Li, Jun A

From: Zhang, Lijuan
Sent: Monday, January 25, 2016 12:22 PM
To: Lin, XiangX; Li, Jun A
Subject: FW: New issue in Meridian Environment

From: Gao, Fengqian
Sent: Monday, January 25, 2016 9:24 AM
To: Cao, Buddy <buddy.cao@intel.com>; Wang, AlbertZQ <albertzq.wang@intel.com>; Zhang, Lijuan <lijuan.zhang@intel.com>; Yi, Lan <lan.yi@intel.com>
Cc: Lou, Ming <ming.lou@intel.com>; Ding, Danny <danny.ding@intel.com>; Lu, Yuan Y <yuan.y.lu@intel.com>; Chen, Cloud <cloud.chen@intel.com>
Subject: RE: New issue in Meridian Environment

Hi, Albert,

After we talked with Soila and Andy, they suggest us to keep the current configuration.
8GB is a good for the yarn.nodemanager.resource.memory-mb because it leaves room for other system and CDH processes to run.
The main reason for increasing the spark driver memory was because Spark's random forest needed more memory at the driver due to the large number of trees.

Thanks
--fengqian

From: Cao, Buddy
Sent: Friday, January 22, 2016 4:53 PM
To: Wang, AlbertZQ; Gao, Fengqian; Kavulya, Soila P; Zhang, Lijuan; Maczuga, Andrzej; Yi, Lan
Cc: Tumialis, Adam; Lou, Ming; Ding, Danny; Lu, Yuan Y; Chen, Cloud
Subject: RE: New issue in Meridian Environment

Albert,

There is no an easy answer, need clearly understand the concept & architecture of Yarn RM/NM/AppMaster/Container and Spark on Yarn worker/driver/manager with Meridian's requirements. Fengqian is offline checking with Andrej and Soila on the key issues and default configurations for more efficient communication.

We will provide you solid answer next week.

Thanks
Wei Cao (Buddy)

From: Wang, AlbertZQ
Sent: Friday, January 22, 2016 3:07 PM
To: Gao, Fengqian <fengqian.gao@intel.com>; Kavulya, Soila P <soila.p.kavulya@intel.com>; Zhang, Lijuan <lijuan.zhang@intel.com>; Maczuga, Andrzej <andrzej.maczuga@intel.com>; Yi, Lan <lan.yi@intel.com>

Cc: Cao, Buddy <buddy.cao@intel.com>; Tumialis, Adam <Adam.Tumialis@intel.com>; Lou, Ming <ming.lou@intel.com>; Ding, Danny <danny.ding@intel.com>; Lu, Yuan Y <yuan.y.lu@intel.com>; Chen, Cloud <cloud.chen@intel.com>
Subject: RE: New issue in Meridian Environment

Thanks Soila for helping clear the issue in time.

Hi Fengqian, if 90% of yarn.nodemanager.resource.memory-mb is good, doesn't mean we can set 90%* 16G = 14.4G? is the maximum value for node manager is 16G? Meridian will actually loading 1M data instead of 10K soon.

Cheers,
Albert

From: Gao, Fengqian
Sent: Friday, January 22, 2016 2:34 PM
To: Kavulya, Soila P; Zhang, Lijuan; Maczuga, Andrzej; Yi, Lan
Cc: Cao, Buddy; Tumialis, Adam; Wang, AlbertZQ; Lou, Ming; Ding, Danny; Lu, Yuan Y; Chen, Cloud
Subject: RE: New issue in Meridian Environment

Thanks you very much, Soila.

It is very helpful to learn about the solution and we might be able to resolve this kind of issue by ourseleves next time. And thanks again for the sharing.

--fengqian

From: Kavulya, Soila P
Sent: Friday, January 22, 2016 12:11 PM
To: Gao, Fengqian; Zhang, Lijuan; Maczuga, Andrzej; Yi, Lan
Cc: Cao, Buddy; Tumialis, Adam; Wang, AlbertZQ; Lou, Ming; Ding, Danny; Lu, Yuan Y; Chen, Cloud
Subject: RE: New issue in Meridian Environment

Random forest is working. I changed the spark.driver.memory to 6gb.

