service-config definition for pubengine publications

In the user-details section you provide general details such as your department name, cost center for charging AWS costs, and an email adress for sending notifications.

**user-details ~~properties~~** **example:**

user-details:

costcenter: 9999

department: bisfin

developer-email: your.email@aegon.nl

team-email: team.email@aegon.nl

**Detailed description** of **user-details:**

|  |  |
| --- | --- |
| setting | costcenter |
| description | four digit cost center |
| type | integer |
| mandatory | yes |
| validation | number between 1 and 9999 |

|  |  |
| --- | --- |
| setting | department |
| description | name of your department |
| type | string |
| mandatory | yes |

|  |  |
| --- | --- |
| setting | developer-email |
| description | email to send important notifications to. |
| mandatory | no |
| validation | should be a valid aegon email |

|  |  |
| --- | --- |
| setting | team-email |
| description | team email to send important notifications to. |
| mandatory | yes |
| validation | should be a valid aegon email |

The service-details section contains some general service information such as the name of your datalab and the version of the data platform to use.

**service-details ~~properties~~**example:

service-details:

datalab-name: bisfin

operation\_mode: on\_demand

dp-version: 2.0.3

data\_redelivery\_notifications\_enabled: true

**Detailed description** of **service-details:**

|  |  |
| --- | --- |
| setting | datalab-name |
| description | the datalab name to publish to from the feature and development stages |
| type | string |
| mandatory | yes |

|  |  |
| --- | --- |
| setting | dp-version |
| description | the version of the data platform to use for your publication pipeline |
| type | string |
| mandatory | no |
| validations | X.Y.Z, X.Y, X and lts-<year> versions are allowed |
| comments | If not provided, the dp-version will be the latest version which is available |

|  |  |
| --- | --- |
| setting | operation\_mode |
| description | The operation mode of the publication. Can be either **on\_demand** or **autopub.**  Unless using Autopub, **on\_demand** should be set as default |
| type | string |
| mandatory | yes |

|  |  |
| --- | --- |
| setting | data\_redelivery\_notifications\_enabled |
| description | Property to enable or disable the sending of an email message whenever a data redelivery happens in datalake with respect to a table which is used in your publication  An email will be sent to team email configured in user details.  By default this setting is enabled |
| type | boolean |
| mandatory | no |

In the publication details section you specify the general settings of your publication pipeline.

**publication-details example**

publication-details:

properties:

publication-name: a publication name

version: the version of the pubengine library to use

target: publication target

target-domain: optional target domain

target-source: optional target source

model-name: a model name

model-function: a model function

export-hub-target: an export hub target

pbi-report-name: a PowerBI report name

pbi-workspace-name: a PowerBI workspace name

schema-name: schema name in the Redshift cluster of the datalab, in which the published data will be made available

reporting\_date\_day\_delta: (optional) relative reporting date delta

event-trigger:

database: database as mentioned in the Glue catalog

table: table of the selected database

domain: (optional) domain of the selected table

source: (optional) source of the selected table

options:

dpu: 5

worker-type: G.1X

number-of-workers: 5

run-mode: default

spark-conf:

executor-memory: 20g

driver-memory: 10g

shuffle-partitions: 25

auto-broadcast-threshold: -1

executor-cores: 8

shuffle-file-buffer: 1m

transferTo: false

legacy-cte-precedence-policy: EXCEPTION

legacy-exponent-literal-as-decimal: true

legacy-from-day-time-string: true

legacy-allow-negative-scale-of-decimal: true

legacy-time-parser-policy: LEGACY

legacy-datetime-to-string: true

legacy-allow-empty-string: true

legacy-statistical-aggregate: true

legacy-integer-grouping-id: true

legacy-cast-complex-types-to-string: true

legacy-char-varchar-as-string: true

     spark.sql.legacy.nullValueWrittenAsQuotedEmptyStringCsv: true (any Spark-defined option allowed)

spark.executor.memory: 1g (a Spark-defined option may overwrite a pre-defined option, e.g. executor-memory: 20g -> 1g)

 glue-version: 3.0

compression-type: gzip (in pubengine v3 this option has moved to the writer options in the job-config.yml)

escape-special-characters: False (in pubengine v3 this option has moved to the writer options in the job-config.yml)

trim-whitespace: False (in pubengine v3 this option has moved to the writer options in the job-config.yml)

generate-md5s: True

use-s3-list-implementation: False

udf\_functions:

- udf\_lib\_general.poc.poc\_functions.poc\_square

- udf\_lib\_general.poc.poc\_functions.poc\_sqrt

details

properties

|  |  |
| --- | --- |
| setting | publication-name |
| description | the name of the publication pipeline. |
| type | string or integer |
| mandatory | yes |
| comments | Should be unique. |

|  |  |
| --- | --- |
| setting | version |
| description | The version of the pubengine to use for the job. |
| type | string |
| mandatory | no |
| comments | Should be a semantic version X.Y.Z |

|  |  |
| --- | --- |
| setting | target |
| description | target to publish to |
| type | string |
| mandatory | mandatory if production target is datalake |
| allowed values | datalake, information portal, export hub, datalab, model, dynamodb (default datalab) |
| comments |  |

|  |  |
| --- | --- |
| setting | target-domain |
| description | optional overwrite for the domain name of data published to datalake |
| type | string |
| mandatory | no |
| comments | requires target to be set, if not given will use default value: "dataplatform" |

|  |  |
| --- | --- |
| setting | target-source |
| description | optional overwrite for the source name of data published to datalake |
| type | string |
| mandatory | no |
| comments | requires target to be set, if not given will use default value: "pubengine" |

|  |  |
| --- | --- |
| setting | model-name |
| description | the name of the model to publish to |
| type | string |
| mandatory | no |
| comments |  |

|  |  |
| --- | --- |
| setting | model-function |
| description | the name of the model's function |
| type | string |
| mandatory | only if model-name is provided |
| allowed values | logic, logical, cognitive, cognition, finance, financial, |
| comments | should be equal the the model's function |

|  |  |
| --- | --- |
| setting | export-hub-target |
| description | the name of the export hub's target |
| type | string |
| mandatory | only if export hub operator is added. |
| allowed values | azurecmre, sad, aptitude, grizzly, lunaprr, canadalifere, peoplesoftoracle, tkp, fabiusoracle |
| comments | The nonprod (acc-run) targets are: aptitude, azurecmre, grizzly |

|  |  |
| --- | --- |
| setting | pbi-report-name |
| description | the name of the PowerBI report |
| type | string |
| mandatory | only if target is information-portal |
| comments |  |

|  |  |
| --- | --- |
| setting | pbi-workspace-name |
| description | the name of the PowerBI workspace |
| type | string |
| mandatory | only if target is information-portal |
| comments | This should be an Information Portal (IP) workspace |

|  |  |
| --- | --- |
| setting | schema-name |
| description | schema name in the Redshift cluster of the datalab, in which the published data will be made available |
| type | string |
| format | max. length 63. Format in regex: "^[a-zA-Z\_][a-zA-Z0-9\_]\*$" Format in plain text:   * must start with one of: lowercase letter, uppercase letter, or underscore * ends with any number of: digit, lowercase letter, uppercase letter, or underscore |
| mandatory | only when the create\_external\_schema operator is added to the workflow |
| comments | The publication will create a schema in the database in your datalab with this particular name, containing your published data. Please make sure that this schema name is not already present in the database for another publication, because changing existing schema's is not supported (if it already exists the schema will then still contain the data of the other publication).  **Note:** most of the datalabs which have an active redshift cluster are configured to shut down the redshift cluster during the night (because of cost savings). If you schedule your publication during the night, the job can succeed but you will receive an error email that it cannot create the external schema. |

|  |  |
| --- | --- |
| setting | event-trigger |
| description | Triggers publication when new data for specified database table gets added to datalake |
| type | Dictionary |
| mandatory | No |
| comments |  |

|  |  |
| --- | --- |
| setting | database |
| description | Database as mentioned in the Glue catalog |
| type | string |
| mandatory | Only when using event-trigger |
| comments |  |

|  |  |
| --- | --- |
| setting | table |
| description | Table of the selected database |
| type | string |
| mandatory | Only when using event-trigger |
| comments |  |

|  |  |
| --- | --- |
| setting | domain |
| description | domain of the selected table |
| type | string |
| mandatory | no |
| comments |  |

