Provenance for in-development pipelines

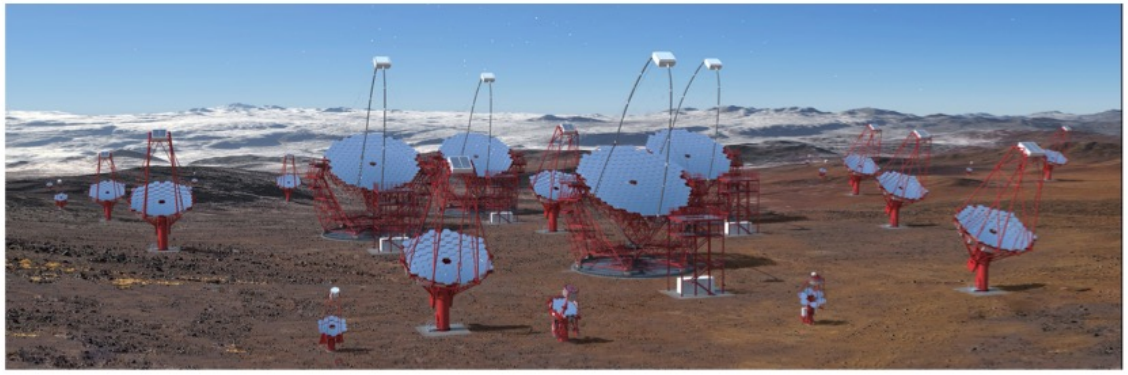
LST1 Large Size Telescope

José Enrique Ruiz (IAA – CSIC)

