Data Injection: MAGIC

Author: Agustin Bruzzese

Date: 5 Oct 2020

Second exercise for the ESCAPE-DIOS project: Data Injection

Aims and Objective

On the one hand, the main objective of this activity is focused on demonstrating the stability of the current RUCIO configuration, injecting data from the different experiments using lightweight rucio-upload command.

On the other hand, it is desirable that all of this report complies with the points established in the WP2 meeting on September 30.

In order to meet the objectives described, the following sections are developed below with the intention of convincingly addressing the incorporation of data from the MAGIC experiments.

Commands used

Set up a rucio-client

```
docker run -e RUCIO_ACCOUNT=bruzzese -it -d --name=rucio-client
projectescape/rucio-client
docker run -e RUCIO_ACCOUNT=bruzzese -e RUCIO_CFG_AUTH_TYPE=userpass -e
RUCIO_CFG_USERNAME=bruzzese_userpass -e RUCIO_CFG_PASSWORD= -it -d
--name=rucio-client projectescape/rucio-client
docker run -e RUCIO_ACCOUNT=bruzzese -v
./usercert.pem:/opt/rucio/etc/usercert.pem -v
./userkey.pem:/opt/rucio/etc/userkey.pem -it -d --name=rucio-client
projectescape/rucio-client
docker exec -it rucio-client /bin/bash
```

Used Rucio commands for data injection

Upload a file, even mock if real data is not available for the time being:

```
rucio upload --scope YOUREXPERIMENT_YOURINSTITUTION_YOURNAME  --rse RSE
FILE_TO_BE_UPLOADED
dd if=/dev/urandom of=FILE_TO_BE_UPLOADED bs=10M count=1
rucio-admin scope add --account bruzzese --scope
YOUREXPERIMENT_YOURINSTITUTION_YOURNAME
```

