EPJ Web of Conferences **214**, 04053 (2019) https://doi.org/10.1051/epjconf/201921404053
CHEP 2018

Final Analysis Sample Metadata

Patrick Meade[1,A]

[1]University of Wisconsin-Madison, WIPAC, 222 W Washington Ave. Suite 500, Madison, WI 53703

Abstract. IceCube is a cubic kilometer neutrino detector located at the South Pole. Data are processed and filtered in a data center at the site. After transfer to a data warehouse in the northern hemisphere, data are further refined through multiple levels of selection and reconstruction to generate analysis samples. So far, the production and curation of these analysis samples has been handled in an ad-hoc way in IceCube. New tools have been developed to capture and validate the metadata associated to these data samples in consistent and machine readable specifications. Development was driven by analysis use-cases and pursuing reproducibility of scientific results.

1 Introduction

The IceCube detector [1] is located at the geographic South Pole and was completed at the end of 2010. It consists of 5160 optical sensors buried between 1450 and 2450 meters below the surface of the South Pole ice sheet and is designed to detect interactions of neutrinos of astrophysical origin.

Data taken from the detector are stored in a data warehouse at the University of Wisconsin-Madison. Scientific Working Groups perform analyses and create derivative data products (i.e. Level 2, Level 3, Level 4) that become analysis samples. Some of these analysis samples contain the data that is used to publish results in scientific papers.

The tracking of analysis samples within working groups has been an ad-hoc process within IceCube. At times, it has been difficult to track down the data and code that were used to generate a particular result. This is not conducive to reproducibility, a key requirement for all results across every scientific discipline.

We created a Final Analysis Sample registration service to provide a central and machine readable record of analysis samples created by working groups. Our goal was to ensure that every result is reproducible by identifying the data and code that were used to generate that result.

2 Methods

We built a web form where physicists can register their analysis data. Some sample level metadata fields are collected, along with metadata for each subsample. The sample level fields collected by the web form are:

[A] Corresponding author: patrick.meade@icecube.wisc.edu

- Sample Name
- Sample Description
- Sample Authors
    - Author Name
    - Author Email
- Code URL (to version control)
- Documentation URL (to wiki)
- Publication DOIs
- Tags

The metadata collected for subsamples are:

- Subsamples
    - Parent Data File List
    - Data Source: Experimental or Simulation
    - Systematics Name
    - Child Data Directory Path

Physicists fill the web form with the appropriate metadata and then click a large button labeled Register at the bottom of the form. After validating the input data, the registration service begins to register the provided analysis sample.

2.1 Subsample Registration

The service ensures that every file specified in the subsample is registered with a File Catalog service in IceCube. After all of the files are registered, the input files are grouped into collections, which are files associated by a query. The service then registers a snapshot of the collection with the File Catalog, to ensure the exact set of input files is always reproducible.

Parent data files represent input to the analysis. These can be experimental or simulation data. It is expected that all such files will already exist in the data warehouse at UW-Madison. The physicist provides a text file containing all the data warehouse paths of the input file set. These are then registered as collections and snapshots as described above.

Child data files represent the output of the analysis. It is assumed that the physicist ran the analysis code on the parent data to produce the child data. The child data is thus assumed to be in a working directory, and not yet registered in the data warehouse. The registration service copies the child data files to the canonical location for analysis samples in the data warehouse. These are then registered as collections and snapshots as described above.

2.2 Metadata Generation

After registering an analysis sample, the child data is stored in the canonical location for analysis samples in the data warehouse. The service also places a metadata file called sample.json in the canonical directory.

An example directory structure looks like this:

```
/mnt/data/fas/canon/my_cool_analysis
.
├── real
│   ├── My-Cool-Result-1.hdf5
│   ├── My-Cool-Result-2.hdf5
│   ├── My-Cool-Result-3.hdf5
├── sample.json
└── sim
    ├── baseline
    │   ├── My-Cool-Result-1.hdf5
    │   ├── My-Cool-Result-2.hdf5
    │   ├── My-Cool-Result-3.hdf5
    └── ice_model_2
        ├── My-Cool-Result-1.hdf5
        ├── My-Cool-Result-2.hdf5
        └── My-Cool-Result-3.hdf5
```

The top level directories real and sim separate experimental and simulation data. Under the sim directory, subdirectories identify the different systematics. In our example above, there is a baseline simulation, along with a simulation that uses a different ice model.

The Final Analysis Sample registration service creates a metadata file named sample.json to describe the registered analysis sample. An example file looks like this:

```
{
  "uuid": "a5cf7435-3e1a-4d0b-a098-2589bcdd1efe",
  "version": 1,
  "sample_name": "my_cool_analysis",
  "summary": "This is a description for the my_cool_analysis sample.",
  "code_url": "https://svn.icecube.wisc.edu",
  "doc_url": "https://wiki.icecube.wisc.edu",
  "quality": "good",
  "tags": [
    "my", “cool”, “analysis”
  ],
  "authors": [
    {
      "authorName": "Patrick Meade",
      "authorEmail": "pmeade@icecube.wisc.edu"
    }
  ],
  "subsamples": [
    {
      "dataType": "real",
      "parent_snapshot_uuid": "231d2744-cdbf-11e8-968d-525400ad3b43",
      "child_snapshot_uuid": "235a88aa-cdbf-11e8-968d-525400ad3b43"
    },
    {
      "dataType": "sim",
      "systematicsName": "baseline",
      "simType": "baseline",
      "parent_snapshot_uuid": "23fd6386-cdbf-11e8-968d-525400ad3b43",
      "child_snapshot_uuid": "2442afc2-cdbf-11e8-968d-525400ad3b43"
    },
    {
      "dataType": "sim",
      "systematicsName": "ice_model_2",
      "simType": "systematics",
      "parent_snapshot_uuid": "24d4891a-cdbf-11e8-968d-525400ad3b43",
      "child_snapshot_uuid": "250e664e-cdbf-11e8-968d-525400ad3b43"
    }
  ],
  "fas_version": "0.0.19"
}
```

3 Discussion

The Final Analysis Sample registration service is still under development. The current version is implemented in 1000 lines of CoffeeScript[2]. Input from physicists was solicited for the development of the form. Further rounds of use and review are expected to occur in the near future.

Acknowledgements

I'd like to thank Vladimir Brik, Gonzalo Merino, and David Schultz for their feedback on draft versions of this paper.

References

[1] The IceCube Neutrino Observatory: Instrumentation and Online Systems - IceCube Collaboration (Aartsen, M.G. et al.) JINST 12 (2017) no.03, P03012 arXiv:1612.05093 [astro-ph.IM]

[2] CoffeeScript project, “CoffeeScript” [software], version 2.3.2, 2018. Available from https://github.com/jashkenas/coffeescript/tarball/2.3.2 [accessed 2018-10-15]