



CYVERSE™

Transforming Science Through Data-driven Discovery

HTCondor, Docker, iRODS and
HPC for every scientist with the
CyVerse Discovery Environment

Ian McEwen – Software Engineer

University of Arizona

mian@cyverse.org



Cold
Spring
Harbor
Laboratory





CYVERSE™

formerly

(just so you know)



iPlant
Collaborative™





Discovery Environment

Main, high-level goals:

- Make storing and analyzing data easy (help scientists focus on science, not IT)
- Use prior work, serve as glue between others' excellent work:



(and more)

Name	Integrated by	Rating
BAM to SAM	Roger Barth...	★★★★★ (1)
Index BAM and ...	Roger Barth...	★★★★★ (1)
Index BAM file	Roger Barth...	☆☆☆☆☆ (0)
Index Fasta file (...)	Roger Barth...	★★★★★ (1)
Merge BAM files	Roger Barth...	☆☆☆☆☆ (0)
SAM to sorted B...	Roger Barth...	★★★★★ (4)
Sort BAM file	Roger Barth...	☆☆☆☆☆ (0)

Name	Owner	App	Start Date	End Date	Status
DE_Word_Coun...	mian@iplant...	DE Word ...	2016 Mar 29 ...	2016 Mar 29 ...	Comp...
DE_Word_Coun...	mian@iplant...	DE Word ...	2016 Mar 22 ...	2016 Mar 22 ...	Comp...
compress_the...	mian@iplant...	compress...	2016 Feb 4 1...	2016 Feb 4 1...	Comp...
compress_the...	mian@iplant...	compress...	2016 Feb 4 1...	2016 Feb 4 1...	Failed



iRODS

Integrated Rule-oriented Database System



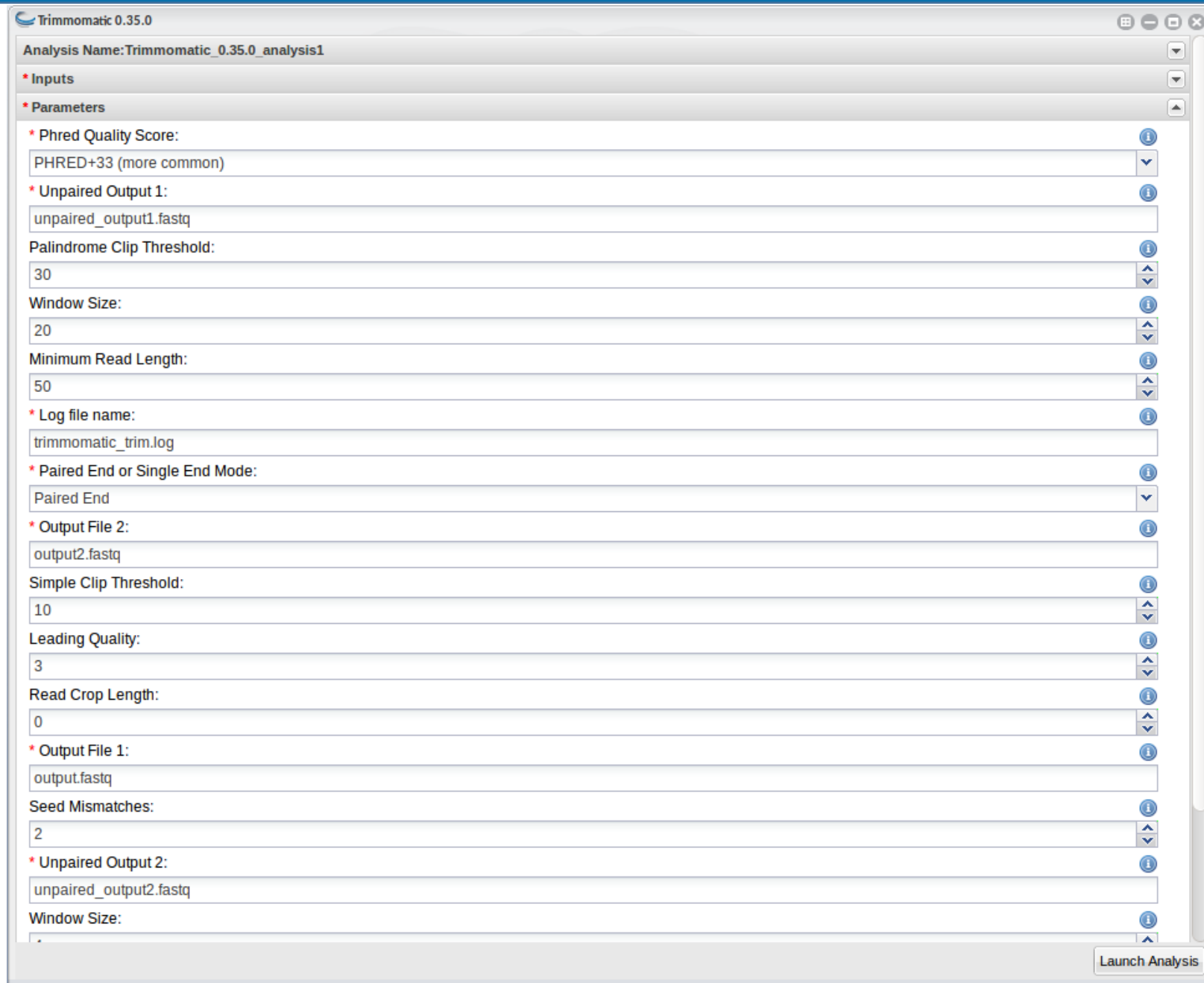
- Data storage & organization for large amounts of data
 - Known and used by scientists already
 - Where mostly-all the data comes from & goes to in our usage
-
- Gateway into High-Performance Computing capabilities through HTTP APIs
 - Used for jobs which need HPC capabilities, but through same interfaces for end users within the DE
-
- Software container system, registry, etc.
 - Known and used by scientists already
 - Used for encapsulation & reproducibility of execution, distribution of tools within cluster, ease of use

