FACULTY OF SCIENCE

UNIVERSITY OF COPENHAGEN

GridPilot - making your grid life easier

Frederik Orellana, Morten Badensø, Jørgen Beck Hansen,
Jacob Debel, Simon Heisterkamp, Ask Emil Jensen

Niels Bohr Institute, University of Copenhagen

More information at
http://www.gridpilot.dk

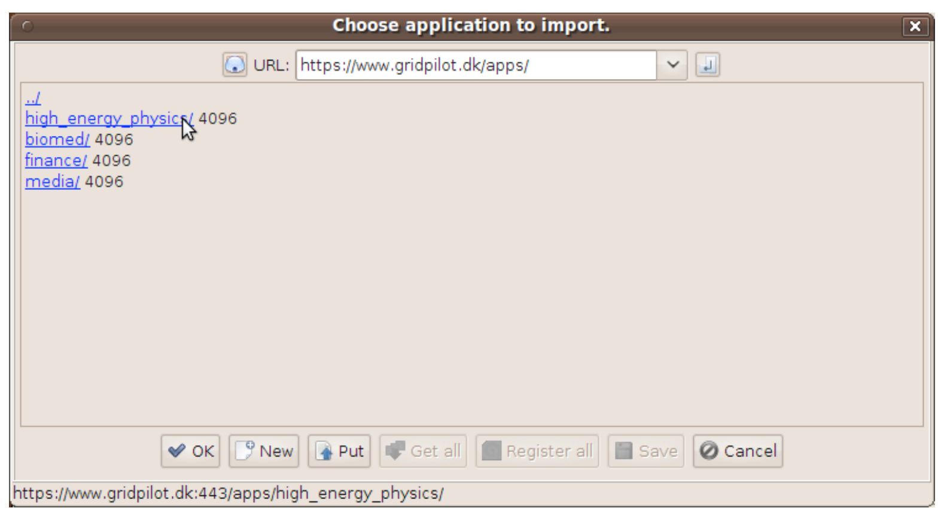
Abstract
We present a novel tool for managing data processing on grid resources. The tool provides a graphical user interface that offers new grid users a quick and gentle start with computing, using a library of applications built up by previous users.

Application library

- get new users started with data processing
- reuse computing knowledge in a group
- ensure data provenance and reproducibility

The computing frameworks of high energy physics experiments can be challenging for new users.

The GridPilot application library provides a collection of template applications that can be run with a few mouse clicks and edited to suit specific needs like different datasets, different executable, different input files etc.



Choose application to import.

URL: https://www.gridpilot.dk/apps/

../
high_energy_physics/ 4096
biomed/ 4096
finance/ 4096
media/ 4096

https://www.gridpilot.dk:443/apps/high_energy_physics/

The default application library.

