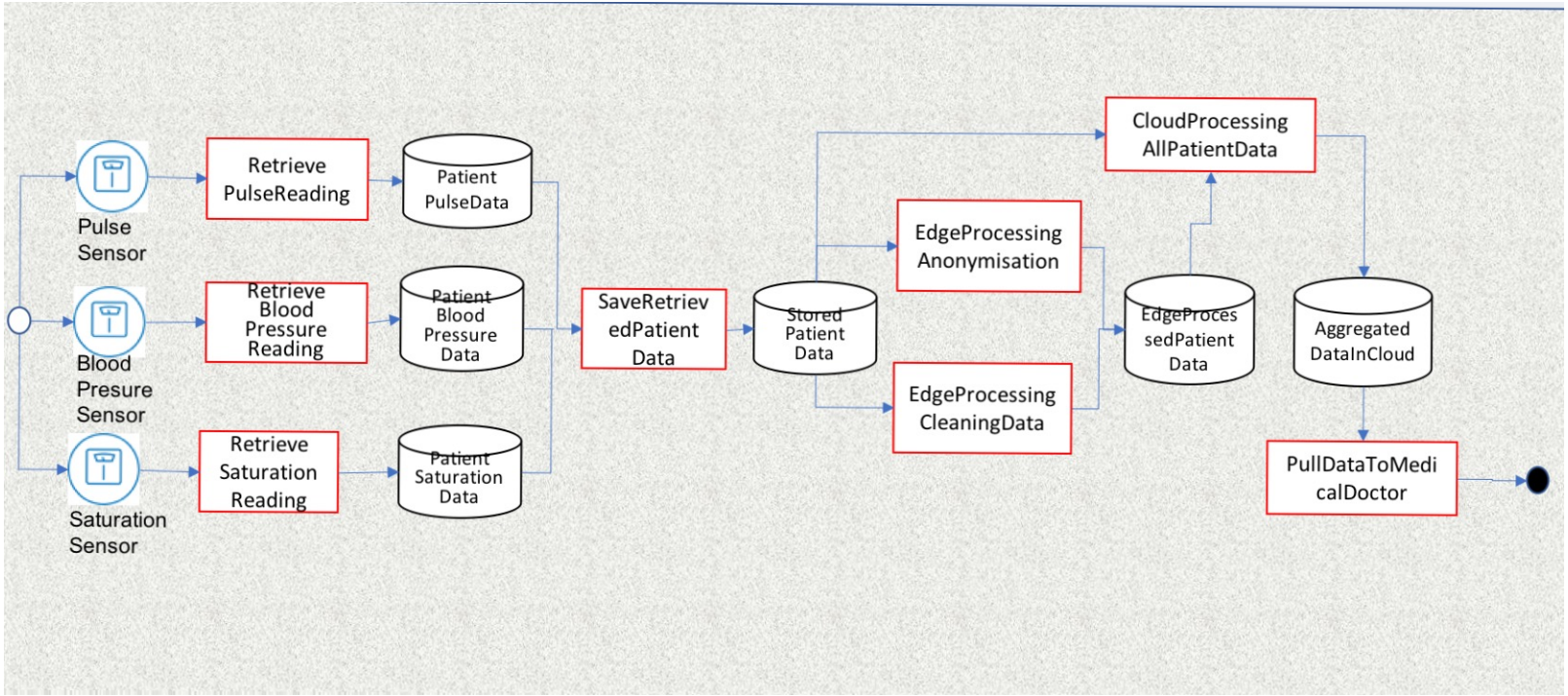


- You can hardcode the pipeline
- You can use some automated tools (Argo, Airflow…)
- You can use fully automated tool-box supporting easy definition of pipelines