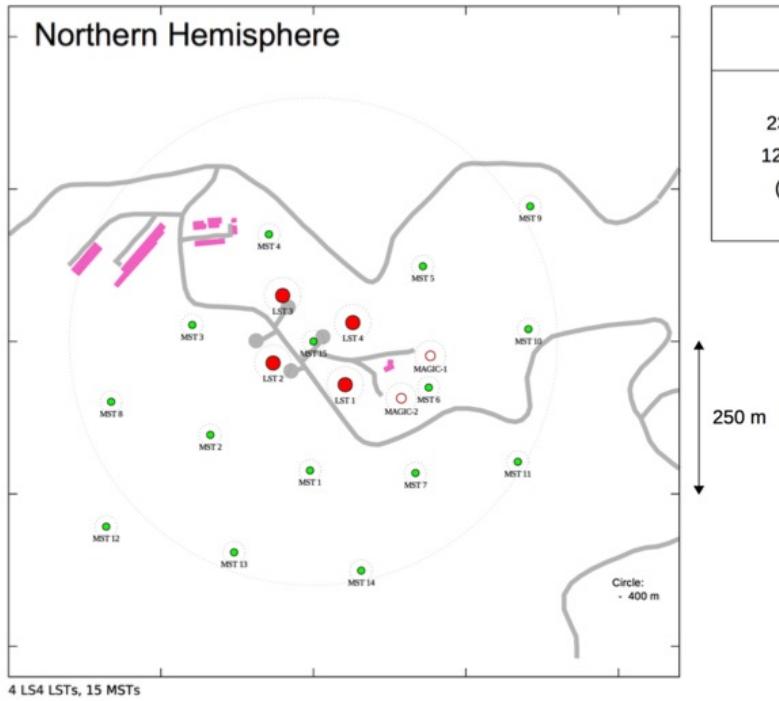
Provenance in Practice – ASOV meeting

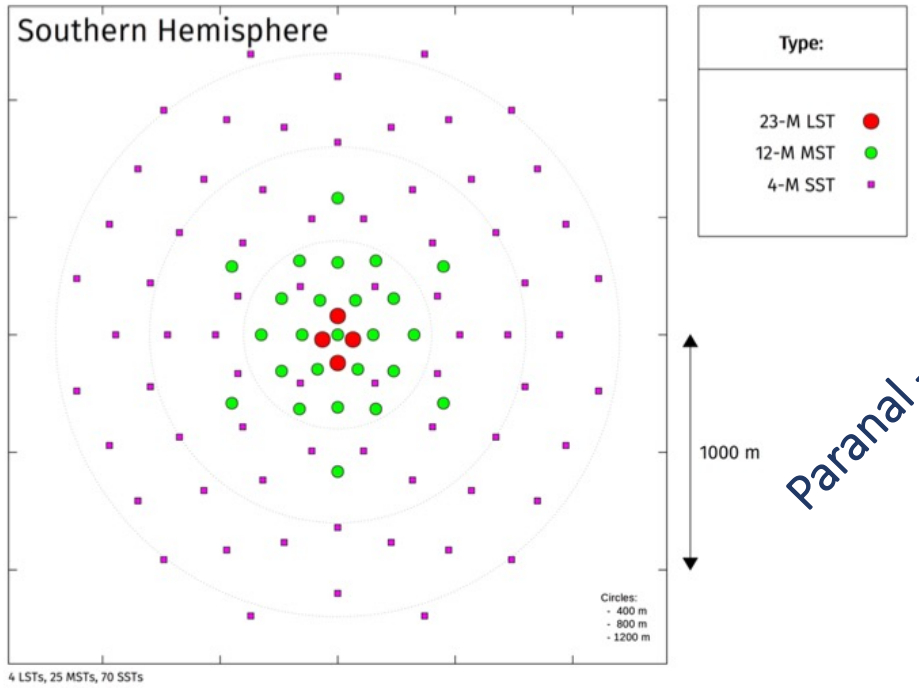
14/12/2022

Observatory sites and arrays



La Palma - Spain





Paranal - Chile

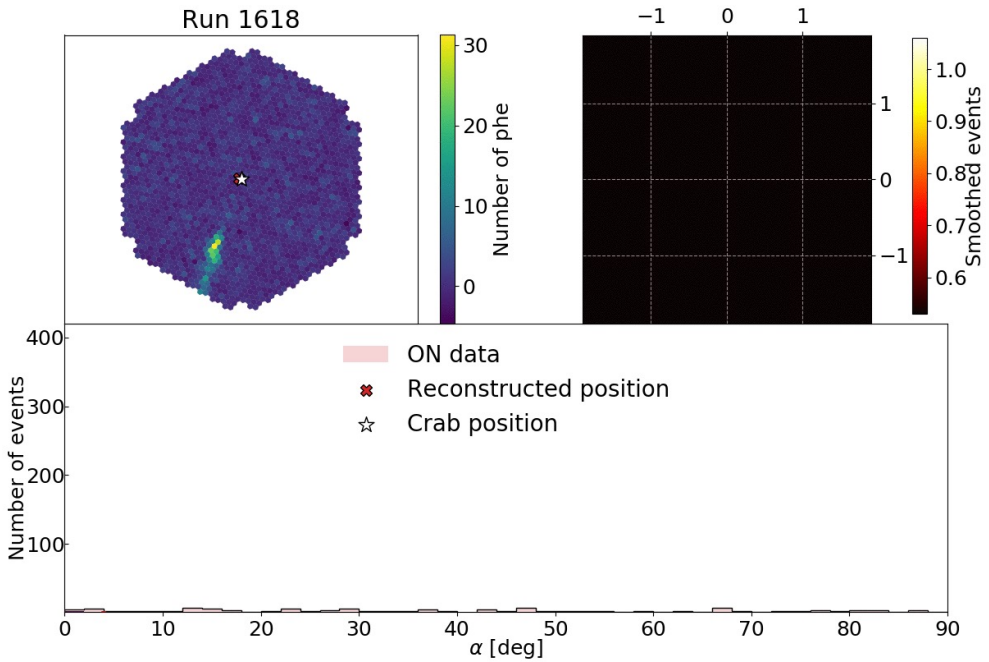
LST1 – The first Large Size Telescope

LST1 – The first Large Size Telescope

Inauguration 10th October 2018

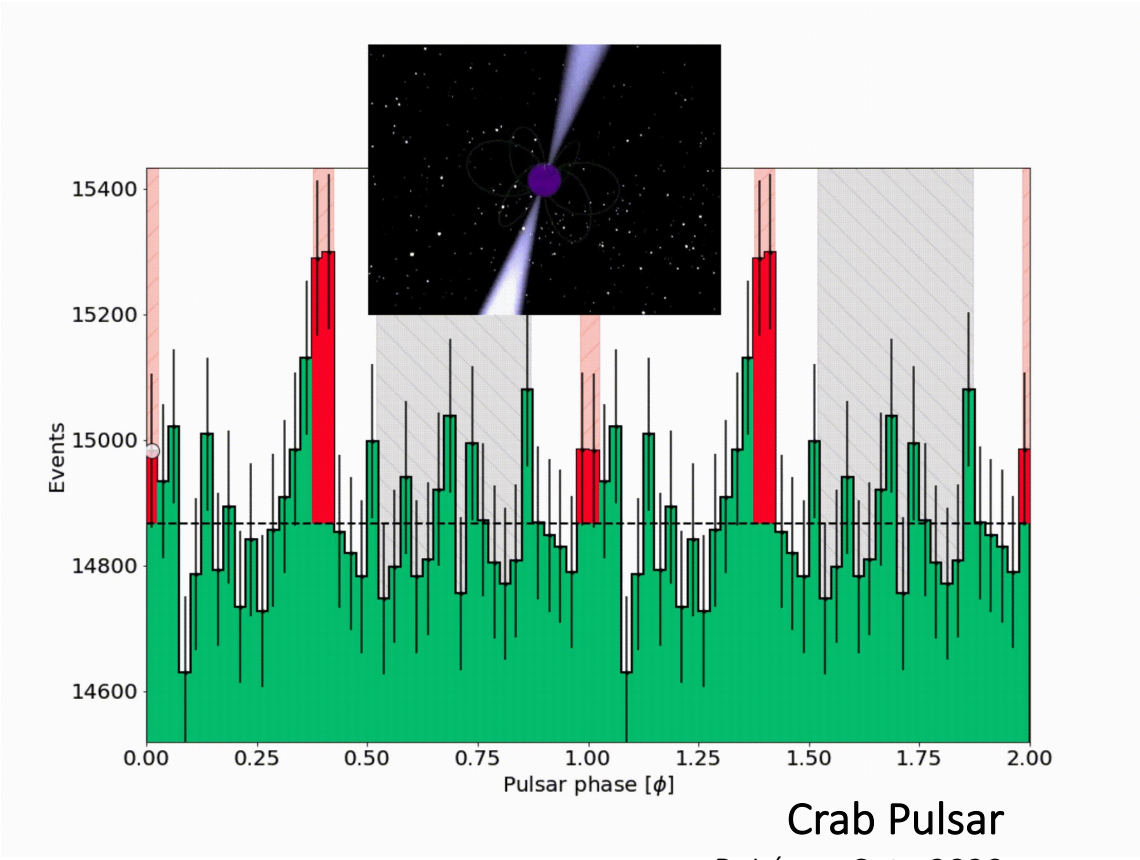
Critical Design Review
Deployment and Commissioning

Crab Campaigns
Extreme Physics i.e. BL Lac, RS Oph Nova



Crab Nebula

R. López-Coto 2019



Crab Pulsar

R. López-Coto 2020

Provenance requirements

CTA General Requirement

A-USER-0110 The CTA Observatory must ensure that data processing is traceable and reproducible

Specific Requirements

Observatory

- Keep the traceability of the data products
- Data quality and reliability checks
- Reproduce or reprocess the data
- Debug a pipeline

Astronomers

- Provide more information to the final user
- Refine final results based on analysis of the provenance info

Provenance requirements

CTA General Requirement

A-USER-0110 The CTA Observatory must ensure that data processing is traceable and reproducible

Specific Requirements

Observatory

- Keep the traceability of the data products
- Data quality and reliability checks
- Reproduce or reprocess the data
- Debug a pipeline

~~Astronomers~~ Developers

- Provide more information to the final user
- Refine final results based on analysis of the provenance info

Terminology

Configuration provenance

- software dependencies and execution environment

Data provenance

- usually gives some context and info on last processing step on the dataset
- mostly recorded in FITS HISTORY headers
- also metadata describing the dataset

Execution provenance

- commands and parameters/values with timestamps
- mostly recorded in log files

Full provenance

- provides a graph view of the whole process
- chain of commands datasets parameters/values

Capturing provenance

On top

- data products collection have been already produced
- extract only information present in data and save it in a structured way elsewhere

Inside

- capture and save provenance at the moment of data processing

Granularity and level of details

- which steps, processes, datasets, ancillary files and parameters
- execution environment and dependencies

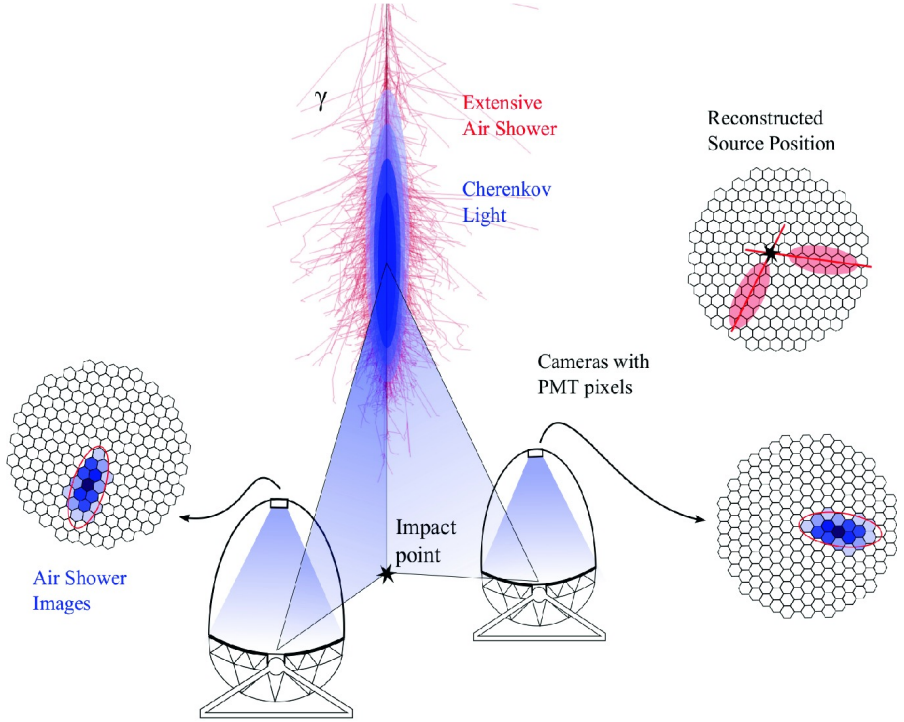
Format

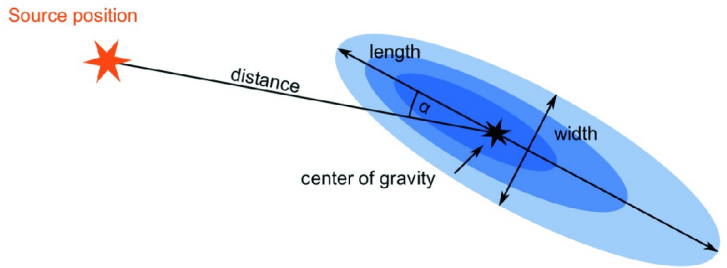
- model and syntax that accounts for describing the pipeline components and relationships
- serializing provenance info as attached metadata, structured logs, graphs, relational DB,..

Data Products in Cherenkov Imaging

1 Observation /run ~ 100 subruns ~ 10 000 events ~ 15 min

1 subrun ~ 10 s





rough file size numbers per Observation in LST 1

R0	RAW	Digital signal acquisition from hardware	2 x 30 x 1855 x events	~ 700 GB
DL0	PRE-CALIBRATED	Real photons and times measured in each telescope	30 x 1855 x events	
DL1	CALIBRATED	Cleaned photons pixel images and geometrical parameters	1855 x events	~ 70 GB
DL2	RECONSTRUCTED	Inferred direction, energy and gammaness	events	~ 1 GB
DL3	REDUCED	Selected event lists and instrumental response	events	~ 10MB
DL4	SCIENCE	Spectra, sky-maps, light-curves		~ 100 KB
DL5	ARCHIVE	Legacy observatory data i.e. catalogs, surveys,..		

- *R0 -> DL0*
  - *gain channel selection and correction*
  - *charge and time calibration*
  - *data volume reducer*

protozfits/hdf5
json/text

- *DL0 -> DL1*
  - *pulse charges integration in time windows*
  - *apply cleaning levels in events islands*
  - *geometrical parametrization of islands*
  - *muon analysis and data quality checks*

60 min

hdf5
json/text

- *DL1 -> DL2*
  - *low count gamma and very high background (gammaness cut)*
  - *direction/energy reconstruction via ML Random Forest algorithms with Montecarlo simulations*

5 min

hdf5
json/text and sav

- *DL2 -> DL3*
  - *tabular list of events selected as gammas*

fits

Software considered

work-in-progress software

lstosa

Python **pipeline** for the **On-site Analysis of low-level** data observations from the LST1 curated and developed by UCM Madrid
https://github.com/cta-observatory/lstosa

lstchain

Python **library** for the **processing of low-level** data observations from the LST1 curated and developed by the LST Collaboration
https://github.com/cta-observatory/cta-lstchain

ctapipe

Python **framework** for prototyping the **low-level data processing** algorithms for the Cherenkov Telescope Array.
https://github.com/cta-observatory/ctapipe

ctapipe

Configuration provenance
- software dependencies and execution environment

Execution provenance
- commands and parameters/values with timestamps
- mostly recorded in log files