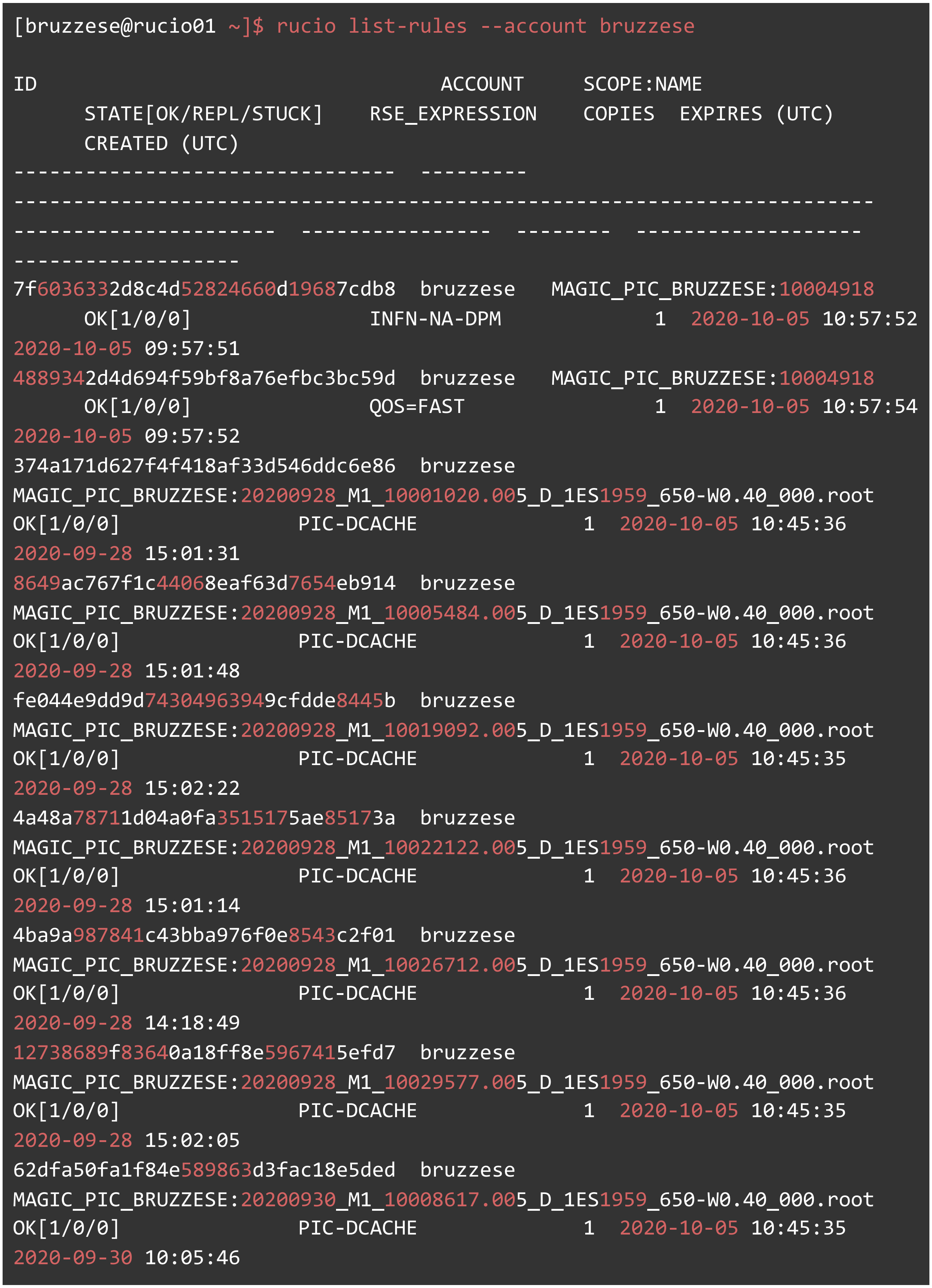
```
rucio upload --rse RSE --scope
YOUREXPERIMENT_YOURINSTITUTION_YOURNAME:FILE_TO_BE_UPLOADED
Add replication rule:
rucio add-rule YOUREXPERIMENT_YOURINSTITUTION_YOURNAME:FILE_TO_BE_UPLOADED
1 RSE
Check rule status:
rucio list-rules
YOUREXPERIMENT_YOURINSTITUTION_YOURNAME:FILE_TO_BE_UPLOADED
Download the file locally:
rucio download YOUREXPERIMENT_YOURINSTITUTION_YOURNAME:FILE_TO_BE_UPLOADED
Create a dataset:
rucio add-dataset YOUREXPERIMENT_YOURINSTITUTION_YOURNAME:DATASET_NAME
Attach a file to a dataset:
rucio attach YOUREXPERIMENT_YOURINSTITUTION_YOURNAME:DATASET_NAME
YOUREXPERIMENT_YOURINSTITUTION_YOURNAME:FILE_TO_BE_UPLOAD
Delete a rule:
rucio delete-rule --purge-replicas --all --rse_expression RSE --account
bruzzese YOUREXPERIMENT_YOURINSTITUTION_YOURNAME:FILE_TO_BE_UPLOAD
Also:
rucio delete-rule --purge-replicas --all --rse_expression RSE --account
bruzzese RULE_ID
Or:
rucio update-rule --lifetime 10 RULE_ID
Change your RSE settings to automatically eliminate your file once its rule
is deleted:
rucio-admin rse set-attribute --key greedyDeletion --value True --rse RSE
Create a simple workflow that accesses the data and performs actions on it
(e.g. get checksum and cross-check it):
mkdir download-files
rucio download --dir ./download-files
YOUREXPERIMENT_YOURINSTITUTION_YOURNAME:FILE_TO_BE_UPLOAD

echo "perform CHECKSUM"

gfal-sum
./download-files/YOUREXPERIMENT_YOURINSTITUTION_YOURNAME/FILE_TO_BE_UPLOAD
```

