

Aggregat - 5 - Operational

- Procedures
 - DataPrep Flow
 - Start
 - Full process
 - Termination
 - Restart
 - Pause
 - Procedure
 - Stop
 - Procedure
 - Alert contacts
 - Reset
 - DataApp Flow
 - Start
 - Full process
- Scheduling
 - Trigger
 - Expected results
 - Intervention
- Monitoring
 - Runtime
 - Run history
 - Resources
 - Additional metrics
 - Logging
- Error handling
 - Alerts
 - Specificity

Procedures

Procedure guide on how to operate the application

DataPrep Flow

Start

Full process

How to start from scratch ?

- Go to the Gitlab Dataprep project in the section Scheduler : https://gitlab.solvay.com/solvay-it-dataops/data-ingestion/ses-agregat-dataprep/environments/dataprep_pipeline_test_env/-/pipeline_schedules
- Click on play for the line Agregat Daily Predict run and Agregat Weekly Retrain run .

The screenshot shows the GitLab Scheduling Pipelines page. The left sidebar contains a navigation menu with 'Schedules' highlighted. The main content area displays a table of scheduled pipelines. The table has the following columns: Description, Target, Last Pipeline, Next Run, and Owner. Two pipelines are listed:

Description	Target	Last Pipeline	Next Run	Owner
Agregat Weekly Retrain run	▼ master	🔗 #9110	in 2 days	👤 Brice Ruzand-ext
Agregat Daily Predict run	▼ master	🔗 #9107	in 19 hours	👤 Brice Ruzand-ext

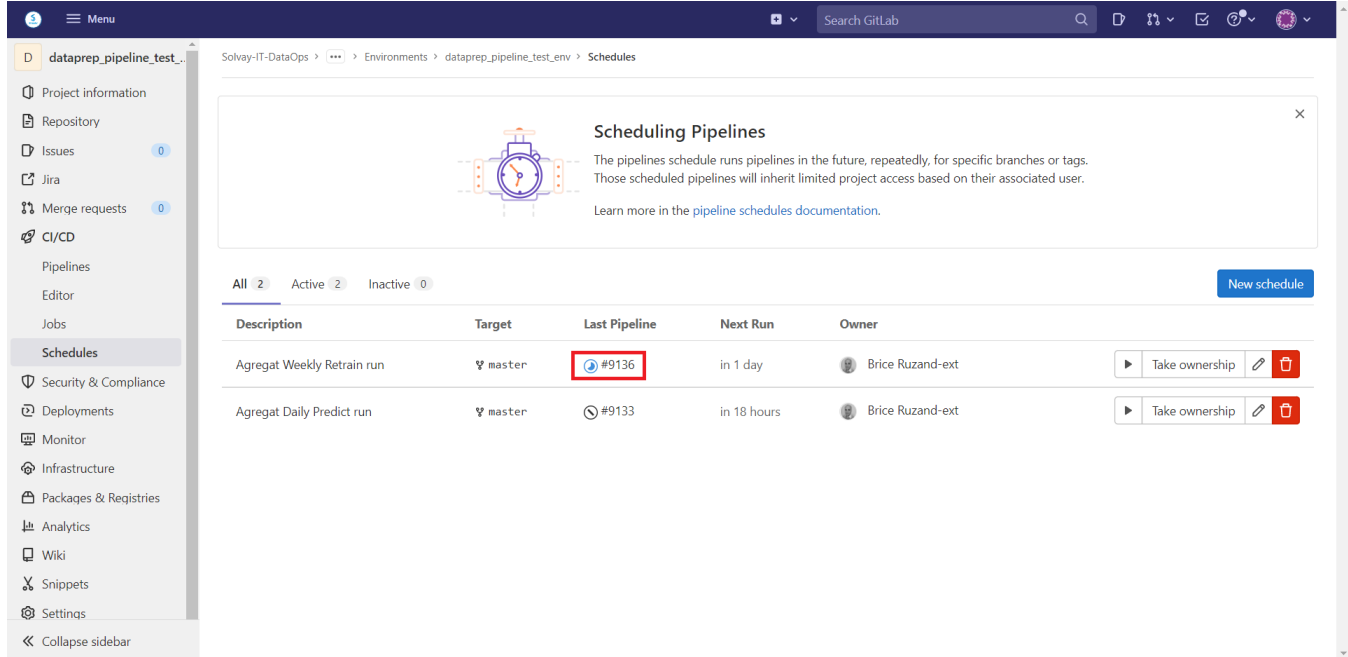
Below the table, there are two rows of actions. The first row has a 'Take ownership' button (highlighted with a red box), an edit icon, and a delete icon. The second row also has a 'Take ownership' button (highlighted with a red box), an edit icon, and a delete icon.

