

smalldata_tools: Analysis tools for aligned data

- [Analysis on S3DF](#)
 - [How do I login to S3Df? How do I open a Jupyter session?](#)
 - [Where do I find the smalldata_tools code and my data?](#)
 - [Make experiment folder accessible from JupyterHub session](#)
 - [Access data](#)
- [Smalldata analysis workflow](#)
- [Old analysis infrastructure \(data taken before 2023\)](#)
 - [Online and offline analysis](#)
 - [Online analysis](#)
 - [Offline analysis](#)
 - [How do I access the relevant computing resources?](#)
 - [Working directories](#)
 - [Access data](#)
 - [JupyterHub](#)
 - [Make experiment folder accessible from JupyterHub session](#)
 - [Advanced topics](#)

This page describes smalldata_tools, a suite of code useful for analysis from the xtc data to small(er) hdf5 files at several stages of analysis. While this page is written with smalldata_tools usage in mind, the information here are also of interest to understand the computing infrastructure, resources and directory structure.

The smalldata_tools code can be found on git at https://github.com/slac-lcls/smalldata_tools.

At XPP or XCS, the code setup is usually taken care off by the beam line staff. For other hutches, please contact the controls POC or pcids-poc-l

At MEC, we have a default setup that will produce single tiff files for each detector in directories in the experiments scratch directory, along with an hdf5 file /run. This is further described [here](#).

Analysis on S3DF

How do I login to S3Df? How do I open a Jupyter session?

LCLS S3DF information: [Running at S3DF](#)

Where do I find the smalldata_tools code and my data?

Analysis code and results should be stored in

```
/sdf/data/lcls/ds/<hutch>/<experiment>/results/.
```

Hence, the smalldata_tools code is setup in:

```
/sdf/data/lcls/ds/<hutch>/<exname>/results/smalldata_tools
```

Make experiment folder accessible from JupyterHub session

In Jupyter hub, you can only navigate within your home folder. It is thus recommended to create shortcuts (soft-links) to the relevant experiment folders, for ease-of-access.

From Jupyter hub, click on the "+" symbol on the top left. Select "terminal" and make a soft-link to the experiment folder:

```
ln -s /sdf/data/lcls/ds/<hutch>/<experiment>/ ./<link_name>
```

Access data

The hdf5 data will be written to:

```
/sdf/data/lcls/ds/<hutch>/<experiment>/hdf5/smalldata
```

Smalldata analysis workflow

The analysis is generally split in two steps, allowing for easy diagnostics and customization of the analysis process. Please contact your controls and data POC to assess the best approach for your experiment.

- The first step is the generation of the "small data" file, the colloquial name for run-based hdf5 files which contain different data arrays where the first dimension is the number of events (shot-to-shot information retained). This production can be run automatically on each new run so that the data is available only a few minutes after the run has ended. It can also be run on request in case you want to tweak the data extraction.

Processing of the area detector can be configured at this stage, performing operation such as extracting a region of interest, azimuthal integration, photon counting, etc. It is non-recommended to save full large area detector data at this step. The following pages describe this in more details:

[Generation of small hdf5 files](#)

[Configuration of SmallData](#)

[Adding Data from area detectors](#)

- The second stage depends much more on the type of experiment. Different options are available:
 - Binning of the full detector images can be performed by setting up the cube analysis, which will return a h5 file of binned data and images, resulting in a relatively light-weight file. While the shot-to-shot information is lost at this point, this approach is generally recommended, as it is more carefree and does not require to delve into the details of the binning procedure. It is also almost mandatory in cases where the analysis of the full image is needed (Q-resolved diffused scattering analysis, for example). Note that the shot-to-shot information remains readily available from the file produced in the first step (without the area detector data). Details on the cube workflow are given here: [Cube production](#)
 - Adapt one of the templated analysis notebooks to suit the current experiment needs. These custom templates have been made for the more common experiments performed at different endstations at LCLS and are available at `/reg/g/psdm/sw/tools/smalldata_tools/example_notebooks` (please refrain from modifying these released notebooks in place). This approach works well for lightweight data analysis, for which the area detector images are reduced to a single (or few) number (integration of a ROI, azimuthal binning, for example) in the first step. It is also suited when detailed shot-to-shot information needs to be examined, and full control over the data binning process is desired. Documentation on the example notebooks can be found here: [Example notebooks](#).

