关于python在yarn上运行时依赖环境的解决方案(R \Tensoflow也可借鉴)

2018年11月15日 9:57

参考文档	https://www.jianshu.com/p/df0a189ff28b
	https://www.jianshu.com/p/9a144c508059
	https://blog.csdn.net/u012328476/article/details/78894669?utm_source=blogxgwz6
	https://blog.csdn.net/hjh00/article/details/64439268
问题解决	https://stackoverflow.com/questions/30518362/how-do-i-set-the-drivers-python-version-in-spark

一、目的

为了pyspark能够直接提交到没有对应python环境的yarn集群上运行，以及对之后的python算子、R算子、TensorFlow、notebook 都有一定的借鉴意义。

二、安装anaconda环境

为了不对运行环境造成影响， 之前我们将anaconda环境安装在了docker里， 版本为Anaconda3-5.2.0-Linux-x86_64，对应的python版本为3.6

不建议按照文档中的采用anaconda3创建虚拟环境的方式，有许多python包需要依赖系统环境，特别是配置TensorFlow，如果需要无网解决yum依赖包的问题，这个问题就变得很棘手。

在docker里安装好python的所有依赖包之后，即可对anaconda3环境进行打包，命令：

```
zip -r -9 -q python.zip ./anaconda3
```

由于安装的依赖包数量较多， 打包后的大小后2.7G， 原生anaconda3打包后一般为300M左右

将打包后python.zip从docker中拿到本地主机

```
docker cp 24e4ba59dde6:/root/python.zip .
```

上传至HDFS上

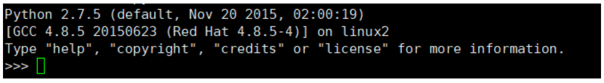
```
hdfs dfs -put python.zip /tmp/spark_market
```

至此Python环境准备完毕

三、准备Spark环境

这里为了验证yarn确实调用了HDFS上的python环境， 我们这里在docker外面上传解压spark-2.3.0-bin-hadoop2.7，

查看docker外的python环境：



```
Python 2.7.5 (default, Nov 20 2015, 02:00:19)
[GCC 4.8.5 20150623 (Red Hat 4.8.5-4)] on linux2
Type "help", "copyright", "credits" or "license" for more information.
>>> 
```

修改spark的配置文件:

```
spark-2.3.0-bin-hadoop2.7/conf/spark-env.sh
HADOOP_CONF_DIR=/etc/hadoop/conf              # hadoop配置
YARN_CONF_DIR=/etc/hadoop/conf.cloudera.yarn/       # yarn配置
export SPARK_HOME=/root/spark-2.3.0-bin-hadoop2.7          # spark配置
```

```
spark-2.3.0-bin-hadoop2.7/conf/spark-defaults.conf
spark.yarn.dist.archives hdfs:///tmp/spark_market/python.zip#ANACONDA
```

# 这里的路径指向python.zip包所在的hdfs路径，#ANACONDA表示解压后的临时目录

```
spark.yarn.appMasterEnv.PYSPARK_DRIVER_PYTHON ANACONDA/anaconda3/bin/python3
```

# python.zip包解压后python所在的位置需要指定

将spark上传至hdfs(若不上传也可, yarn会读取本地的spark环境打包上传至hdfs临时目录再进行分发, 但是会消耗一定的时间)

```
hdfs dfs -put spark-2.3.0-bin-hadoop2.7  /tmp/spark-2.3.0-bin-hadoop2.7
```

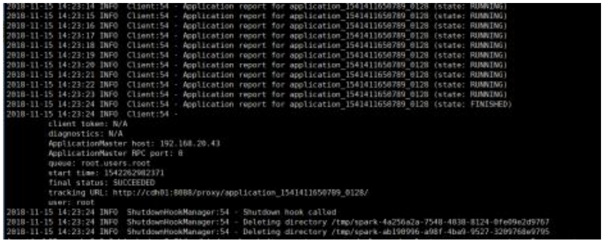
至此spark环境准备完毕

四、测试

进入spark-2.3.0-bin-hadoop2.7目录执行

```
./bin/spark-submit --master yarn --deploy-mode cluster   --num-executors 2 --executor-cores 3 --executor-memory 2G /root/spark-2.3.0-bin-hadoop2.7/examples/src/main/python/mllib/correlations_example.py
```

