Arvados - Bug #13330

Clean up container input collection naming and properties

04/04/2018 08:11 PM - Tom Morris

Status:	Resolved		**Start date:**	06/04/2018
Priority:	Normal		**Due date:**	
Assigned To:	Fuad Muhic		**% Done:**	100%
Category:			**Estimated time:**	0.00 hour
Target version:	2018-07-18 Sprint			

Description

Currently container outputs and logs have recognizable names and a specific set of properties, e.g.

```
"properties": {
    "type": "output",
    "container_request": "dhhck-xvhdp-yyrg8t00dwajq45"
},
```

Similar naming and properties should be provided for the temporary input collections which are created by arvados-cwl-runner which currently have no properties and have names of the form "New collection (2018-03-08T08:34:40.636Z)". They should also inherit the intermediate output time-to-live setting.

Subtasks:

Task # 13352: Review 13330-collection-save	**Resolved**
Task # 13754: Review 13330-cwl-intermediate-collections-cleanup	**Resolved**

Associated revisions

Revision 1dc6c511 - 06/07/2018 03:20 PM - Fuad Muhic

Merge branch '13330-collection-save'

refs #13330

Arvados-DCO-1.1-Signed-off-by: Fuad Muhic <fmuhic@capeannenterprises.com>

Revision b478b8a0 - 07/10/2018 02:20 PM - Fuad Muhic

Merge branch '13330-cwl-intermediate-collections-cleanup'

closes #13330

Arvados-DCO-1.1-Signed-off-by: Fuad Muhic <fmuhic@capeannenterprises.com>

History

#1 - 04/04/2018 08:14 PM - Tom Morris

- Subject changed from Clean up container input naming and properties to Clean up container input collection naming and properties

#2 - 04/11/2018 03:46 PM - Peter Amstutz

- Tracker changed from Bug to Feature

- Description updated

#3 - 04/11/2018 04:11 PM - Peter Amstutz

- Assigned To set to Peter Amstutz

#4 - 04/11/2018 04:17 PM - Peter Amstutz

- Assigned To changed from Peter Amstutz to Fuad Muhic

#5 - 04/17/2018 02:11 PM - Peter Amstutz

This ticket is about Collection objects which are created by arvados-cwl-runner:

arvados/sdk/cwl/arvados_cwl/

```
grep --color -nHr -e Collection\( *.py
arvcontainer.py:117:                    vwd = arvados.collection.Collection(api_client=self.arvrunner.api,
arvjob.py:55:              vwd = arvados.collection.Collection(api_client=self.arvrunner.api,
__init__.py:293:        final = arvados.collection.Collection(api_client=self.api,
pathmapper.py:119:          collection = arvados.collection.Collection(api_client=self.arvrunner.api,
pathmapper.py:160:             c = arvados.collection.Collection(api_client=self.arvrunner.api,
pathmapper.py:175:             c = arvados.collection.Collection(api_client=self.arvrunner.api,
runner.py:288:   collection = arvados.collection.Collection(api_client=arvrunner.api,
```

In each case, the collection should:

- Get an auto-generated name that describes its purpose
- Set the "trash_at" time if the intermediate output time-to-live (ttl) is greater than zero (trash_at set to (now + output ttl))
- Set "type" and "container_request" in the "properties" field of the collection.
    - Use "type": "intermediate"
    - Use "container": current_container["uuid"] (get the current container using self.api.containers().current().execute(num_retries=self.num_retries), see ArvCwlRunner.set_crunch_output for an example)
- You should add "trash_at" and "properties" parameters to arvados.collection.Collection.save_new()
- Remove the existing behavior of searching for and reusing existing collections. It should always create a new collection with the correct name, trash_at time, and properties.

#6 - 04/25/2018 02:52 PM - Peter Amstutz

- Target version changed from 2018-04-25 Sprint to 2018-05-09 Sprint

#7 - 04/25/2018 03:00 PM - Tom Morris

- Tracker changed from Feature to Bug

#8 - 04/27/2018 03:19 PM - Peter Amstutz

13330-intermediates-test has a test workflow (and a bugfix), run it with:

arvados-cwl-runner --api=containers --local run_in_single.cwl

#9 - 05/09/2018 08:35 AM - Fuad Muhic

- Status changed from New to In Progress

#10 - 05/09/2018 03:21 PM - Fuad Muhic

- Target version changed from 2018-05-09 Sprint to 2018-05-23 Sprint

#11 - 05/22/2018 05:47 PM - Peter Amstutz

Since this involves updating both the Python SDK and arvados-cwl-runner, can you put the Python SDK changes and a-c-r changes in separate branches, so we can merge the Python SDK changes first? We will need to update the minimum Python SDK version in a-c-r setup.py, and we don't know what that version is until it is merged.