A brief tour of the interfaces around jobs



App interface

(for final users who just want to do some science)



Trimmomatic 0.35.0

Analysis Name: Trimmomatic_0.35.0_analysis1

* Inputs

* Parameters

* Phred Quality Score:
PHRED+33 (more common)

* Unpaired Output 1:
unpaired_output1.fastq

Palindrome Clip Threshold:
30

Window Size:
20

Minimum Read Length:
50

* Log file name:
trimmomatic_trim.log

* Paired End or Single End Mode:
Paired End

* Output File 2:
output2.fastq

Simple Clip Threshold:
10

Leading Quality:
3

Read Crop Length:
0

* Output File 1:
output.fastq

Seed Mismatches:
2

* Unpaired Output 2:
unpaired_output2.fastq

Window Size:
.

Launch Analysis

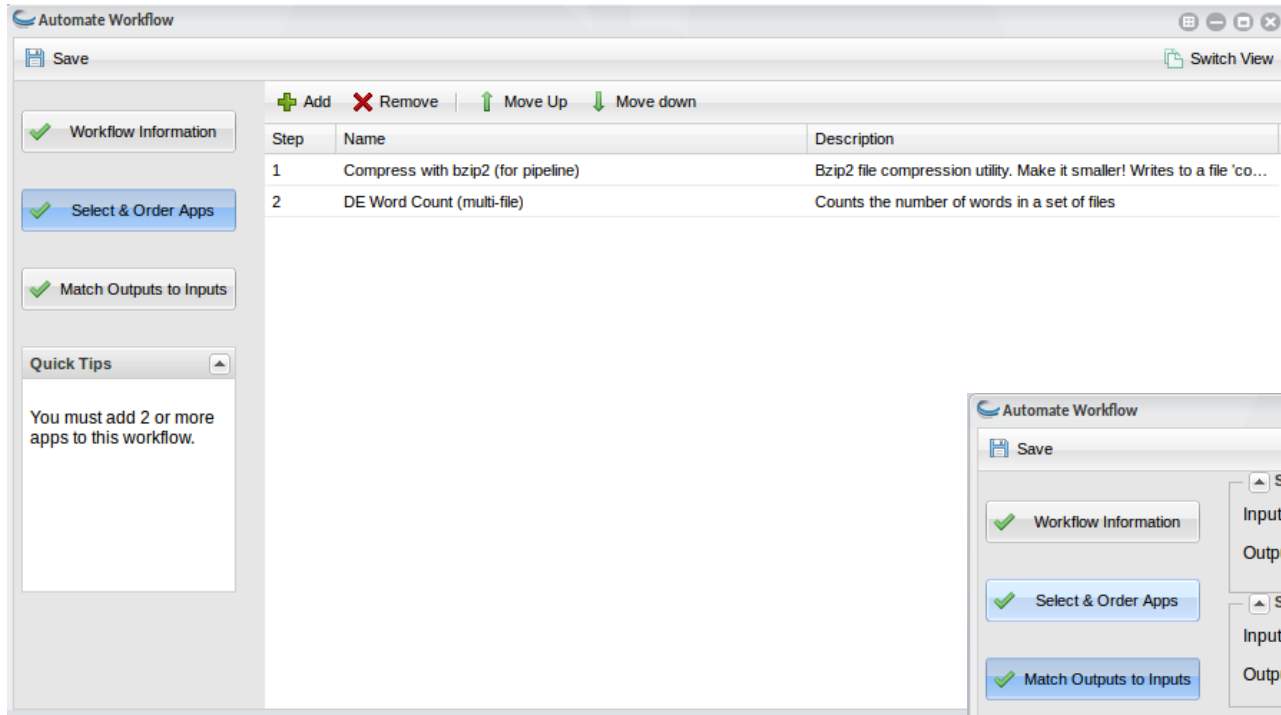


App editor interface (for intermediate users or internal people, creating interfaces around tools)

The screenshot shows the 'DE Word Count (multi-file)' app editor interface. The window title is 'DE Word Count (multi-file)'. At the top, there are buttons for 'Save', 'Preview', and 'Argument order'. The interface is divided into several sections:

- App Items:** A sidebar on the left with a 'Section' field containing 'Section' and a 'Files/Folders' section with 'Multiple Input Files' (containing 'Files') and 'Input File' (with 'File' and 'Browse' buttons). Below that is 'Input Folder' (with 'Folder' and 'Browse' buttons) and 'Text/Numerical Input' (with 'Info Text', 'Single-line Text' (containing 'Text'), and 'Multi-line Text' (containing 'Text')).
- DE Word Count (multi-file) Details:**
 - Tool used:** 'wc 0.0.1'
 - App name:** 'DE Word Count (multi-file)'
 - App description:** 'Counts the number of words in a set of files'
 - Parameters:** A list of checkboxes: 'count lines' (unchecked), 'count words' (checked), 'count characters' (checked), and 'count bytes' (checked). Below is 'Input Files:' with an empty list and '+ Add' and '- Delete' buttons.
- Details: count lines:** A panel on the right for the 'count lines' parameter.
 - Checkbox label:** 'count lines'
 - Checked:** '-1' (with 'Value when checked' field)
 - Not Checked:** 'Argument option when NOT ch' (with 'Value when NOT checked' field)
 - Check item by default
 - Do not display this item in the app.
 - Tool tip text:** 'Enter tool tip here'
- Command line view:** A text area at the bottom containing 'WC -w, -m, -c,'.

Workflow editor (a bit rudimentary though...)



Automate Workflow

Save

Switch View

+ Add - Remove ↑ Move Up ↓ Move down

Step	Name	Description
1	Compress with bzip2 (for pipeline)	Bzip2 file compression utility. Make it smaller! Writes to a file 'co...
2	DE Word Count (multi-file)	Counts the number of words in a set of files