这里需要调整 --num-executors --executor-cores 以及查看yarn上的任务工作状态， 如果yarn上任务较多或者申请的资源较多，可以会使任务一直处于ACCEPT状态



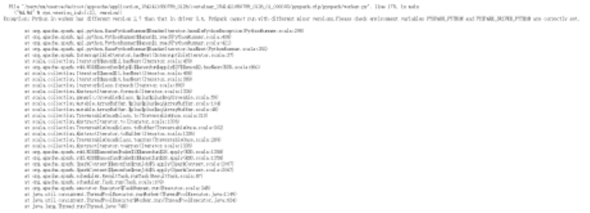
```
2018-11-15 14:23:14 INFO  Client:54 - Application report for application_1541411650789_0128 (state: RUNNING)
2018-11-15 14:23:15 INFO  Client:54 - Application report for application_1541411650789_0128 (state: RUNNING)
2018-11-15 14:23:16 INFO  Client:54 - Application report for application_1541411650789_0128 (state: RUNNING)
2018-11-15 14:23:17 INFO  Client:54 - Application report for application_1541411650789_0128 (state: RUNNING)
2018-11-15 14:23:18 INFO  Client:54 - Application report for application_1541411650789_0128 (state: RUNNING)
2018-11-15 14:23:19 INFO  Client:54 - Application report for application_1541411650789_0128 (state: RUNNING)
2018-11-15 14:23:20 INFO  Client:54 - Application report for application_1541411650789_0128 (state: RUNNING)
2018-11-15 14:23:21 INFO  Client:54 - Application report for application_1541411650789_0128 (state: RUNNING)
2018-11-15 14:23:22 INFO  Client:54 - Application report for application_1541411650789_0128 (state: RUNNING)
2018-11-15 14:23:23 INFO  Client:54 - Application report for application_1541411650789_0128 (state: RUNNING)
2018-11-15 14:23:24 INFO  Client:54 - Application report for application_1541411650789_0128 (state: FINISHED)
2018-11-15 14:23:24 INFO  Client:54 - 
         client token: N/A
         diagnostics: N/A
         ApplicationMaster host: 192.168.20.43
         ApplicationMaster RPC port: 0
         queue: root.users.root
         start time: 1542262982371
         final status: SUCCEEDED
         tracking URL: http://cdh01:8088/proxy/application_1541411650789_0128/
         user: root
2018-11-15 14:23:24 INFO  ShutdownHookManager:54 - Shutdown hook called
2018-11-15 14:23:24 INFO  ShutdownHookManager:54 - Deleting directory /tmp/spark-4a256a2a-7548-4838-8124-0fe09e2d9767
2018-11-15 14:23:24 INFO  ShutdownHookManager:54 - Deleting directory /tmp/spark-ab190996-a98f-4ba9-9527-3209768a9795
```

可以成功执行

为了提高session的启动速度，程序里在配置SparkConf时可以指定"spark.yarn.jars":"hdfs://cdh01:8020/tmp/spark-2.3.0-bin-hadoop2.7/jars/*"， 这样yarn就不会每次启动时都把本地文件重新上传至hdfs一份

