Arvados - Story #10111

[Workbench][Crunch2] Provenance graph for Container Request

09/20/2016 06:44 PM - Tom Morris

Status:	Resolved	**Start date:**	04/25/2017
Priority:	Normal	**Due date:**	
Assigned To:	Lucas Di Pentima	**% Done:**	67%
Category:		**Estimated time:**	0.00 hour
Target version:	2017-05-24 sprint		

Description

Extract dependency graph from Container Request and pass to existing code which knows how to use GraphViz to format and reuse the rest of the existing infrastructure.

Subtasks:

Task # 11628: Rwview 10111-collection-labels	**Resolved**
Task # 11381: Review 10111-cr-provenance-graph	**Resolved**
Task # 11619: Review 10111-cr-prov-regression-fixes	**Resolved**

Associated revisions

Revision b075d1be - 04/26/2017 06:56 PM - Lucas Di Pentima

Merge branch '10111-cr-provenance-graph'
Closes #10111

Revision 6fd54bd2 - 05/04/2017 06:30 PM - Lucas Di Pentima

Merge branch '10111-cr-prov-regression-fixes'
Refs #10111

Revision f5fbc488 - 05/16/2017 02:32 PM - Lucas Di Pentima

Merge branch '10111-collection-labels'
Closes #10111

History

#1 - 09/20/2016 06:45 PM - Tom Morris

- Subject changed from Provenance graph for Container Request to [Crunch2] Provenance graph for Container Request

#2 - 10/26/2016 07:47 PM - Tom Morris

- Story points set to 2.0

Extract dependency graph from Container Request and pass to existing code which knows how to use GraphViz to format and reuse the rest of the existing infrastructure.

#3 - 11/08/2016 07:34 PM - Tom Morris

- Target version set to 2016-11-23 sprint

#4 - 11/09/2016 08:52 PM - Tom Clegg

- Assigned To set to Tom Clegg

#5 - 11/09/2016 09:20 PM - Tom Clegg

- Assigned To deleted (Tom Clegg)

#6 - 11/09/2016 09:21 PM - Tom Clegg

- Target version changed from 2016-11-23 sprint to Arvados Future Sprints

#7 - 12/09/2016 02:57 PM - Tom Morris

- Description updated

#8 - 03/20/2017 08:42 PM - Tom Morris

- Target version changed from Arvados Future Sprints to 2017-04-12 sprint

#9 - 03/29/2017 07:29 PM - Tom Morris

- Tracker changed from Bug to Story

- Subject changed from [Crunch2] Provenance graph for Container Request to [Workbench][Crunch2] Provenance graph for Container Request

- Assigned To set to Lucas Di Pentima

#10 - 04/11/2017 02:38 PM - Lucas Di Pentima

- Status changed from New to In Progress

#11 - 04/12/2017 07:02 PM - Lucas Di Pentima

- Target version changed from 2017-04-12 sprint to 2017-04-26 sprint

#12 - 04/24/2017 08:11 PM - Lucas Di Pentima

Updates at 93c92875aaebe5b06f8dbfe2822b59a772895c08 (branch 10111-cr-provenance-graph)

Radhika: While I start working on the tests, I would like to check with you if this is the correct approach, and if there are missing elements on the graph to be included.

#13 - 04/25/2017 03:52 AM - Lucas Di Pentima

Merged master && added some tests: 4ccbea9ef
Test run: https://ci.curoverse.com/job/developer-run-tests/247/

#14 - 04/25/2017 02:17 PM - Radhika Chippada

For a CR -> Use Inputs from Mounts + output uuid + log uuid as the nodes

And then need to get all child CRs for this CR and repeat the above.

#15 - 04/25/2017 07:45 PM - Lucas Di Pentima

Updates at 30146198f
Test run: https://ci.curoverse.com/job/developer-run-tests/251/

- Removed container_image and requesting_container from the graph.
- Added child CRs with their own mounts/output/log.

#16 - 04/25/2017 10:41 PM - Radhika Chippada

The graph is looking pretty good. A few observations:

- As we discussed, I think we should only use input collections (plus output and log uuids) for the CR in “find_collections cr[:mounts]” and exclude any other collections, if any. I think you probably need to look for collection uuids / pdhs in the segment returned by application_helper.get_cwl_inputs?