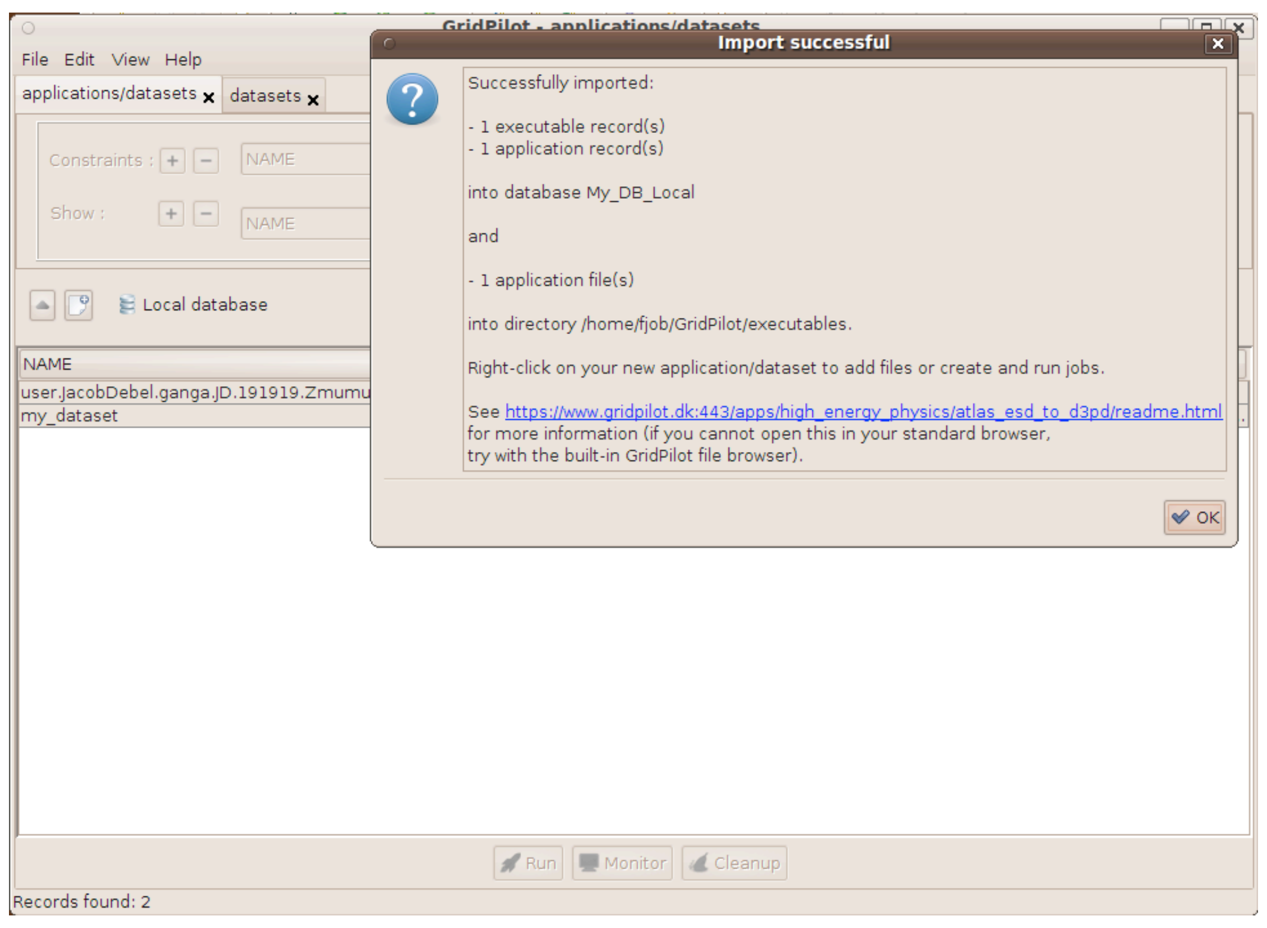
The GridPilot web site provides a collection of example applications that can be directly imported an run from GridPilot. You can also easily set up your own examples library.

We have put in place a library of applications with examples from finance, bioinformatics, media processing and high energy physics with particular attention to ATLAS. New ATLAS users can simply find an application that matches his needs and only modify e.g. the input dataset and/or other input files (e.g. Athena jobOptions or Root macros).

Since importing and exporting applications is next to trivial, the library is expected to grow if and when students need different templates than those available.

Currently the library includes:

- Simulation with a standard event generation transformation and a custom jobOptions file
- ESD to D3PD conversion with standard ATLAS reconstruction transformation
- ESD to ntuple conversion with standard ATLAS reconstruction transformation
- RDO to ESD conversion with a custom jobOptions file and extra Athena tags
- "Boildown" of ntuple files using a Root macro



GridPilot - applications/datasets

Import successful

Successfully imported:

- 1 executable record(s)
- 1 application record(s)

into database My_DB_Local

and

- 1 application file(s)

into directory /home/fjob/GridPilot/executables.

Right-click on your new application/dataset to add files or create and run jobs.

See https://www.gridpilot.dk:443/apps/high_energy_physics/atlas_esd_to_d3pd/readme.html for more information (if you cannot open this in your standard browser, try with the built-in GridPilot file browser).

Records found: 2

Confirmation of application import.

After choosing an application, GridPilot downloads and unpacks files, creates corresponding database records and displays a confirmation that the application is ready to be used.

