Archiving Large Galaxy Histories

Minnesota Supercomputing Institute
University of Minnesota
03/15/2013
1 Introduction.................................................................................................................................3
  1.1 Outline ........................................................................................................................................3

2 Exporting large histories..................................................................................................................4
  2.1 Generating public key ..................................................................................................................5
  2.2 Uploading public key to your MSI home directory.......................................................................6
  2.3 Configuring public key ................................................................................................................7
  2.4 Create directory to store data......................................................................................................8
  2.5 Exporting input datasets ............................................................................................................9
  2.6 Extracting workflows ...................................................................................................................11
  2.7 Exporting input datasets............................................................................................................13

3 Re-creating large histories .............................................................................................................14
  3.1 Import input datasets from Galaxy library...................................................................................15
  3.2 Download workflow from MSI home directory to your local machine......................................16
  3.3 Import workflow .......................................................................................................................17
  3.4 Run workflow ............................................................................................................................18
  3.5 Deleting old histories ................................................................................................................19
1 Introduction
When running large workflows in Galaxy, data usage resulting from intermediate files can rapidly increase storage usage to an unsustainable level. Due to the continued increase in number of Galaxy users, we would like to request that you be considerate to other users and delete old data and histories no longer being used. We have a finite amount of resources (disk space and compute nodes) on the Galaxy system. Deleting old data and histories will significantly improve performance.

Even though Galaxy provides a tool for exporting your workflows and re-importing them at a later date, this option will not work for large histories (> 2GB when exported) on MSIs local installation of Galaxy.

This document outlines an alternative for exporting large histories on MSI’s local installation of Galaxy that will enable re-creation of these histories at a later date, essentially archiving the history. Additionally, it also shows how to re-create the history using the input datasets and exported workflow. Users can then safely delete these large histories to reduce their data usage in Galaxy. Section 2 outlines steps for exporting input datasets from a galaxy history together with a workflow extracted from the same history directly to a users MSI home directory. This extracted workflow contains all the analysis steps used generate the history and together with the input datasets it, can be used to re-create the history. Section 3 outlines steps for re-creating the history. We HIGHLY recommend re-creating at least one history using the exported input datasets and extracted workflows to ensure you will be able to re-create the history at a later stage. Section 3 then shows how to permanently delete histories.

Note: These instructions were generated on a Macintosh machine and can be followed step-by-step if using Macintosh machine. However, some of these steps are operating system specific. Users using a machine with any other operating system e.g., PC users will need to slight modification to this protocol on some of the system specific steps e.g., changing file names. Such users can log into the MSI system using the NX client. For details on logging into MSI machines using NX client see https://www.msi.umn.edu/support/nx.

1.1 Outline
1 Introduction
2 Exporting large histories
3 Re-creating large histories
2 Exporting large histories

**Generating public key** *(Sect. 2.1 page 5)*

The tool for exporting input datasets directly from Galaxy histories to a users MSI home directory is meant for users with some familiarity with use of file transfer protocols such as sftp and WinSCP, MSI systems and Unix shells. Each semester MSI offers tutorials on using MSI systems and UNIX. For more information on tutorials visit [https://www.msi.umn.edu/tutorial/](https://www.msi.umn.edu/tutorial/). To use this tool, one will need to first need to generate a public key that Galaxy will use to upload files on your behalf. See [http://en.wikipedia.org/wiki/Public-key_cryptography](http://en.wikipedia.org/wiki/Public-key_cryptography) for more information on public keys. Galaxy has a tool for generating a public key. We will show how to use it to generate a public key to be used by Galaxy for data upload. NOTE: You will need to generate this public key only once. The first time you export data from Galaxy directly to your home directory.

**Uploading public key to your MSI home directory** *(Sect 2.2 page 6)*

Once Galaxy has generated a public key, you will need to download the generated file to your local machine, rename it and then copy it to your MSI home directory. You will then need to log into your home directory and run a series of commands that will enable Galaxy to use the public key to copy files directly to your home directory.

**Exporting input datasets** *(Sect 2.5 page 9)*

Exporting input datasets is relatively straightforward once the public key has been generated, copied to your home directory and configured using a series of UNIX commands. NOTE: This tool for exporting input data files directly into your home directory could be used for any other datasets in your history such as final output or results files.

**Extracting workflows** *(Sect 2.6 page 11)*

Galaxy workflows provide an easy method to automate an analysis pipeline. These workflows are very useful when the same analysis needs to be performed on multiple datasets. For our purposes, we will use this workflow to re-create the large history that was generated using the input datasets. We will demonstrate how to extract a workflow from your working history and export it to your home directory.