|  |  |
| --- | --- |
| setting | source |
| description | source of the selected table |
| type | string |
| mandatory | no |
| comments |  |

|  |  |
| --- | --- |
| setting | reporting\_date\_day\_delta |
| description | Amount of days to add or substract of event triggered or CRON scheduled reporting date  example: using value -2 will run the pipeline with a reporting date of 2 days ago |
| type | integer |
| mandatory | no |
| comments |  |

options

|  |  |
| --- | --- |
| setting | dpu |
| descriptions | amount of data processing units to use by AWS Glue |
| type | integer |
| mandatory | no, |
| validations | number between 2 and 99 |
| comments | A single Data Processing Unit (DPU) provides 4 vCPU and 16 GB of memory. For more details and pricing visit <https://aws.amazon.com/glue/pricing/>.  **Deprecated:** Use number-of-workers instead. |

|  |  |
| --- | --- |
| setting | worker-type |
| descriptions | AWS Glue comes with 3 worker types: Standard, G.1X and G.2X. The standard worker consists of 16 GB memory, 4 vCPUs of compute capacity, and 50 GB of attached EBS storage with two Spark executors. The G.1X worker consists of 16 GB memory, 4 vCPUs, and 64 GB of attached EBS storage with one Spark executor. The G.2X worker allocates twice as much memory, disk space, and vCPUs as G.1X worker type with one Spark executor. For more details on AWS Glue Worker types, see the documentation on [AWS Glue Jobs](https://docs.aws.amazon.com/glue/latest/dg/add-job.html). |
| type | integer |
| mandatory | no, but must be set if number-of-workers is. |
| validations | Must be Standard, G.1X or G.2X |
| comments | For more details and pricing visit <https://aws.amazon.com/glue/pricing/> .  If you are using Glue version 3.0 you can not make use of Standard as Glue version 3.0 does not support the Standard Worker type. |

|  |  |
| --- | --- |
| setting | number-of-workers |
| descriptions | The number of workers that are maximally allocated when a job runs.   Glue Autoscaling is enabled for Glue version 3.0 and higher (glue-version parameter): it will automatically scale down to a lower number of workers when less capacity is needed, thus reducing costs.  The parameter number-of-workers will indicate the number of workers statically allocated when glue-version is set to 2.0! |
| type | integer |
| mandatory | no, but must be set if worker-type is. |
| validations | between 2 and 20 (defaults to 2) |
| comments | A single Data Processing Unit (DPU) provides 4 vCPU and 16 GB of memory. For more details and pricing visit <https://aws.amazon.com/glue/pricing/>. Overrides dpu option. |

|  |  |
| --- | --- |
| setting | run-mode |
| descriptions | the run mode of the glue etl job |
| type | string |
| mandatory | no |
| allowed values | default or scenario |
| comments | In the scenario option, you have to supply scenario sql filters in the api call. |

|  |  |
| --- | --- |
| setting | glue-version |
| descriptions | the glue version of the glue etl job |
| type | string |
| mandatory | no |
| allowed values | 3.0, 4.0 (defaults to 3.0 after 1st of February 2023, 2.0 deprecated since June 2023) |
| comments |  |

|  |  |
| --- | --- |
| setting | compression-type |
| description | the compression codec to use for output data |
| type | string |
| mandatory | no |
| allowed values | one of case-insensitive shorten names (none, bzip2, gzip, lz4, and snappy). Defaults to gzip compression when a codec is not specified |

|  |  |
| --- | --- |
| setting | escape-special-characters |
| description | a computation heavy operation that escapes special characters in a string field such as quotes, backslashes and comma's |
| type | boolean |
| mandatory | no |
| allowed values | False or True. Defaults to False |
| comments | only use this option if necessary. A publication follows all CSV standards, but some systems like Redshift can have problems with unescaped quotes or backslashes in a string field |

|  |  |
| --- | --- |
| setting | trim-whitespace |
| description | a computation heavy operation that trims white spaces from all string fields |
| type | boolean |
| mandatory | no |
| allowed values | False or True. Defaults to False |
| comments | only use this option if necessary. A publication follows all CSV standards, but some systems like Redshift can have problems with whitespace only string field |