Monitoring Commands

Gather list of rules:



```
[bruzzese@rucio01 ~]$ rucio list-rules --account bruzzese

ID                                    ACCOUNT     SCOPE:NAME
      STATE[OK/REPL/STUCK]    RSE_EXPRESSION    COPIES  EXPIRES (UTC)
      CREATED (UTC)
--------------------------------  ---------
------------------------------------------------------------------------
----------------------  ----------------  --------  -------------------
-------------------
7f6036332d8c4d52824660d19687cdb8  bruzzese   MAGIC_PIC_BRUZZESE:10004918
      OK[1/0/0]               INFN-NA-DPM             1  2020-10-05 10:57:52
2020-10-05 09:57:51
4889342d4d694f59bf8a76efbc3bc59d  bruzzese   MAGIC_PIC_BRUZZESE:10004918
      OK[1/0/0]               QOS=FAST                1  2020-10-05 10:57:54
2020-10-05 09:57:52
374a171d627f4f418af33d546ddc6e86  bruzzese
MAGIC_PIC_BRUZZESE:20200928_M1_10001020.005_D_1ES1959_650-W0.40_000.root
OK[1/0/0]               PIC-DCACHE              1  2020-10-05 10:45:36
2020-09-28 15:01:31
8649ac767f1c44068eaf63d7654eb914  bruzzese
MAGIC_PIC_BRUZZESE:20200928_M1_10005484.005_D_1ES1959_650-W0.40_000.root
OK[1/0/0]               PIC-DCACHE              1  2020-10-05 10:45:36
2020-09-28 15:01:48
fe044e9dd9d74304963949cfdde8445b  bruzzese
MAGIC_PIC_BRUZZESE:20200928_M1_10019092.005_D_1ES1959_650-W0.40_000.root
OK[1/0/0]               PIC-DCACHE              1  2020-10-05 10:45:35
2020-09-28 15:02:22
4a48a78711d04a0fa3515175ae85173a  bruzzese
MAGIC_PIC_BRUZZESE:20200928_M1_10022122.005_D_1ES1959_650-W0.40_000.root
OK[1/0/0]               PIC-DCACHE              1  2020-10-05 10:45:36
2020-09-28 15:01:14
4ba9a987841c43bba976f0e8543c2f01  bruzzese
MAGIC_PIC_BRUZZESE:20200928_M1_10026712.005_D_1ES1959_650-W0.40_000.root
OK[1/0/0]               PIC-DCACHE              1  2020-10-05 10:45:36
2020-09-28 14:18:49
12738689f83640a18ff8e5967415efd7  bruzzese
MAGIC_PIC_BRUZZESE:20200928_M1_10029577.005_D_1ES1959_650-W0.40_000.root
OK[1/0/0]               PIC-DCACHE              1  2020-10-05 10:45:35
2020-09-28 15:02:05
62dfa50fa1f84e589863d3fac18e5ded  bruzzese
MAGIC_PIC_BRUZZESE:20200930_M1_10008617.005_D_1ES1959_650-W0.40_000.root
OK[1/0/0]               PIC-DCACHE              1  2020-10-05 10:45:35
2020-09-30 10:05:46
```

```
b5a3df431f344b4da117603de104e212  bruzzese
MAGIC_PIC_BRUZZESE:20200930_M1_10008617.005_D_1ES1959_650-W0.40_000.root
OK[1/0/0]               QOS=FAST                1  2020-10-05 10:45:35
2020-09-30 10:05:48
067ee68e07b74b099659b280f8762572  bruzzese
MAGIC_PIC_BRUZZESE:20200930_M1_10019032.005_D_1ES1959_650-W0.40_000.root
OK[1/0/0]               PIC-DCACHE              1  2020-10-05 10:45:35
2020-09-30 10:05:36
6d3a742ac21548999b4b37456dfd19ab  bruzzese
MAGIC_PIC_BRUZZESE:20201001_M1_10003703.005_D_1ES1959_650-W0.40_000.root
OK[1/0/0]               PIC-DCACHE              1  2020-10-05 10:45:35
2020-10-01 13:14:30
22c47c669d2f40f6a8ea9630aef3fe43  bruzzese
MAGIC_PIC_BRUZZESE:20201001_M1_10005718.005_D_1ES1959_650-W0.40_000.root
OK[1/0/0]               PIC-DCACHE              1  2020-10-05 10:45:35
2020-10-01 13:17:33
a2bef7f5d7664d59b779047755c55230  bruzzese
MAGIC_PIC_BRUZZESE:20201001_M1_10020055.005_D_1ES1959_650-W0.40_000.root
OK[1/0/0]               INFN-NA-DPM             1  2020-10-05 10:45:35
2020-10-01 13:24:56
7d06bf3282764a0080c647e6785d61ff  bruzzese
MAGIC_PIC_BRUZZESE:20201001_M1_10020055.005_D_1ES1959_650-W0.40_000.root
OK[1/0/0]               PIC-DCACHE              1  2020-10-05 10:45:35
2020-10-01 13:24:54
55f723107bc04a0b9d0594cdab3c07d2  bruzzese
MAGIC_PIC_BRUZZESE:20201005_M1_10004918.005_D_1ES1959_650-W0.40_000.root
OK[1/0/0]               PIC-DCACHE              1  2020-10-05 10:57:48
2020-10-05 09:57:42
```

Gather usage of a particular RSE:

```
[bruzzese@rucio01 ~]$ rucio list-rse-usage PIC-DCACHE
USAGE:
------
  files: 39875
  used: 1.762 TB
  rse: PIC-DCACHE
  updated_at: 2020-10-05 10:07:23
  source: rucio
  rse_id: dcee605ef2734b9ba295ba6eb0cc6be9
------
```