Note: Once you’ve extracted the workflow, it is important to ensure all the tools in the analysis workflow are linked properly and parameters for each tool correspond to those used in the analysis. Galaxy sometimes re-sets parameters to defaults. It will also be useful to annotate inputs in the workflow with the expected input data type. It will make it easier to verify the right input datasets have been selected later on when running the workflow. RISS provides a number of Galaxy tutorials with instructions on how to edit and annotate workflows. For an example on how to edit and annotate workflows see Section **2.11** in [https://www.msi.umn.edu/sites/default/files/RISS_Hsapiens_variant_Detection_v2.0a.pdf](https://www.msi.umn.edu/sites/default/files/RISS_Hsapiens_variant_Detection_v2.0a.pdf)

**Exporting workflows**

The tool describe above can only export data from histories. To export a “workflow” extracted from a working history to your home directory, you will need to use the same approach used to upload your public key. Workflows are relatively small text files and will be easily uploaded using sftp or WinSCP in a few seconds.
2.1 Generating public key

a) At the top of the Tools Pane, select "Send Data -> Generate Public Key (Beta)"

b) Click "Execute"
2.2 Uploading public key to your MSI home directory

a) In the history pane click on the name of the public key file just generated “Public Key”
b) Click the floppy disk icon to download a copy of the file to your local computer

c) Open up a Terminal and change directory to the location of the downloaded file (e.g., Desktop).
d) Change the name of the downloaded file to “galaxy-key.pub”
e) Login into your MSI home directory using sftp
f) Upload the file “galaxy-key.pub” to your MSI home directory
g) Exit from sftp

NOTE: If you are using a different system, change the name of the downloaded file to “galaxy-key.pub”. Copy this file to your MSI home directory and skip to Section 2.3. Otherwise proceed to c).
2.3 Configuring public key

a) SSH into your MSI home directory
b) Type the command `mkdir ~/.ssh` and press Enter
c) Type the command `chmod 700 ~/.ssh` and press Enter
d) Type the command `touch ~/.ssh/authorized_keys2` and press Enter
e) Type the command `chmod 600 ~/.ssh/authorized_keys2` and press Enter
f) Type the command `cat ~/galaxy-key.pub >> ~/.ssh/authorized_keys2` and press Enter
2.4 Create directory to store data

a) Type the command `mkdir galaxy_data` and create a directory that will be used to store galaxy data and press Enter. **Note:** You can substitute the directory name “`galaxy_data`” with whatever name you would like to call your directory.

b) Type the command `cd galaxy_data` and press Enter. This command will change your active directory to the directory just created (“`galaxy_data`”)

c) Type the command `pwd` and press Enter. This command will give you the full path to the directory just created. In this case “/home/support/onsongog/galaxy_data”. The tool for uploading data directly from Galaxy will need this path.

d) Copy this path and go back to Galaxy

```
onsongog@login02 [~] % mkdir galaxy_data
onsongog@login02 [~] % cd galaxy_data
onsongog@login02 [~/galaxy_data] %
onsongog@login02 [~/galaxy_data] % pwd
/home/support/onsongog/galaxy_data
```
2.5 Exporting input datasets

a) At the top of the Tools Pane, select “Send Data -> Upload Files to MSI (Beta)”
b) In the input field under “Destination Path:”, enter path copied in Section 2.4(d) (“/home/support/onsongog/galaxy_data”) in this case.
c) Click “Add new File” to have drop down menus equaling the number of files you want to copy
d) Select the files you want to copy
e) Click “Execute”
f) Once the "Upload Summary" file name turns green, click on the “eye” to display a summary of the load in the center pane

g) Review upload summary to confirm files were uploaded successfully

h) You can additionally confirm files were uploaded successfully by going back to your SSH terminal and typing ls

i) Uploaded files should be displayed
2.6 Extracting workflows

a) At the top of the history pane click on the small gear icon and select “Extract Workflow” from the pop-up menu
b) Scroll down the center pane and uncheck boxes next to tools you do not want to re-run later. In our case we will uncheck the two tools we used to upload data (“Generate Public Key (Beta) and “Upload Files to MSI (Beta)

c) Change the name of the workflow to the same name as the history
d) Click “Create Workflow”
e) Click the *Workflow* tab
f) On the drop down menu next to the workflow that was just created, click “Download or Export”
g) Click the link below “Download to File”
2.7 Exporting input datasets

a) Open a terminal, change directory to the same directory the exported workflow was downloaded to and sftp to your MSI home directory

b) Upload workflow to your home directory using the command `put Galaxy-Workflow-PercOfFragOnTarget.ga`

