Arvados - Bug #13201

arvados-cwl-runner workflow checks are expensive and repeated twice when run with --submit

03/09/2018 08:41 AM - Joshua Randall

Status:	Resolved	**Start date:**	03/30/2018
Priority:	Normal	**Due date:**	
Assigned To:	Peter Amstutz	**% Done:**	100%
Category:	Crunch	**Estimated time:**	0.00 hour
Target version:	2018-04-11 Sprint		

Description

When run with `--submit`, arvados-cwl-runner (a-c-r) first performs a validation ("workflow checker") step, which in the case of complex workflows can be costly. On one of our current workflows we have to wait close to 40m before the runner container request is even created:

```
$ arvados-cwl-runner --eval-timeout 10 --submit-runner-ram 16384 --submit-runner-image mercury/arv
ados-jobs:sandbox
jsdebug21 --submit --no-wait --api=containers workflows/gatk-4.0.0.0-genotype-gvcf.cwl 20180303-11
00-a-c-r-submit-no-wait-21.gvcf-output.vcf_input.json
2018-03-08 23:08:31 cwltool INFO: /usr/bin/arvados-cwl-runner 1.0.20180216164101, arvados-python-c
lient 0.1.20171211211613, cwltool 1.0.20180130110340
2018-03-08 23:08:31 cwltool INFO: Resolved 'workflows/gatk-4.0.0.0-genotype-gvcf.cwl' to 'file:///
home/mercury/checkouts/arvados-pipelines/cwl/workflows/gatk-4.0.0.0-genotype-g
vcf.cwl'
2018-03-08 23:09:39 cwltool WARNING: Workflow checker warning:
workflows/gatk-4.0.0.0-genotype-gvcf.cwl:90:9: Source 'variant-index' of type {"items": ["null", "
File"], "type": "array"} is partially incompatible
workflows/gatk-4.0.0.0-genotype-gvcf.cwl:99:7:   with sink 'secondary_files' of type {"items": ["F
ile", {"type": "array", "items": "File"}], "type": "array"}
2018-03-08 23:09:43 cwltool WARNING: Workflow checker warning:
workflows/gatk-4.0.0.0-genotype-gvcf.cwl:90:9: Source 'variant-index' of type {"items": ["null", "
File"], "type": "array"} is partially incompatible
workflows/gatk-4.0.0.0-genotype-gvcf.cwl:99:7:   with sink 'secondary_files' of type {"items": ["F
ile", {"type": "array", "items": "File"}], "type": "array"}
2018-03-08 23:45:14 arvados.cwl-runner INFO: [container gatk-4.0.0.0-genotype-gvcf.cwl] submitted
container ncucu-xvhdp-as06624s5xo96gn
ncucu-xvhdp-as06624s5xo96gn
2018-03-08 23:45:14 cwltool INFO: Final process status is success
```

Once a-c-r on the RunnerContainer starts, it performs those same workflow checks again. It might make sense to do that if these checks are cheap, as I guess the submitted workflow may have been slightly modified (e.g. to map inputs to keep locators, etc) and this could catch errors. However, since they apparently are not that cheap, perhaps an optimisation could be made so that a-c-r in the runnercontainer trusts that the workflow has already been validated before submission?

In the case of the above example, the runner container stderr from a-c-r begins with:

```
2018-03-08T23:46:18.223467681Z cwltool INFO: /usr/bin/arvados-cwl-runner 1.0.20180221170604, arvad
os-python-client 0.1.20180221170604, cwltool 1.0.20180130110340
2018-03-08T23:46:18.228709270Z cwltool INFO: Resolved '/var/lib/cwl/workflow.json#main' to 'file:/
//var/lib/cwl/workflow.json#main'
2018-03-08T23:47:17.319055335Z cwltool WARNING: Workflow checker warning:
2018-03-08T23:47:17.319055335Z ../../lib/cwl/workflow.json:1:59914: Source 'variant-index' of type
 {"items": ["null", "File"], "type": "array"} is partially incompatible
2018-03-08T23:47:17.319055335Z ../../lib/cwl/workflow.json:1:60174:   with sink 'secondary_files'
of type {"items": ["File", {"items": "File", "type": "array"}], "type": "array"}
2018-03-08T23:47:27.223296520Z cwltool WARNING: Workflow checker warning:
2018-03-08T23:47:27.223296520Z ../../lib/cwl/workflow.json:1:59914: Source 'variant-index' of type
 {"items": ["null", "File"], "type": "array"} is partially incompatible
2018-03-08T23:47:27.223296520Z ../../lib/cwl/workflow.json:1:60174:   with sink 'secondary_files'
of type {"items": ["File", {"items": "File", "type": "array"}], "type": "array"}
2018-03-09T00:23:17.752082746Z cwltool INFO: [workflow workflow.json#main] start
```

Subtasks:	
Task # 13291: Add "already validated" flag	**Resolved**
Task # 13292: Review 13201-less-validate	**Resolved**

Associated revisions

Revision c4bd314f - 04/05/2018 05:25 PM - Peter Amstutz

Merge branch '13201-less-validate' closes #13201

Arvados-DCO-1.1-Signed-off-by: Peter Amstutz <pamstutz@veritasgenetics.com>

History

#1 - 03/22/2018 03:35 PM - Tom Clegg

(From discussion offline)
https://github.com/curoverse/arvados/blob/fe346bd892c3bcd1ec18adabfe9c6cfa179fa8f6/sdk/cwl/arvados_cwl/pathmapper.py#L96-L99 seems to be expensive and (at least in some cases) unnecessary to do even once, let alone twice.

#2 - 03/22/2018 03:36 PM - Tom Morris

- Target version set to To Be Groomed

#3 - 03/28/2018 03:50 PM - Tom Morris

- Target version changed from To Be Groomed to 2018-04-11 Sprint

#4 - 03/28/2018 03:53 PM - Tom Morris

Perhaps as a starting point at least skip the second set of checks when already done once.

#5 - 03/28/2018 04:20 PM - Peter Amstutz

- Assigned To set to Peter Amstutz

#6 - 04/05/2018 02:54 PM - Peter Amstutz

13201-less-validate @ 67a9e63d83a429eadddfa1424a37e010f7c2c365

Skip 2nd redundant validation reloading with updated file locations.

Skip validation on the submitted runner entirely.

(validation includes: schema validation of the document, checking linked files exist, checking types of input/output ports, javascript syntax checking).

#7 - 04/05/2018 05:21 PM - Lucas Di Pentima

Checked this update against the new cwltool, it LGTM.

#8 - 04/05/2018 06:05 PM - Peter Amstutz

- Status changed from New to Resolved

- % Done changed from 50 to 100

Applied in changeset arvados|c4bd314ff8fe1cab2283cca9e09de55706da9606.

#9 - 07/23/2018 06:52 PM - Tom Morris

- Release set to 13