Count the number of files or datasets :

```
[bruzzese@rucio01 ~]$ rucio list-dids MAGIC_PIC_BRUZZESE:* --filter
type="ALL"
+---------------------------------------------------------------------------
------------+--------------+
| SCOPE:NAME
      | [DID TYPE]    |
|---------------------------------------------------------------------------
------------+--------------|
| MAGIC_PIC_BRUZZESE:20200203_M1_98179007.005_I_Perseus-MA-W0.26_2B288.root
      | FILE       |
| MAGIC_PIC_BRUZZESE:98179007
      | DATASET    |
+---------------------------------------------------------------------------
------------+------------
```

Data injection script output

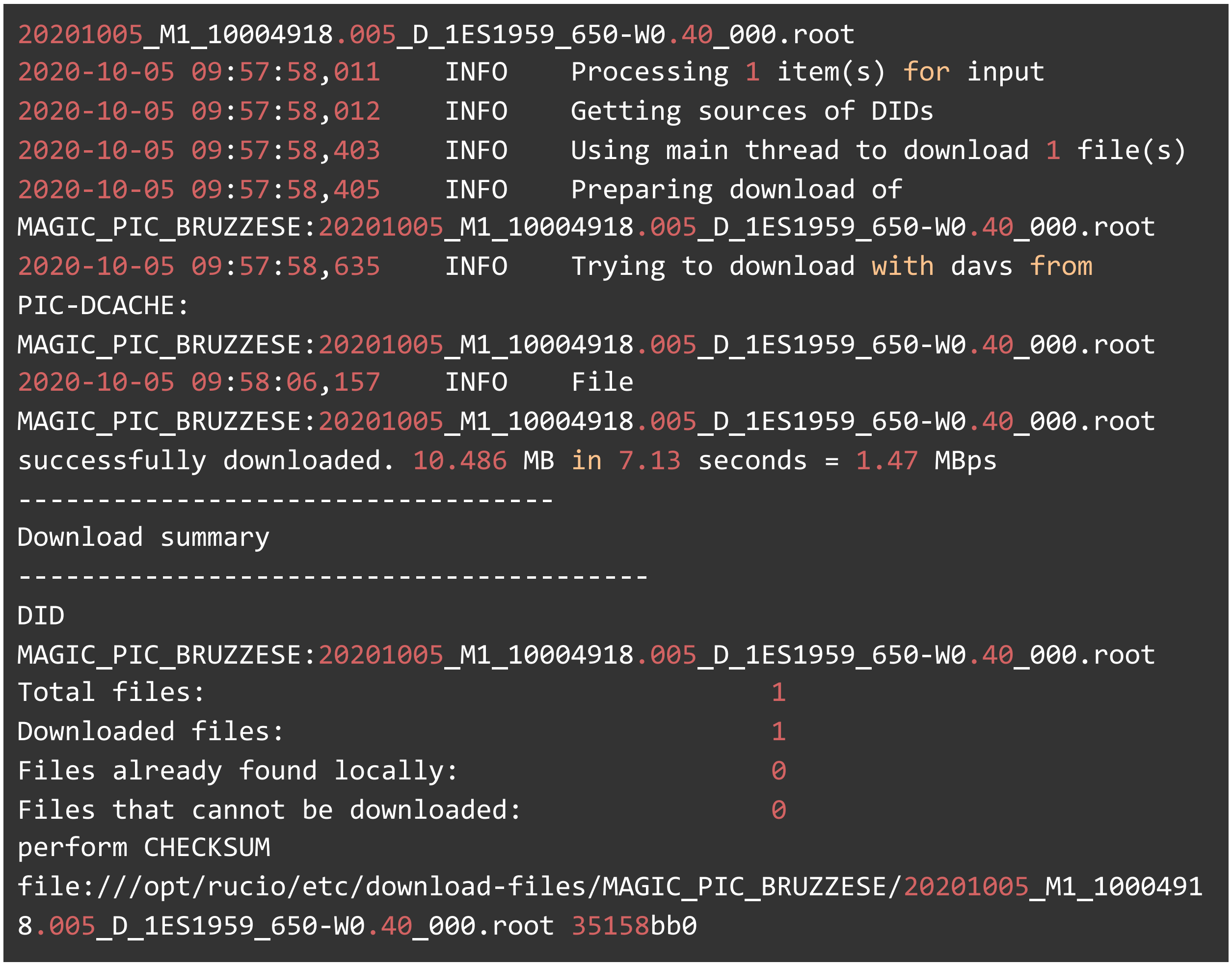
Below we include the output of the script that can be found at https://github.com/pic-es/rucio-client/blob/master/ESCAPE-workflow.sh. Briefly, the script is designed to execute the lightweight rucio command discussed in the section rucio commands for data injection. Please see the next section for more details on performance and data injection over time.

```
[root@rucio-client etc]# ./scripts/ESCAPE-dataflow.sh
set up your RSEs to eliminate files once the rules has been deleted
Added new RSE attribute for PIC-DCACHE: greedyDeletion-True
this file will be generated
20201005_M1_10004918.005_D_1ES1959_650-W0.40_000.root
1+0 records in
1+0 records out
10485760 bytes (10 MB) copied, 0.717038 s, 14.6 MB/s
20201005_M1_10004918.005_D_1ES1959_650-W0.40_000.root will be replicated
under the scope MAGIC_PIC_BRUZZESE to the RSE
2020-10-05 09:57:41,973    DEBUG    uploadclient.py    upload    Num. of
```

```
files that upload client is processing: 1
2020-10-05 09:57:42,224    DEBUG    uploadclient.py    upload    Input
validation done.
2020-10-05 09:57:42,224    INFO    Preparing upload for file
20201005_M1_10004918.005_D_1ES1959_650-W0.40_000.root
2020-10-05 09:57:42,494    DEBUG    uploadclient.py    upload    wan domain
is used for the upload
2020-10-05 09:57:42,494    DEBUG    uploadclient.py    _register_file
Registering file
2020-10-05 09:57:42,556    DEBUG    uploadclient.py    _register_file
Skipping dataset registration
2020-10-05 09:57:42,629    DEBUG    uploadclient.py    _register_file
File DID does not exist
2020-10-05 09:57:42,788    INFO    Successfully added replica in Rucio
catalogue at PIC-DCACHE
2020-10-05 09:57:42,930    INFO    Successfully added replication rule at
PIC-DCACHE
2020-10-05 09:57:42,949    DEBUG    gfal.py    connect    connecting
2020-10-05 09:57:43,000    DEBUG    gfal.py    exists    path None
2020-10-05 09:57:43,001    DEBUG    gfal.py    __gfal2_exist    path None
2020-10-05 09:57:43,171    DEBUG    rsemanager.py    exists    Checking if
root://xrootd.pic.es:1094//pnfs/pic.es/data/escape/rucio/pic_dcache/MAGIC_P
IC_BRUZZESE/8f/b0/20201005_M1_10004918.005_D_1ES1959_650-W0.40_000.root
exists
2020-10-05 09:57:43,177    DEBUG    gfal.py    exists    path
root://xrootd.pic.es:1094//pnfs/pic.es/data/escape/rucio/pic_dcache/MAGIC_P
IC_BRUZZESE/8f/b0/20201005_M1_10004918.005_D_1ES1959_650-W0.40_000.root
2020-10-05 09:57:43,177    DEBUG    gfal.py    __gfal2_exist    path
root://xrootd.pic.es:1094//pnfs/pic.es/data/escape/rucio/pic_dcache/MAGIC_P
IC_BRUZZESE/8f/b0/20201005_M1_10004918.005_D_1ES1959_650-W0.40_000.root
