Using SAM to get to the Data and using the Grid to process jobs

a.k.a.

How to give up your happiness to find out just how deep the rabbit hole goes

Roberto Acciarri

LArIAT Working Meeting
September 18th, 2015

Disclaimers

- Don't take this as a definitive set of instructions. We are finding and fixing problems on the way. Things evolve rapidly, so this talk shows how to get around the problems we know now

- Matter of fact, I'm not a “software expert”. I just happened to be in need of using these tools and I shamelessly hassled half of Fermilab to help me get unstuck. So, it would be nice to share with you what I've learned to prevent you getting stuck in the same problems

- Most of the information in this talk come from DocDB and great help from Jason, Pawel, Johnny, Tingjun and Mike Kirby

Setting up the environment

- I will assume you all know how to setup your working environment. If that's not the case, you can look at Jonathan's tutorial in DocDB 1590-v1 and use the script outlined in his slides

- We are going to do few additions to his script and, since some important modification to Reco.fcl and into lariatutil and larbatch happened just yesterday (Thursday), it is better to work on the latest and most updated version of lariatsoft

Setting up the environment

- In the first part of the script, among the setup commands, add this lines:

  setup sam_web_client

  export SAM_EXPERIMENT=lariat

- At the end of the script add instead these lines:

  cd $MRB_TOP

  setup lariatutil v01_24_01 -q e7:debug

  get-cert

- In this was you will have both SAM and project.py ready to go and the latest versions of lariatsoft, lariatutil and larbatch

Using SAM

- Complete overview on how to use SAM and create a dataset are both on DocDB 1570-v2 (Jason's talk) and on the wiki page: https://cdcvs.fnal.gov/redmine/projects/lardbt/wiki/LArIAT_Data_Handling

- The following couple of slides give a quick list of useful commands to retrieve information from SAM and create a SAM definition, which can be used later in the process of job submission to the Grid

Using SAM

- samweb list-files "run_number 6366"

returns list of file names (no path) beloning to the run

- samweb locate-file lariat_r006326_sr0907.root

return the location of the specified file

- samweb list-definitions

return the list of all definitions created

- samweb describe-definition Good64GevPosSecondary

- samweb count-files "defname: Good64GevPosSecondary and tertiary.magnet_current > 95"

Useful commands as the definition you plan to create may already exist!

Create a SAM definition

- samweb create-definition MyDefinitionName "MyDefinitionConditions"

For istance:

samweb create-definition Good64GeVPosSecondary "secondary.intensity > 2000 and secondary.momentum > 60 and secondary.momentum < 68 and secondary.polarity Positive"

It is also possible to create definition of definitions:

- samweb create-definition TPC_OnAndReadOutNominal "defname: TPC_voltages_nominal and defname: TPC_MaxGainAndFilter and defname: TPC_nominal_read_out"

And if you don't remember a command, no problem!!!

- samweb help-commands

Job submission to Grid

- The wiki page contains instruction on how to submit and check jobs on the Grid: https://cdcvs.fnal.gov/redmine/projects/lardbt/wiki/Running_Grid_Jobs_using_projectpy

- Project.py is a phyton script which takes care of the grid setup for you. It requires a configuration xml file

- An example of xml file you can use is /lariat/app/users/acciarri/Example/TestBatch.xml

Job submission to Grid

TestBatch.xml

```xml
<?xml version="1.0"?>

<!-- Production Project -->

<!DOCTYPE project [
<!ENTITY release "v01_08_02">
<!ENTITY file_type "data">
<!ENTITY run_type "physics">
<!ENTITY name "Test">
<!ENTITY tag "datatest">
]>

<project name="&name;">

  <!-- Project size -->
  <numevents>1000000</numevents>

  <!-- Operating System -->
  <os>SL6</os>

  <!-- Batch resources -->
  <resource>DEDICATED,OPPORTUNISTIC</resource>

  <!-- Larsoft information -->
  <larsoft>
    <tag>&release;</tag>
    <qual>e7:debug</qual>
    <local>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml/localproducts.tar</local>
  </larsoft>

  <!-- Project stages -->

  <stage name="slicer">
```

The lariatsoft version set up

"data" or "MC"

Max number of event reconstructed. Make it big!

The qualifier of the set up lariatsoft version

The full path to your localProduct directory or, BETTER, the full path to a tarball of such directory

Job submission to Grid

```
    <qual>e7:debug</qual>
    <local>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml/localproducts.tar</local>
  </larsoft>
```

- The <local> field should contain the path to your LocalProducts area. Unfortunately, not all the Grid nodes can access that area, with the result that some of the jobs will fail if they land on the wrong node.