This will start the scheduling of the Dataprep each day at 7:00 am. The DataApp is triggered once the dataPrep is done so this is also the way to start the full process of DataApp as well.

Termination

How to assess the application's process has terminated

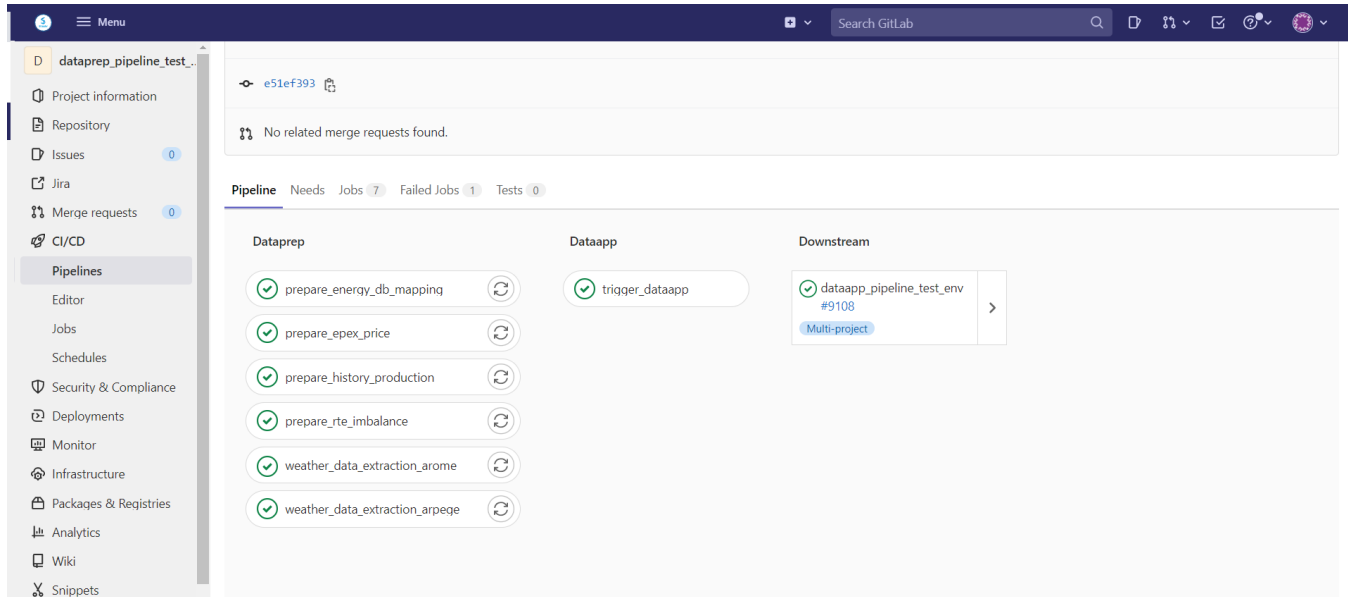
Once the process is launched you can follow the status by clicking on the pipeline number



The screenshot shows the GitLab interface for the 'Schedules' page. A modal window titled 'Scheduling Pipelines' is open, explaining that scheduled pipelines run in the future and inherit limited project access. Below the modal, a table lists the scheduled pipelines:

Description	Target	Last Pipeline	Next Run	Owner	Actions
Agregat Weekly Retrain run	▼ master	#9136	in 1 day	Brice Ruzand-ext	▶ Take ownership, ✎, 🗑️
Agregat Daily Predict run	▼ master	#9133	in 18 hours	Brice Ruzand-ext	▶ Take ownership, ✎, 🗑️

Then you can see a list of all pipelines running with there current status. This is the expected view of a pipeline that has terminated without errors. On the left are all the pipeline from the DataPrep, on the left this is the trigger toward the DataApp Pipelines.



