Kepler - Bug #5403

Improve the performance of the workflow archiving data from dataturbine to metacat

05/12/2011 02:29 PM - Jing Tao

Status:	Resolved	**Start date:**	05/12/2011
Priority:	Normal	**Due date:**	
Assignee:	Derik Barseghian	**% Done:**	0%
Category:	sensor-view	**Estimated time:**	0.00 hour
Target version:	sensor-view-0.9.0	**Spent time:**	0.00 hour
Bugzilla-Id:	5403		

Description

Now the workflow iterates each sensor sequentially to archiving data. Probably we can parallel the process.

History

#1 - 07/14/2011 12:30 PM - Derik Barseghian

Dan made significant improvements to this workflow's performance.
Good enough to close this bug?

#2 - 07/14/2011 03:21 PM - Matt Jones

What is the performance difference? What is the difference in workflow run times now with provenance on compared to provenance off?

#3 - 06/15/2012 04:52 PM - Daniel Crawl

Modified EcogridWriter actor:

- input port for object tokens as the source data
- retries if uploading fails
- guaranteed unique doc ids

#4 - 06/15/2012 05:01 PM - Daniel Crawl

The sensor data read from DT is no longer written to file system, but written directly to EcogridWriter via ObjectTokens.

There are parameters to control the concurrency of reading from DT and writing to Metacat. In my tests, I found that increasing the number of writers to Metacat improved performance the best. However, as the data sizes and number of writers increased, I started getting errors from EcogridWriter, such as permission denied...

#5 - 06/19/2012 02:19 PM - Derik Barseghian

I'm retargeting for consideration for kepler-2.4.0, and reassigning. The actor changes could ideally be included in kepler 2.4.0, and after that, the remaining workflow portion of the bug could then be targeted to sensor-view-1.x.y. The actor changes need review, ideally by Jing since he's familiar w/ the actor, and the above attached faster version of the workflow needs to be debugged to find the source of the errors Dan mentions in comment#4. We also need to make sure all the improvements Jing and I have made to the current version of the workflow after publication of the paper are merged into the faster version. I believe the current version has been more thoroughly tested and doesn't have the errors mentioned, so I'd rather stick with it for the initial sensor-view release, even if it's slower.

#6 - 08/17/2012 07:46 PM - Derik Barseghian

Just ran this workflow again (albeit w/ provenance on) and it took 13:54 to run, for not a great deal of data. That's pretty bad.
After discussion w/ Dan, getting at least one of his performance changes in probably won't be hard, and will help a lot: don't use the filesystem when transferring date between dataturbine and ecogrid writer -- swap CreateDataSets actors and use newer EcogridWriter actor. Temporarily retarget to me from Jing, and sensor-view-0.9.0 for this change.

#7 - 08/18/2012 02:28 AM - Derik Barseghian

I got the EcogridWriter changes merged into trunk, finished at r30478.

After this I tried to swap in CreateDataSetsFromDataTurbine for CreateDataSets in the existing archive workflow, but it's not that simple, offhand it seemed like some non-trivial changes would have to be made to the workflow.

So then I tried to run Dan's version of the workflow attached to this bug (with url parameters changed to my test DT, and dev2). I got some Kepler lockup errors initially due to memory. Then I got some authentication errors. I found the EcogridWriter username using dc=sdsc, so I changed to dc=ecoinformatics , but I still get authentication issues. The workflow proceeds til complete despite the error, but I don't think it submits anything (each time, debug msgs on console report dealing w/ the same number of data points). I then ran the original workflow to make sure it wasn't a network/dev2 issue. That succeeded.

I expect I'm missing making a simple config change to the workflow.
After that's done, the summary report composite should be adapted and merged into this workflow, and any other changes Jing and I made since the fork as well. Then some testing.

#9 - 08/24/2012 05:49 PM - Derik Barseghian

Today I did a bunch of testing of the new (archiveModified.xml -- attached) and the existing (archiveDataturbineDataToMetacat.kar) workflows, with and without the EcogridWriter changes, trying to track down why run times were dismally long.

The summary is uploading to dev2 seems to take a minimum of ~125s per datapackage right now, no matter the workflow, or EcogridWriter actor changes. Uploading to chico1, which is further away and a slower machine, for similar amounts of data, takes only ~30s to 40s per datapackage. So we'll have to look into why dev2 is performing so badly later. Jing restarted the tomcat, it didn't help.

So i'm planning to include the EcogridWriter changes and the new version of the workflow in the sensor-view release. I've spliced in the nice end display results that Jing created for the old workflow.

#10 - 08/24/2012 07:35 PM - Derik Barseghian

I forgot to add that the auth error was due to dc=edu instead of dc=org, when trying to save to dev2.

I committed the changes in the comment above, re-saved the workflow beneath sensor-view-0.9.0, and tested it beneath 0.9.0, and it worked. Closing. Will open a different bug for dev2's metacat performance.

#11 - 03/27/2013 02:30 PM - Redmine Admin

Original Bugzilla ID was 5403

Files

EcogridWriter.java	19.5 KB	06/15/2012	Daniel Crawl
archive-8.xml	401 KB	06/16/2012	Daniel Crawl
archiveModified.xml	377 KB	08/18/2012	Derik Barseghian