五、问题解决

上述测试代码较为简单， 实际使用过程中我们发现如果任务较大，yarn调用到不同work时会报以下错误：



```
  File "/yarn/nm/usercache/root/appcache/application_[illegible]_0126/container_[illegible]_0126_01_000002/pyspark.zip/pyspark/worker.py", line 176, in main
    ("%d.%d" % sys.version_info[:2], version))
Exception: Python in worker has different version 2.7 than that in driver 3.6, PySpark cannot run with different minor versions.Please check environment variables PYSPARK_PYTHON and PYSPARK_DRIVER_PYTHON are correctly set.

        at org.apache.spark.api.python.BasePythonRunner$ReaderIterator.handlePythonException(PythonRunner.scala:298)
        at org.apache.spark.api.python.PythonRunner$$anon$1.read(PythonRunner.scala:438)
        at org.apache.spark.api.python.PythonRunner$$anon$1.read(PythonRunner.scala:421)
        at org.apache.spark.api.python.BasePythonRunner$ReaderIterator.hasNext(PythonRunner.scala:252)
        at org.apache.spark.InterruptibleIterator.hasNext(InterruptibleIterator.scala:37)
        at scala.collection.Iterator$$anon$1[illegible].hasNext(Iterator.scala:[illegible])
        at org.apache.spark.rdd.[illegible].hasNext(RDD.scala:[illegible])
        at scala.collection.Iterator$$anon$11.hasNext(Iterator.scala:408)
        at scala.collection.Iterator$$anon$1[illegible].hasNext(Iterator.scala:[illegible])
        at scala.collection.Iterator$class.foreach(Iterator.scala:893)
        at scala.collection.AbstractIterator.foreach(Iterator.scala:1336)
        at scala.collection.generic.Growable$class.$plus$plus$eq(Growable.scala:59)
        at scala.collection.mutable.ArrayBuffer.$plus$plus$eq(ArrayBuffer.scala:104)
        at scala.collection.mutable.ArrayBuffer.$plus$plus$eq(ArrayBuffer.scala:48)
        at scala.collection.TraversableOnce$class.to(TraversableOnce.scala:310)
        at scala.collection.AbstractIterator.to(Iterator.scala:1336)
        at scala.collection.TraversableOnce$class.toBuffer(TraversableOnce.scala:302)
        at scala.collection.AbstractIterator.toBuffer(Iterator.scala:1336)
        at scala.collection.TraversableOnce$class.toArray(TraversableOnce.scala:289)
        at scala.collection.AbstractIterator.toArray(Iterator.scala:1336)
        at org.apache.spark.rdd.RDD$$anonfun$[illegible].apply(RDD.scala:1358)
        at org.apache.spark.rdd.RDD$$anonfun$[illegible].apply(RDD.scala:1358)
        at org.apache.spark.SparkContext$$anonfun$runJob$5.apply(SparkContext.scala:2067)
        at org.apache.spark.SparkContext$$anonfun$runJob$5.apply(SparkContext.scala:2067)
        at org.apache.spark.scheduler.ResultTask.runTask(ResultTask.scala:87)
        at org.apache.spark.scheduler.Task.run(Task.scala:109)
        at org.apache.spark.executor.Executor$TaskRunner.run(Executor.scala:345)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1149)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
        at java.lang.Thread.run(Thread.java:748)
```

说明在调用python环境时， worker并没有使用hdfs上的， 而是使用了系统的默认python

解决方案：

直接在程序里设置PYTHON的系统环境变量即可

```
import os
os.environ["PYSPARK_PYTHON"]="ANACONDA/anaconda3/bin/python3"
os.environ["PYSPARK_DRIVER_PYTHON"]="ANACONDA/anaconda3/bin/python3"
```