The screenshot shows the GitLab Pipeline view for a specific pipeline. The pipeline is successful, with all jobs completed. The jobs are categorized into three stages:

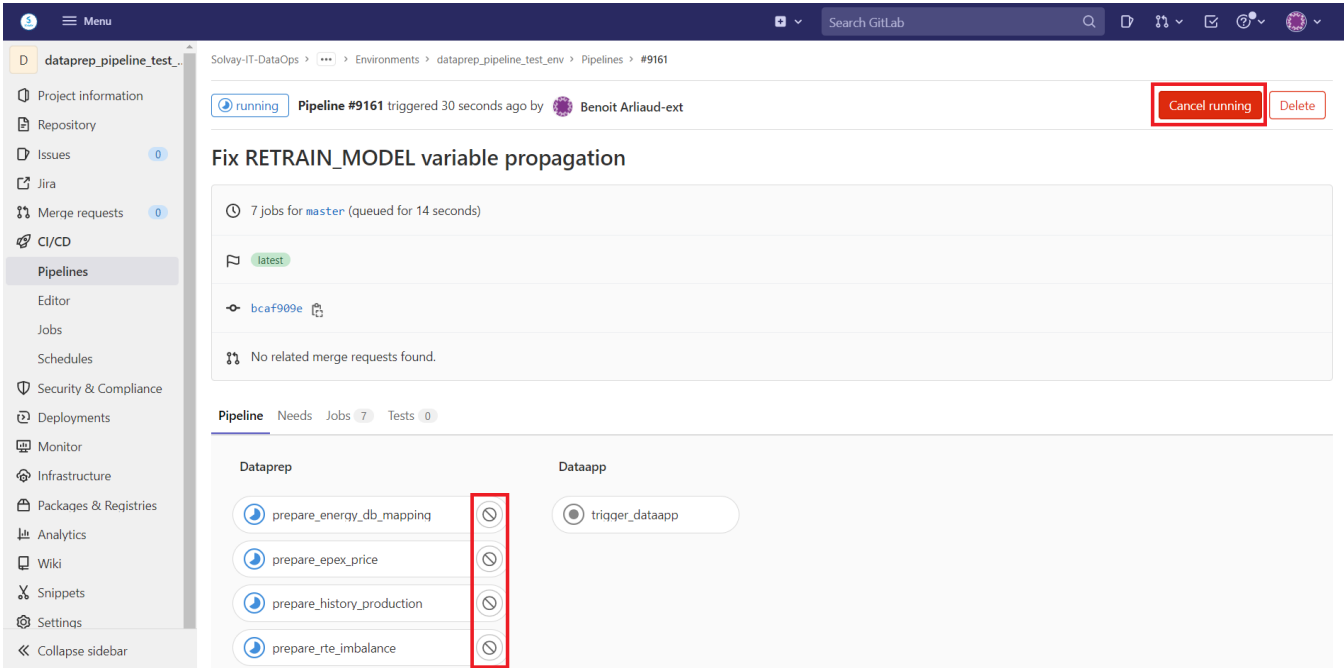
- Dataprep:** prepare_energy_db_mapping, prepare_epex_price, prepare_history_production, prepare_rte_imbalance, weather_data_extraction_arome, weather_data_extraction_arpege.
- Dataapp:** trigger_dataapp.
- Downstream:** dataapp_pipeline_test_env #9108 (Multi-project).

Restart

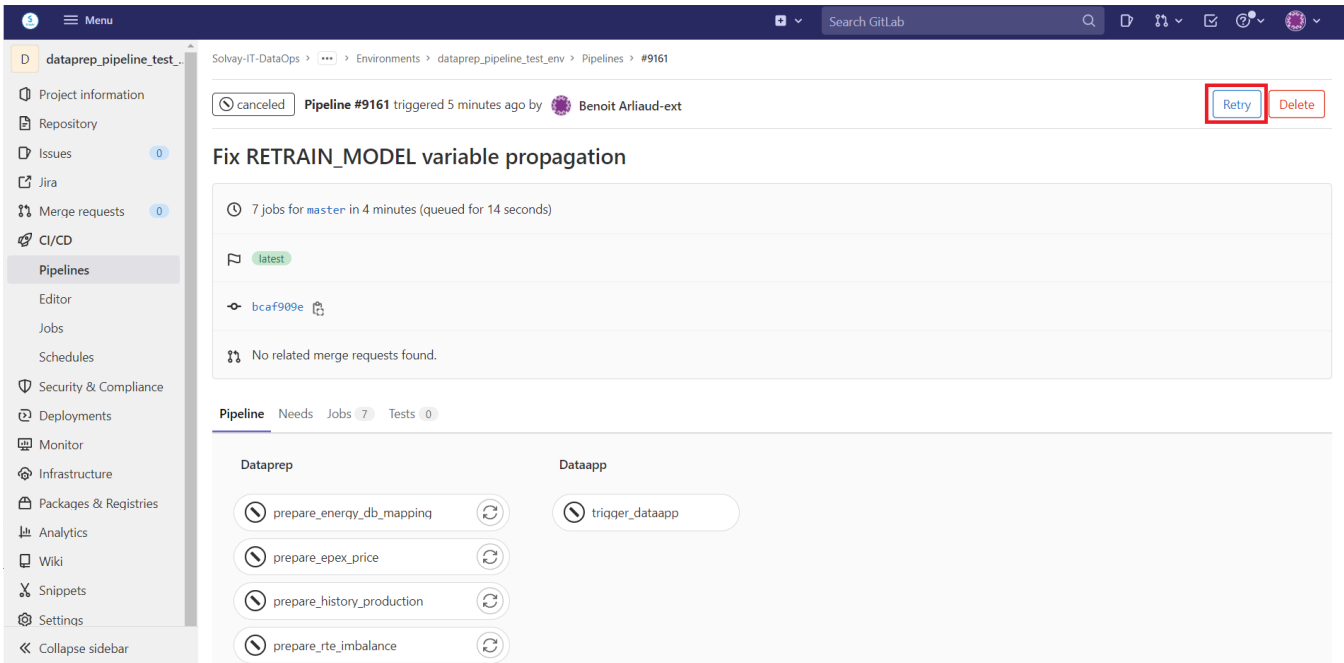
Is the process the same ?