When jobs fail immediately with "cannot submit to cluster", the most common reason is that the application is requesting more resources than are available on Yarn. A rule of thumb would be to set the maximum container size to about 90% of yarn.nodemanager.resource.memory-mb to cater for Spark memory overhead. It is not advisable to set container sizes that use up all the memory on the machine because you need to leave space for system processes and other hadoop processes to run on the node.

Below are links to a number of good blog posts about how to allocate resources on Yarn:

- http://hortonworks.com/blog/how-to-plan-and-configure-yarn-in-hdp-2-0/
- http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.0.6.0/bk_installing_manually_book/content/rpm-chap1-11.html
- http://www.cloudera.com/documentation/enterprise/latest/topics/cdh_ig_yarn_tuning.html

I also made some changes to the firefox and DNS configurations so that you can view the Yarn application logs to figure out the reasons for failed jobs.

- Modified Firefox about:config on 218.241.151.248 to resolve DNS using the socks proxy by setting "network.proxy.socks_remote_dns" to True
- Modified /etc/hosts on cdh-launcher as follows to allow hostname resolution:

- Disabled anonymous usage data collection on CDH manager since I was experiencing some delays loading pages (possibly due to DNS issues) http://www.cloudera.com/documentation/enterprise/latest/topics/cm_ag_data_collection.html

Thanks,

Soila

From: Gao, Fengqian
Sent: Thursday, January 21, 2016 12:32 AM
To: Kavulya, Soila P; Zhang, Lijuan; Maczuga, Andrzej; Yi, Lan
Cc: Cao, Buddy; Tumialis, Adam; Wang, AlbertZQ; Lou, Ming; Ding, Danny; Lu, Yuan Y; Chen, Cloud
Subject: RE: New issue in Meridian Environment

Hi, Soila,

We tried to change the configuration as you told us. But looks it is not working.
And it is even report the error at the beginning of frame creation. Another weird thing is that we could find any applications log in yarn. Looks the tasks are not submit to yarn.

Maybe we redeploy the app in a wrong way. You would be better to check it. We are using the atk_push.sh script and make some changes due to the error. The app we trying to re-deploy is atk-262f437b-f57b-419b-a2a2.

Thanks
--fengqian

From: Kavulya, Soila P
Sent: Thursday, January 21, 2016 12:25 PM
To: Zhang, Lijuan; Gao, Fengqian; Maczuga, Andrzej; Yi, Lan
Cc: Cao, Buddy; Tumialis, Adam; Wang, AlbertZQ; Lou, Ming; Ding, Danny; Lu, Yuan Y; Chen, Cloud
Subject: RE: New issue in Meridian Environment

It looks like the failures are due to running out of memory in the spark driver since the training phase succeeds if the number of trees is small. Spark's random forest algorithm needs sufficient memory on the driver to store large trees. I

have given Fengqian some tips on configuration parameters to modify, e.g., increasing driver memory to 8gb. Hopefully 8gb will be sufficient to hold 200 trees since the maximum amount of memory available on the instances is 16gb.

Thanks,

Soila

From: Zhang, Lijuan
Sent: Wednesday, January 20, 2016 8:00 PM
To: Gao, Fengqian; Kavulya, Soila P; Maczuga, Andrzej; Yi, Lan
Cc: Cao, Buddy; Tumialis, Adam; Wang, AlbertZQ; Lou, Ming; Ding, Danny; Lu, Yuan Y; Chen, Cloud
Subject: RE: New issue in Meridian Environment

+Yi, Lan

Thanks
Lijuan

From: Gao, Fengqian
Sent: Wednesday, January 20, 2016 9:57 AM
To: Kavulya, Soila P <soila.p.kavulya@intel.com>; Maczuga, Andrzej <andrzej.maczuga@intel.com>
Cc: Cao, Buddy <buddy.cao@intel.com>; Tumialis, Adam <Adam.Tumialis@intel.com>; Wang, AlbertZQ <albertzq.wang@intel.com>; Lou, Ming <ming.lou@intel.com>; Ding, Danny <danny.ding@intel.com>; Lu, Yuan Y <yuan.y.lu@intel.com>; Zhang, Lijuan <lijuan.zhang@intel.com>; Chen, Cloud <cloud.chen@intel.com>
Subject: RE: New issue in Meridian Environment

Hi, Soila,

Please see my comments and thanks for your help.