SOLUTION: create a tarball file of such directory and pass the full path to it to <local> </local>

- Create the tarball using the command

  /grid/fermiapp/products/lariat/lariatutil/<*YourVersion*>/bin/make_tar_lariat.sh -d <*YourLocalProductDirectory*> <*name_of_the_output.tar*>

Job submission to Grid

TestBatch.xml

```
31  <!-- Project stages -->
32
33  <stage name="slicer">
34    <fcl>SlicerInput.fcl</fcl>
35    <outdir>/pnfs/lariat/scratch/users/acciarri/&release;/&tag;/slicer</outdir>
36    <workdir>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml/&tag;_slicer_wordir</workdir>
37    <inputlist>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml/Filelist/test_list.txt</inputlist>
38    <!--    <inputdef>Run6366</inputdef> -->
39    <numjobs>2</numjobs>
40    <!--    <datatier>2</datatier> -->
41    <!--    <defname>&name;_&tag;_sliced</defname> -->
42  </stage>
43
44  <stage name="filter">
45    <fcldir>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml</fcldir>
46    <fcl>Filter.fcl</fcl>
47    <outdir>/pnfs/lariat/scratch/users/acciarri/&release;/&tag;/filter</outdir>
48    <workdir>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml/&tag;_filter_wordir</workdir>
49    <inputlist>/pnfs/lariat/scratch/users/acciarri/&release;/&tag;/slicer/files.list</inputlist>
50    <numjobs>3</numjobs>
51    <!--        <defname>&name;_&tag;_sliced</defname> -->
52  </stage>
53
54  <stage name="reco">
55    <fcldir>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml</fcldir>
56    <fcl>Reco_Slicer.fcl</fcl>
57    <outdir>/pnfs/lariat/scratch/users/acciarri/&release;/&tag;/reco</outdir>
58    <workdir>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml/&tag;_reco_wordir</workdir>
59    <inputlist>/pnfs/lariat/scratch/users/acciarri/&release;/&tag;/slicer/files.list</inputlist>
60    <numjobs>2</numjobs>
61    <!--        <defname>&name;_&tag;_sliced</defname> -->
62  </stage>
63
```

It's good practice to divide the full reconstruction in stages. Stages are run consecutively with the output of the first being passed to the second one and so on

Job submission to Grid

TestBatch.xml

```
31  <!-- Project stages -->
32
33  <stage name="slicer">
34    <fcl>SlicerInput.fcl</fcl>
35    <outdir>/pnfs/lariat/scratch/users/acciarri/&release;/&tag;/slicer</outdir>
36    <workdir>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml/&tag;_slicer_wordir</workdir>
37    <inputlist>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml/Filelist/test_list.txt</inputlist>
38    <!--    <inputdef>Run6366</inputdef> -->
39    <numjobs>2</numjobs>
40    <!--    <datatier>2</datatier> -->
41    <!--    <defname>&name;_&tag;_sliced</defname> -->
42  </stage>
43
44  <stage name="filter">
45    <fcldir>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml</fcldir>
46    <fcl>Filter.fcl</fcl>
47    <outdir>/pnfs/lariat/scratch/users/acciarri/&release;/&tag;/filter</outdir>
48    <workdir>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml/&tag;_filter_wordir</workdir>
49    <inputlist>/pnfs/lariat/scratch/users/acciarri/&release;/&tag;/slicer/files.list</inputlist>
50    <numjobs>3</numjobs>
51    <!--        <defname>&name;_&tag;_sliced</defname> -->
52  </stage>
53
54  <stage name="reco">
55    <fcldir>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml</fcldir>
56    <fcl>Reco_Slicer.fcl</fcl>
57    <outdir>/pnfs/lariat/scratch/users/acciarri/&release;/&tag;/reco</outdir>
58    <workdir>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml/&tag;_reco_wordir</workdir>
59    <inputlist>/pnfs/lariat/scratch/users/acciarri/&release;/&tag;/slicer/files.list</inputlist>
60    <numjobs>2</numjobs>
61    <!--        <defname>&name;_&tag;_sliced</defname> -->
62  </stage>
63
```

Number of parallel jobs (hence, output files) for each stage.