附录、

依赖包：

```
# Name                    Version                   Build  Channel
_ipyw_jlab_nb_ext_conf    0.1.0            py36he11e457_0
alabaster                 0.7.10           py36h306e16b_0
anaconda                  5.2.0                    py36_3
anaconda-client           1.6.14                   py36_0
anaconda-navigator        1.8.7                    py36_0
anaconda-project          0.8.2            py36h44fb852_0
asn1crypto                0.24.0                   py36_0
astroid                   1.6.3                    py36_0
astropy                   3.0.2            py36h3010b51_1
attrs                     18.1.0                   py36_0
babel                     2.5.3                    py36_0
backcall                  0.1.0                    py36_0
backports                 1.0              py36hfa02d7e_1
backports.shutil_get_terminal_size 1.0.0           py36hfea85ff_2
beautifulsoup4            4.6.0            py36h49b8c8c_1
bitarray                  0.8.1            py36h14c3975_1
bkcharts                  0.2              py36h735825a_0
blas                      1.0                         mkl
blaze                     0.11.3           py36h4e06776_0
bleach                    2.1.3                    py36_0
blosc                     1.14.3               hdbcaa40_0
bokeh                     0.12.16                  py36_0
boto                      2.48.0           py36h6e4cd66_1
bottleneck                1.2.1            py36haac1ea0_0
bzip2                     1.0.6                h14c3975_5
ca-certificates           2018.03.07                    0
cairo                     1.14.12              h7636065_2
certifi                   2018.4.16                py36_0
cffi                      1.11.5           py36h9745a5d_0
chardet                   3.0.4            py36h0f667ec_1
click                     6.7              py36h5253387_0
cloudpickle               0.5.3                    py36_0
clyent                    1.2.2            py36h7e57e65_1
colorama                  0.3.9            py36h489cec4_0
conda                     4.5.9                    py36_0
conda-build               3.10.5                   py36_0
conda-env                 2.6.0                h36134e3_1
conda-verify              2.0.0            py36h98955d8_0
contextlib2               0.5.5            py36h6c84a62_0
cryptography              2.2.2            py36h14c3975_0
cssselect                 1.0.3                     <pip>
curl                      7.60.0               h84994c4_0
cycler                    0.10.0           py36h93f1223_0
cython                    0.28.2           py36h14c3975_0
cytoolz                   0.9.0.1          py36h14c3975_0
dask                      0.17.5                   py36_0
dask-core                 0.17.5                   py36_0
datashape                 0.5.4            py36h3ad6b5c_0
dbus                      1.13.2               h714fa37_1
decorator                 4.3.0                    py36_0
distributed               1.21.8                   py36_0
docopt                    0.6.2                    py36_0
docutils                  0.14             py36hb0f60f5_0
dukpy                     0.2.0                     <pip>
entrypoints               0.2.3            py36h1aec115_2
et_xmlfile                1.0.1            py36hd6bccc3_0
expat                     2.2.5                he0dffb1_0
fastavro                  0.19.7           py36h14c3975_0
fastcache                 1.0.2            py36h14c3975_2
filelock                  3.0.4                    py36_0
flask                     1.0.2                    py36_1
flask-cors                3.0.4                    py36_0
fontconfig                2.12.6               h49f89f6_0
freetype                  2.8                  hab7d2ae_1
future                    0.16.0                    <pip>
get_terminal_size         1.0.0                haa9412d_0
gevent                    1.3.0            py36h14c3975_0
glib                      2.56.1               h000015b_0
```

```
glob2                 0.6            py36he249c77_0
gmp                   6.1.2              h6c8ec71_1
gmpy2                   2.0.8          py36hc8893dd_2
graphite2               1.3.11             h16798f4_2
greenlet                0.4.13         py36h14c3975_0
gst-plugins-base        1.14.0             hbbd80ab_1
gstreamer               1.14.0             hb453b48_1
h5py                  2.7.1          py36ha1f6525_2
harfbuzz                1.7.6              h5f0a787_1
hdf5                  1.10.2             hba1933b_1
heapdict                1.0.0                  py36_2
html5lib                1.0.1          py36h2f9c1c0_0
icu                   58.2               h9c2bf20_1
idna                  2.6            py36h82fb2a8_1
imageio                 2.3.0                  py36_0
imagesize               1.0.0                  py36_0
intel-openmp            2018.0.0                    8
ipykernel               4.8.2                  py36_0