Provenance in Tools

- provides a very simple API to **code provenance capture features in ctapipe Tools**
- provides serialization in JSON format and as Python dictionaries
- captures execution environment and profiling by default
- allows capture of processes, input and output datasets
- compliant with IVOA Provenance model

```python
from ctapipe.core import Provenance

prov = Provenance()
 # prov a singleton, so this gives you the same provenance class

prov.start_activitity("some_activity")

... # do things
prov.add_input_file("test.txt")
prov.add_output_file("out.txt")

prov.start_activity("some_sub_activity")

# do more things
prov.add_output_file("out2.txt")

prov.finish_activity() # finish some_activity
prov.finish_activity() # finish some_sub_activity
```

Cons

- tracks provenance execution **within a single Tool**
- needs the use of ctapipe Tools

Up to now it **does not allow chaining** provenance products of different Tools but allows **concatenation** of provenance records for the same Tool into a single log file

git

Tracks the execution history of a Tool producing a dataset across successive reprocessing linked to different pipeline versions

lstchain

Data provenance

- usually gives some context and info on last processing step on the dataset
- mostly recorded in FITS HISTORY headers
- also metadata describing the dataset

Store configuration execution metadata in HDF5 files #747

Merged rlopezcoto merged 18 commits into `master` from `metadata` 16 days ago

Conversation 22 | Commits 18 | Checks 8 | Files changed 10

Write | Preview

This PR stores the config execution of scripts as metadata attributes of the HDF5 files and tables produced in the R0->calibration->DL1->DL1AB->DL2 workflow.

The scripts involved in this workflow are:

- `/scripts/lstchain_data_create_time_calibration_file.py`
- `/tools/lstchain_create_calibration_file.py`
- `/scripts/lstchain_data_r0_to_dl1.py`
- `/scripts/lstchain_dl1ab.py`
- `/scripts/lstchain_dl1_to_dl2.py`
- `/scripts/lstchain_add_source_dependent_parameters.py`

The global metadata stored at the root level of the HDF5 files, as well as into each of the tables/containers produced, account for the versions of the software used in the execution of the script that produces de file i.e. `CTAPIPE_VERSION`, `LSTCHAIN_VERSION`, `CTAPIPE_IO_LST`

The configuration parameters used for each script/tool are stored as an attribute `config` of the tables/containers produced.

lstchain

Data provenance

- usually gives some context and info on last processing step on the dataset
- mostly recorded in FITS HISTORY headers
- also metadata describing the dataset

These metadata may be easily retrieved with `pytables` library and `vitables` software, as described in the python snippets and screenshots below.

```
with tables.open_file(filename) as file:
   ...:     print(file.root._v_attrs["LSTCHAIN_VERSION"])

0.7.4.post16+git0deb44e
```

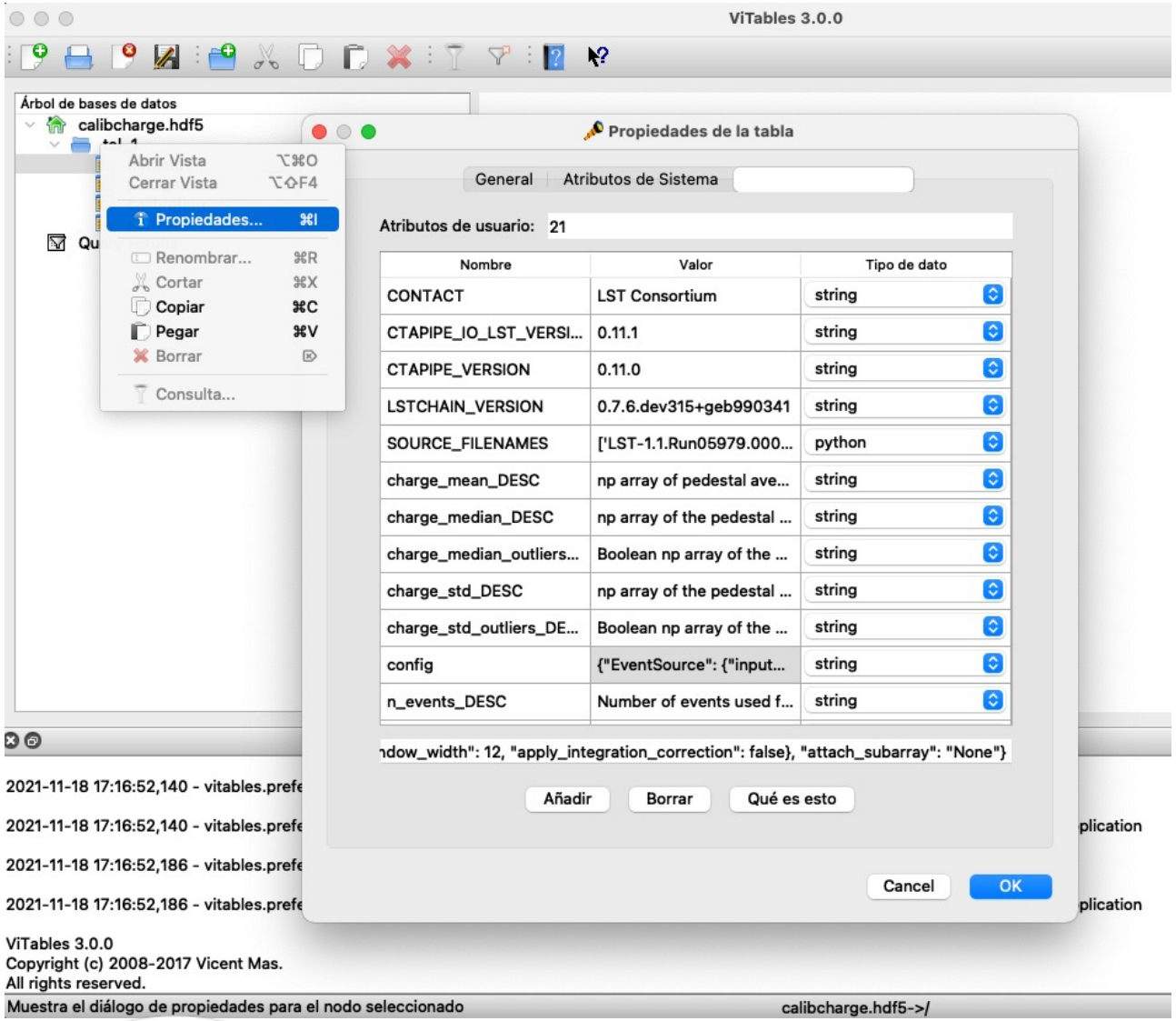
```
with tables.open_file(filename) as file:
   ...:     config = yaml.safe_load(file.root.dl1.event.telescope.parameters.LST_LSTCam.attrs["config"])
print(config["tailcut"])

{'picture_thresh': 6,
 'boundary_thresh': 3,
 'keep_isolated_pixels': False,
 'min_number_picture_neighbors': 2,
 'use_only_main_island': False,
 'delta_time': 2}
```

In the case where, during one of the workflow steps, a specific table is directly retrieved from one previous existing file and copied into the new produced HDF5 file, then the software versions and the config execution metadata are kept untouched.

lstchain

Data provenance

- usually gives some context and info on last processing step on the dataset
- mostly recorded in FITS HISTORY headers
- also metadata describing the dataset




```
CalibrationHDF5Writer:
  output_file: "calibcharge.hdf5"
  config_file: "/Users/jer/git/cta-observatory/cta-lstchain/lstchain/data/onsite_camera_calibration_param.json"
  one_event: true
  events_to_skip: 1000
  calibration_product: "LSTCalibrationCalculator"
  log_level: "DEBUG"
  log_config: {…}
  EventSource:
    input_url: "/Users/jer/DATA/LST/R0/20210902/LST-1.1.Run05979.0000.fits.fz"
LSTEventSource:
  LSTR0Corrections:
    drs4_pedestal_path: "/Users/jer/DATA/LST/calibration/20210902/v0.7.3/drs4_pedestal.Run05978.0000.fits"
    drs4_time_calibration_path: "/Users/jer/DATA/LST/calibration/20210902/v0.7.3/time_calibration.Run05979.0000.hdf5"
    select_gain: false
  EventTimeCalculator:
    run_summary_path: "/Users/jer/DATA/LST/monitoring/RunSummary/RunSummary_20210902.ecsv"
  allowed_tels:
    0: 1
LSTCalibrationCalculator:
  squared_excess_noise_factor: 1.222
  flatfield_product: "FlasherFlatFieldCalculator"
  pedestal_product: "PedestalIntegrator"
PedestalIntegrator:
  sample_duration: 100000
  tel_id: 1
  charge_median_cut_outliers:
    0: -10
    1: 10
  charge_std_cut_outliers:
    0: -10
    1: 10
  charge_product: "FixedWindowSum"
FlasherFlatFieldCalculator:
  sample_duration: 100000
  tel_id: 1
  charge_product: "LocalPeakWindowSum"
  charge_median_cut_outliers:
    0: -0.5
    1: 0.5
  charge_std_cut_outliers:
    0: -10
    1: 10
  time_cut_outliers:
    0: 2
    1: 38
LocalPeakWindowSum:
  window_shift: 5
  window_width: 12
  apply_integration_correction: false
FixedWindowSum:
```

lstosa

Collection of daily scheduled **scripts** that are **run in parallel in a grid environment**

Full provenance

- provides a graph view of the whole process
- chain of commands datasets parameters/values

Provenance capture

structured logging / provlog

How?