#12 - 05/23/2018 02:50 PM - Tom Morris

- Target version changed from 2018-05-23 Sprint to 2018-06-06 Sprint

#13 - 06/04/2018 09:10 PM - Peter Amstutz

- I found a typo, should be if trash_at: (this appears in two places)

```
    if storage_classes:
        body["trash_at"] = trash_at
```

- Please mention in the documentation string for "properties" that it will replace any existing properties.

- Could you add access methods get_properties() and get_trash_at() to the Collection class which return the values of those respective fields from self._api_response?
  If self._api_response is None then get_properties() should return {} and get_trash_at() should return None.

trash_at should return/accept a Python datetime object instead of/in addition to a string. (use ciso8601 for datetime parsing and conversion).

#14 - 06/05/2018 01:28 PM - Peter Amstutz

Running tests here https://ci.curoverse.com/job/developer-run-tests/736/

Assuming that passes 13330-collection-save @ 4f0c3d501d19bed5915d5d188598d3a7f1dec7f8 LGTM

#15 - 06/05/2018 01:30 PM - Peter Amstutz

To follow up, the next step after merging will be to determine the timestamped package version (one way to do that is to look at the build-packages logs on jenkins) and then update the setup.py for arvados-cwl-runner so that the minimum version is the one you just merged.

#16 - 06/06/2018 02:49 PM - Fuad Muhic

- Target version changed from 2018-06-06 Sprint to 2018-06-20 Sprint

#17 - 06/07/2018 01:02 PM - Fuad Muhic

I made a change to 13330-collection-save. I updated failing test, and its passing locally when merged with newest master. Could you please rerun tests on Jenkins to verify that.

#18 - 06/20/2018 03:23 PM - Fuad Muhic

- Target version changed from 2018-06-20 Sprint to 2018-07-03 Sprint

#19 - 07/02/2018 05:02 PM - Fuad Muhic

My initial idea for tests was to patch save_new method and check if it's called with corrects parameters (in run method of ArvadosJob, ArvadosContainer and ArvPathMapper). That would require a lot of mocks for each run method and tests would be very fragile to any code change, so I ended up just testing get_intermediate_collection_info method instead. If you have any better ideas please let me know.

#20 - 07/03/2018 03:18 PM - Tom Morris

- Target version changed from 2018-07-03 Sprint to 2018-07-18 Sprint

#21 - 07/05/2018 02:48 PM - Peter Amstutz

The branch is 13330-cwl-intermediate-collections-cleanup

#22 - 07/05/2018 03:10 PM - Peter Amstutz

Could you please merge latest master

#23 - 07/05/2018 03:39 PM - Peter Amstutz

Comments:

- When computing trash_at, please use datetime.datetime.utcnow() instead of datetime.datetime.now()
- Please merge master, then add current_container to the RuntimeContext object so that you don't have to request the current container on every instance of run()
- Can we avoid defining get_collection_attributes() twice? It could be implemented as a plain function which is called by both ArvPathMapper and ArvadosContainer.
- {"type": "Intermediate"} should be "intermediate" (lower case).

#24 - 07/09/2018 01:48 PM - Fuad Muhic

13330-cwl-intermediate-collections-cleanup is ready for another look.
test-run: https://ci.curoverse.com/job/developer-run-tests/788/

#25 - 07/09/2018 02:41 PM - Peter Amstutz

arvcontainer.py:172 vwd.save_new() is missing owner_uuid=self.arvrunner.project_uuid (this looks like it was a bug in the previous code)

arvjob.py:83   same thing with vwd.save_new()

arvcontainer.py:251 should be setting container_request["output_name"] using the name of the workflow step, something like "output for step %s" % (self.name)

get_intermediate_collection_info() should include the name of the step and generate a name like intermediate collection for step %s" % (self.name)

util.py:28

    return current_container;

Don't need ; in python

#26 - 07/10/2018 12:41 PM - Fuad Muhic

13330-cwl-intermediate-collections-cleanup is updated and ready for review.
test-run: https://ci.curoverse.com/job/developer-run-tests/794/

#27 - 07/10/2018 02:18 PM - Peter Amstutz

Fuad Muhic wrote:

> 13330-cwl-intermediate-collections-cleanup is updated and ready for review.
> test-run: https://ci.curoverse.com/job/developer-run-tests/794/

This LGTM, thanks!

#28 - 07/10/2018 02:41 PM - Fuad Muhic

- Status changed from In Progress to Resolved

#29 - 07/23/2018 06:52 PM - Tom Morris

- Release set to 13