|  |  |
| --- | --- |
| setting | udf\_functions |
| description | a setting to enable enduser to use UDF functions from their department's UDF Libraries from Artifactory in glue publication pipeline |
| type | list |
| mandatory | no |
| allowed values | should start with prefix 'udf\_lib\_general' and should follow below format:  udf\_lib\_general.<package\_name>.<module\_name>.<udf\_name> |
| comments | only use this option if necessary to use udf libraries from artifactory in glue publication pipeline |

|  |  |
| --- | --- |
| setting | generate-md5s |
| description | generate an MD5 and add it to the metadata. |
| type | boolean |
| mandatory | no |
| comments | depending on the file size it will add to the processing time of the publication engine |

|  |  |
| --- | --- |
| setting | use-s3-list-implementation |
| description | AWS Glue doesn't cache the list of files in memory all at once. Instead, AWS Glue caches the list in batches. This means that the driver is less likely to run out of memory. |
| type | boolean |
| mandatory | no |
| comments | enabling this option is best practice when using the bookmark option  <https://aws.amazon.com/premiumsupport/knowledge-center/glue-oom-java-heap-space-error/> |

spark\_conf

This feature is not available in LTS-2021.

Changing the mentioned Spark configurations can either improve but also DECREASE the performance and costs of a publication. So, tread carefully and experiment at your own risk.

When no spark-conf is defined in the service-config, it will default to the defaults of Glue.

Legacy options are also provided to support a smoother migration between Spark SQL versions upon upgrading the Glue version.

Besides the information provided below, please also refer to the documentation of Spark:

* [Configuration - Spark 3.4.0 Documentation (apache.org)](https://spark.apache.org/docs/latest/configuration.html)
* [Migration Guide: SQL, Datasets and DataFrame - Spark 3.4.0 Documentation (apache.org)](https://spark.apache.org/docs/latest/sql-migration-guide.html)

For selecting your relevant migration sections, see the following table:

| **AWS Glue 4.0** | **AWS Glue 3.0** | **AWS Glue 2.0** |
| --- | --- | --- |
| Amazon EMR-optimized Spark 3.3.0 | Amazon EMR-optimized Spark 3.1.1 | Amazon EMR-optimized Spark 2.4 |

VERSION: 1.5.0

It is now possible to tweak a select few Spark configurations through the service-config.yml. We've enabled these options because we've found out that it can help with the performance of a publication.

For example: it helped us deal with the infamous OutOfMemory issue with respect to some IFRS17 publications. This publication processes change tables and benefits from setting the 'auto-broadcast-threshold' to **-1** and the 'shuffle-partitions' to 25 for this specific publication. All the possible pre-defined configurations under the spar-conf section are:

* executor-memory
* driver-memory
* shuffle-partitions
* auto-broadcast-threshold
* executor-cores
* shuffle-file-buffer
* transferTo

|  |  |
| --- | --- |
| setting | executor-memory |
| descriptions | Default is 10g. Amount of memory to use per executor process, in the same format as JVM memory strings with a size unit suffix ("k", "m", "g" or "t") (e.g. 512m, 2g). |
| type | string |
| mandatory | false |
| comments |  |

|  |  |
| --- | --- |
| setting | driver-memory |
| descriptions | Default is 10g. Amount of memory to use for the driver process, i.e. where SparkContext is initialized, in the same format as JVM memory strings with a size unit suffix ("k", "m", "g" or "t") (e.g. 512m, 2g). |
| type | string |
| mandatory | false |
| comments |  |

|  |  |
| --- | --- |
| setting | shuffle-partitions |
| descriptions | Default is 200. Configures the number of partitions to use when shuffling data for joins or aggregations. |
| type | integer |
| mandatory | false |
| comments |  |

|  |  |
| --- | --- |
| setting | auto-broadcast-threshold |
| descriptions | Default is none. Configures the maximum size in bytes for a table that will be broadcast to all worker nodes when performing a join. By setting this value to -1 broadcasting can be disabled. For more information |
| type | integer |
| mandatory | false |
| comments | When processing a publication with increments, thus a lot of change tables, it is advised to put this on -1. |