The DataCloud project



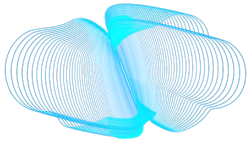
Example pipeline



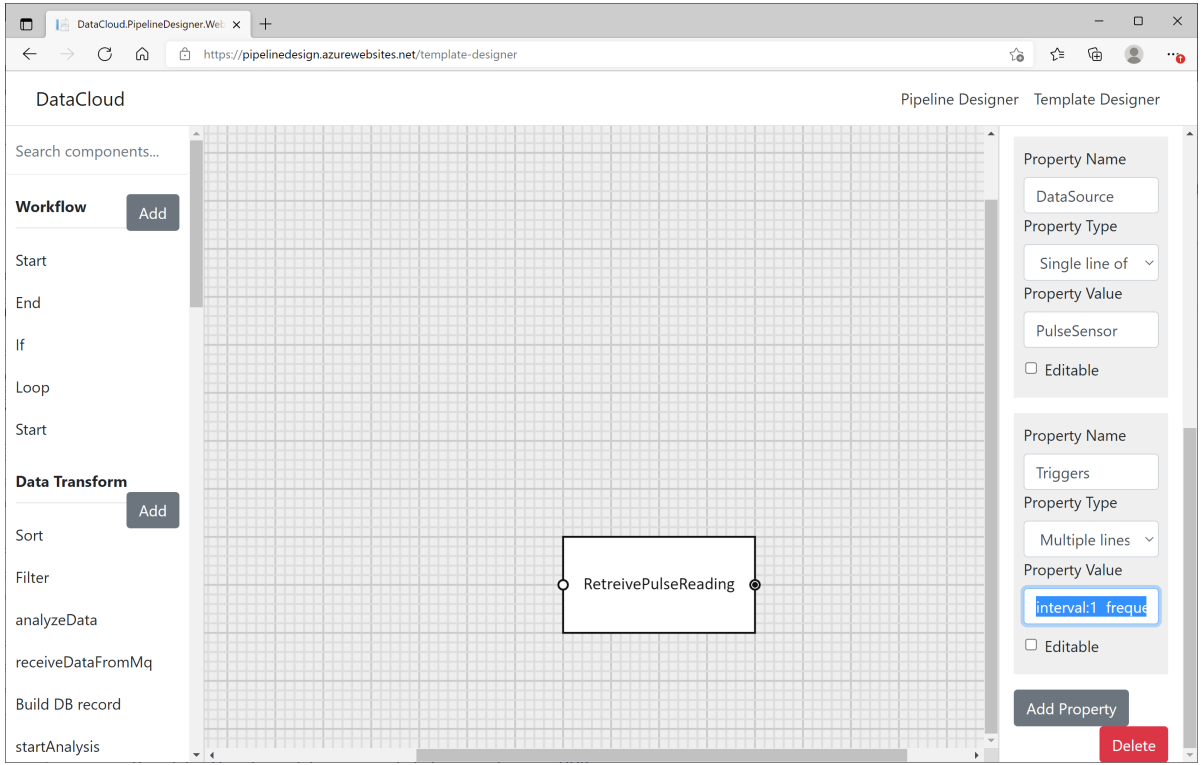
Model serialization (grammar)

```
Pipeline DigitalHealthUsecase {
   communicationMedium: MessageQueue,
   environmentParameters: {
      "MQ_HOST": kubeMQ,
      "MQ_USER": user1
   },
   steps:
      data-source step: RetreivePulseReading
      implementation: docker-implementation
      image: 'tellucare-api:latest'
      dataSource: PulseSensor,
      triggers: interval-schedule interval: 1
      frequency: MINUTE
      startTime: '30.10.2021 12:00:00'

      data-source step: RetrieveBloodPressure
      implementation: docker-implementation
      image: 'registry.sintef.cloud/tellucare-edge'
      dataSource: BloodPressureSensor,
      triggers: interval-schedule interval: 1
      frequency: MINUTE
      startTime: '30.10.2021 12:00:00'
...

      data-sink step: SaveRetrievedPatientData
      implementation: docker-implementation
      image: 'tellucare-application:latest'
      environmentParameters: {
            RABBITMQ_HOST=oslo.sintef.no:5672
            RABBITMQ_USERNAME=tellucareapi
            RABBITMQ_PASSWORD=???

            }
```

