USER DOCUMENTATION

1. Introduction

The purpose of this tool is to manage pipeline functions.The emphasis is on providing a functional solution with a minimal working example (MVP) that prioritizes functionality over the User Interface/User Experience (UI/UX). As of now this project is in a raw state and should not be used in production.

2. Requirements

As the integration work for this project is extensive we need some prerequisites that should be set up before.

- Amazon S3 bucket, this is where the file storage is.
- Amazon EC2 instances frontend, server, Apache Airflow are set up separately in EC2 instances or however you like. Make sure to have a big enough instances for Apache Airflow (The minimum memory required we recommend Airflow to run with is 4GB)

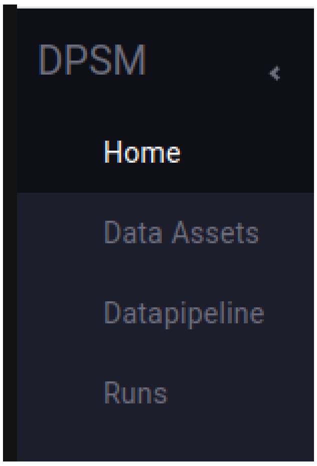
3. Building the project

The build description is given on the github page. And in the various project README-files /documentation.

- Frontend Documentation
- Backend Documentation
- Data pipeline Documentation

4. DPSM Usage Overview

Sidebar Component

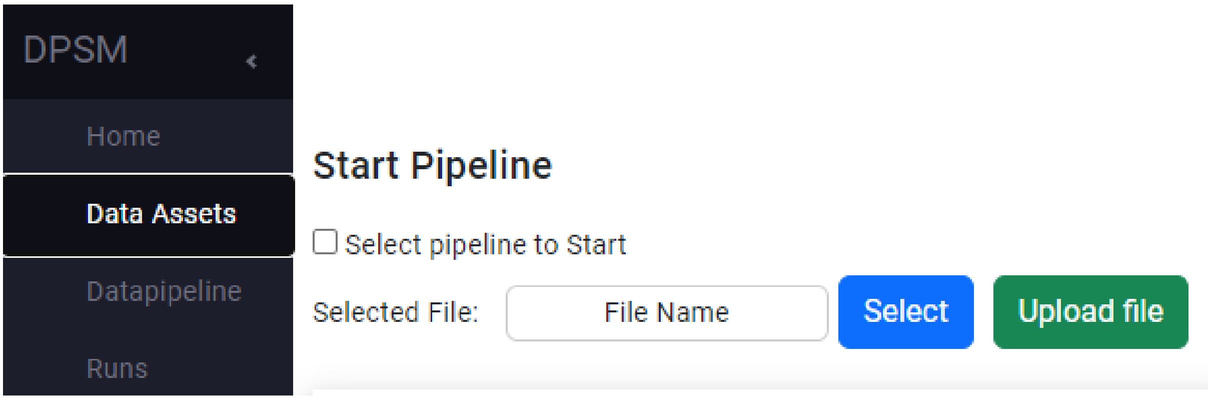


The sidebar is parted into 4 different selections

- **Home:**
  Here is just a small overview of the used technologies and a link to the github repository
- **Data Assets:**
  Data Assets is a management overview for the file assets, one can use for the data pipeline.
- **Data Pipeline:**
  This page is used to show an overview for the active and ready data pipelines.
  You can also start a data pipeline on this page.
- **Run:**
  Here one can find an overview of all the data pipeline runs / executions, that were triggered.

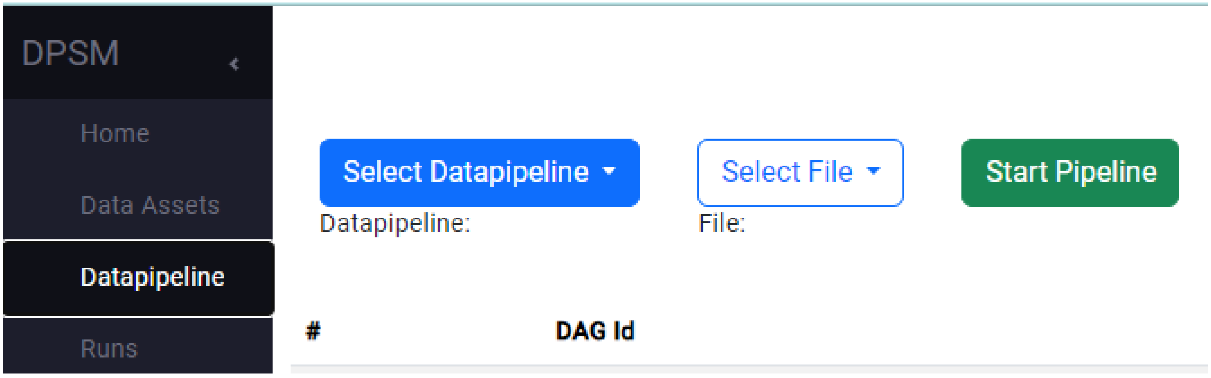
5. Workflow

a. Upload files:
To upload a file, press on the 'Select'-Button, you will see a file input window, where you can select the file. After that to upload it to the S3 bucket, press 'Upload file'.