Job submission to Grid

TestBatch.xml

```
31  <!-- Project stages -->
32
33  <stage name="slicer">
34    <fcl>SlicerInput.fcl</fcl>
35    <outdir>/pnfs/lariat/scratch/users/acciarri/&release;/&tag;/slicer</outdir>
36    <workdir>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml/&tag;_slicer_wordir</workdir>
37    <inputlist>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml/Filelist/test_list.txt</inputlist>
38    <!--    <inputdef>Run6366</inputdef> -->
39    <numjobs>2</numjobs>
40    <!--    <datatier>2</datatier> -->
41    <!--    <defname>&name;_&tag;_sliced</defname> -->
42  </stage>
43
44  <stage name="filter">
45    <fcldir>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml</fcldir>
46    <fcl>Filter.fcl</fcl>
47    <outdir>/pnfs/lariat/scratch/users/acciarri/&release;/&tag;/filter</outdir>
48    <workdir>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml/&tag;_filter_wordir</workdir>
49    <inputlist>/pnfs/lariat/scratch/users/acciarri/&release;/&tag;/slicer/files.list</inputlist>
50    <numjobs>3</numjobs>
51    <!--        <defname>&name;_&tag;_sliced</defname> -->
52  </stage>
53
54  <stage name="reco">
55    <fcldir>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml</fcldir>
56    <fcl>Reco_Slicer.fcl</fcl>
57    <outdir>/pnfs/lariat/scratch/users/acciarri/&release;/&tag;/reco</outdir>
58    <workdir>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml/&tag;_reco_wordir</workdir>
59    <inputlist>/pnfs/lariat/scratch/users/acciarri/&release;/&tag;/slicer/files.list</inputlist>
60    <numjobs>2</numjobs>
61    <!--        <defname>&name;_&tag;_sliced</defname> -->
62  </stage>
63
```

Fcl file to process. If it is not one of those in the localProduct directory, the full path needs to be specified in the <fcldir> </fcldir> field

Job submission to Grid

TestBatch.xml

```xml
<!-- Project stages -->

<stage name="slicer">
  <fcl>SlicerInput.fcl</fcl>
  <outdir>/pnfs/lariat/scratch/users/acciarri/&release;/&tag;/slicer</outdir>
  <workdir>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml/&tag;_slicer_wordir</workdir>
  <inputlist>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml/Filelist/test_list.txt</inputlist>
  <!--    <inputdef>Run6366</inputdef> -->
  <numjobs>2</numjobs>
  <!--    <datatier>2</datatier> -->
  <!--    <defname>&name;_&tag;_sliced</defname> -->
</stage>

<stage name="filter">
  <fcldir>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml</fcldir>
  <fcl>Filter.fcl</fcl>
  <outdir>/pnfs/lariat/scratch/users/acciarri/&release;/&tag;/filter</outdir>
  <workdir>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml/&tag;_filter_wordir</workdir>
  <inputlist>/pnfs/lariat/scratch/users/acciarri/&release;/&tag;/slicer/files.list</inputlist>
  <numjobs>3</numjobs>
  <!--        <defname>&name;_&tag;_sliced</defname> -->
</stage>

<stage name="reco">
  <fcldir>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml</fcldir>
  <fcl>Reco_Slicer.fcl</fcl>
  <outdir>/pnfs/lariat/scratch/users/acciarri/&release;/&tag;/reco</outdir>
  <workdir>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml/&tag;_reco_wordir</workdir>
  <inputlist>/pnfs/lariat/scratch/users/acciarri/&release;/&tag;/slicer/files.list</inputlist>
  <numjobs>2</numjobs>
  <!--        <defname>&name;_&tag;_sliced</defname> -->
</stage>
```

Output directory, where root files (and others) are stored.

EACH STAGES NEEDS A DIFFERENT DIRECTORY!.

Recommended to create it in the /pnfs/lariat/scratch/user area rather than /lariat/data user

Job submission to Grid

TestBatch.xml

```
31  <!-- Project stages -->
32
33  <stage name="slicer">
34    <fcl>SlicerInput.fcl</fcl>
35    <outdir>/pnfs/lariat/scratch/users/acciarri/&release;/&tag;/slicer</outdir>
36    <workdir>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml/&tag;_slicer_wordir</workdir>
37    <inputlist>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml/Filelist/test_list.txt</inputlist>
38    <!--    <inputdef>Run6366</inputdef> -->
39    <numjobs>2</numjobs>
40    <!--    <datatier>2</datatier> -->
41    <!--    <defname>&name;_&tag;_sliced</defname> -->
42  </stage>
43
44  <stage name="filter">
45    <fcldir>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml</fcldir>
46    <fcl>Filter.fcl</fcl>
47    <outdir>/pnfs/lariat/scratch/users/acciarri/&release;/&tag;/filter</outdir>
48    <workdir>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml/&tag;_filter_wordir</workdir>
49    <inputlist>/pnfs/lariat/scratch/users/acciarri/&release;/&tag;/slicer/files.list</inputlist>
50    <numjobs>3</numjobs>
51    <!--        <defname>&name;_&tag;_sliced</defname> -->
52  </stage>
53
54  <stage name="reco">
55    <fcldir>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml</fcldir>
56    <fcl>Reco_Slicer.fcl</fcl>
57    <outdir>/pnfs/lariat/scratch/users/acciarri/&release;/&tag;/reco</outdir>
58    <workdir>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml/&tag;_reco_wordir</workdir>
59    <inputlist>/pnfs/lariat/scratch/users/acciarri/&release;/&tag;/slicer/files.list</inputlist>
60    <numjobs>2</numjobs>
61    <!--        <defname>&name;_&tag;_sliced</defname> -->
62  </stage>
63
```

Working directory, used by project.py to store information while jobs run on the grid.