- Using standard Python logging mechanism and a **provenance model defined in a YAML file**
- Non-intrusive implementation with **function/class decoration** in existing code
- Info within decorated functions can be extracted and **mapped** into W3C/IVOA Prov syntax
- Python **logging configuration** is set in an independent configuration file

Which info?

- Used and produced datasets, input parameters, config files and processes

What do we get?

- **Post-processed** log files as merged/filtered logs with info in W3C/IVOA Prov syntax
- W3C provenance **JSON** files and **PDF** graphs as final provenance products

lstosa

Collection of daily scheduled **scripts** that are **run in parallel in a grid environment**

Full provenance

- provides a graph view of the whole process
- chain of commands datasets parameters/values

Detailed considerations

- Captured *subrun-wise* and **post-processed** to *run-wise*
- Post-processing/merging yields provenance products at **different levels of granularity on demand**
  - calibration
  - r0_to_dl1
  - dl1_to_dl2
  - all levels

different processing levels and intermediate steps
python provprocess.py -- args

- Execution **environment** is captured and stored as a session property
- Info *hidden* in **configuration files**
  - compared with hash-content algorithm and **copied** for reproducibility purposes
- Montecarlo simulated training datasets are not copied but **referenced**
- **Dry execution** mechanism allows provenance capture and merging avoiding data processing
- **Session properties**: observation date, run number, lstosa config file, software versions

config/definition.yaml

```yaml
activities:
    r0_to_dl1:
        description:
            "Create DL1 files for an observation run and subrun"
        parameters:
            - name: ObservationRun
              description: "Observation run number"
              value: ObservationRun
            - name: ObservationSubRun
              description: "Observation subrun number"
              value: ObservationSubRun
            - name: CalibrationRun
              description: "Calibration run number"
              value: CalibrationRun
            - name: PedestalRun
              description: "Pedestal run number"
              value: PedestalRun
            - name: ProdID
              description: "Production ID"
              value: ProdID
        usage:
            - role: "Observation subrun"
              description: "Observation subrun used"
              entityName: R0SubrunDataset
              value: R0SubrunDataset
              # filepath: /fefs/aswg/data/real/R0/20200218/LST1.1Run02006.0001.fits.fz
            - role: "Pedestal file"
              description: "Pedestal file used"
              entityName: PedestalFile
              value: PedestalFile
              # filepath: /fefs/aswg/data/real/calibration/20200218/v00/drs4_pedestal.Run02005.0000.fits
            - role: "Coefficients calibration file"
              description: "Coefficients calibration file"
              entityName: CoefficientsCalibrationFile
              value: CoefficientsCalibrationFile
              # filepath: /fefs/aswg/data/real/calibration/20200218/v00/calibration.Run02006.0000.hdf5
            - role: "Time calibration file"
              description: "Time calibration file"
              entityName: TimeCalibrationFile
              value: TimeCalibrationFile
              # filepath: /fefs/aswg/data/real/calibration/20191124/v00/time_calibration.Run1625.0000.hdf5
            - role: "Pointing file"
              description: "Pointing filename for DL1"
              entityName: PointingFile
              value: PointingFile
              # filepath: /fefs/home/lapp/DrivePositioning/drive_log_20_02_18.txt
```

config/logger.yaml

```yaml
version: 1
formatters:
    simple:
        format: '%(levelname)s %(name)s %(message)s'
        #format: '%(asctime)s.%(msecs)03d%(message)s'
        datefmt: '%Y-%m-%dT%H:%M:%S'
handlers:
    provHandler:
        class: logging.handlers.WatchedFileHandler
        level: INFO
        formatter: simple
        filename: prov.log
loggers:
    provLogger:
        level: INFO
        handlers: [provHandler]
        propagate: False
disable_existing_loggers: False
PREFIX: __PROV__
HASH_METHOD: md5
HASH_BUFFER: path
capture: True
```

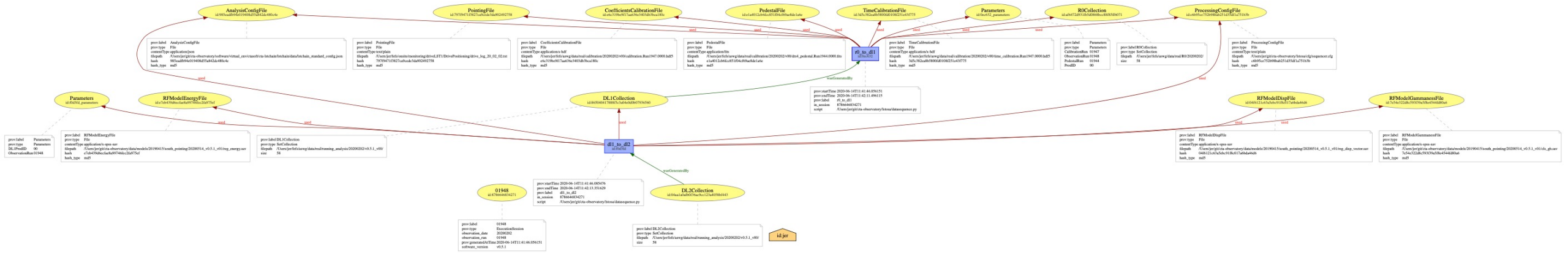
config/environment.yaml

```yaml
# Conda environment for provenance package
# conda env update -f environment.yaml

channels:
  - conda-forge

dependencies:
  - python
  - pyyaml
  - prov
  - pydot
  - pydotplus
  # dev dependencies
  - pytest
  - pytest-cov
  - black
  - isort
```