|  |  |
| --- | --- |
| setting | executor-cores |
| descriptions | The number of cores to use on each executor. |
| type | integer |
| mandatory | false |
| comments |  |

|  |  |
| --- | --- |
| setting | shuffle-file-buffer |
| descriptions | 32b by default. Size of the in-memory buffer for each shuffle file output stream, in KiB unless otherwise specified. These buffers reduce the number of disk seeks and system calls made in creating intermediate shuffle files. |
| type | string |
| mandatory | false |
| comments |  |

|  |  |
| --- | --- |
| setting | transferTo |
| descriptions | True by default. Spark can be forced to use more file buffer before writing to disk by setting this parameter to False. The benefit is to reduce I/O activity. |
| type | boolean |
| mandatory | false |
| comments |  |

|  |  |  |
| --- | --- | --- |
| **Pre-defined configurations** | **Allowed values** | **Details from Spark documentation** |
| legacy-cte-precedence-policy | LEGACY, CORRECTED, EXCEPTION | In Spark 3.0, spark.sql.legacy.ctePrecedencePolicy is introduced to control the behavior for name conflicting in the nested WITH clause. By default value EXCEPTION, Spark throws an AnalysisException, it forces users to choose the specific substitution order they wanted. If set to CORRECTED (which is recommended), inner CTE definitions take precedence over outer definitions. For example, set the config to false, WITH t AS (SELECT 1), t2 AS (WITH t AS (SELECT 2) SELECT \* FROM t) SELECT \* FROM t2 returns 2, while setting it to LEGACY, the result is 1 which is the behavior in version 2.4 and below. |
| legacy-exponent-literal-as-decimal | true or false | In Spark 3.0, numbers written in scientific notation(for example, 1E2) would be parsed as Double. In Spark version 2.4 and below, they’re parsed as Decimal. To restore the behavior before Spark 3.0, you can set spark.sql.legacy.exponentLiteralAsDecimal.enabled to true. |
| legacy-from-day-time-string | true or false | In Spark 3.0, day-time interval strings are converted to intervals with respect to the from and to bounds. If an input string does not match to the pattern defined by specified bounds, the ParseException exception is thrown. For example, interval '2 10:20' hour to minute raises the exception because the expected format is [+|-]h[h]:[m]m. In Spark version 2.4, the from bound was not taken into account, and the to bound was used to truncate the resulted interval. For instance, the day-time interval string from the showed example is converted to interval 10 hours 20 minutes. To restore the behavior before Spark 3.0, you can set spark.sql.legacy.fromDayTimeString.enabled to true. |
| legacy-allow-negative-scale-of-decimal | true or false | In Spark 3.0, negative scale of decimal is not allowed by default, for example, data type of literal like 1E10BD is DecimalType(11, 0). In Spark version 2.4 and below, it was DecimalType(2, -9). To restore the behavior before Spark 3.0, you can set spark.sql.legacy.allowNegativeScaleOfDecimal to true. |
| legacy-time-parser-policy | LEGACY | Parsing/formatting of timestamp/date strings. This effects on CSV/JSON datasources and on the unix\_timestamp, date\_format, to\_unix\_timestamp, from\_unixtime, to\_date, to\_timestamp functions when patterns specified by users is used for parsing and formatting. In Spark 3.0, we define our own pattern strings in Datetime Patterns for Formatting and Parsing, which is implemented via DateTimeFormatter under the hood. New implementation performs strict checking of its input. For example, the 2015-07-22 10:00:00 timestamp cannot be parse if pattern is yyyy-MM-dd because the parser does not consume whole input. Another example is the 31/01/2015 00:00 input cannot be parsed by the dd/MM/yyyy hh:mm pattern because hh supposes hours in the range 1-12. In Spark version 2.4 and below, java.text.SimpleDateFormat is used for timestamp/date string conversions, and the supported patterns are described in SimpleDateFormat. The old behavior can be restored by setting spark.sql.legacy.timeParserPolicy to LEGACY. |
| legacy-datetime-to-string | true or false | In Spark 3.0, Spark casts String to Date/Timestamp in binary comparisons with dates/timestamps. The previous behavior of casting Date/Timestamp to String can be restored by setting spark.sql.legacy.typeCoercion.datetimeToString.enabled to true. |
| legacy-allow-empty-string | true or false | In Spark version 2.4 and below, the parser of JSON data source treats empty strings as null for some data types such as IntegerType. For FloatType, DoubleType, DateType and TimestampType, it fails on empty strings and throws exceptions. Spark 3.0 disallows empty strings and will throw an exception for data types except for StringType and BinaryType. The previous behavior of allowing an empty string can be restored by setting spark.sql.legacy.json.allowEmptyString.enabled to true. |
| legacy-statistical-aggregate | true or false | In Spark 3.1, statistical aggregation function includes std, stddev, stddev\_samp, variance, var\_samp, skewness, kurtosis, covar\_samp, corr will return NULL instead of Double.NaN when DivideByZero occurs during expression evaluation, for example, when stddev\_samp applied on a single element set. In Spark version 3.0 and earlier, it will return Double.NaN in such case. To restore the behavior before Spark 3.1, you can set spark.sql.legacy.statisticalAggregate to true. |
| legacy-integer-grouping-id | true or false | In Spark 3.1, grouping\_id() returns long values. In Spark version 3.0 and earlier, this function returns int values. To restore the behavior before Spark 3.1, you can set spark.sql.legacy.integerGroupingId to true. |
| legacy-cast-complex-types-to-string | true or false | In Spark 3.1, structs and maps are wrapped by the {} brackets in casting them to strings. For instance, the show() action and the CAST expression use such brackets. In Spark 3.0 and earlier, the [] brackets are used for the same purpose. To restore the behavior before Spark 3.1, you can set spark.sql.legacy.castComplexTypesToString.enabled to true. |
| In Spark 3.1, NULL elements of structures, arrays and maps are converted to “null” in casting them to strings. In Spark 3.0 or earlier, NULL elements are converted to empty strings. To restore the behavior before Spark 3.1, you can set spark.sql.legacy.castComplexTypesToString.enabled to true. |
| legacy-char-varchar-as-string | true or false | Since Spark 3.1, CHAR/CHARACTER and VARCHAR types are supported in the table schema. Table scan/insertion will respect the char/varchar semantic. If char/varchar is used in places other than table schema, an exception will be thrown (CAST is an exception that simply treats char/varchar as string like before). To restore the behavior before Spark 3.1, which treats them as STRING types and ignores a length parameter, e.g. CHAR(4), you can set spark.sql.legacy.charVarcharAsString to true. |