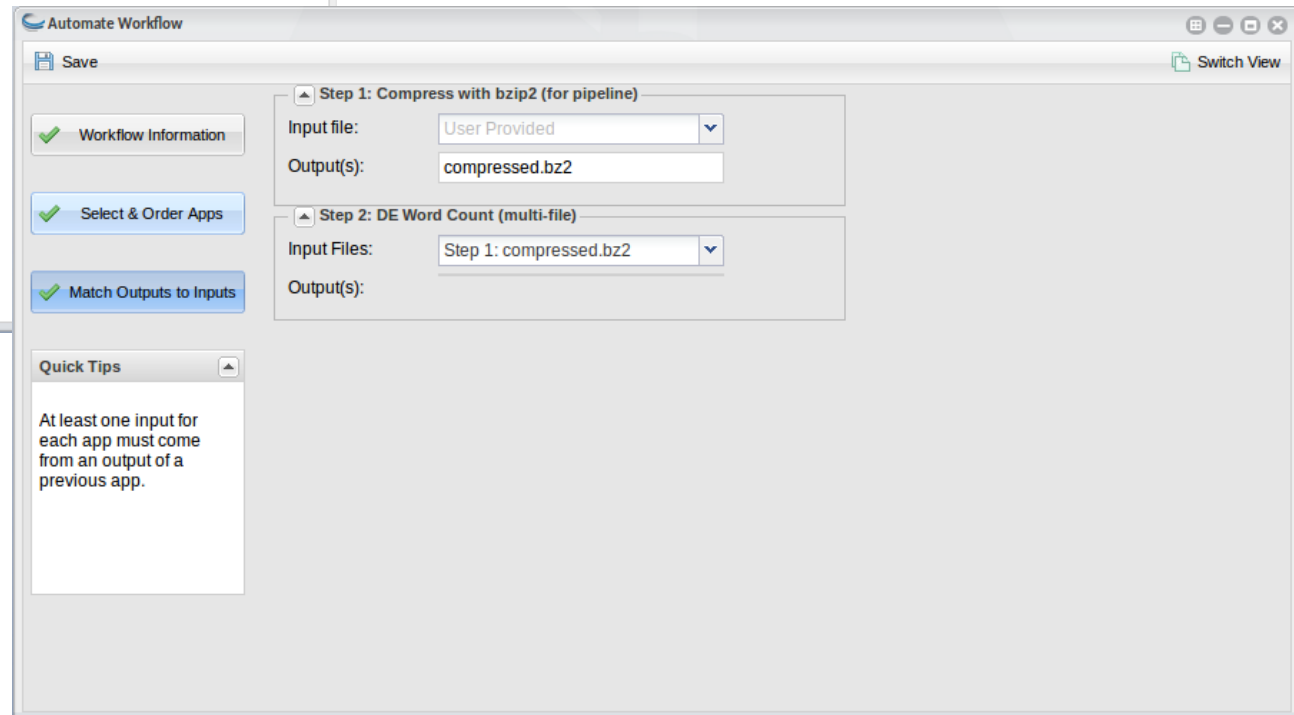
Workflow Information

Select & Order Apps

Match Outputs to Inputs

Quick Tips

You must add 2 or more apps to this workflow.



Automate Workflow

Save

Switch View

Step 1: Compress with bzip2 (for pipeline)

Input file: User Provided

Output(s): compressed.bz2

Step 2: DE Word Count (multi-file)

Input Files: Step 1: compressed.bz2

Output(s):

Quick Tips

At least one input for each app must come from an output of a previous app.

Tool interface

(for software authors or internal people, who build images that get sent to internal Docker registry)

Request New Tool Installation

Discovery Environment **tools** are executables or binaries from which apps are built.

Before a new app can be created, the tool that app is based on must be made available in the DE.

This form contains the required information needed for our team to install your new tool.

Once your request is submitted, you will receive status updates as the installation progresses.

Tool Information

* What is the name of the tool ?

* Please briefly describe the tool:

Tool Admin

* Tool Name:

Description:

* Type:

Attribution:

Version:

* Location:

Tool Implementation

* Implementor:

* Implementor Email:

Sample Input Files:

File Name(s):

Sample Output Files:

File Name(s):

Container Image

* Name:

Tag:

URL:

Container Details

Name:

Working Directory:

WARNING: Do not add a tool without an Entry Point setting if its Docker image also does not have a default 'ENTRYPOINT'. If a tool like this is required, then its Network Mode setting should be set to 'none' to contain any risky scripts run by this tool.