2020-10-05 09:57:43,490    DEBUG    gfal.py    close    closing protocol
connection
2020-10-05 09:57:43,490    DEBUG    uploadclient.py    upload
[{u'extended_attributes': None, u'hostname': u'door05.pic.es', u'prefix':
u'//rucio/pic_dcache', u'domains': {u'wan': {u'read': 1, u'write': 2,
u'third_party_copy': 1, u'delete': 1}, u'lan': {u'read': 1, u'write': 2,
u'delete': 2}}, u'scheme': u'davs', u'port': 8452, u'impl':
u'rucio.rse.protocols.gfal.Default'}, {u'extended_attributes': None,
u'hostname': u'xrootd.pic.es', u'prefix':
u'//pnfs/pic.es/data/escape/rucio/pic_dcache', u'domains': {u'wan':
{u'read': 1, u'write': 1, u'third_party_copy': 1, u'delete': 1}, u'lan':
{u'read': 1, u'write': 1, u'delete': 1}}, u'scheme': u'root', u'port':
1094, u'impl': u'rucio.rse.protocols.gfal.Default'}]
```

```
2020-10-05 09:57:43,490    INFO    Trying upload with root to PIC-DCACHE
2020-10-05 09:57:43,490    DEBUG    uploadclient.py    upload    Processing
upload with the domain: wan
2020-10-05 09:57:43,492    DEBUG    uploadclient.py    _upload_item    The
PFN created from the LFN:
root://xrootd.pic.es:1094//pnfs/pic.es/data/escape/rucio/pic_dcache/MAGIC_P
IC_BRUZZESE/8f/b0/20201005_M1_10004918.005_D_1ES1959_650-W0.40_000.root
2020-10-05 09:57:43,684    DEBUG    utils.py    __call__    put: Attempt 1
2020-10-05 09:57:46,505    INFO    Successful upload of temporary file.
root://xrootd.pic.es:1094//pnfs/pic.es/data/escape/rucio/pic_dcache/MAGIC_P
IC_BRUZZESE/8f/b0/20201005_M1_10004918.005_D_1ES1959_650-W0.40_000.root.ruc
io.upload
2020-10-05 09:57:46,505    DEBUG    uploadclient.py    _upload_item
skip_upload_stat=False
2020-10-05 09:57:46,505    DEBUG    uploadclient.py    _retry_protocol_stat
stat:
pfn=root://xrootd.pic.es:1094//pnfs/pic.es/data/escape/rucio/pic_dcache/MAG
IC_PIC_BRUZZESE/8f/b0/20201005_M1_10004918.005_D_1ES1959_650-W0.40_000.root
.rucio.upload
2020-10-05 09:57:46,548    DEBUG    uploadclient.py    _upload_item
Filesize: Expected=10485760 Found=10485760
2020-10-05 09:57:46,548    DEBUG    uploadclient.py    _upload_item
Checksum: Expected=35158bb0 Found=35158bb0
2020-10-05 09:57:46,548    DEBUG    uploadclient.py    _upload_item
Renaming file
root://xrootd.pic.es:1094//pnfs/pic.es/data/escape/rucio/pic_dcache/MAGIC_P
IC_BRUZZESE/8f/b0/20201005_M1_10004918.005_D_1ES1959_650-W0.40_000.root.ruc
io.upload to
root://xrootd.pic.es:1094//pnfs/pic.es/data/escape/rucio/pic_dcache/MAGIC_P
IC_BRUZZESE/8f/b0/20201005_M1_10004918.005_D_1ES1959_650-W0.40_000.root
2020-10-05 09:57:46,578    DEBUG    uploadclient.py    upload    Upload
done.
2020-10-05 09:57:46,578    INFO    Successfully uploaded file
20201005_M1_10004918.005_D_1ES1959_650-W0.40_000.root
2020-10-05 09:57:47,094    DEBUG    uploadclient.py    upload    Summary
will be available at rucio_upload.json
10004918 dataset 10004918 will be created
Added MAGIC_PIC_BRUZZESE:10004918
20201005_M1_10004918.005_D_1ES1959_650-W0.40_000.root will be listed the
scope MAGIC_PIC_BRUZZESE
ID                                    ACCOUNT     SCOPE:NAME
      STATE[OK/REPL/STUCK]    RSE_EXPRESSION    COPIES   EXPIRES (UTC)
      CREATED (UTC)
```

```
--------------------------------  ---------
------------------------------------------------------------------------