Data management

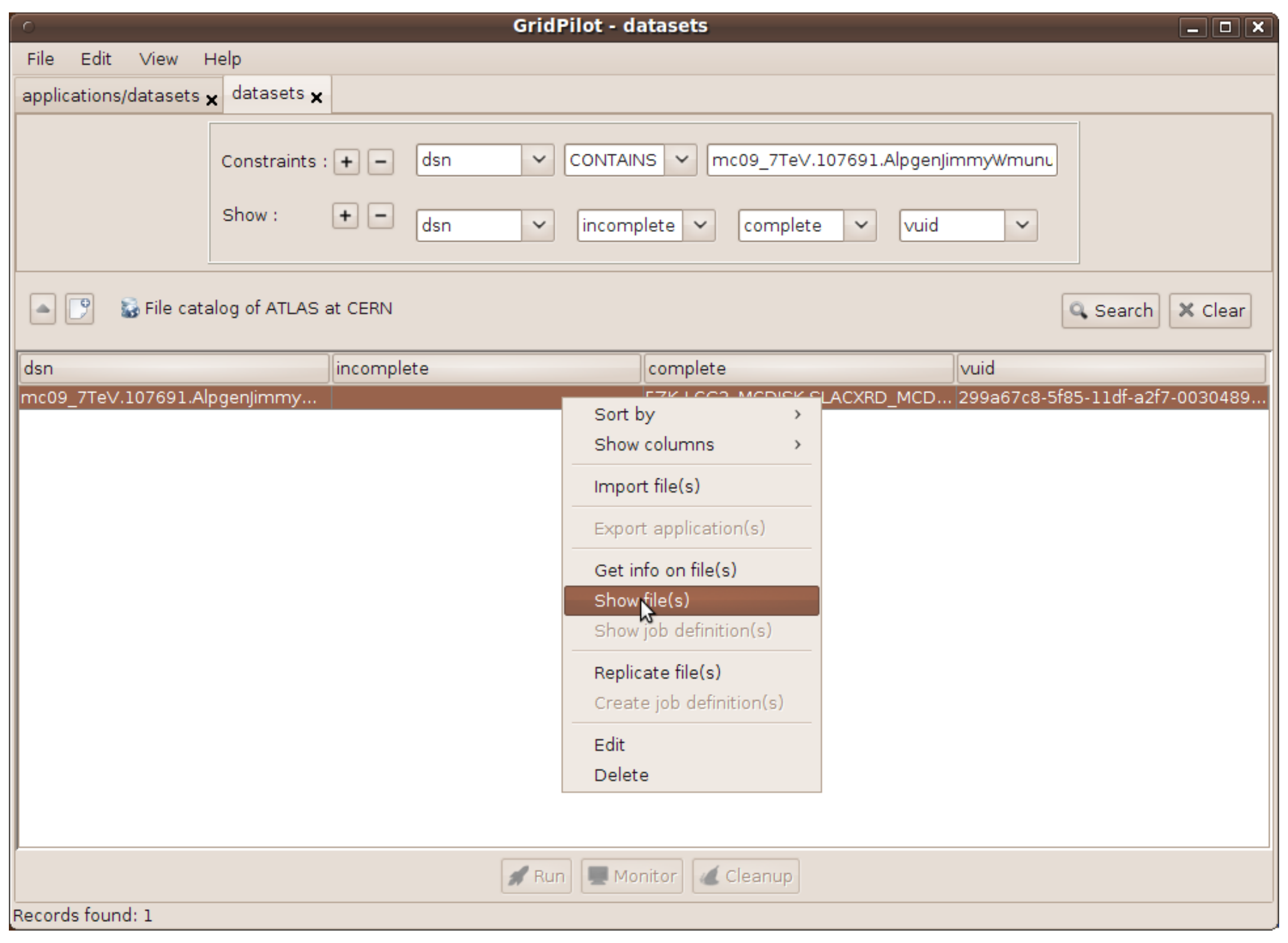
- easily search for datasets
- download files
- use datasets as input when creating jobs
- replicate datasets to other locations

In the grid world, data files are typically registered in the LFC file catalog. The CERN experiments use this and moreover each implement their own dataset catalog (where datasets are collections of files).