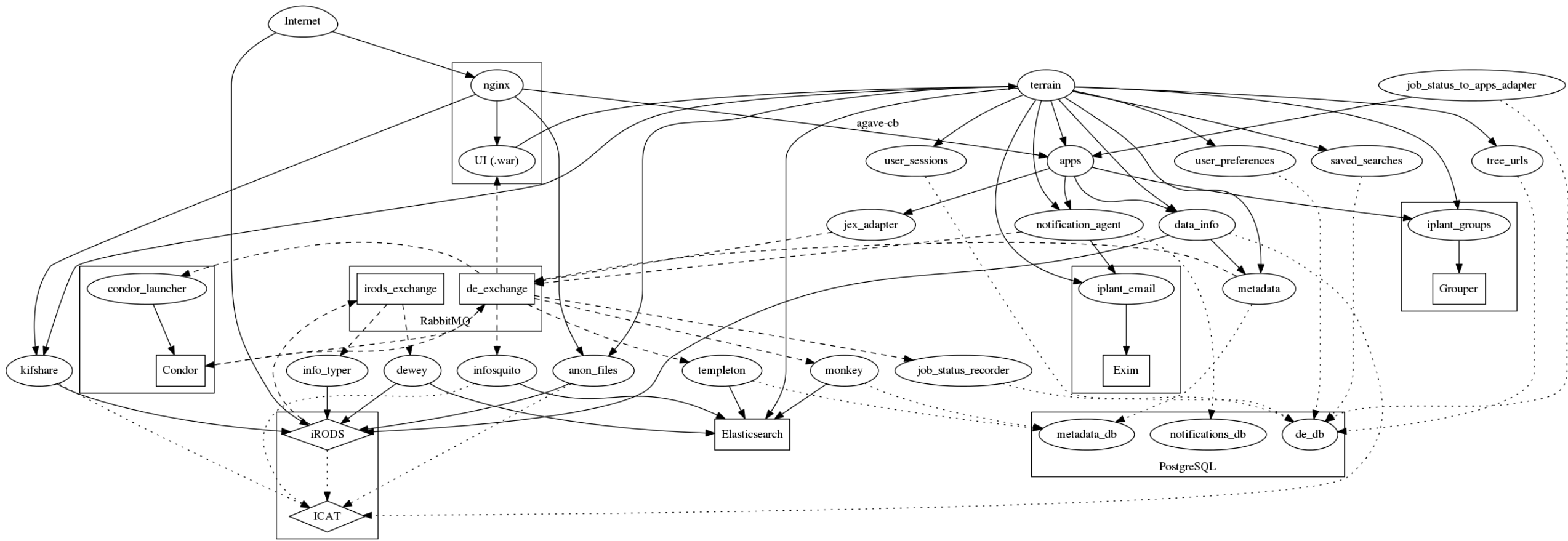
Entry Point:

Memory Limit:

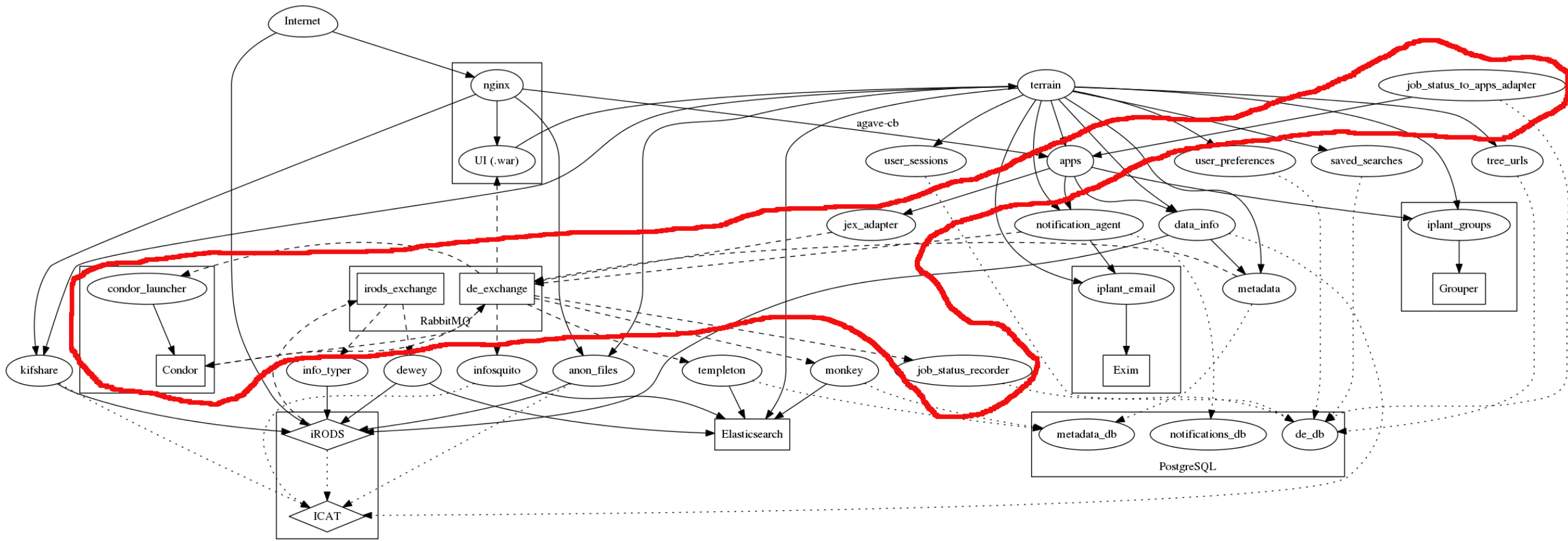


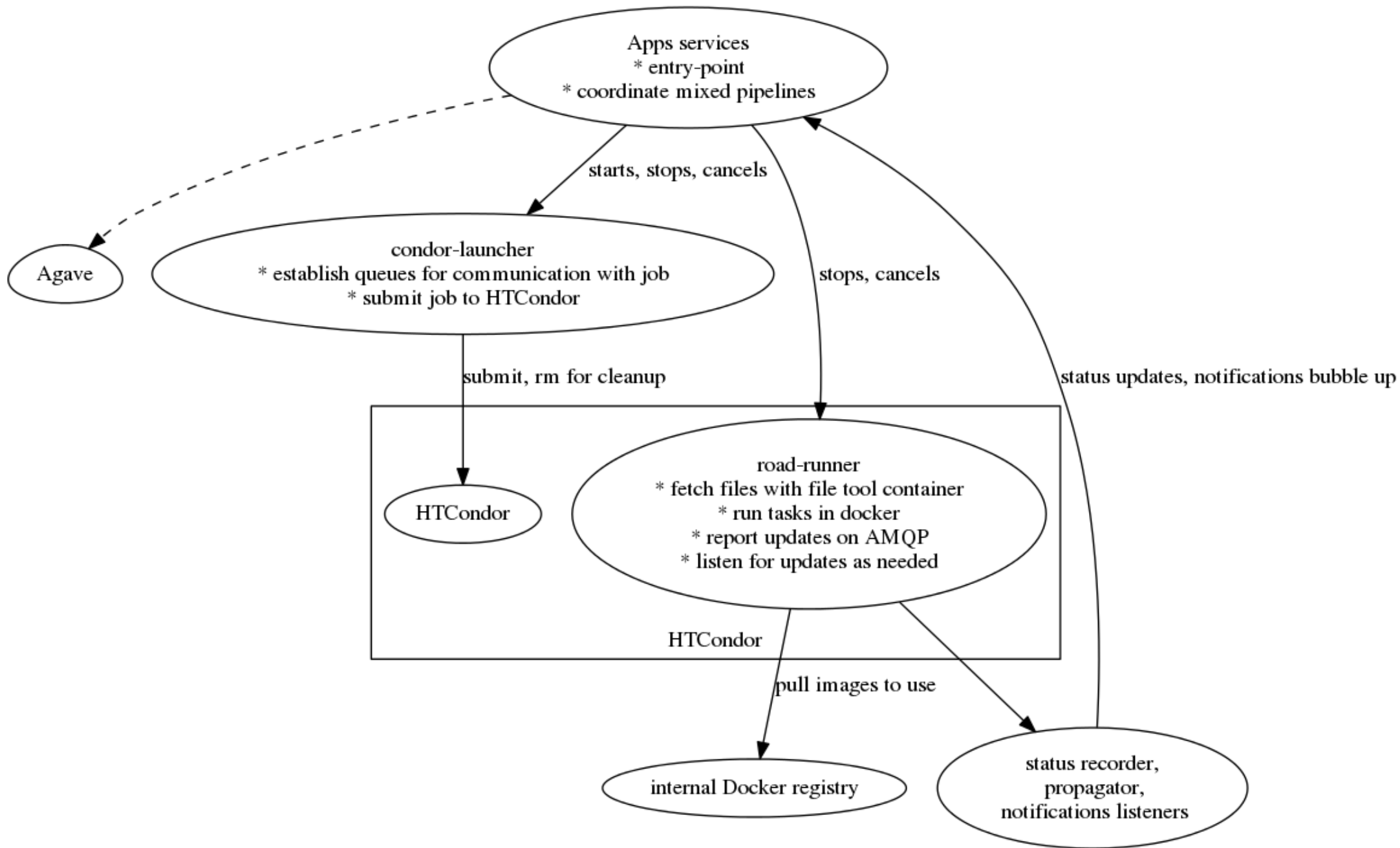
Now let's talk about just the HTCondor part, behind the scenes





(maybe I should focus this a bit)





Future (potential) Work & Ideas

- Time limitation (e.g., 48h per job, then request more)
- Other computational limits (number of processes (no forkbombs!), memory, CPU, network)
- Interactive jobs (iPython etc., or even just for exploration)
- More complex (non-linear, more flexible) workflows
- Public APIs and command-line tools (expand into more technical users)
- “connect your own compute” – spin up machines with appropriate software, put keys/other credentials into DE and have your own jobs go to non-limited, no-waiting compute you control

(question time!)

<https://de.iplantcollaborative.org/>
(try it out yourself)

more to say/ask?

Ian McEwen
University of Arizona
mian@cyverse.org

