Arvados - Feature #10081

[CWL] Run several steps in single job

09/15/2016 04:23 PM - Peter Amstutz

Status:	Resolved		Start date:	09/16/2016
Priority:	Normal		Due date:	
Assigned To:	Peter Amstutz		% Done:	100%
Category:			Estimated time:	0.00 hour
Target version:	2016-09-28 sprint			

Description
Add workflow hint "arv:RunInSingleContainer" which uses cwltool to run a subworkflow as a single job in order to amortize the overhead of spinning up new jobs.

Subtasks:	
Task # 10086: Support RunInSingleContainer hint	**Resolved**
Task # 10087: Review 10081-cwl-run-same-job	**Resolved**

Associated revisions

Revision 523dadeb - 09/26/2016 12:42 PM - Peter Amstutz

Merge branch '10081-cwl-run-same-job' closes #10081

Revision 69972d44 - 09/26/2016 09:17 PM - Peter Amstutz

Merge branch '10081-update-cwl-runner' refs #10081

History

#1 - 09/15/2016 04:25 PM - Peter Amstutz

- *Description updated*

#2 - 09/15/2016 04:29 PM - Peter Amstutz

- *Description updated*

#3 - 09/16/2016 03:09 PM - Peter Amstutz

- *Assigned To set to Peter Amstutz*

#4 - 09/16/2016 06:57 PM - Radhika Chippada

- TestWorkflow failing with run-tests because scatter2.cwl is not in tests/wf dir (it is in tests dir)

- I copied it into the tests/wf dir, but still failing (I did a reinstall as well)

```
======================================================================
ERROR: test_run (tests.test_job.TestWorkflow)
----------------------------------------------------------------------
Traceback (most recent call last):
  File "/home/radhika/arvados/sdk/cwl/tests/test_job.py", line 220, in test_run
    it.next().run()
  File "/home/radhika/arvados/sdk/cwl/arvados_cwl/arvjob.py", line 48, in run
    n.write(p.resolved.encode("utf-8"))
  File "/tmp/tmp.VV3pxv7gTR/VENVDIR/local/lib/python2.7/site-packages/arvados/arvfile.py", line 59, in __exit_
_
    self.close()
  File "/tmp/tmp.VV3pxv7gTR/VENVDIR/local/lib/python2.7/site-packages/arvados/arvfile.py", line 1101, in close
    self.flush()
  File "/tmp/tmp.VV3pxv7gTR/VENVDIR/local/lib/python2.7/site-packages/arvados/arvfile.py", line 51, in before_
close_wrapper
    return orig_func(self, *args, **kwargs)
  File "/tmp/tmp.VV3pxv7gTR/VENVDIR/local/lib/python2.7/site-packages/arvados/arvfile.py", line 1097, in flush
    self.arvadosfile.flush()
  File "/tmp/tmp.VV3pxv7gTR/VENVDIR/local/lib/python2.7/site-packages/arvados/arvfile.py", line 238, in synchr
onized_wrapper
    return orig_func(self, *args, **kwargs)
```

```
  File "/tmp/tmp.VV3pxv7gTR/VENVDIR/local/lib/python2.7/site-packages/arvados/arvfile.py", line 936, in flush
    self.parent._my_block_manager().commit_bufferblock(self._current_bblock, sync=sync)
  File "/tmp/tmp.VV3pxv7gTR/VENVDIR/local/lib/python2.7/site-packages/arvados/arvfile.py", line 587, in commit
_bufferblock
    loc = self._keep.put(block.buffer_view[0:block.write_pointer].tobytes(), copies=self.copies)
  File "/tmp/tmp.VV3pxv7gTR/VENVDIR/local/lib/python2.7/site-packages/arvados/retry.py", line 158, in num_retr
ies_setter
    return orig_func(self, *args, **kwargs)
  File "/tmp/tmp.VV3pxv7gTR/VENVDIR/local/lib/python2.7/site-packages/arvados/keep.py", line 1096, in put
    data_hash, copies, writer_pool.done()), service_errors, label="service")
KeepWriteError: failed to write 0c17b076db9ae2ee0b7250d3db394952 (wanted 2 copies but wrote 0): service http:/
/keep1.zzzzz.arvadosapi.com:25107/ responded with 0 (7, 'Failed to connect to keep1.zzzzz.arvadosapi.com port
25107: Connection refused'); service http://keep0.zzzzz.arvadosapi.com:25107/ responded with 0 (28, 'Connectio
n timed out after 2002 milliseconds')
```

#5 - 09/22/2016 02:18 PM - Peter Amstutz

The tests are fixed, thanks for catching that. Please take another look.

#6 - 09/22/2016 07:21 PM - Radhika Chippada

- Any concern about hardcoding this url: https://w3id.org/cwl/arv-cwl-schema.yml? (I couldn't access it using my browser though)

- "raise Exception("Uh oh %s" % obj["location"])" -- may be you can clarify that the location be keep locator with so and so format?

- Does this update result in any unwanted "sequential" ordering of running jobs (instead of parallelization) resulting in longer test run times?

#7 - 09/23/2016 06:20 PM - Peter Amstutz

Radhika Chippada wrote:

> - Any concern about hardcoding this url: https://w3id.org/cwl/arv-cwl-schema.yml? (I couldn't access it using my browser though)

No, it is just pre-populating a cache, so it won't ever try to download from that URL. However I realize I should should probably change the URI to
http://arvados.org/cwl to be consistent with the namespacing of the Arvados hints.

> - "raise Exception("Uh oh %s" % obj["location"])" -- may be you can clarify that the location be keep locator with so and so format?

Ooops, that was a debugging check that should be removed.

> - Does this update result in any unwanted "sequential" ordering of running jobs (instead of parallelization) resulting in longer test run times?

This feature intentionally runs a series of steps in a single job using cwltool. Currently cwltool doesn't parallelize, so it will run those jobs sequentially.
However much more time is saved by avoiding the overhead of spinning up additional crunch jobs than the lost opportunities for parallelism when
each step only runs for a few minutes.

This has no effect on test times.

I'll update the ticket when I've addressed the first two items.

#8 - 09/23/2016 06:28 PM - Peter Amstutz

Actually, while the check was for debugging, it should stay. Improved the exception text.

Now at 8b7d63024652c112973d4dd82f9a5d89cc624fc7

#9 - 09/23/2016 06:53 PM - Radhika Chippada

LGTM

#10 - 09/26/2016 12:45 PM - Peter Amstutz

- Status changed from New to Resolved

- % Done changed from 50 to 100

Applied in changeset arvados|commit:523dadebfbee9a73a21c3f78c7b4af329930d393.