The actual data files are typically accessed with SRM, gridftp and/or https.

Standard grid tools do not include graphical browsing and managing of files.

GridPilot provides a graphical dataset/file manager. A built-in or external database can be used as both file and dataset catalog, and also, LFC and the ATLAS dataset catalog are explicitly supported. The main aim of this is to make it easy to select input datasets/files for data processing and register produced datasets.



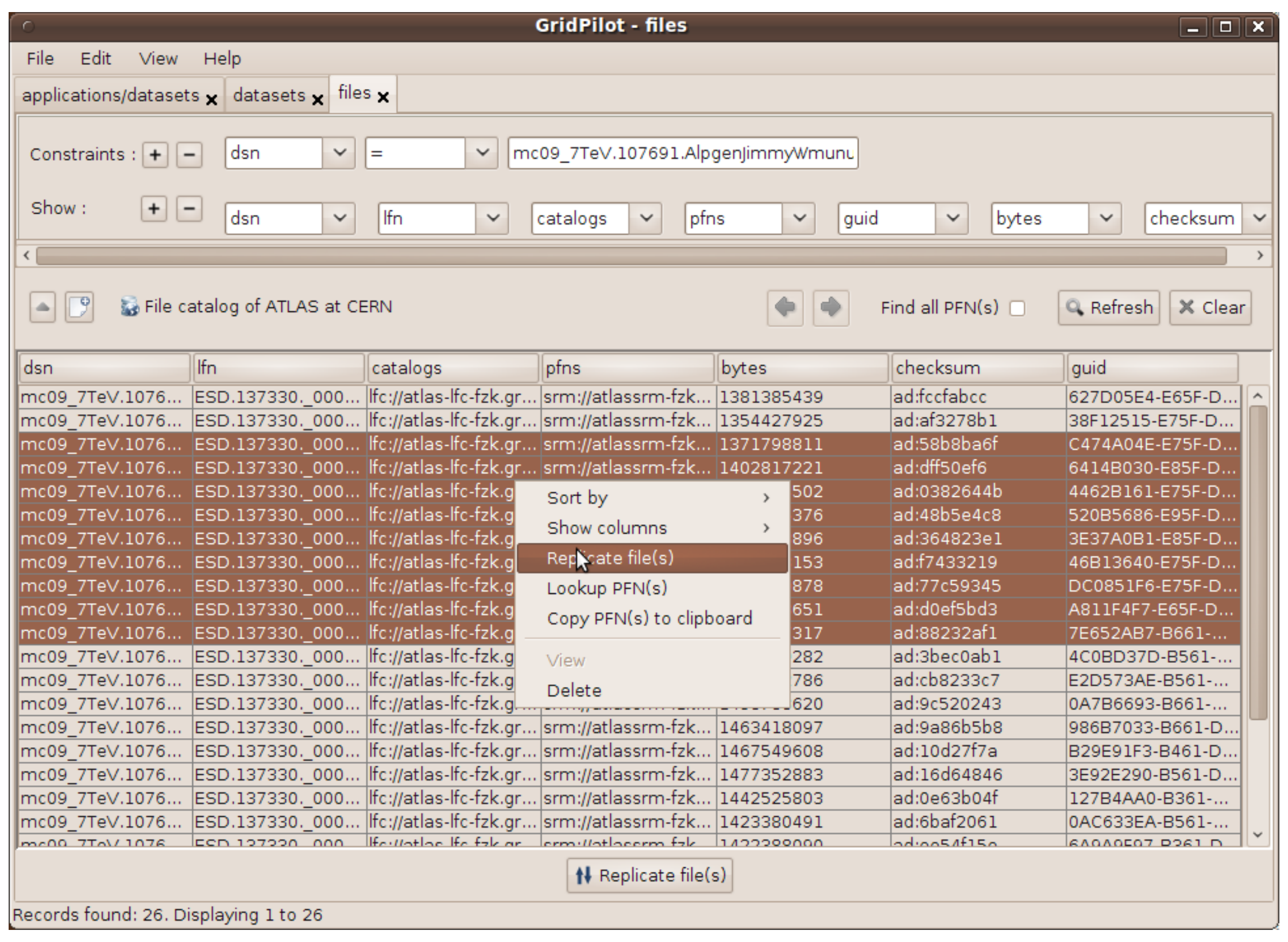
GridPilot - datasets

Records found: 1

A dataset located by searching the ATLAS dataset catalog.