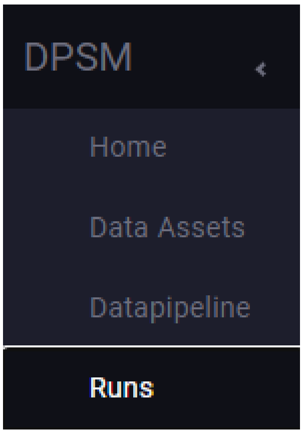


b. Adding data pipelines:
To add data pipelines, one has to put the *pipeline file in the dags directory* of the Apache Airflow Service, you can do this via ssh, Airflow will recognize it automatically after a while, when it's running.

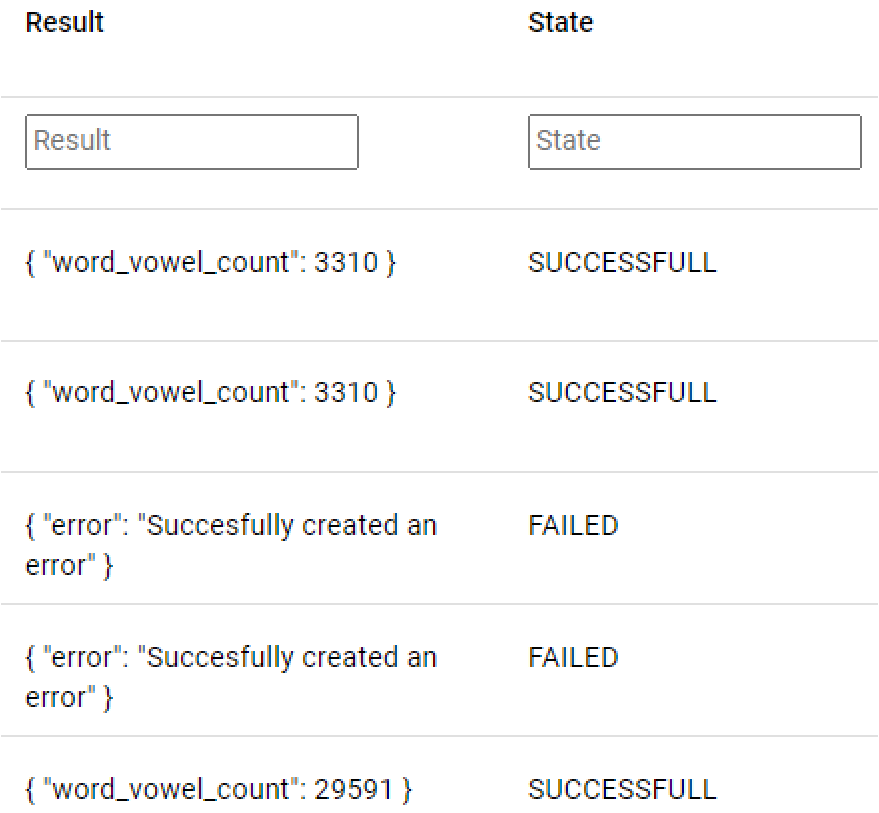
c. Start data pipeline:
To initiate a data pipeline, you can choose an existing or previously added pipeline from the "Select pipeline" dropdown, then select a file that has already been uploaded from the "Select File" dropdown, and finally click the "Start pipeline" button to activate the pipeline.



d. Check the result of the data pipeline:
After the data pipeline starts running, its progress will be synchronized to the "Runs" page.



The results of the data pipeline's execution will be displayed in the "Result" column, while the "State" column will simply state whether the data pipeline ran successfully or not.



Result	State
Result	State
{ "word_vowel_count": 3310 }	SUCCESSFULL
{ "word_vowel_count": 3310 }	SUCCESSFULL
{ "error": "Succesfully created an error" }	FAILED
{ "error": "Succesfully created an error" }	FAILED
{ "word_vowel_count": 29591 }	SUCCESSFULL