----------------------  ----------------  --------  -------------------
-------------------
55f723107bc04a0b9d0594cdab3c07d2  bruzzese
MAGIC_PIC_BRUZZESE:20201005_M1_10004918.005_D_1ES1959_650-W0.40_000.root
OK[1/0/0]                PIC-DCACHE              1  2020-10-05 10:57:48
2020-10-05 09:57:42
20201005_M1_10004918.005_D_1ES1959_650-W0.40_000.root will be discovered
under the paths :
+---------------------------------------------------------------------------
+--------------+
| SCOPE:NAME                                                            |
[DID TYPE]   |
|---------------------------------------------------------------------------
+--------------|
| MAGIC_PIC_BRUZZESE:20201005_M1_10004918.005_D_1ES1959_650-W0.40_000.root
| FILE      |
+---------------------------------------------------------------------------
+--------------+
20201005_M1_10004918.005_D_1ES1959_650-W0.40_000.root will be located under
the paths :
+---------------------+-----------------------------------------------------
--+------------+-----------+------------------------------------------------
----------------------------------------------------------------------------
--------------+
| SCOPE              | NAME                                                 |
FILESIZE    | ADLER32    | RSE: REPLICA
      |
|---------------------+-----------------------------------------------------
--+------------+-----------+------------------------------------------------
----------------------------------------------------------------------------
--------------|
| MAGIC_PIC_BRUZZESE |
20201005_M1_10004918.005_D_1ES1959_650-W0.40_000.root | 10.486 MB  |
35158bb0  | PIC-DCACHE:
davs://door05.pic.es:8452//rucio/pic_dcache/MAGIC_PIC_BRUZZESE/8f/b0/202010
05_M1_10004918.005_D_1ES1959_650-W0.40_000.root |
+---------------------+-----------------------------------------------------
--+------------+-----------+------------------------------------------------
----------------------------------------------------------------------------
--------------+
the following file will be downloaded:
```



```
20201005_M1_10004918.005_D_1ES1959_650-W0.40_000.root
2020-10-05 09:57:58,011    INFO    Processing 1 item(s) for input
2020-10-05 09:57:58,012    INFO    Getting sources of DIDs
2020-10-05 09:57:58,403    INFO    Using main thread to download 1 file(s)
2020-10-05 09:57:58,405    INFO    Preparing download of
MAGIC_PIC_BRUZZESE:20201005_M1_10004918.005_D_1ES1959_650-W0.40_000.root
2020-10-05 09:57:58,635    INFO    Trying to download with davs from
PIC-DCACHE:
MAGIC_PIC_BRUZZESE:20201005_M1_10004918.005_D_1ES1959_650-W0.40_000.root
2020-10-05 09:58:06,157    INFO    File
MAGIC_PIC_BRUZZESE:20201005_M1_10004918.005_D_1ES1959_650-W0.40_000.root
successfully downloaded. 10.486 MB in 7.13 seconds = 1.47 MBps
----------------------------------
Download summary
----------------------------------------
DID
MAGIC_PIC_BRUZZESE:20201005_M1_10004918.005_D_1ES1959_650-W0.40_000.root
Total files:                                    1
Downloaded files:                               1
Files already found locally:                    0
Files that cannot be downloaded:                0
perform CHECKSUM
file:///opt/rucio/etc/download-files/MAGIC_PIC_BRUZZESE/20201005_M1_1000491
8.005_D_1ES1959_650-W0.40_000.root 35158bb0
```