ipython                 6.4.0                  py36_0
ipython_genutils        0.2.0          py36hb52b0d5_0
ipywidgets              7.2.1                  py36_0
isort                 4.3.4                  py36_0
itsdangerous            0.24           py36h93cc618_1
javascripthon           0.10                    <pip>
jbig                  2.1                hdba287a_0
jdcal                 1.4                    py36_0
jedi                  0.12.0                 py36_1
jinja2                2.10           py36ha16c418_0
jpeg                  9b                 h024ee3a_2
jsonschema              2.6.0          py36h006f8b5_0
jupyter                 1.0.0                  py36_4
jupyter-echarts-pypkg   0.1.2                   <pip>
jupyter_client          5.2.3                  py36_0
jupyter_console         5.2.0          py36he59e554_1
jupyter_core            4.4.0          py36h7c827e3_0
jupyterlab              0.32.1                 py36_0
jupyterlab_launcher     0.10.5                 py36_0
kiwisolver              1.0.1          py36h764f252_0
krb5                  1.16.1             hc83ff2d_6
lazy-object-proxy       1.3.1          py36h10fcdad_0
libcurl               7.60.0             h1ad7b7a_0
libedit               3.1.20170329       h6b74fdf_2
libffi                3.2.1              hd88cf55_4
libgcc-ng               7.2.0              hdf63c60_3
libgfortran-ng          7.2.0              hdf63c60_3
libpng                1.6.34             hb9fc6fc_0
libsodium               1.0.16             h1bed415_0
libssh2               1.8.0              h9cfc8f7_4
libstdcxx-ng            7.2.0              hdf63c60_3
libtiff               4.0.9              he85c1e1_1
libtool               2.4.6              h544aabb_3
libxcb                1.13               h1bed415_1
libxml2               2.9.8              h26e45fe_1
libxslt               1.1.32             h1312cb7_0
llvmlite              0.23.1         py36hdbcaa40_0
lml                   0.0.2                   <pip>
locket                0.2.0          py36h787c0ad_1
lxml                  4.2.1          py36h23eabaa_0
lzo                   2.10               h49e0be7_2
macropy3                1.1.0b2                 <pip>
markupsafe              1.0            py36hd9260cd_1
matplotlib              2.2.2          py36h0e671d2_1
mccabe                  0.6.1          py36h5ad9710_1
mistune                 0.8.3          py36h14c3975_1
mkl                   2018.0.2                    1
mkl-service             1.1.2          py36h17a0993_4
mkl_fft               1.0.1          py36h3010b51_0
mkl_random              1.0.1          py36h629b387_0
more-itertools          4.1.0                  py36_0
mpc                   1.0.3              hec55b23_5
mpfr                  3.1.5              h11a74b3_2
mpmath                  1.0.0          py36hfeacd6b_2
msgpack-python          0.5.6          py36h6bb024c_0
multipledispatch        0.5.0                  py36_0
navigator-updater       0.2.1                  py36_0
nbconvert               5.3.1          py36hb41ffb7_0
nbformat                4.4.0          py36h31c9010_0
ncurses               6.1                hf484d3e_0
networkx                2.1                    py36_0
nltk                  3.3.0                  py36_0
nose                  1.3.7          py36hcdf7029_2
notebook                5.5.0                  py36_0
numba                   0.38.0         py36h637b7d7_0
numexpr                 2.6.5          py36h7bf3b9c_0
numpy                   1.14.3         py36hcd700cb_1
numpy-base              1.14.3         py36h9be14a7_1
numpydoc                0.8.0                  py36_0
odo                   0.5.1          py36h90ed295_0
olefile               0.45.1                 py36_0
openpyxl                2.5.3                  py36_0
openssl                 1.0.2o             h20670df_0
packaging               17.1                   py36_0
```

```
pandas                        0.23.0           py36h637b7d7_0
pandoc                        1.19.2.1         hea2e7c5_1
pandocfilters                 1.4.2            py36ha6701b7_1
pango                         1.41.0           hd475d92_0
parso                         0.2.0            py36_0
partd                         0.3.8            py36h36fd896_0
patchelf                      0.9              hf79760b_2
path.py                       11.0.1           py36_0
pathlib2                      2.3.2            py36_0
patsy                         0.5.0            py36_0
pcre                          8.42             h439df22_0
pep8                          1.7.1            py36_0
pexpect                       4.5.0            py36_0
pickleshare                   0.7.4            py36h63277f8_0
pillow                        5.1.0            py36h3deb7b8_0
pip                           10.0.1           py36_0
pixman                        0.34.0           hceecf20_3
pkginfo                       1.4.2            py36_1