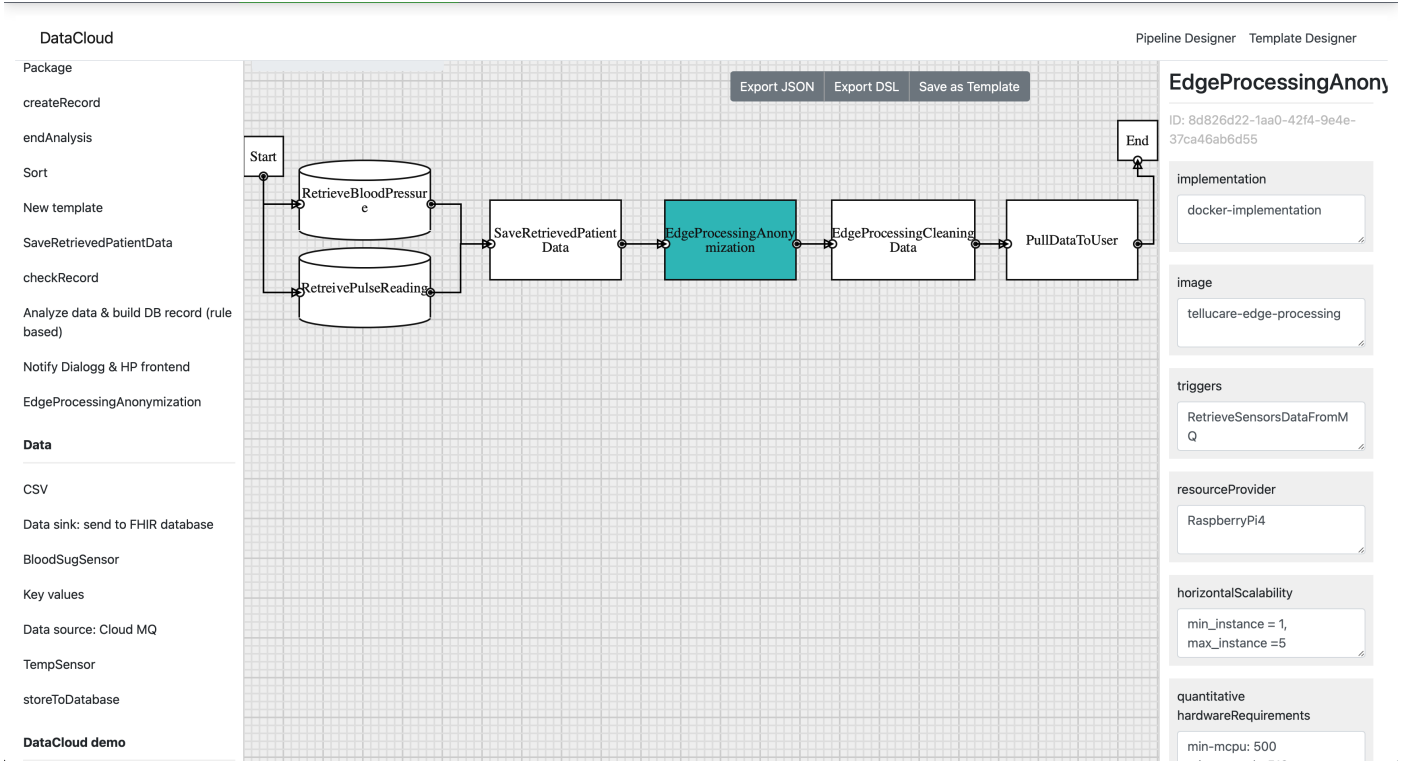
```
dataSource: StoredPatientData,
preCondition:
      [{condition: CheckInputINotEmpty,
       inputStep: retrieveDataOne},
       {condition: CheckInputIINotEmpty,
       inputStep: retrieveDataTwo},
       {condition: CheckInputIIINotEmpty,
       inputStep: retrieveDataThree},],
      operator: OR
...
data-processing step:EdgeProcessingAnonymization
implementation: docker-implementation
image: 'tellucare-edge-processing'
executionRequirements:
      horizontalScalability {
      min-instance : 1
      max-instance: 5
      },
      quantitative hardwareRequirements {
            min-mcpu: 500
            min-ram-mb: 512
            min-storage-mb: 1024
      }
triggers: RetrieveSensorsDataFromMQ
resourceProvider: RaspberryPi4

data-processing step: EdgeProcessingCleaningData
implementation: docker-implementation
image: 'tellucare-edge-cleaning'
triggers: RetrieveSensorsDataFromMQ
resourceProvider: RaspberryPi4

data-processing step: PullDataToUser
implementation: docker-implementation
image: 'tellucare-user'
resourceProvider: AWS
```