--fengqian

From: Kavulya, Soila P
Sent: Wednesday, January 20, 2016 7:52 AM
To: Gao, Fengqian; Maczuga, Andrzej
Cc: Cao, Buddy; Tumialis, Adam; Wang, AlbertZQ; Lou, Ming; Ding, Danny; Lu, Yuan Y; Zhang, Lijuan; Chen, Cloud
Subject: RE: New issue in Meridian Environment

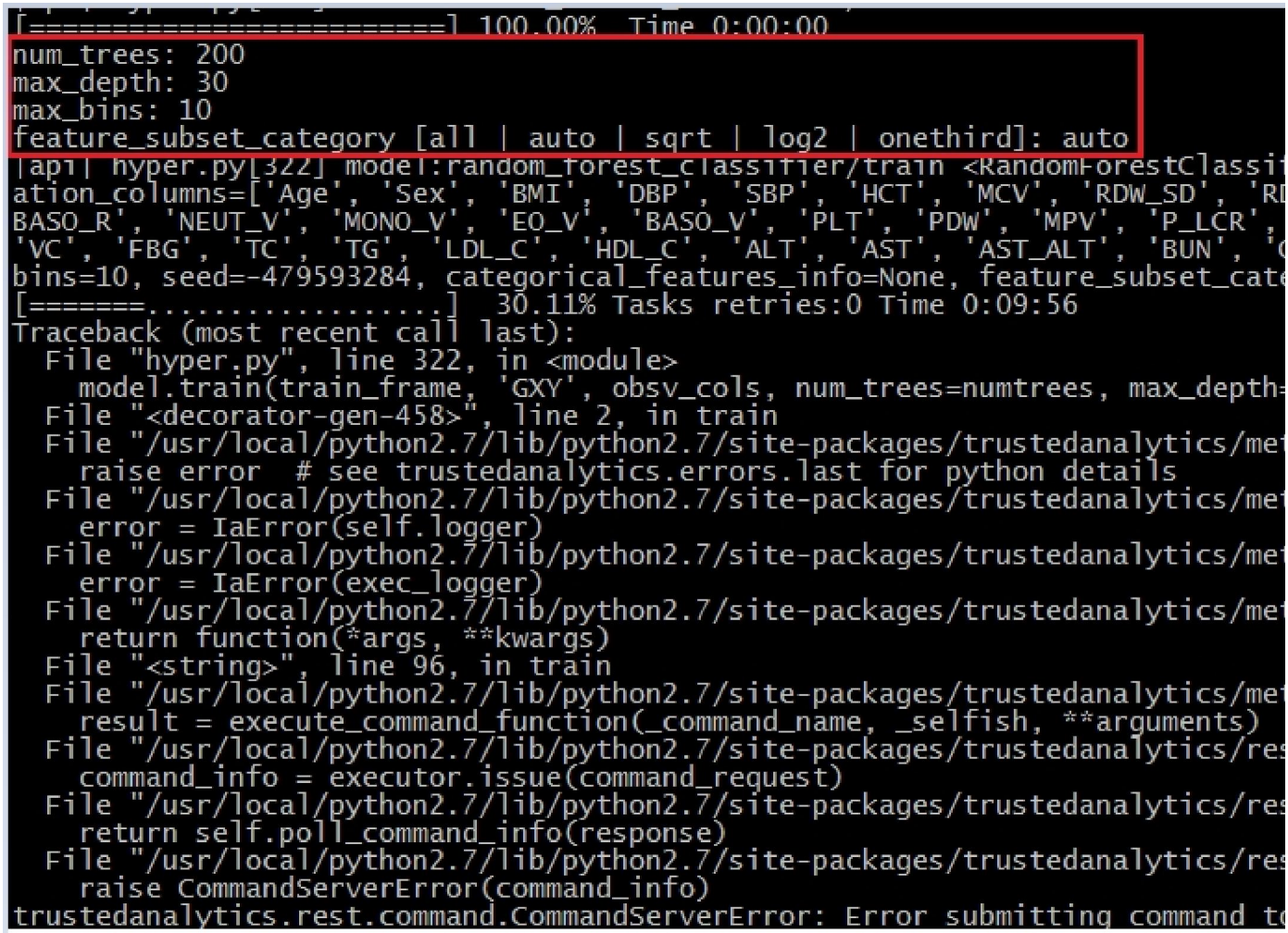
Hi Fengqian,

The most common reasons for the error is if the yarn application is requesting for more resources than available, or permission issues. Could you send me instructions on:

- how to reproduce the error

[Fengqian:] After you connect to 218.241.151.248 via vnc, run the /home/yliu/hyper.py program.
There are some parameters that need to input while the program is running. It shows in the red box.

```
[root@localhost yliu]# python hyper.py
/usr/local/python2.7/lib/python2.7/site-packages/trustedanalytics/meta/insta
--------------------------------------------------------------------------------
WARNING - Client version '0.4.2-2016010810495' does not match server version
--------------------------------------------------------------------------------
  handle_client_server_version_mismatch(client_version, server_version)
Connected.  This client instance connected to server http://atk-262f437b-f57
1-19 16:46:44.794234.
|api| hyper.py[228] globals/get_frame_names get_frame_names
|api| hyper.py[231] frame:/__init__ <Frame:7f8e040b7990>.__init__(source={'f
a9-715a5f48ac25/7816b1c9-0ec1-430d-98b8-9fb86924a057/000000_1', 'delimiter':
e 'numpy.int32'>), ('Sex', <type 'numpy.int32'>), ('BMI', <type 'numpy.int32
mpy.int32'>), ('MCV', <type 'numpy.int32'>), ('RDW_SD', <type 'numpy.int32'>
umpy.int32'>), ('MCHC', <type 'numpy.int32'>), ('RBC', <type 'numpy.int32'>)
 'numpy.int32'>), ('MONO_R', <type 'numpy.int32'>), ('EO_R', <type 'numpy.in
', <type 'numpy.int32'>), ('EO_V', <type 'numpy.int32'>), ('BASO_V', <type '
, <type 'numpy.int32'>), ('P_LCR', <type 'numpy.int32'>), ('PCT', <type 'num