The contents of the smallData files are described here

[smallData Contents](#)

```
#####  
#####  
  
#####  
#####  
  
#####  
#####
```

Old analysis infrastructure (data taken before 2023)

Online and offline analysis

Two analysis infrastructures comprising of various queues and interactive nodes, are available to use depending on the status of the experiment.

Online analysis

Ongoing experiment are generally using the online analysis infrastructure, the fast feedback system (ffb). More info on the system here: [Fast Feedback System](#)

This system is faster and provides prioritization to ongoing experiments. Some time after the experiment is over, the access to the data will be locked and only the offline system will be available.

Offline analysis

After the experiment is over, the data and smalldata production code are moved to the offline system, the anafs. This system available for analysis indefinitely and can be used to reprocess or refine the data.

How do I access the relevant computing resources?

Often one can work exclusively from the JupyterHub interface (see below). At times it can nonetheless be useful to be able to access the relevant computing system and directories via a terminal.

```
ssh -X <ACCOUNT>@pslogin.slac.stanford.edu
```

If using NoMachine, login to `psnxserv.slac.stanford.edu`

For the online analysis:

```
ssh -X psffb
source /reg/g/psdm/etc/psconda.sh -py3 # Environment to use psana, etc
```

And for the offline analysis:

```
ssh -X psana
source /reg/g/psdm/etc/psconda.sh -py3 # Environment to use psana, etc
```

Working directories

The working directory structure can be confusing, as some of the offline folders are mounted and accessible from the online system. **As a rule of thumb, until things are moved away from the online system, one should exclusively work on the ffb.**

The results folder should hosts most of users' code, notebooks, etc. This folder lives on the offline system but is mounted on the ffb:

```
/cds/data/psdm/<hutch>/<experiment>/results/.
```

The smalldata_tools working directory generally are:

```
/cds/data/psdm/<hutch>/<expname>/results/smalldata_tools
```

for the offline system (psana) and

```
/cds/data/drpsrcf/<hutch>/<expname>/scratch/smalldata_tools
```

for the fast feedback system (psffb).

Access data

The data will be written to:

When using the FFB processing, data are written to:

```
/cds/data/drpsrcf/<hutch>/<experiment>/scratch/hdf5/smalldata
```

and to:

```
/cds/data/psdm/<hutch>/<experiment>/hdf5/smalldata
```

for the processing using the 'SLAC' endpoint / the psana system. Data will be moved from the FFB system to this directory within 3-4 weeks after the experiment has ended.

JupyterHub

General information about JupyterHub at LCLS: [JupyterHub](#)

When starting a JupyterHub server, one can choose to run the server either on psana or on the ffb.

If you have an error 511 when trying to access the server, please run

```
/reg/g/psdm/sw/jupyterhub/psjhub/jhub/generate-keys.sh
```

Make experiment folder accessible from JupyterHub session

In Jupyter hub, you can only navigate within your home folder. It is thus recommended to create shortcuts (soft-links) to the relevant experiment folders, for ease-of-access.

From Jupyter hub, click on the "+" symbol on the top left. Select "terminal" and make a soft-link to the experiment folder:

```
ln -s /cds/data/psdm/<hutch>/<experiment>/ ./<link_name>
```

If the experiment is going to make use of the FFB, make a second soft-link:

```
ln -s /cds/data/drpsrcf/<hutch>/<experiment>/ ./<link_name>
```

Advanced topics

The results folder is backed-up and for that reason can only hold-up to 10'000 files after which a quota exceeded error will pop up. Users' who wish to build code with more files should do it on the scratch folder (online: /cds/data/drpsrcf/<hutch>/<experiment>/scratch), where there is no file limits (but no back-up).