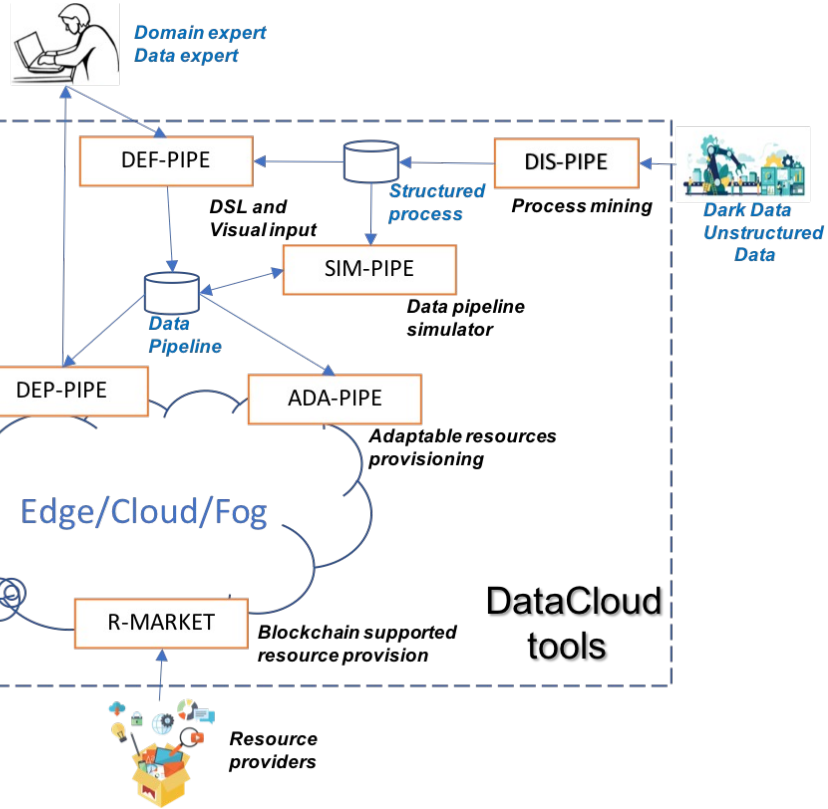
Visual tool (step description)



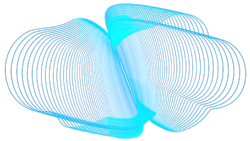
Visual tool (pipeline description)



The DataCloud project

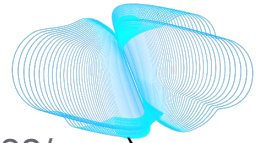


The Goal of Evaluation



- We would like validate integration of DEF-PIPE and SIM-PIPE tools of DataCloud tool-box.
- Validation must be done by external users (students of the ID2209 course)

Your tasks



1. You must login to the DEF-PIPE tool (http://crowdserv.sys.kth.se:8082/repo)

   -- Username: testuser and Password: 0AsK31lQaYd

2. Your must select a pipeline DEF-SIM-PIPE in the list of pipelines of the DEF-PIPE tool (see category SE Course).

   --It appears as a box in canvas with … on the upper right corner. Click on … and open the pipeline

3. You must generate YAML text corresponding to the graphical presentation of the DEF-SIM-PIPE pipeline (see the previous item).

   -- You do that by clicking “Export YAML” button

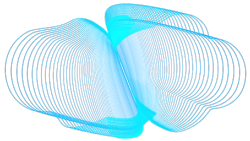
   -- The generated text will be downloaded into the browser

4. You must invoke API with parameters from http://crowdserv.sys.kth.se:8082/docs

   -- find Get API: /api/repo/exportyaml/, click on it and on the “Try it out” button

   -- fill in user: testuser and pipeline: DEF-SIM-PIPE and execute it

Your tasks



5. You must make API Invocation and YAML Export from a program

-- Write a program in your preferred programming language to invoke a REST API and retrieve a YAML export from a remote server.

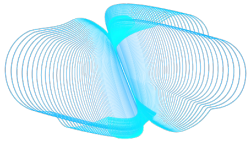
API Details:

HTTP Method: GET
API Endpoint:
http://crowdserv.sys.kth.se:8082/api/repo/exportyaml/{username}/{pipeline}
Username: testuser Pipeline: DEF-SIM-PIPE
Header:
accept: text/plain

-- Requirements:

- Use your preferred programming language to create a script or program. The script must take the username and name of the pipeline as inputs and make an HTTP GET request to the provided API endpoint. Finally, save the YAML export received from the API into file.

Your tasks



- You are not expected to execute the generated Yaml text
- You have to deliver:
  - Screenshot with DEF-SIM-PIPE pipeline graphical view
  - Generated Yaml text from the graphical view of the pipeline (2 bonus points)
  - Protocol of invocation and output of invocation of the API from http://crowdserv.sys.kth.se:8082/docs (2 bonus points)
  - Code for invocation of API from your program and invocation result (2 bonus points)
  - (optional but appreciated) Any comments you would like to send us
- In order to get bonus points you must upload your deliverable by December 4, 19:00.

THANK YOU!

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016835, the DataCloud Project.

https://datacloudproject.eu/