ype 'numpy.int32'>), ('KET', <type 'numpy.int32'>), ('BLD', <type 'numpy.int
numpy.int32'>), ('SG', <type 'numpy.int32'>), ('LEU', <type 'numpy.int32'>),
int32'>), ('TC', <type 'numpy.int32'>), ('TG', <type 'numpy.int32'>), ('LDL_
32'>), ('AST', <type 'numpy.int32'>), ('AST_ALT', <type 'numpy.int32'>), ('B
'>), ('Na', <type 'numpy.int32'>)]}, name='RawFrame', _info=None)
[==========] 100.00% Tasks retries:0 Time 0:00:25
input your seed number: 332987
input your training percent: 0.9
|api| hyper.py[243] frame/assign_sample <Frame:7f8e040b7990>.assign_sample(s
t_column='sample_bin', random_seed=332987)
[==========] 100.00% Tasks retries:0 Time 0:00:18
|api| hyper.py[245] frame/copy <Frame:7f8e040b7990>.copy(columns=None, where
[==========] 100.00% Tasks retries:0 Time 0:00:18
|api| hyper.py[246] frame/copy <Frame:7f8e040b7990>.copy(columns=None, where
[==========] 100.00% Tasks retries:0 Time 0:00:16
|api| hyper.py[262] globals/get_model_names get_model_names
please select the model you prefer........
1: svm model
2: linear regression model
3: logistic regression model
4: naive bayes model
5: principal component model
6: random forest classifier model
0: stop and exit
type your choice:
6
|api| hyper.py[311] model:random_forest_classifier/new <RandomForestClassifi
[==========] 100.00%  Time 0:00:00
num_trees: 200
```



- how to access the machine where you are running and deploying the cloud-foundry app

[Fengqian:] for the openstack dashboard access, please open http://10.1.1.35 with user 'admin' password 'admin'
You can ssh to the controller node from 218.241.151.248 using ssh root@10.1.1.35 password: openstack
To access the bastion VM, you need to ssh to controller node first, then ssh –i /root/.ssh/id_rsa ubuntu@10.1.1.75

- how to access the Cloudera manager web UI from 218.241.151.248

[Fengqian:] you can login the CDH UI via http://10.1.1.35:8080/cmd/home with user 'admin' password 'admin'.