prov.log


```
INFO provLogger __PROV__2020-05-18T14:18:30.445713__PROV__{'session_id': 8739478486569, 'name': '01618', 'startTime': '2020-05-18T14:18:06.362321', 'system': {'executable':
'/fefs/aswg/software/virtual_env/anaconda3/envs/osa/bin/python', 'platform': {'architecture_bits': '64bit', 'architecture_linkage': '', 'machine': 'x86_64', 'processor':
'x86_64', 'node': 'cp15', 'version': '#1 SMP Thu Nov 8 23:39:32 UTC 2018', 'system': 'Linux', 'release': '3.10.0-957.el7.x86_64', 'libcver': "('glibc', '2.10')", 'num_cpus':
32, 'boot_time': '2020-03-24T03:48:43'}, 'python': {'version_string': '3.7.6 | packaged by conda-forge | (default, Mar 23 2020, 23:03:20) \n[GCC 7.3.0]', 'version': '3.7.6',
'compiler': 'GCC 7.3.0', 'implementation': 'CPython'}, 'environment': {'CONDA_DEFAULT_ENV': 'osa', 'CONDA_PREFIX': '/fefs/aswg/software/virtual_env/anaconda3/envs/osa',
'CONDA_PYTHON_EXE': '/fefs/aswg/software/virtual_env/anaconda3/bin/python', 'CONDA_EXE': '/fefs/aswg/software/virtual_env/anaconda3/bin/conda', 'CONDA_PROMPT_MODIFIER':
'(osa) ', 'CONDA_SHLVL': '2', 'PATH': '/local/home/lstanalyzer/usr/bin:/local/home/lstanalyzer/.local/bin:/fefs/aswg/software/virtual_env/anaconda3/envs/osa/bin:/fefs/aswg/
software/virtual_env/anaconda3/condabin:/usr/lib64/qt-3.3/bin:/usr/local/bin:/usr/bin:/usr/local/sbin:/usr/sbin:/opt/ibutils/bin:/local/home/lstanalyzer/.local/bin:/local/
home/lstanalyzer/bin', 'LD_LIBRARY_PATH': '/local/home/lstanalyzer/usr/lib:', 'DYLD_LIBRARY_PATH': None, 'USER': 'lstanalyzer', 'HOME': '/local/home/lstanalyzer', 'SHELL':
'/bin/bash'}, 'arguments': ['/fefs/aswg/lstosa/datasequence.py', '-c', 'cfg/sequencer_Nov2019_dragontime_v03.cfg', '-d', '2019_11_23', '--prod_id', 'v0.5.1_v03', '/fefs/
aswg/data/real/calibration/20191123/v03/calibration.Run1614.0000.hdf5', '/fefs/aswg/data/real/calibration/20191123/v03/drs4_pedestal.Run1611.0000.fits', '/fefs/aswg/data/
real/calibration/20191123/v03/time_calibration.Run1614.0000.hdf5', '/fefs/aswg/scripts-osa/corrected_drive_logs_Nov19/drive_log_19_11_23.txt', '0', '0', '0', '0', '--
stderr=sequence_LST1_01618_2432982.err', '--stdout=sequence_LST1_01618_2432982.out', '01618.0008', 'LST1'], 'start_time_utc': '2020-05-18T14:18:30.445684'},
'software_version': 'v0.5.1', 'observation_date': '20191123', 'observation_run': '01618', 'session_tag': 'r0_to_dl1:01618'}
INFO provLogger __PROV__2020-05-18T14:18:30.447360__PROV__{'activity_id': '621ca2', 'name': 'r0_to_dl1', 'startTime': '2020-05-18T14:18:06.362321', 'in_session':
8739478486569, 'agent_name': 'lstanalyzer', 'script': '/fefs/aswg/lstosa/datasequence.py', 'session_tag': 'r0_to_dl1:01618'}
INFO provLogger __PROV__2020-05-18T14:18:30.447499__PROV__{'activity_id': '621ca2', 'parameters': {'ObservationRun': '01618', 'ObservationSubRun': '0008', 'CalibrationRun':
'01614', 'PedestalRun': '01611', 'ProdID': '03'}, 'session_tag': 'r0_to_dl1:01618'}
INFO provLogger __PROV__2020-05-18T14:18:30.447628__PROV__{'entity_id': '446f45dd1c878559585395eedce5bc7a', 'name': 'R0SubrunDataset', 'filepath': '/fefs/aswg/data/real/
R0/20191123/LST-1.1.Run01618.0008.fits.fz', 'hash': '446f45dd1c878559585395eedce5bc7a', 'hash_type': 'md5', 'type': 'File', 'contentType': 'application/fits', 'session_tag':
'r0_to_dl1:01618'}
INFO provLogger __PROV__2020-05-18T14:18:30.447752__PROV__{'activity_id': '621ca2', 'used_id': '446f45dd1c878559585395eedce5bc7a', 'used_role': 'Observation subrun',
'session_tag': 'r0_to_dl1:01618'}
INFO provLogger __PROV__2020-05-18T14:18:30.447859__PROV__{'entity_id': '7404bb00748454d63badfe247c774a13', 'name': 'PedestalFile', 'filepath': '/fefs/aswg/data/real/
calibration/20191123/v03/drs4_pedestal.Run1611.0000.fits', 'hash': '7404bb00748454d63badfe247c774a13', 'hash_type': 'md5', 'type': 'File', 'contentType': 'application/fits',
'session_tag': 'r0_to_dl1:01618'}
INFO provLogger __PROV__2020-05-18T14:18:30.447966__PROV__{'activity_id': '621ca2', 'used_id': '7404bb00748454d63badfe247c774a13', 'used_role': 'Pedestal file',
'session_tag': 'r0_to_dl1:01618'}
INFO provLogger __PROV__2020-05-18T14:18:30.448065__PROV__{'entity_id': 'd8077e3bbdcb371f688ae65b0972e212', 'name': 'CoefficientsCalibrationFile', 'filepath': '/fefs/aswg/
data/real/calibration/20191123/v03/calibration.Run1614.0000.hdf5', 'hash': 'd8077e3bbdcb371f688ae65b0972e212', 'hash_type': 'md5', 'type': 'File', 'contentType':
'application/x-hdf', 'session_tag': 'r0_to_dl1:01618'}
INFO provLogger __PROV__2020-05-18T14:18:30.447752__PROV__{'session_id': 8773094569769, 'name': '01618', 'startTime': '2020-05-18T14:18:06.362323', 'system': {'executable':
'/fefs/aswg/software/virtual_env/anaconda3/envs/osa/bin/python', 'platform': {'architecture_bits': '64bit', 'architecture_linkage': '', 'machine': 'x86_64', 'processor':
'x86_64', 'node': 'cp15', 'version': '#1 SMP Thu Nov 8 23:39:32 UTC 2018', 'system': 'Linux', 'release': '3.10.0-957.el7.x86_64', 'libcver': "('glibc', '2.10')", 'num_cpus':
32, 'boot_time': '2020-03-24T03:48:43'}, 'python': {'version_string': '3.7.6 | packaged by conda-forge | (default, Mar 23 2020, 23:03:20) \n[GCC 7.3.0]', 'version': '3.7.6',
'compiler': 'GCC 7.3.0', 'implementation': 'CPython'}, 'environment': {'CONDA_DEFAULT_ENV': 'osa', 'CONDA_PREFIX': '/fefs/aswg/software/virtual_env/anaconda3/envs/osa',
'CONDA_PYTHON_EXE': '/fefs/aswg/software/virtual_env/anaconda3/bin/python', 'CONDA_EXE': '/fefs/aswg/software/virtual_env/anaconda3/bin/conda', 'CONDA_PROMPT_MODIFIER':
'(osa) ', 'CONDA_SHLVL': '2', 'PATH': '/local/home/lstanalyzer/usr/bin:/local/home/lstanalyzer/.local/bin:/fefs/aswg/software/virtual_env/anaconda3/envs/osa/bin:/fefs/aswg/
software/virtual_env/anaconda3/condabin:/usr/lib64/qt-3.3/bin:/usr/local/bin:/usr/bin:/usr/local/sbin:/usr/sbin:/opt/ibutils/bin:/local/home/lstanalyzer/.local/bin:/local/
home/lstanalyzer/bin', 'LD_LIBRARY_PATH': '/local/home/lstanalyzer/usr/lib:', 'DYLD_LIBRARY_PATH': None, 'USER': 'lstanalyzer', 'HOME': '/local/home/lstanalyzer', 'SHELL':
'/bin/bash'}, 'arguments': ['/fefs/aswg/lstosa/datasequence.py', '-c', 'cfg/sequencer_Nov2019_dragontime_v03.cfg', '-d', '2019_11_23', '--prod_id', 'v0.5.1_v03', '/fefs/
aswg/data/real/calibration/20191123/v03/calibration.Run1614.0000.hdf5', '/fefs/aswg/data/real/calibration/20191123/v03/drs4_pedestal.Run1611.0000.fits', '/fefs/aswg/data/
real/calibration/20191123/v03/time_calibration.Run1614.0000.hdf5', '/fefs/aswg/scripts-osa/corrected_drive_logs_Nov19/drive_log_19_11_23.txt', '0', '0', '0', '0', '--
stderr=sequence_LST1_01618_2432980.err', '--stdout=sequence_LST1_01618_2432980.out', '01618.0006', 'LST1'], 'start_time_utc': '2020-05-18T14:18:30.447727'},
'software_version': 'v0.5.1', 'observation_date': '20191123', 'observation_run': '01618', 'session_tag': 'r0_to_dl1:01618'}
INFO provLogger __PROV__2020-05-18T14:18:30.448191__PROV__{'activity_id': '621ca2', 'used_id': 'd8077e3bbdcb371f688ae65b0972e212', 'used_role': 'Coefficients calibration
file', 'session_tag': 'r0_to_dl1:01618'}
INFO provLogger __PROV__2020-05-18T14:18:30.448243__PROV__{'activity_id': '7e8065', 'name': 'r0_to_dl1', 'startTime': '2020-05-18T14:18:06.362323', 'in_session':
8773094569769, 'agent_name': 'lstanalyzer', 'script': '/fefs/aswg/lstosa/datasequence.py', 'session_tag': 'r0_to_dl1:01618'}
INFO provLogger __PROV__2020-05-18T14:18:30.448296__PROV__{'entity_id': 'fcf52d425d9033504c154f25986978c2', 'name': 'TimeCalibrationFile', 'filepath': '/fefs/aswg/data/real/
calibration/20191123/v03/time_calibration.Run1614.0000.hdf5', 'hash': 'fcf52d425d9033504c154f25986978c2', 'hash_type': 'md5', 'type': 'File', 'contentType': 'application/x-
hdf', 'session_tag': 'r0_to_dl1:01618'}
```

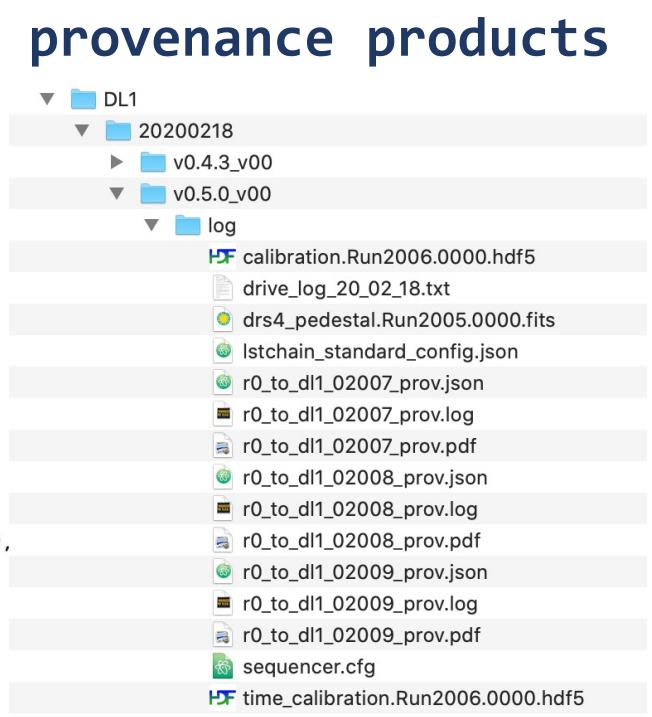
R0->DL1 data processing provenance graph