VERSION: 2.6.0

From now on, any Spark-defined configurations are supported through the service-config.yml. This will give the user the most liberty in configurating and optimizing their publication runs.  
However, due to the lack of input validation, the user also has the responsibility to make sure that all set configuration options are actually valid.

All pre-defined configuration options are still usable and valid, but any overlap in Spark-defined configuration options will result in overwriting the pre-defined configuration options.

|  |  |
| --- | --- |
| setting | Any property mentioned in [Configuration - Spark 3.4.0 Documentation (apache.org)](https://spark.apache.org/docs/latest/configuration.html) and [Migration Guide: SQL, Datasets and DataFrame - Spark 3.4.0 Documentation (apache.org)](https://spark.apache.org/docs/latest/sql-migration-guide.html), i.e. spark.sql.legacy.interval.enabled |
| descriptions | Because any property is allowed, see the links provided for more information. |
| type | string, boolean, integer |
| mandatory | false |
| comments | No validation is run on these properties! |

The workflow-details section contains information about the tasks that your model or publication flow needs to execute.

**workflow-details properties**

workflow-details:

properties:

workflow-name: your\_workflow\_name

options:

retries: 3

email-settings: email\_disabled

operator-chain:

- operator-name: operator\_name

properties:

arg: value

details

|  |  |
| --- | --- |
| setting | workflow-name |
| description | name of your workflow |
| type | string |
| mandatory | yes |
| comments |  |

|  |  |
| --- | --- |
| setting | retries |
| description | amount of retries to use when the workflow fails |
| type | integer |
| mandatory | yes |
| comments |  |

|  |  |
| --- | --- |
| setting | email-settings |
| description | only send mails on error, or not at all |
| type | string |
| mandatory | no |
| comments | Allowed values:   * email\_on\_error * email\_disabled   Default behaviour: Not defined, always send emails |

|  |  |
| --- | --- |
| setting | operator-chain |
| description | operators to execute in your workflow |
| type | list |
| mandatory | yes |
| validation | should have at least one operator |
| comments | the operators will be executed sequentially from top to bottom |

|  |  |
| --- | --- |
| setting | operator-name |
| description | name of the operator to use |
| type | string |
| validation | should be an existing operator |
| comments | see operators block for available operators |