Also for restarting after an error ?

To restart the DataPrep pipeline one must stop the running pipeline by either clicking on `Cancel running` or cancelling each pipeline one by one.



Then when once all the pipelines are stopped you can select the `Retry` option. This option also appear if everything succeeded.



Pause

Procedure

How ? (Resume)

Procedure can't be paused but it can be restarted.

Stop

Procedure

[How ?](#)

See section Restart

Alert contacts

[Who should be alerted ?](#)

Reset

[How ?](#)

DataApp Flow

Start

Full process

[How to start from scratch ?](#)

DataApp is automatically launched after DataPrep has finished thank to a trigger, so the whole process DataPrep and DataApp is launched from DataPrep Pipeline.

However the DataApp Pipeline can be launched manually in [DataApp pipelines](#)

Solvay-IT-DataOps > ... > Environments > dataapp_pipeline_test_env > Pipelines

All 47 Finished Branches Tags

Clear runner caches CI lint **Run pipeline**

Filter pipelines Show Pipeline ID ▾

Status	Pipeline	Triggerer	Stages
passed 01:15:03 1 hour ago	Document RETRAIN_MODEL variables #9283 master -> 8da9db38 latest		! → ✓ ✓
passed 01:17:15 2 days ago	Document RETRAIN_MODEL variables #9210 master -> 8da9db38 latest		! → ✓ ✓
passed 01:15:00 3 days ago	Document RETRAIN_MODEL variables #9179 master -> 8da9db38 latest		! → ✓ ✓
failed 02:10:29 4 days ago	Document RETRAIN_MODEL variables #9177 master -> 8da9db38 latest		! → ✗ » »

Just click on on blue button "Run Pipeline".

Solvay-IT-DataOps > ... > Environments > dataapp_pipeline_test_env > Pipelines

Run pipeline

Run for branch name or tag

master

Variables

Variable RETRAIN_MODEL false

Run dataiku dataapp RETRAIN or not (value: 'true'/'false', default: 'false')

Variable Input variable key Input variable value

Specify variable values to be used in this run. The values specified in [CI/CD settings](#) will be used by default.

Run pipeline Cancel

Then choose to set `RETRAIN_MODEL` to `true` if you want to retrain model.

If a pipeline fail you can select the pipeline then the job to see the logs like for the DataPrep.

Scheduling

Trigger

What is the start trigger ? Event based ? Time based ?

Are there differences between DataPrep and DataApp ?

The DataPrep is the first Trigger to set up. It launched the DataApp automatically based on a scheduler.

DataApp doesn't have a scheduler, all scheduler must be set up from DataPrep side.

Expected results

For each brick, what is the expected output ?

All Pipeline must be succeeded as a resulting status which means that in dataiku the scenario has succeeded in every step.

It is recommended to click on the link in the logs window to check in the Dataiku instance that the jobs has well succeeded.

Intervention

When is the time frame to intervene ? (when downtime is acceptable or scheduled)

Monitoring

Runtime

Where and how can we see the application status (Stopped, waiting, running, etc) ?

The status can be seen in the logs window when selecting a specific pipeline.

Run history

Where are the run actions historic ?

What form does it take ? Logs ?

The history of pipeline execution can be seen here :

- [DataPrep pipelines](#)
- [DataApp pipelines](#)

Resources

Memory / Disk / CPU used by application

Additional metrics

According to operational requirements, detail application metrics (Processed Volume, Process duration, ...)