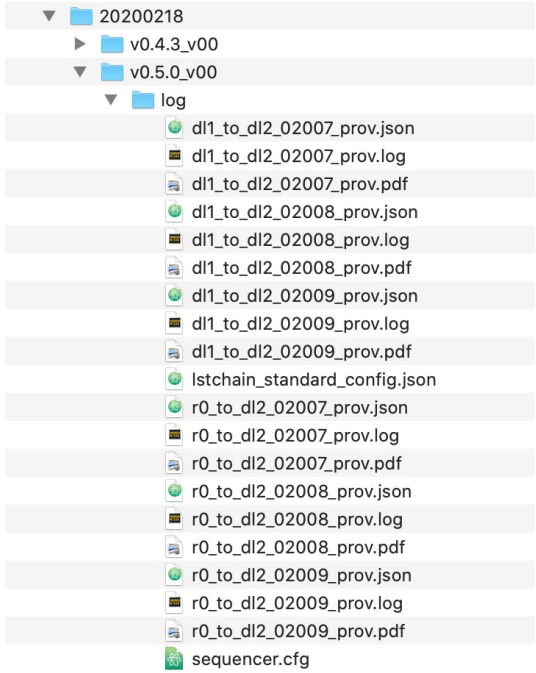


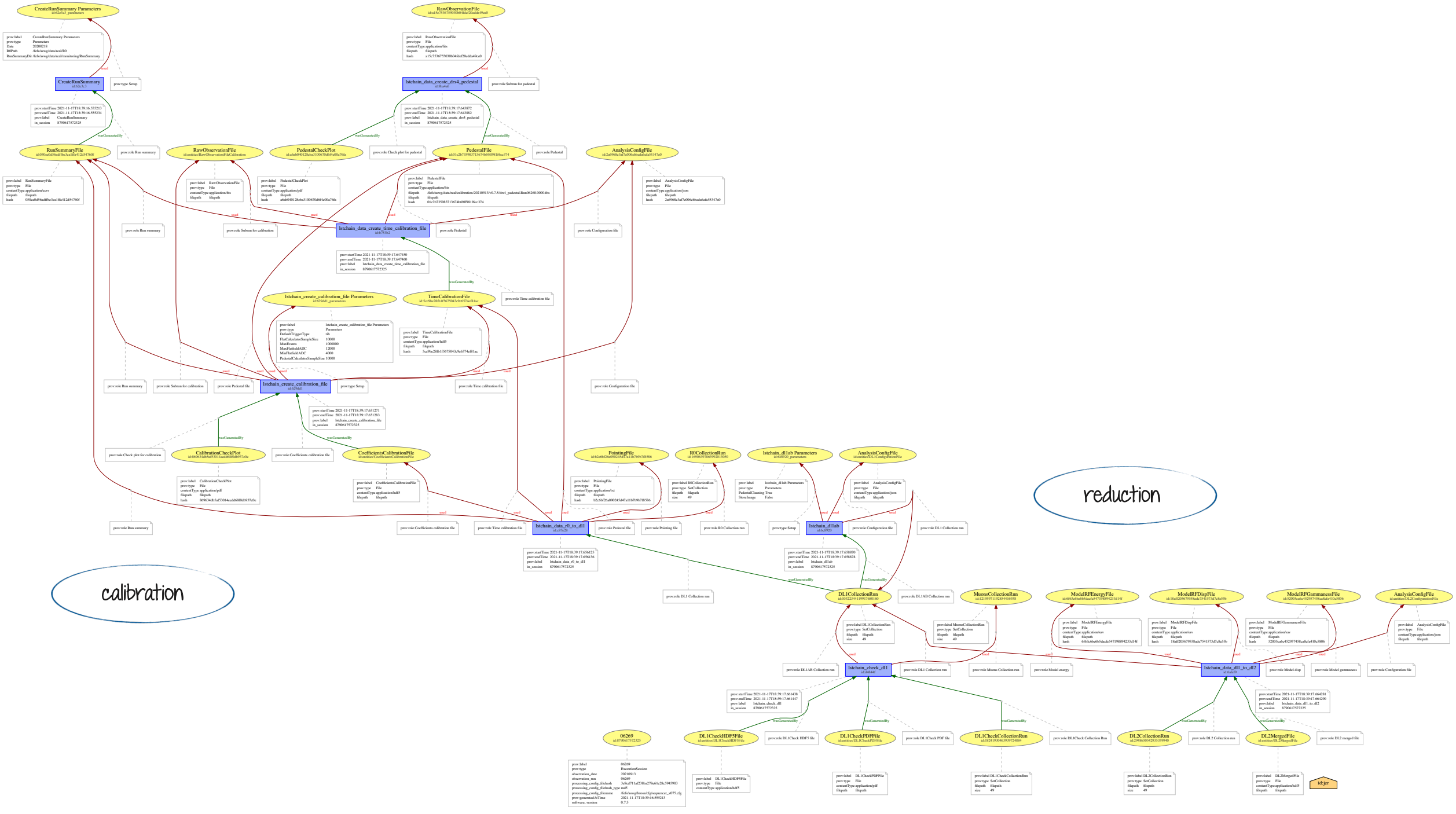
json serialization

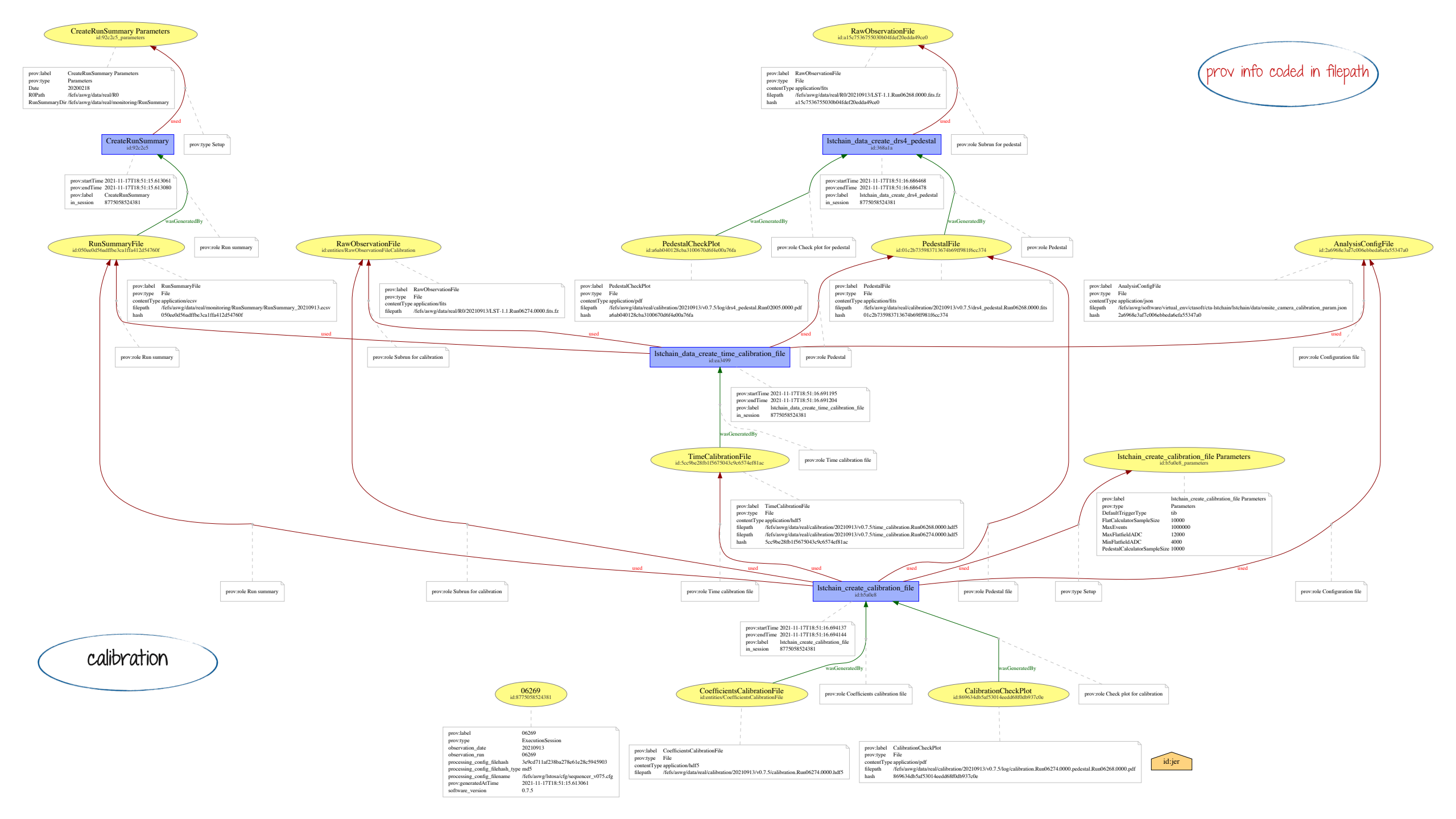

```
{
    "prefix": {
        "id": "id:",
        "default": "param:"
    },
    "entity": {
        "id:8756102484710": {
            "prov:label": "r0_to_dl1",
            "prov:type": "ExecutionSession",
            "prov:generatedAtTime": "2020-05-08T11:51:47.051025",
            "script": "/home/daniel.morcuende/lstosa/datasequence.py",
            "software_version": "v0.4.5.post566+gite395ef2",
            "observation_date": "20200118",
            "observation_run": "01832"
        },
        "id:4b1e22_parameters": {
            "ObservationRun": "01832",
            "CalibrationRun": "01831",
            "PedestalRun": "01830",
            "ProdID": "02",
            "prov:type": "Parameters",
            "prov:label": "Parameters"
        },
        "id:6d23620c4d92ef92b642a0116908fad1": {
            "prov:label": "PedestalFile",
            "prov:type": "File",
            "filepath": "/fefs/aswg/workspace/daniel.morcuende/data/real/calibration/20200118/v02/drs4_pedestal.Run1830.0000.fits",
            "hash": "6d23620c4d92ef92b642a0116908fad1",
            "hash_type": "md5",
            "contentType": "application/fits"
        },
        "id:58f0d6b71f12a4e8a73eaad09345d64c": {
            "prov:label": "CoefficientsCalibrationFile",
            "prov:type": "File",
            "filepath": "/fefs/aswg/workspace/daniel.morcuende/data/real/calibration/20200118/v02/calibration.Run1831.0000.hdf5",
            "hash": "58f0d6b71f12a4e8a73eaad09345d64c",
            "hash_type"
            "conte                    /x-hdf"
        },
        "id:a                                    5b": {
```