## Scheduled workflow

**scheduled-workflow-details properties**

workflow-details:

properties:

workflow-name: your\_workflow\_name

options:

retries: 3

scheduled-workflow:

schedule-cron: "0 09 01 \* ? \*"

operator-chain:

- operator-name: operator\_name

properties:

arg: value

For a scheduled operation, you want to use Cron (Chronus; Greek word for time).   
To this end you have to specify 6 elements, as explained below.   
If you want to allow any value for a given element (say any hour) then use \*.

**Cron notation**

# ┌───────────── minute (0 - 59)

# │ ┌───────────── hour (0 - 23)

# │ │ ┌───────────── day of the month (1 - 31)

# │ │ │ ┌───────────── month (1 - 12 or JAN-DEC)

# │ │ │ │ ┌───────────── day of the week (1-7 or SUN-SAT)

# │ │ │ │ │ ┌───────────── year (1970 - 2099)

# │ │ │ │ │ │

# │ │ │ │ │ │

# \* \* \* \* \* \* <command to execute>

**Note:**

* **All scheduled events use UTC time zone, which differs from the time in the Netherlands by 1 or 2 hours.**
* **If you use a day of the month in your cron, set the day of the week value to '?' and vise versa. E.g. "\* \* 5 \* ? \*" and "\* \* ? \* FRI \*" This is an Amazon Web Services Standard restriction.**→ **You cannot use \* in both the Day-of-month and Day-of-week fields. If you use it in one, you must use ? in the other.**

Cron examples:

I want to schedule my workflow every Monday morning at 8:00:  
schedule-cron: "0 08 ? \* MON \*" (In practice the hour would differ, because the hours are in UTC).  
  
I want to schedule my workflow every 1st of each month at 09:00:  
schedule-cron: "0 09 01 \* ? \*"  
  
I want to schedule my workflow every 15 and 45 minutes every hour on every Wednesday:  
schedule-cron: "15,45 \* ? \* WED \*"

For more cron examples, please visit crontab.guru (but keep in mind the restriction within Amazon with the ? symbol described above). And for the full documentation on what is possible in Amazon Web Services, please visit [this webpage](https://docs.aws.amazon.com/AmazonCloudWatch/latest/events/ScheduledEvents.html)

## Available operators

**run\_publication\_pipeline operator**

The run publication pipeline operator requests and revokes required permissions, and publishes data according to your publication configuration.

**run\_publication\_pipeline example:**

operator-chain:

- operator-name: run\_publication\_pipeline

properties:

type: run\_glue\_job

**Detailed description** of **run\_publication\_pipeline settings:**

|  |  |
| --- | --- |
| Setting | type |
| Description | the type of publication pipeline to run |
| type | string |
| allowed values | run\_glue\_job |
| mandatory | yes |
| comments | currently only running a glue job is supported. You need to fill in the publication-details section to use this operator. |

**callback operator**

The callback operator sends the run status of a publication or model run to an API  
This operator will replace the default behavior of sending emails on workflow failure.  
(This operator does not run for feature branches)

operator-chain:

- operator-name: callback

properties:

skip\_in\_nonprod: "no"

url: api\_url

method: POST

credentials\_secret\_arn: secret\_arn

details

|  |  |  |
| --- | --- | --- |
| setting | skip\_in\_nonprod | |
| description | specifies whether to send a callback in non production stages | |
| type | string | |
| mandatory | yes |
| allowed values | "yes" or "no" | |

|  |  |
| --- | --- |
| setting | url |
| description | the api url |
| type | string |
| mandatory | yes |
| validation | should be a valid url |

|  |  |
| --- | --- |
| setting | method |
| description | the request method to use |
| type | string |
| mandatory | yes |
| allowed values | POST or GET |

|  |  |
| --- | --- |
| setting | credentials\_secret\_arn |
| description | the AWS arn to the secret stored in a secret manager in the BI account |
| type | string |
| allowed values | valid arn |
| comments | this secret contains the authentication parameters for the api |

**Export hub operator**

The export\_hub operator exports data from export\_hub to external targets.