GridPilot has an plugin architecture and supports various database back-ends, including HSQLDB and MySQL. Support for the ATLAS dataset catalog and file catalogs have been implemented in a single database plugin.

GridPilot has some support for data management: Files can be downloaded from, uploaded to and replicated between grid file servers and the ATLAS DQ2 dataset catalog and LFC file catalog are specifically supported. Notice that this is meant to allow a user to easily find and download a few files to study before submitting a large-scale production to a grid back-end. For large-scale transfers, the PanDA web interface for replication requests or the DQ2 suite of command-line tools should be used.



GridPilot - files

Records found: 26. Displaying 1 to 26

Files of a dataset.

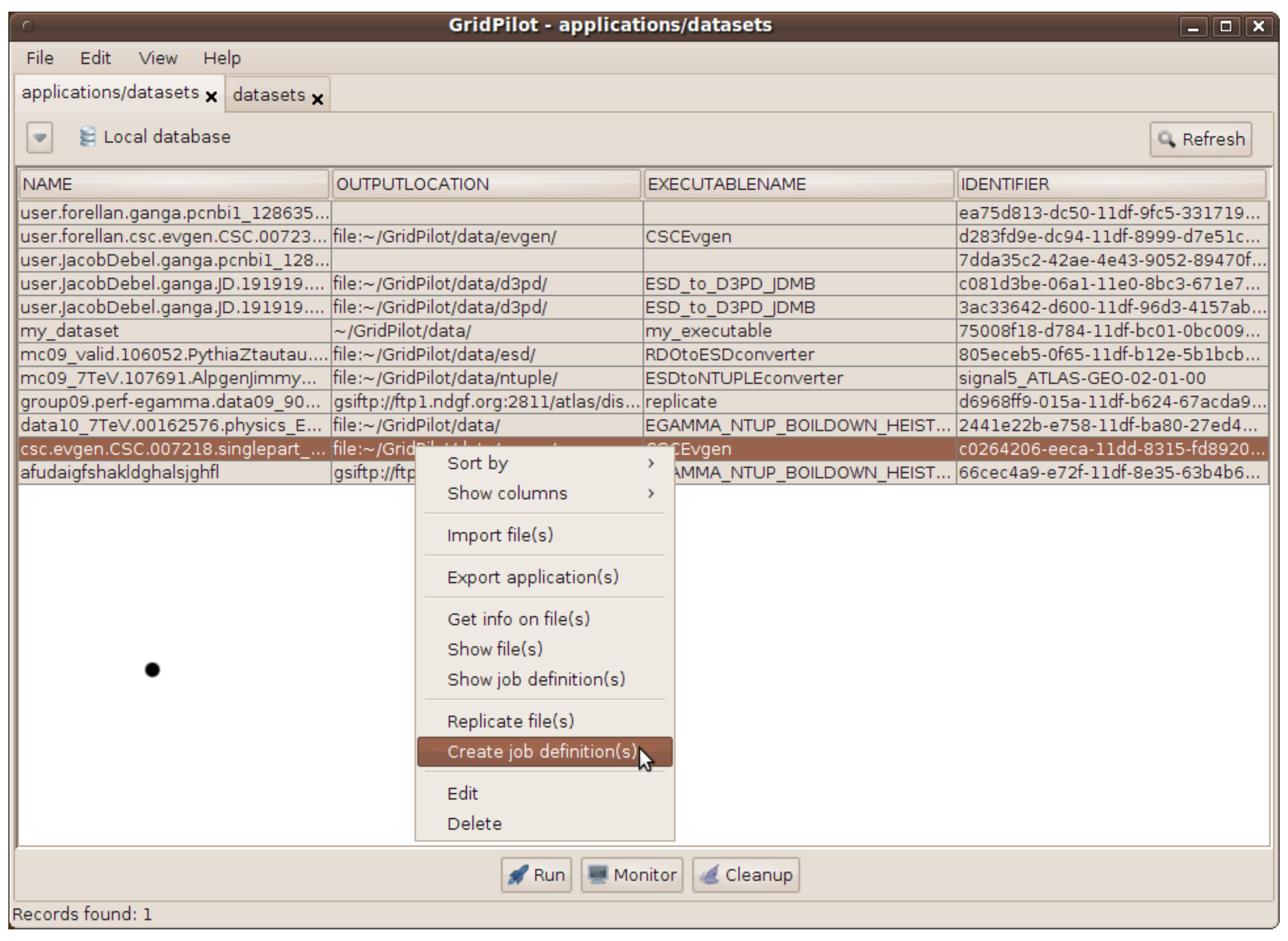
The right-click menu allows downloading or replicating both entire datasets and individual files,