Performance and Failed output messages

In this section, we describe the dataflow for MAGIC mock data that was run in the ESCAPE Data Lake for 3 days. From Friday, October 2 to Monday, October 5, files that mimicked real data from MAGIC experiments were automatically created and injected into the data lake. During this time, the data follows a specific workflow that is described below, which meets the aims described in previous sections.

First, a file is generated locally and it is uploaded to the destination RSE PIC-DCACHE. From this upload two replication rules are created organized in datasets that contain the file. The replication rules destination are the RSE of INFN-NA-DPM and PIC QOS. Then, we proceed to make some interactions with the uploaded file. The file is downloaded using rucio commands, and then a checksum is carried out to fulfill the point of accessing the data and performing actions on it. Finally, the elimination of all three rules is established and, therefore, the erase of the files that they protect. This is carried out by means of the "rucio delete-rules" command, which gives a grace period of 1h until the rule elimination takes effect, and in conjunction with the "purge-replicas" option, so that also associated files are deleted.To automate these processes, cron is configured to run the workflow every hour.

In Fig. 1.1, the fluctuations in the creation / elimination of rules can be seen. We use rucio commands for monitoring rules creation and deletion. Specifically, since Friday, October 2 at 14:00 pm to Monday, October 5 at 14:00 pm, an hourly increase in the number of rules is seen, which corresponds to the uploads of a file to PIC-DCACHE (red) and two in the form of datasets (green)