**properties**

operator-chain:

- operator-name: export\_hub

properties:

use\_datetime\_path: 'true'

details

|  |  |
| --- | --- |
| setting | use\_datetime\_path |
| description | This option will enforce that the datetime/partition dates will be in the filename (e.g. 20220719) of the exported file. |
| type | string |
| mandatory | yes |
| details | It is important to enclose the value in quotes. |

**Check\_data\_availability operator**

The check\_data\_availability operator will check the availability of the data for the selected database(s).

**Use 1:**  
  
If you only provide the database name then all tables of that database will be read from the glue catalog, and all will be checked for your reporting date

operator-chain:

- operator-name: check\_data\_availability

properties:

databases\_to\_check:

- "database1"

- "database2"

**Use 2:**  
  
By making the database name a dictionary you can specify the tables to be checked per database.  
In the following example for database1 only "table1" will be checked, for database2 tables "table1" and "table2" will be checked,  
while table 3 will behave the same as the example above and all tables of this database will be checked.

operator-chain:

- operator-name: check\_data\_availability

properties:

databases\_to\_check:

- "database1":

- "table1"

- "database2":

- "table1"

- "table2"

- "database3"

details

|  |  |
| --- | --- |
| setting | databases\_to\_check |
| description | specifies the databases that are being checked on data availability |
| type | string or dictionary |
| mandatory | yes |
| allowed values | database names |

**Refresh Report operator**

The refresh\_report operator refreshes your power-bi reports in respective power-bi workspace.

**properties**

operator-chain:

- operator-name: refresh\_report

details

You need to fill in the power-bi report and workspace names in the publication details while running pubconfig

**execute\_workflow operator**

The execute\_workflow operator allows you to start a new workflow from within the current workflow. For example, you might want to start a batchrun model container after the publication for that container is finished.

Note: It is only possible to trigger a new workflow that is in the same stage as the current workflow. For example, a 'prd' publication workflow can trigger a 'prd' model workflow but not an 'acc' model workflow. It is possible to provide the feature branch name of the new workflow (workflow-stage-override), which will only be used if the current workflow is also based on a feature branch.

**properties**

operator-chain:

- operator-name: execute\_workflow

properties:

workflow-name: my-model

workflow-type: cgn

workflow-department: anp

workflow-stage-override: featinitial

options:

sync-status: false

details

|  |  |
| --- | --- |
| setting | workflow-name |
| description | The name of the workflow to start |
| type | string |
| allowed values |  |
| mandatory | yes |
| comments |  |

|  |  |
| --- | --- |
| setting | workflow-type |
| description | The type of the workflow to start |
| type | string |
| allowed values | ps, cgn, lgc |
| mandatory | yes |
| comments |  |

|  |  |
| --- | --- |
| setting | workflow-department |
| description | The department of the workflow to start |
| type | string |
| allowed values |  |
| mandatory | no |
| comments | If not provided, the department of the current workflow is used (as provided in user-details) |

|  |  |
| --- | --- |
| setting | workflow-stage-override |
| description | The workflow stage name to use when the workflow is triggered from a feature branch |
| type | string |
| allowed values |  |
| mandatory | no |
| comments | If not provided, the stage of the current workflow is used |

|  |  |
| --- | --- |
| setting | sync-status |
| description | A boolean to indicate whether the current workflow should wait for the new workflow to finish |
| type | string |
| allowed values | true, false |
| mandatory | no |
| comments | The default is 'false' |

**Create external schema operator**

Automatically creates the external schema for you in the Redshift cluster in the datalab

**properties**

operator-chain:

- operator-name: create\_external\_schema

details

To use this you must fill in the **schema-name** property in the  [publication details](file:///C:\display\DP\publication-details)

**Note:** most of the datalabs which have an active redshift cluster are configured to shut down the redshift cluster during the night (because of cost savings). If you schedule your publication during the night, the job can succeed but you will receive an error email that it cannot create the external schema.