Computing - job management

- process entire dataset collections
- run on various back-ends, including WLCG/gLite and NorduGrid/ARC, local virtual machines and Amazon's EC2
- monitor jobs and have failed jobs resubmitted automatically

Standard grid tools do not provide a GUI to manage large amounts of compute jobs.

With GridPilot, once an application has been imported and customized and input datasets have been selected, compute jobs are created, submitted and controlled with a few mouse clicks. GridPilot hides peculiarities of the different back-ends and simply presents the user with choice of which back-end to submit to.

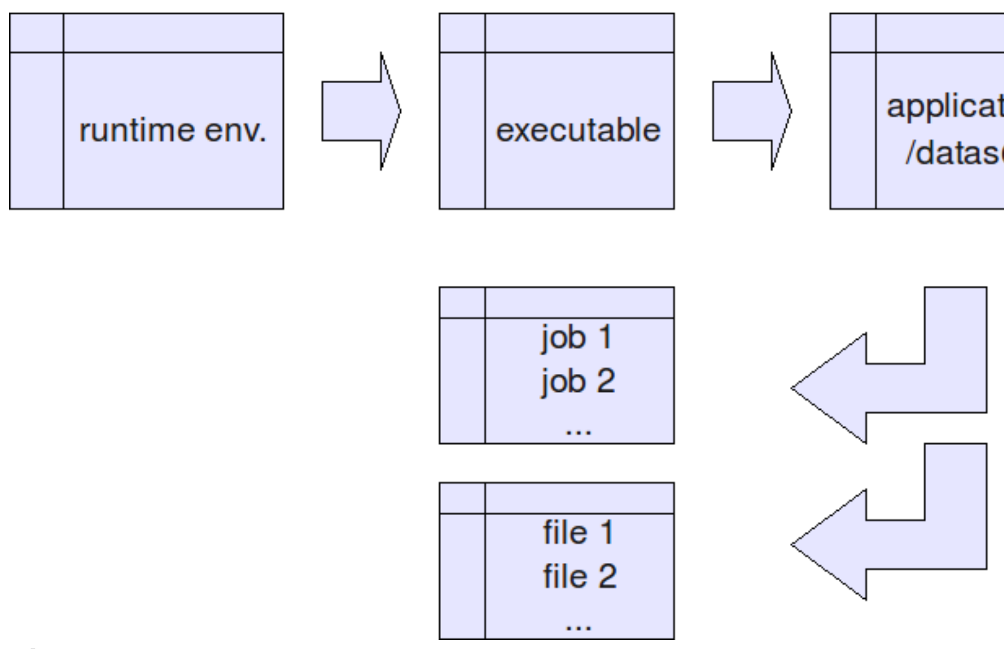


GridPilot - applications/datasets