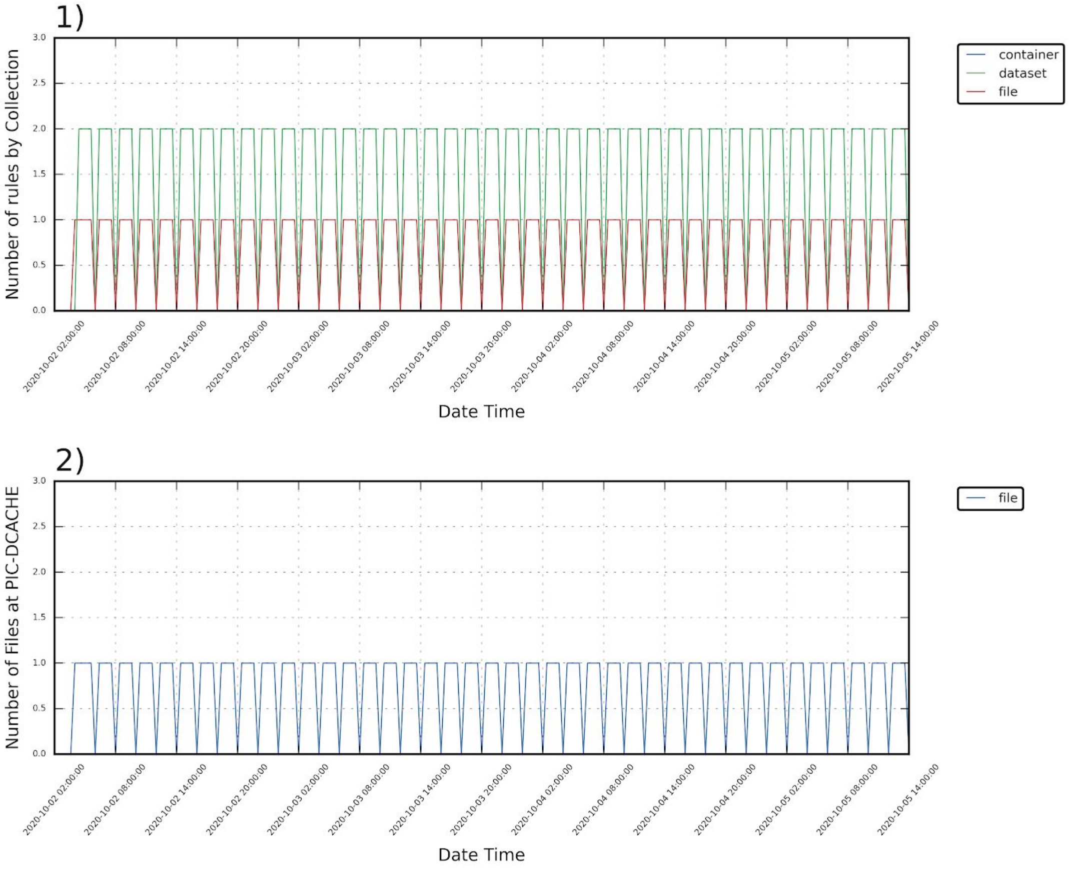


Figure 1. Dataflow cycle of MAGIC data injection

As an additional task, a file counting plugin was installed to monitor the number of files at the RSE PIC-DCACHE. This way, we can see that the elimination of a rule (Fig. 1.1) does indeed trigger the deletion of its corresponding file in the RSE. The results show consistency in the upload fluctuations of a file and the elimination of a rule (blue, Fig. 1.2), so it can be suggested that the daemon that is responsible for the elimination of files in the ESCAPE project (daemon-reaper) is functional.

After 72 h of continuous running of this workflow, the uploading and deletion of 71 files with the MAGIC namespace injected into the data lake were recorded.

Difficulties encountered if any

-

Personal feedback

Here are some suggestions intended to improve the current workflow.

- First, we still think that certain characters should be allowed in the data lake schema ("+").Here is a brief explanation of how it could be done.

```
agus@desktop:~/Desktop/Rucio-Server/rucio$ nano
lib/rucio/common/schema/generic.py
here you could add the "+" character (FYI :
https://github.com/rucio/rucio/blob/master/lib/rucio/common/schema/generic.
py#L71)
It should look something like this:
NAME = {"description": "Data Identifier name",
            "type": "string",
"pattern": "^[A-Za-z0-9][A-Za-z0-9\\.\\-\\_+]{1,%s}$" % NAME_LENGTH}
```

- At the docker image level, I think it would be advantageous to include the GFAL2 tool by default.
- For a next exercise I propose to move from the basic command line of the RUCIO API to python. We believe that the python-RUCIO platform offers a more sophisticated and easy mastery over the automation of replicas through data lake.

In turn, we would like to propose the following points for future exercises:

- Include an exercise with non-deterministic RSE. From comments from data lake users, many of them would use the "add_replicas" command to register a file that already exists in a local path.
- The configuration of a RSE in Roque de los Muchachos Observatory (ORM) to fully automate with Rucio the MAGIC ORM → PIC data flow. This would be very useful to implement the test dataflow in a realistic network path.