Lab 15.1: Dataset mirroring using Falcon

|  |  |
| --- | --- |
| **Objective:** | To replicate data from one Cluster to another.  In this case we the source and target cluster will be the same. |
| **Successful Outcome:** | You successfully replicate data from the source cluster to the target cluster. |
| **Before You Begin:** | SSH into node1. |

1. Configure falcon User Permissions
   1. Go the HDFS page in Ambari, then scroll down and expand the Custom core-site.xml section.
   2. The Falcon job you defined is going to be executed by the root user, so root needs permission to communicate with the falcon server. Add root to the hadoop.proxyuser.falcon.groups property:

Description: Macintosh HD:Users:rich:Desktop:Screen Shot 2013-11-11 at 8.37.54 AM.png

* 1. Click the Save button to save your changes to the HDFS config.
  2. Restart all dependent services as marked in Services tab.
  3. **Make sure to restart ‘oozie’ service as well.**

1. Kick off the Dataset mirroring pipeline
   1. Login to node1 and change directory to /root/labs/falcon:

# cd ~/labs/falcon

* 1. Let’sreview cluster and data-feed configuration files:

# cat horton.xml

# cat horton-bkp.xml

# cat data-feed.xml

* 1. Submit above configuration files to Falcon server:

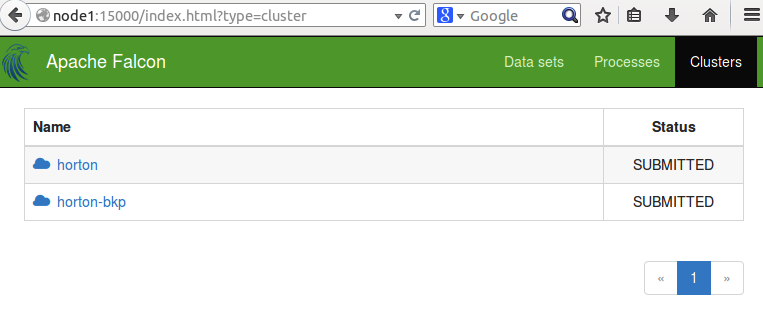
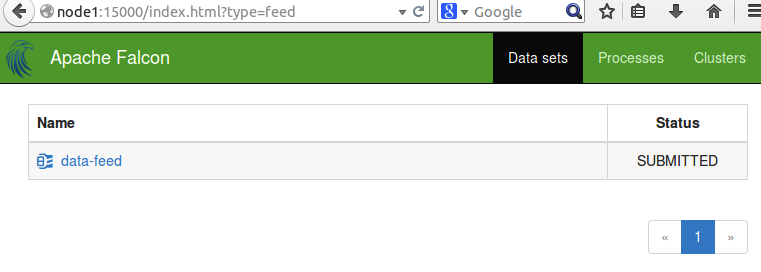
# falcon entity -type cluster -submit -file horton.xml

# falcon entity -type cluster -submit -file horton-bkp.xml

# falcon entity -type feed -submit -file data-feed.xml

* 1. Navigate to following URL and check the status of above entities. They must be in ‘SUBMITTED” state:

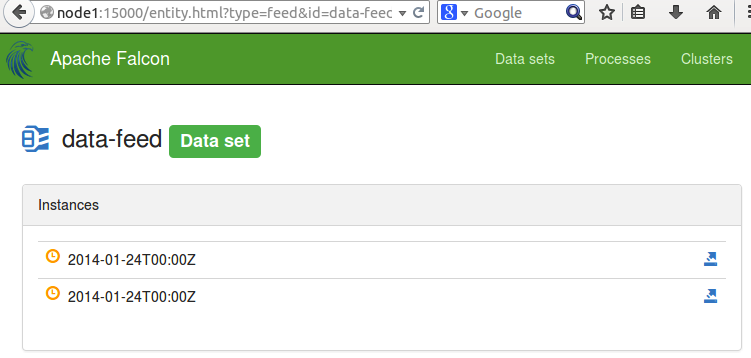
http://node1:15000



* 1. Click on ‘data-feed’. It will show its configuration/definition. It also shows that it is dependent upon ‘horton’ and ‘horton-bkp’ clusters.
  2. Now, schedule the data-feed. Once it is done successfully, it will wait for new feed to process in /user/root/horton/email-logs directory:

# falcon entity -type feed -schedule -name data-feed

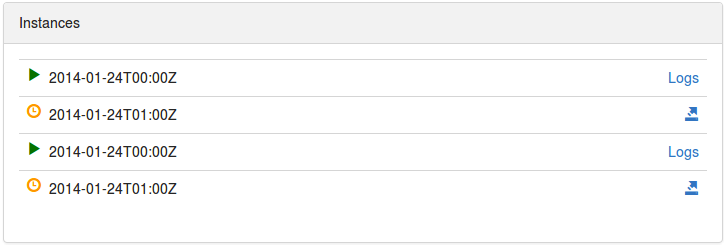
* 1. Now check the status on Falcon UI and you will see 2 instances are running, which are waiting for the data feed:

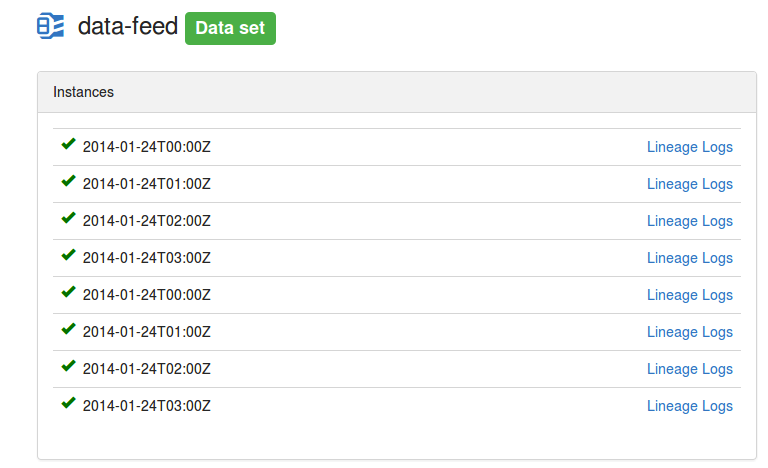


* 1. Now, let’s create an input directory ‘horton’ in HDFS and copy some data (email log):

# hadoop fs -mkdir horton

# hadoop fs -put email-logs horton

* 1. Reload Falcon UI after few seconds. You will see a new set of instances running and they started processing data: 



* 1. Check HDFS for new files and you will see that a new directory ‘horton-bkp’ is created by Falcon instances, which has the same contents as ‘horton’ directory. If you add a new file in ‘horton’ folder now. It will be automatically recognized by Falcon instances after 1 hour. Please review ‘data-feed.xml’ file and check the ‘frequency’ parameter.

1. Delete the data-feed and cluster definition
   1. If you do not want to run this job any more then you can do the cleanup by running following commands:

# falcon entity -type feed -delete -name data-feed

# falcon entity -type cluster -delete -name horton

# falcon entity -type cluster -delete -name horton-bkp