Records found: 1

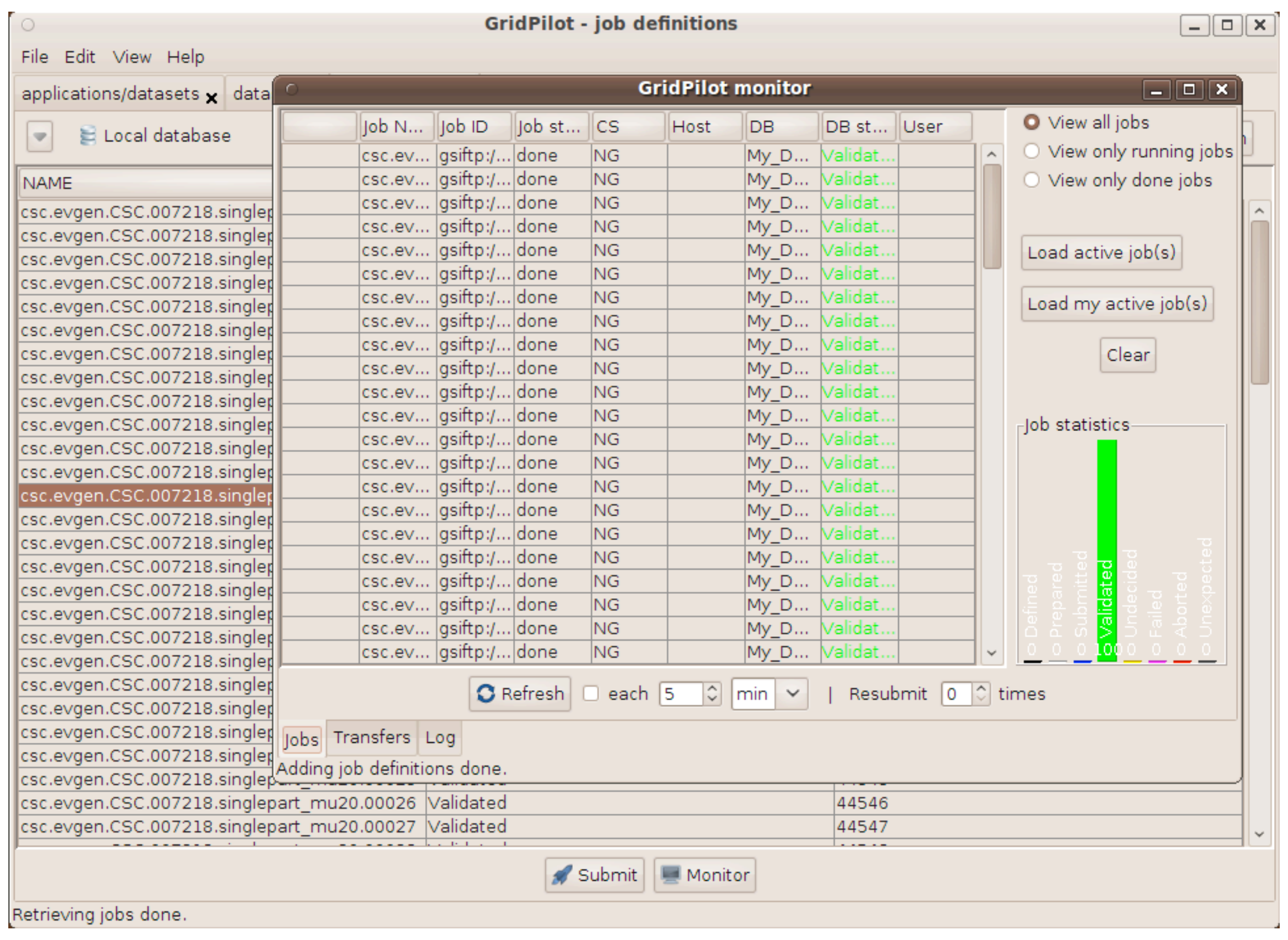
Creating the jobs to create a new ATLAS dataset (by processing another dataset).

To GridPilot an application and a dataset is the same thing: an application creates a collection of files, i.e. a dataset, that are labelled by the name of the application.



GridPilot data structures.

Internally, GridPilot organizes bookkeeping data in 6 kinds of tables: runtime environments, executables, applications/datasets, jobs and files.



GridPilot - job definitions

GridPilot monitor

Adding job definitions done.

Retrieving jobs done.

Job monitoring.

Submitted jobs are monitored and managed (killed, resubmitted) from the job monitor.