- extra white space at line ends in _show_provenance.html.erb

Thanks.

#17 - 04/26/2017 12:13 PM - Lucas Di Pentima

Updates: 260e85a9d

> - As we discussed, I think we should only use input collections (plus output and log uuids) for the CR in “find_collections cr[:mounts]” and exclude any other collections, if any. I think you probably need to look for collection uuids / pdhs in the segment returned by application_helper.get_cwl_inputs?

Ah yes, I thought that all mounts that specified a UUID/PDH were implicitly an input.

I have changed the code so it searches only for collections inside the /var/lib/cwl/cwl.input.json mount, that as I understand by reading the get_cwl_inputs() helper, it's the object describing the input mounts.
The result is that if a CR step is created and have mounts that aren’t from a CWL definition (example: bwa command execution that uses FUSE), those mounts on the child CR won’t be included in the graph (ie: 9tee4’s /container_requests/9tee4-xvhdp-29wnyz1npk9bycs#Provenance), is that ok or should I search for “any collection” inside mounts when not using arvados-cwl-runner on the command?

> - extra white space at line ends in _show_provenance.html.erb

Oops! done.

Another question: Currently I'm showing the Provenance tab on those CRs with state != Uncommitted, should I change that to only CR in Final state?

#18 - 04/26/2017 02:43 PM - Radhika Chippada

> I have changed the code so it searches only for collections inside the /var/lib/cwl/cwl.input.json mount, that as I understand by reading the get_cwl_inputs() helper, it's the object describing the input mounts. The result is that if a CR step is created and have mounts that aren't from a CWL definition (example: bwa command execution that uses FUSE), those mounts on the child CR won't be included in the graph (ie: 9tee4's /container_requests/9tee4-xvhdp-29wnyz1npk9bycs#Provenance), is that ok or should I search for "any collection" inside mounts when not using arvados-cwl-runner on the command?

Yes, this seems problematic. I think we should check for /keep/<pdh> format instead in mounts to get the input collections. Please confirm with Peter. Thanks.

> Another question: Currently I'm showing the Provenance tab on those CRs with state != Uncommitted, should I change that to only CR in Final state?

Comparing with pipeline_instances and jobs, this seems correct (to show graph for Queued etc)

#19 - 04/26/2017 04:33 PM - Lucas Di Pentima

Update at: 88c241d7c
Test run will be on: https://ci.curoverse.com/job/developer-run-tests/257/

Search for all PDHs on "mounts" on cases when cwl.input.json is not included. As talked with Radhika & Bryan, outputs aren't listed using PDHs, just paths. So there's no possibility of including an output as an input.

#20 - 04/26/2017 06:31 PM - Lucas Di Pentima

Update at: edfc619e6

As requested on the sprint review meeting, changed the graph edges from "cr" to "child" and "mounts" to "input".

#21 - 04/26/2017 06:48 PM - Radhika Chippada

LGTM @ edfc619

#22 - 04/26/2017 07:00 PM - Lucas Di Pentima

- Status changed from In Progress to Resolved

- % Done changed from 0 to 100

Applied in changeset arvados|commit:b075d1be1377760f5d8497a29f63c8e416cd5378.

#23 - 05/04/2017 02:28 PM - Peter Amstutz

- Status changed from Resolved to Feedback

- Target version changed from 2017-04-26 sprint to 2017-05-10 sprint

Additional comments:

- Needs to use PDH so that inputs and output match up. For example, in this graph the output of "rev" is supposed to be an input of "sort": https://workbench.9tee4.arvadosapi.com/container_requests/9tee4-xvhdp-w382mn52hn18oad#Provenance
- Label collection inputs by name. If the collection shows up under multiple different names, prefer the name of the collection in the current project. Otherwise pick any name and render it something like "HWI-ST1027_129_D0THKACXX for CWL tutorial + 4 more"
- Don't render "log" outputs. They are just clutter.
- I'm not sure if its a good idea to render "child" links. If you have 300 child containers it is just a lot of lines providing very little information. Consider using a graphviz "subgraph" or "cluster".
- To determine the inputs of a container request, recursively search "mounts" for JSON fields that look like "portable_data_hash": "abc+123" and "location": "keep:abc+123"

#24 - 05/04/2017 02:58 PM - Peter Amstutz

In the interests of time, let's limit it to these changes:

- Needs to use PDH so that inputs and output match up. For example, in this graph the output of "rev" is supposed to be an input of "sort": https://workbench.9tee4.arvadosapi.com/container_requests/9tee4-xvhdp-w382mn52hn18oad#Provenance (this is the most important change, because the current behavior it is effectively a regression from the equivalent functionality for jobs)
- Don't render "log" outputs. They are just clutter.
- To determine the inputs of a container request, recursively search "mounts" for JSON fields that look like "portable_data_hash": "abc+123" and "location": "keep:abc+123" (this should ensure that nothing is missed)

#25 - 05/04/2017 05:37 PM - Lucas Di Pentima

Fixes at a4a8d41f6 - branch 10111-cr-prov-regression-fixes
Test run: https://ci.curoverse.com/job/developer-run-tests/273/

#26 - 05/04/2017 09:40 PM - Lucas Di Pentima

Branch 10111-collection-labels - commit 39755f764
Test run: https://ci.curoverse.com/job/developer-run-tests/274/

Added better collection labelling on CR provenance graph.

#27 - 05/05/2017 01:51 PM - Lucas Di Pentima

An integration test was failing, updated fix at e01823785
New test run: https://ci.curoverse.com/job/developer-run-tests/275/

#28 - 05/05/2017 02:08 PM - Peter Amstutz

Additional comment: where we have a container request with an explicit output_uuid, make sure to use the label corresponding to the name of the collection in output_uuid, before falling back on the logic outlined in note-24

#29 - 05/05/2017 02:47 PM - Lucas Di Pentima

Updates at 4259263d2
Test run: https://ci.curoverse.com/job/developer-run-tests/276/

Addressed issue about naming output collections after the cr's output_uuid collection reference.

#30 - 05/05/2017 03:15 PM - Peter Amstutz

How hard would it be to fix the hyperlinks so that when you have a specific UUID associated with a collection, clicking on it takes you directly to it and not to the "this PDH has multiple collections" page?

#31 - 05/05/2017 03:32 PM - Peter Amstutz

Another note. For labeling, if there are multiple collections but they have the same name, you don't need the "+N more"

It's making a separate API call for every collection. That adds a lot of latency. It should find all the PDHs in the graph, make a batch request for them all, and then filter on the workbench side.

#32 - 05/08/2017 03:27 PM - Lucas Di Pentima

Updates at b29ca38e4
Test run: https://ci.curoverse.com/job/developer-run-tests/278/

- Refactored the graph creation code for CR so that it minimizes the amount of API calls when looking for information about outputs, inputs and childs.
- For input collections, when there are more than one with the same name, don't add the "+N more" to the name label.
- For output collections, added an option on describe_node() helper function so that the graph node is referenced by PDH, but link urls are rendered by UUID so they take the user to the specific collection page when clicking on it.

#33 - 05/10/2017 07:04 PM - Lucas Di Pentima

- Target version changed from 2017-05-10 sprint to 2017-05-24 sprint

#34 - 05/11/2017 03:37 PM - Peter Amstutz

Ok, for large workflows, it still takes forever to load, but it seems that "dot" is the bottleneck now. We need to rethink representation, but not for this story (I'm putting that on a new ticket, #11680).

On the implementation:

The intended way to call GenerateGraph() was with pdata to contain all the nodes that will be used in the graph. In order to have better separation of

concerns, would it make sense for the new code in generate_provenance_edges() that does the batch queries to move to container_requests_controller#generate_provenance ?

#35 - 05/15/2017 09:46 PM - Lucas Di Pentima

Updates at 795bf007c
Test run: https://ci.curoverse.com/job/developer-run-tests/284/

Moved the code related to API requests to the CR controller.

#36 - 05/16/2017 02:08 PM - Peter Amstutz

Thanks. This is a much better separation of concerns.

I'm unhappy with how it behaves with large graphs, but instead of continuing to go around back and forth I think we should merge 795bf007cbe24775bd348fb40fc5c28d93c8f23d and schedule a grooming session to figure out how rendering can be improved.

LGTM.

#37 - 05/16/2017 02:35 PM - Lucas Di Pentima

- Status changed from Feedback to Resolved

- % Done changed from 33 to 100

Applied in changeset arvados|commit:f5fbc48810d1397df9e6244c16cf07c05162d36a.