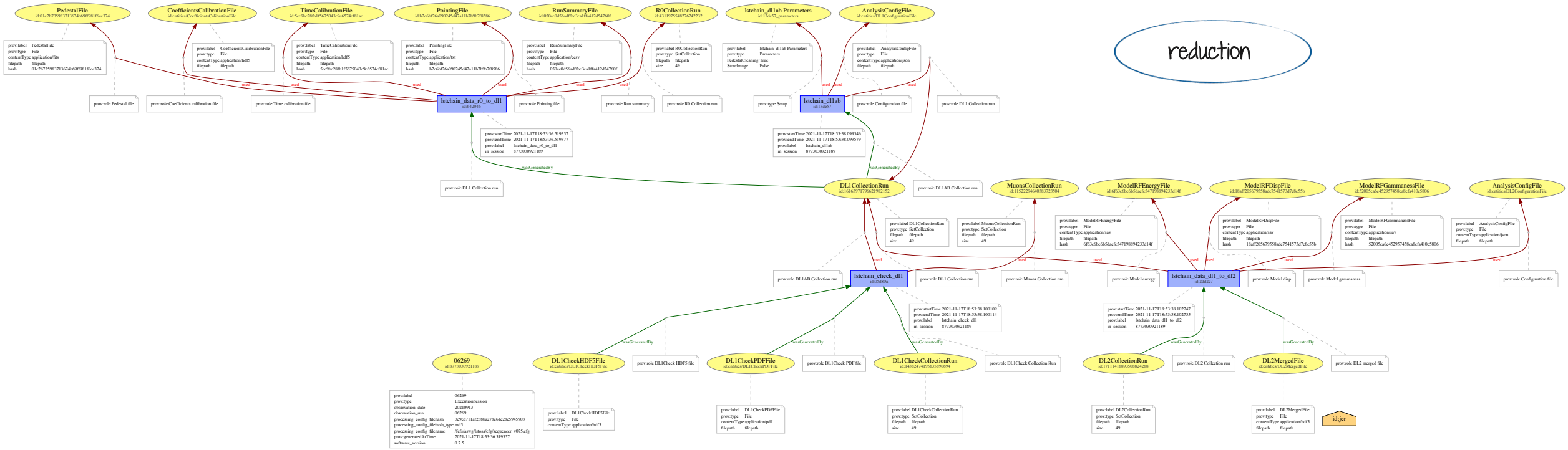

provenance products






CreateRunSummary Parameters
RawObservationFile
CreateRunSummary
lstchain_data_create_drs4_pedestal
RunSummaryFile
RawObservationFile
PedestalCheckPlot
PedestalFile
AnalysisConfigFile
lstchain_data_create_time_calibration_file
lstchain_create_calibration_file Parameters
TimeCalibrationFile
lstchain_create_calibration_file
CalibrationCheckPlot
CoefficientsCalibrationFile
PointingFile
R0CollectionRun
lstchain_dl1ab Parameters
AnalysisConfigFile
lstchain_data_r0_to_dl1
lstchain_dl1ab
DL1CollectionRun
MuonsCollectionRun
ModelRFEnergyFile
ModelRFDispFile
ModelRFGammanessFile
AnalysisConfigFile
lstchain_check_dl1
lstchain_data_dl1_to_dl2
06269
DL1CheckHDF5File
DL1CheckPDFFile
DL1CheckCollectionRun
DL2CollectionRun
DL2MergedFile
id:jer
calibration
reduction


CreateRunSummary Parameters
id:92c2c5_parameters
prov:label CreateRunSummary Parameters
prov:type Parameters
Date 20200218
R0Path /fefs/aswg/data/real/R0
RunSummaryDir /fefs/aswg/data/real/monitoring/RunSummary
used
CreateRunSummary
id:92c2c5
prov:type Setup
prov:startTime 2021-11-17T18:51:15.613061
prov:endTime 2021-11-17T18:51:15.613080
prov:label CreateRunSummary
in_session 8775058524381
wasGeneratedBy
RunSummaryFile
id:050ee0d56adffbe3ca1ffa412d54760f
prov:role Run summary
prov:label RunSummaryFile
prov:type File
contentType application/ecsv
filepath /fefs/aswg/data/real/monitoring/RunSummary/RunSummary_20210913.ecsv
hash 050ee0d56adffbe3ca1ffa412d54760f
prov:role Run summary
used
RawObservationFile
id:entities/RawObservationFileCalibration
prov:label RawObservationFile
prov:type File
contentType application/fits
filepath /fefs/aswg/data/real/R0/20210913/LST-1.1.Run06274.0000.fits.fz
prov:role Subrun for calibration
used
RawObservationFile
id:a15c7536755030b04fdef20edda49ce0
prov:label RawObservationFile
prov:type File
contentType application/fits
filepath /fefs/aswg/data/real/R0/20210913/LST-1.1.Run06268.0000.fits.fz
hash a15c7536755030b04fdef20edda49ce0
used
lstchain_data_create_drs4_pedestal
id:368a1a
prov:role Subrun for pedestal
prov:startTime 2021-11-17T18:51:16.686468
prov:endTime 2021-11-17T18:51:16.686478
prov:label lstchain_data_create_drs4_pedestal
in_session 8775058524381
wasGeneratedBy
PedestalCheckPlot
id:a6ab040128cba3100670d6f4e00a76fa
prov:role Check plot for pedestal
prov:label PedestalCheckPlot
prov:type File
contentType application/pdf
filepath /fefs/aswg/data/real/calibration/20210913/v0.7.5/log/drs4_pedestal.Run02005.0000.pdf
hash a6ab040128cba3100670d6f4e00a76fa
wasGeneratedBy
PedestalFile
id:01c2b735983713674b69ff981f6cc374
prov:role Pedestal
prov:label PedestalFile
prov:type File
contentType application/fits
filepath /fefs/aswg/data/real/calibration/20210913/v0.7.5/drs4_pedestal.Run06268.0000.fits
hash 01c2b735983713674b69ff981f6cc374
AnalysisConfigFile
id:2a6968e3af7c006ebbeda6efa55347a0
prov:label AnalysisConfigFile
prov:type File
contentType application/json
filepath /fefs/aswg/software/virtual_env/ctasoft/cta-lstchain/lstchain/data/onsite_camera_calibration_param.json
hash 2a6968e3af7c006ebbeda6efa55347a0
used
lstchain_data_create_time_calibration_file
id:ea3499
prov:role Pedestal
prov:role Configuration file
prov:startTime 2021-11-17T18:51:16.691195
prov:endTime 2021-11-17T18:51:16.691204
prov:label lstchain_data_create_time_calibration_file
in_session 8775058524381
wasGeneratedBy
TimeCalibrationFile
id:5cc9be28fb1f5675043c9c6574ef81ac
prov:role Time calibration file
prov:label TimeCalibrationFile
prov:type File
contentType application/hdf5
filepath /fefs/aswg/data/real/calibration/20210913/v0.7.5/time_calibration.Run06268.0000.hdf5
filepath /fefs/aswg/data/real/calibration/20210913/v0.7.5/time_calibration.Run06274.0000.hdf5
hash 5cc9be28fb1f5675043c9c6574ef81ac
lstchain_create_calibration_file Parameters
id:b5a0e8_parameters
prov:label lstchain_create_calibration_file Parameters
prov:type Parameters
DefaultTriggerType tib
FlatCalculatorSampleSize 10000
MaxEvents 1000000
MaxFlatfieldADC 12000
MinFlatfieldADC 4000
PedestalCalculatorSampleSize 10000
used
lstchain_create_calibration_file
id:b5a0e8
prov:role Run summary
prov:role Subrun for calibration
prov:role Time calibration file
prov:role Pedestal file
prov:type Setup
prov:role Configuration file
prov:startTime 2021-11-17T18:51:16.694137
prov:endTime 2021-11-17T18:51:16.694144
prov:label lstchain_create_calibration_file
in_session 8775058524381
wasGeneratedBy
CoefficientsCalibrationFile
id:entities/CoefficientsCalibrationFile
prov:role Coefficients calibration file
prov:label CoefficientsCalibrationFile
prov:type File
contentType application/hdf5
filepath /fefs/aswg/data/real/calibration/20210913/v0.7.5/calibration.Run06274.0000.hdf5
wasGeneratedBy
CalibrationCheckPlot
id:869634db5af53014eedd68f0dh937c0e
prov:role Check plot for calibration
prov:label CalibrationCheckPlot
prov:type File
contentType application/pdf
filepath /fefs/aswg/data/real/calibration/20210913/v0.7.5/log/calibration.Run06274.0000.pedestal.Run06268.0000.pdf
hash 869634db5af53014eedd68f0dh937c0e
06269
id:8775058524381
prov:label 06269
prov:type ExecutionSession
observation_date 20210913
observation_run 06269
processing_config_filehash 3e9cd711af238ba278e61e28c5945903
processing_config_filehash_type md5
processing_config_filename /fefs/aswg/lstosa/cfg/sequencer_v075.cfg
prov:generatedAtTime 2021-11-17T18:51:15.613061
software_version 0.7.5
prov info coded in filepath
calibration
id:jer