Thanks,

Soila

From: Gao, Fengqian
Sent: Tuesday, January 19, 2016 2:30 AM
To: Maczuga, Andrzej; Kavulya, Soila P
Cc: Cao, Buddy; Tumialis, Adam; Wang, AlbertZQ; Lou, Ming; Ding, Danny; Lu, Yuan Y; Zhang, Lijuan; Chen, Cloud
Subject: New issue in Meridian Environment
Importance: High

Hi, Andy and Soila,

Now the Meridian is accounting another 'Yarn-cluster' error. Please see below picture for more details.
We did some debug by ourselves, trying to change some spark configuration parameters, but still got no luck, so your help is needed.

You can access the Meridian environment by vncviewer from 10.239.82.171 username: Administrator password: zaq1@WSX
Connect VNVViewer to 218.241.151.248:5901 with password meridian1234!@#$
Please aware that the ssh port is blocked now and you only could access the environment via VNC.

```
|api| hyper.py[311] model:random_forest_classifier/new <RandomForestClassifierMo
[========================] 100.00%  Time 0:00:00
num_trees: 200
max_depth: 30
max_bins: 10
feature_subset_category [all | auto | sqrt | log2 | onethird]: auto
|api| hyper.py[322] model:random_forest_classifier/train <RandomForestClassifier
ation_columns=['Age', 'Sex', 'BMI', 'DBP', 'SBP', 'HCT', 'MCV', 'RDW_SD', 'RDW_C
BASO_R', 'NEUT_V', 'MONO_V', 'EO_V', 'BASO_V', 'PLT', 'PDW', 'MPV', 'P_LCR', 'PC
'VC', 'FBG', 'TC', 'TG', 'LDL_C', 'HDL_C', 'ALT', 'AST', 'AST_ALT', 'BUN', 'Cr',
bins=10, seed=-479593284, categorical_features_info=None, feature_subset_categor
[=======.................]  30.11% Tasks retries:0 Time 0:09:56
Traceback (most recent call last):
  File "hyper.py", line 322, in <module>
    model.train(train_frame, 'GXY', obsv_cols, num_trees=numtrees, max_depth=max
  File "<decorator-gen-458>", line 2, in train
  File "/usr/local/python2.7/lib/python2.7/site-packages/trustedanalytics/meta/c
    raise error  # see trustedanalytics.errors.last for python details
  File "/usr/local/python2.7/lib/python2.7/site-packages/trustedanalytics/meta/c
    error = IaError(self.logger)
  File "/usr/local/python2.7/lib/python2.7/site-packages/trustedanalytics/meta/c
    error = IaError(exec_logger)
  File "/usr/local/python2.7/lib/python2.7/site-packages/trustedanalytics/meta/c
    return function(*args, **kwargs)
  File "<string>", line 96, in train
  File "/usr/local/python2.7/lib/python2.7/site-packages/trustedanalytics/meta/m
    result = execute_command_function(_command_name, _selfish, **arguments)
  File "/usr/local/python2.7/lib/python2.7/site-packages/trustedanalytics/rest/c
    command_info = executor.issue(command_request)
  File "/usr/local/python2.7/lib/python2.7/site-packages/trustedanalytics/rest/c
    return self.poll_command_info(response)
  File "/usr/local/python2.7/lib/python2.7/site-packages/trustedanalytics/rest/c
    raise CommandServerError(command_info)
trustedanalytics.rest.command.CommandServerError: Error submitting command to ya
[root@localhost yliu]#
```

Thanks,

Fengqian Gao
Intel Information Technology | Flex Service