Each stage needs to have its own.

Job submission to Grid

```
33  <stage name="slicer">
34    <fcl>SlicerInput.fcl</fcl>
35    <outdir>/pnfs/lariat/scratch/users/acciarri/&release;/&tag;/slicer</outdir>
36    <workdir>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml/&tag;_slicer_wordir</workdir>
37    <inputlist>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml/Filelist/test_list.txt</inputlist>
38    <!--    <inputdef>Run6366</inputdef> -->
39    <numjobs>2</numjobs>
40    <!--    <datatier>2</datatier> -->
41    <!--    <defname>&name;_&tag;_sliced</defname> -->
42  </stage>
43
```

INPUT FILE

There are two ways to tell project.py which files it has to work on: through a file list or through a SAM definition.

- **FILES LIST** ***(red arrow)*** create a text file containing the list of root files to be processed, one entry per line, full path to the file. You can inquire SAM to find out the list of files matching the criteria you are interested in, then ask Johnny for his nice script converting it into a handy text file :)

Job submission to Grid

```
33  <stage name="slicer">
34    <fcl>SlicerInput.fcl</fcl>
35    <outdir>/pnfs/lariat/scratch/users/acciarri/&release;/&tag;/slicer</outdir>
36    <workdir>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml/&tag;_slicer_wordir</workdir>
37    <inputlist>/lariat/app/users/acciarri/analysis/reconstruction/&release;/xml/Filelist/test_list.txt</inputlist>
38    <!--    <inputdef>Run6366</inputdef> -->
39    <numjobs>2</numjobs>
40    <!--    <datatier>2</datatier> -->
41    <!--    <defname>&name;_&tag;_sliced</defname> -->
42  </stage>
43
```

SAM DEFINITION *(BLUE arrows)* use a SAM definition (inserted in the <inputdef> </inputdef> field) to let project.py know which files to pick. If you use a SAM definition the <inputlist> line must be commented or absent and the <datatier> and <defname> lines have to be uncommented. In addition the fcl file must have “dataTier” and “compressionLevel” entries in the “outputs” field like shown in this example

```
outputs:
{
 out1:
 {
   module_type: RootOutput
   #fileName:    "%ifb_%tc_reco2D.root"
   fileName:    "reco_data.root"
   dataTier:    "reconstructed-2d"
   compressionLevel: 1
 }
}
```

Job submission to Grid

RUN PROJECT.PY

project.py --xml TestBatch.xml –stage slicer –submit

You can check the status of the jobs with the command

jobsub_q –user=<your user name>

Once the jobs are all completed, the directory defined in <outdir> will contain one subdirectory for each job. Inside each subdirectory there will be the processed root file. Before moving to the next stage you have to check the result of the jobs by giving the command:

project.py --xml TestBatch.xml –stage slicer –check

If you have the latest latest (not older than yesterday) version of larutils and larbatch, it should work. If your version is older or you get an error like this:

Job submission to Grid

Warning: File standard_slicer.root in directory
/pnfs/lariat/scratch/users/acciarri/v01_08_02/datatest/slicer/4198499_0 is not a valid root file.
Warning: File slicer_hist.root in directory
/pnfs/lariat/scratch/users/acciarri/v01_08_02/datatest/slicer/4198499_0 is not a valid root file.
Problem with root file(s) in subdirectory 4198499_0.
Bad subdirectory 4198499_0.
0 total good events.
0 total good root files.
0 total good histogram files.
2 processes with errors.
2 missing files.

Then you can't use the check command. Clean up the directory giving the command

project.py --xml TestBatch.xml –stage slicer –clean

And submit the first stage once again.

project.py --xml TestBatch.xml –stage slicer –submit

Job submission to Grid

Once the jobs are completed **DO NOT** give the command check. Instead, create a file list for the root files just created and pass it as input list to the second stage.

ls <outdir path>/<subdirectory created>_*/<rootfile> > <outdir path>/files.list

Of course in this case the second stage needs to have a <inputlist> </inputlist> field as well.

Reapeat the same for the next stages.

What if

the first stages go well but the track reconstruction does not work in Grid while it works interactively?

You realize what frustration really means.

What if

the first stages go well but the track reconstruction does not work in Grid while it works interactively?

But you can fix that adding to your Reco.fcl this line:

services.scheduler.SkipEvent: [ProductNotFound]

After and outside the services field.

Such fix has been implemented by Tingjun and pushed into develop yesterday afternoon, so if you pulled down your lariatsoft version later than yesterday, this line should already be there.