reduction
PedestalFile
id:01c2b7359837136748e9ff9818fcc374
prov:label PedestalFile
prov:type File
contentType application/fits
filepath filepath
hash 01c2b7359837136748e9ff9818fcc374
prov:role Pedestal file
CoefficientsCalibrationFile
id:entities/CoefficientsCalibrationFile
prov:label CoefficientsCalibrationFile
prov:type File
contentType application/hdf5
filepath filepath
prov:role Coefficients calibration file
TimeCalibrationFile
id:5cc9be28fb1f5675043c9c6574ef81ac
prov:label TimeCalibrationFile
prov:type File
contentType application/hdf5
filepath filepath
hash 5cc9be28fb1f5675043c9c6574ef81ac
prov:role Time calibration file
PointingFile
id:b2c6bf26a090245d47a11b7b9b7f8586
prov:label PointingFile
prov:type File
contentType application/txt
filepath filepath
hash b2c6bf26a090245d47a11b7b9b7f8586
prov:role Pointing file
RunSummaryFile
id:050eed0d56adffbe3ca1ffa412d54760f
prov:label RunSummaryFile
prov:type File
contentType application/ecsv
filepath filepath
hash 050eed0d56adffbe3ca1ffa412d54760f
prov:role Run summary
R0CollectionRun
id:4311975548276242232
prov:label R0CollectionRun
prov:type SetCollection
filepath filepath
size 49
prov:role R0 Collection run
lstchain_data_r0_to_dl1
id:b42046
prov:startTime 2021-11-17T18:53:36.519357
prov:endTime 2021-11-17T18:53:36.519377
prov:label lstchain_data_r0_to_dl1
in_session 877303092189
prov:role DL1 Collection run
lstchain_dl1ab Parameters
id:13de57_parameters
prov:label lstchain_dl1ab Parameters
prov:type Parameters
PedestalCleaning True
StoreImage False
prov:type Setup
lstchain_dl1ab
id:13de57
prov:startTime 2021-11-17T18:53:38.099546
prov:endTime 2021-11-17T18:53:38.099579
prov:label lstchain_dl1ab
in_session 877303092189
AnalysisConfigFile
id:entities/DL1ConfigurationFile
prov:label AnalysisConfigFile
prov:type File
contentType application/json
filepath filepath
prov:role Configuration file
prov:role DL1 Collection run
DL1CollectionRun
id:161639717966219821521 52
prov:role DL1 Collection run
prov:role DL1AB Collection run
MuonsCollectionRun
id:115222946403837235 04
prov:label DL1CollectionRun
prov:type SetCollection
filepath filepath
size 49
prov:label MuonsCollectionRun
prov:type SetCollection
filepath filepath
size 49
ModelRFEnergyFile
id:6f63c6be6b5dacfc54719889423 3d14f
prov:label ModelRFEnergyFile
prov:type File
contentType application/xav
filepath filepath
hash 6f63c6be6b5dacfc54719889423 3d14f
ModelRFDispFile
id:18aff205679558ade7541573d7c8c55b
prov:label ModelRFDispFile
prov:type File
contentType application/xav
filepath filepath
hash 18aff205679558ade7541573d7c8c55b
ModelRFGammanessFile
id:52005ca6c452957458ca8cfa410c5806
prov:label ModelRFGammanessFile
prov:type File
contentType application/xav
filepath filepath
hash 52005ca6c452957458ca8cfa410c5806
AnalysisConfigFile
id:entities/DL2ConfigurationFile
prov:label AnalysisConfigFile
prov:type File
contentType application/json
filepath filepath
prov:role DL1AB Collection run
lstchain_check_dl1
id:058f0b
prov:role DL1 Collection run
prov:role Muons Collection run
prov:role Model energy
lstchain_data_dl1_to_dl2
id:2dd2c7
prov:role Model disp
prov:role Model gammaness
prov:role Configuration file
prov:startTime 2021-11-17T18:53:38.100109
prov:endTime 2021-11-17T18:53:38.100114
prov:label lstchain_check_dl1
in_session 877303092189
prov:startTime 2021-11-17T18:53:38.102747
prov:endTime 2021-11-17T18:53:38.102755
prov:label lstchain_data_dl1_to_dl2
in_session 877303092189
06269
id:877303092189
prov:label 06269
prov:type ExecutionSession
observation_date 20210913
observation_run 06269
processing_config_filehash 3e9cdf711af238ba278e61e28c5945903
processing_config_filehash_type md5
processing_config_filename /lfs/ismrp/htoxa/cfg/sequencer_v075.cfg
prov:generatedAtTime 2021-11-17T18:53:36.519357
software_version 0.7.5
DL1CheckHDF5File
id:entities/DL1CheckHDF5File
prov:label DL1CheckHDF5File
prov:type File
contentType application/hdf5
prov:role DL1Check HDF5 file
DL1CheckPDFFile
id:entities/DL1CheckPDFFile
prov:label DL1CheckPDFFile
prov:type File
contentType application/pdf
filepath filepath
prov:role DL1Check PDF file
DL1CheckCollectionRun
id:143824741958358966 94
prov:label DL1CheckCollectionRun
prov:type SetCollection
filepath filepath
size 49
prov:role DL1Check Collection Run
DL2CollectionRun
id:171114188935088242 88
prov:label DL2CollectionRun
prov:type SetCollection
filepath filepath
size 49
prov:role DL2 Collection run
DL2MergedFile
id:entities/DL2MergedFile
prov:label DL2MergedFile
prov:type File
contentType application/hdf5
filepath filepath
prov:role DL2 merged file
id:jer
used
wasGeneratedBy

Lessons learnt

- Continuous modifications on workflow and additions of captured info – **flexible model and implementation**
- **Configuration** may be stored as metadata attached to datasets (***fits headers/hdf5 attributes***)
- **Capturing relationships** among automatized activities and entities (*datasets*) needs a provenance model
- **Independent prov capture for the different dependent software** packages

  lstosa requires lstchain requires ctapipe

  lstosa requires calibration products produced in containerized sub pipeline

- Massive logging subrun-wise makes a very poor performance because of **huge usage of disk**
- **Post-processing** of prov info is needed to produce run-wise info and different **levels of granularity on demand**
- **Structured logging** in *small* text files may be a solution for provenance sessions storage
- Use of ***graphs with access-links in browser*** for provenance inspection and data access

- Considering storing provenance in a RDBMS or *temptatively* in a noSQL **database** (i.e. *mongo + json)*
- Improvement of capture mechanism for **execution environment**/grid nodes configuration is needed
- Development of a **provenance query mechanism** for detailed analysis and inspection is needed
- Why IVOA syntax and model for low-level processing provenance capture?