Process duration can be seen from the logs window on the right panel

The screenshot displays the GitLab CI/CD interface. On the left, a sidebar shows navigation options like 'Jobs', 'Schedules', and 'Security & Compliance'. The main area shows a terminal window with the following log output:

```
19 Executing "step_script" stage of the job script
20 $ cd /dataiku_api_python/
21 $ echo "dataiku_api_python commit ${cat .commit.log}"
22 dataiku_api_python commit 8f56f73ee926a8039bceb61af65569e0bb6649c8
23 $ python ./trigger_scenario/trigger_scenario.py
24 [2022-04-19T06:14:09.628702] Scenario 'PREDICT' on project 'SESAGGREGAT_PROD' on server https://dss-automation-test.solvay.com : Triggered ...
25 Traceback (most recent call last):
26   File "./trigger_scenario/trigger_scenario.py", line 100, in <module>
27     trigger_scenario()
28   File "./trigger_scenario/trigger_scenario.py", line 58, in trigger_scenario
29     raise dataikuapi.utils.DataikuException(message)
30 dataikuapi.utils.DataikuException:
31 [2022-04-19T06:15:10.136988] Scenario 'PREDICT' on project 'SESAGGREGAT_PROD' on server https://dss-automation-test.solvay.com : Failed after 51.798s, Error was:
32 {
33   "target": {
34     "projectKey": "SESAGGREGAT_PROD",
35     "datasetName": "History_Production_DATABANK",
36     "partition": "NP",
37     "type": "DATASET_PARTITION"
38   },
39   "error": "DATASET CHECKS"
40 }
41 See details at https://dss-automation-test.solvay.com/projects/SESAGGREGAT_PROD/scenarios/PREDICT/runs/list/2022-04-19-06-14-09-715
42
43 Cleaning up file based variables
44
45 ERROR: Job failed: command terminated with exit code 1
```

On the right side, the 'predict' pipeline details are shown. A red box highlights the following information:

- Duration: 1 minute 7 seconds
- Finished: 1 week ago
- Timeout: 3h (from job)

Below this, the commit ID 'a5ed2a72' and the pipeline ID '#8768 for master' are visible.

Logging

Where to find each step logs ?

On the logs window.

Error handling

As a general guideline, application should stop as soon as possible.

In case of an error you must select the erroneous pipeline to see the logs. At the end of the logs you will have a link to the dataiku's job error.

By following link to dataiku, you will be able to see error details.

- ✓ Computed metrics on dataset History_Production_DATABANK in SESAGGREGAT_PROD (partition: NP)
- ✓ Computed metrics on dataset Park_Specifications in SESAGGREGAT_PROD (partition: NP)
- ✓ Computed metrics on dataset Energy_DB_Mapping_DATABANK in SESAGGREGAT_PROD (partition: NP)
- ✓ Computed metrics on dataset Arpege in SESAGGREGAT_PROD (partition: NP)
- ✓ Computed metrics on dataset Arpege_History in SESAGGREGAT_PROD (partition: NP)
- ✓ Computed metrics on dataset EpexPrice_DATABANK in SESAGGREGAT_PROD (partition: NP)
- ✓ Computed metrics on dataset Park_Maintenance in SESAGGREGAT_PROD (partition: NP)
- ✓ Computed metrics on dataset Hours_Limit_Mapping_DATABANK in SESAGGREGAT_PROD (partition: NP)
- ✗ **Run checks — INPUT_CHECKS** View step log at 08:14 10s
 - ✓ Checked dataset RTE_ImbalanceData_DATABANK in SESAGGREGAT_PROD (partition: NP)
 - ✓ Checked dataset RTE_ImbalanceUnitCost_DATABANK in SESAGGREGAT_PROD (partition: NP)
 - ✗ Checked dataset History_Production_DATABANK in SESAGGREGAT_PROD (partition: NP)
 - ⚠ Checked dataset Park_Specifications in SESAGGREGAT_PROD (partition: NP)
 - ✓ Checked dataset Energy_DB_Mapping_DATABANK in SESAGGREGAT_PROD (partition: NP)
 - ✓ Checked dataset Arpege in SESAGGREGAT_PROD (partition: NP)
 - ✓ Checked dataset Arpege_History in SESAGGREGAT_PROD (partition: NP)

Alerts

- *Contacts*
- *Meaningful message (timestamps, description, criticality)*

Meaningful message is displayed on the right panel.

Specificity

Detail procedure for specific error cases

Each error has a link to the dataiku job related. If the error message isn't clear, it is recommended to go access dataiku job and have more details.