c) Return to your SSH terminal and copy this workflow to the same directory containing the input datasets using the command "mv ~/Galaxy-Workflow-PercOfFragOnTarget.ga". The double quotes are not part of the command

```
onson001@x-128-101-135-198 ~/Desktop Thu Mar 14 $ sftp onsong@login.msi.umn.edu
onson001@x-128-101-135-198 ~/Desktop Thu Mar 14 $ sftp onsong@login.msi.umn.edu's password:
Connected to login.msi.umn.edu.
sftp> put Galaxy-Workflow-PercOfFragOnTarget.ga
Uploading Galaxy-Workflow-PercOfFragOnTarget.ga to /panfs/roc/groups/2/support/onsong/ Galaxy-Workflow-PercOfFragOnTarget.ga
Galaxy-Workflow-PercOfFragOnTarget.ga 100% 18KB 18.0KB/s 00:00
sftp>
onson001@x-128-101-135-198 ~/galaxy_data %
onsong@login01 [~/galaxy_data] %
onsong@login01 [~/galaxy_data] %
onson001@x-128-101-135-198 ~/galaxy_data %
onson001@x-128-101-135-198 ~/galaxy_data %
onson001@x-128-101-135-198 ~/galaxy_data %
onson001@x-128-101-135-198 ~/galaxy_data %
onson001@x-128-101-135-198 ~/galaxy_data % mv ~/Galaxy-Workflow-PercOfFragOnTarget.ga .
onson001@x-128-101-135-198 ~/galaxy_data %
onson001@x-128-101-135-198 ~/galaxy_data %
onson001@x-128-101-135-198 ~/galaxy_data %
onson001@x-128-101-135-198 ~/galaxy_data %
onson001@x-128-101-135-198 ~/galaxy_data %
onson001@x-128-101-135-198 ~/galaxy_data %
onson001@x-128-101-135-198 ~/galaxy_data %
onson001@x-128-101-135-198 ~/galaxy_data %
```
3 Re-creating large histories

Before deleting your history, it is important to verify you will be able to re-create the history. In this section we will show you how to re-create the history using the input datasets and workflow in your MSI home directory. You do not have to re-create histories for all your archived histories but we suggest you try re-creating at least one history to verify you can successfully “archive” your histories.

☆ Getting data from your MSI home directory into Galaxy
Unlike uploading data from Galaxy to MSI, currently users have to send a request to help@msi.umn.edu to have their data uploaded to Galaxy. Send an email to help@msi.umn.edu and request the input data you just uploaded to your home directory (or one dataset if you exported data from multiple histories) be uploaded to Galaxy. Your data will be uploaded to a Galaxy library from which you will be able to access your data. For the purposes of demonstrating how to re-create histories, we saved the exported data in a library named "re-create history".

☆ Re-creating history
To be thorough, delete the local copy of your workflow. Get the copy of this workflow saved in your MSI home directory using sftp or any other file transfer protocol you prefer e.g., WinSCP. Create a new history and copy input datasets from the Galaxy library ("re-create history" for this demonstration). Import the workflow that was stored in your MSI home directory. Run this workflow using datasets copied form the Galaxy library. Compare results to verify you were able to re-create the workflow.

☆ Deleting histories
Once you have successfully exported your input datasets and workflow to your home directory, and verified you can re-create the history, you can delete your history from Galaxy. We will demonstrate how to permanently delete your old history.
3.1 Import input datasets from Galaxy library

a) At the top of the history pane click on the small gear icon and select “Create New”
b) Name the new history e.g., test-archiving
c) Navigate to the library containing the input datasets and select the datasets
d) Click import
e) Navigate back to the history by clicking “Analyze data”
3.2 Download workflow from MSI home directory to your local machine

a) Make sure you no longer have a local copy of the workflow you extracted. Open up a terminal and sftp to your MSI home directory using the command “sftp username@login.msi.umn.edu”

b) Change directory to the directory containing the exported workflow e.g., cd galaxy_data

c) Get a copy of the workflow using the get command “get Galaxy-Workflow-PercOfFragOnTarget.ga”

```bash
sonon001@x-128-101-135-198:/DesktopThu Mar 14
sonon001@x-128-101-135-198:/DesktopThu Mar 14$ sftp onsongog@login.msi.umn.edu
Connected to login.msi.umn.edu.
sftp> cd galaxy_data
sftp> get Galaxy-Workflow-PercOfFragOnTarget.ga
Fetching /panfs/roc/groups/2/support/onsongog/galaxy_data/Galaxy-Workflow-PercOfFragOnTarget.ga
to Galaxy-Workflow-PercOfFragOnTarget.ga
/panfs/roc/groups/2/support/onsongog/galaxy_data/Galaxy-Workflow 100% 18KB 18.0KB/s 00:00
sftp>
```
3.3 Import workflow

a) At the top click “Workflow”
b) Click “Upload or import workflow”
c) Click “Choose file” then navigate to the location of the workflow downloaded from your MSI home directory and select the workflow
d) Click “Import”
3.4 Run workflow

a) Click in the drop down arrow next to the just imported workflow and click "Run"
b) Run the workflow
3.5 Deleting old histories

a) Navigate to the history that needs to be deleted and at the top of the history pane click on the small gear icon and select “Delete Permanently” from the pop-up menu.