pluggy                        0.6.0            py36hb689045_0
ply                           3.11             py36_0
prompt_toolkit                1.0.15           py36h17d85b1_0
psutil                        5.4.5            py36h14c3975_0
ptyprocess                    0.5.2            py36h69acd42_0
py                            1.5.3            py36_0
py4j                          0.10.7           py36_0
pycodestyle                   2.4.0            py36_0
pycosat                       0.6.3            py36h0a5515d_0
pycparser                     2.18             py36hf9f622e_1
pycrypto                      2.6.1            py36h14c3975_8
pycurl                        7.43.0.1         py36hb7f436b_0
pyDes                         2.0.1            <pip>
pyecharts                     0.5.8            <pip>
pyecharts-javascripthon       0.0.6            <pip>
pyecharts-jupyter-installer   0.0.3            <pip>
pyflakes                      1.6.0            py36h7bd6a15_0
pygments                      2.2.0            py36h0d3125c_0
PyHDFS                        0.2.1            <pip>
PyHive                        0.6.0            <pip>
pyhocon                       0.3.44           <pip>
pykerberos                    1.1.14           py36h84109d8_2
pylint                        1.8.4            py36_0
pyodbc                        4.0.23           py36hf484d3e_0
pyopenssl                     18.0.0           py36_0
pyparsing                     2.2.0            py36hee85983_1
pyqt                          5.9.2            py36h751905a_0
pyquery                       1.4.0            <pip>
pysocks                       1.6.8            py36_0
pytables                      3.4.3            py36h02b9ad4_2
pytest                        3.5.1            py36_0
pytest-arraydiff              0.2              py36_0
pytest-astropy                0.3.0            py36_0
pytest-doctestplus            0.1.3            py36_0
pytest-openfiles              0.3.0            py36_0
pytest-remotedata             0.2.1            py36_0
python                        3.6.5            hc3d631a_2
python-dateutil               2.7.3            py36_0
python-hdfs                   2.1.0            py36_0
pytz                          2018.4           py36_0
pywavelets                    0.5.2            py36he602eb0_0
pyyaml                        3.12             py36hafb9ca4_1
pyzmq                         17.0.0           py36h14c3975_0
qt                            5.9.5            h7e424d6_0
qtawesome                     0.4.4            py36h609ed8c_0
qtconsole                     4.3.1            py36h8f73b5b_0
qtpy                          1.4.1            py36_0
readline                      7.0              ha6073c6_4
requests                      2.18.4           py36he2e5f8d_1
requests-kerberos             0.12.0           py36_0
rope                          0.10.7           py36h147e2ec_0
ruamel_yaml                   0.15.35          py36h14c3975_1
sasl                          0.2.1            <pip>
scikit-image                  0.13.1           py36h14c3975_1
scikit-learn                  0.19.1           py36h7aa7ec6_0
scipy                         1.1.0            py36hfc37229_0
seaborn                       0.8.1            py36hfad7ec4_0
send2trash                    1.5.0            py36_0
setuptools                    39.1.0           py36_0
sh                            1.12.14          py36_0
simplegeneric                 0.8.1            py36_2
simplejson                    3.16.0           <pip>
singledispatch                3.4.0.3          py36h7a266c3_0
sip                           4.19.8           py36hf484d3e_0
six                           1.11.0           py36h372c433_1
snappy                        1.1.7            hbae5bb6_3
snowballstemmer               1.2.1            py36h6febd40_0
sortedcollections             0.6.1            py36_0
sortedcontainers              1.5.10           py36_0
sphinx                        1.7.4            py36_0
sphinxcontrib                 1.0              py36h6d0f590_1
sphinxcontrib-websupport      1.0.1            py36hb5cb234_1
spyder                        3.2.8            py36_0
sqlalchemy                    1.2.7            py36h6b74fdf_0
```

```
sqlite                    3.23.1              he433501_0
statsmodels               0.9.0           py36h3010b51_0
sympy                     1.1.1           py36hc6d1c1c_0
tblib                     1.3.2           py36h34cf8b6_0
terminado                 0.8.1                   py36_1
testpath                  0.3.1           py36h8cadb63_0
thrift                    0.11.0                   <pip>
thrift-sasl               0.3.0                    <pip>
tk                        8.6.7               hc745277_3
toolz                     0.9.0                   py36_0
tornado                   5.0.2                   py36_0
traitlets                 4.3.2           py36h674d592_0
typing                    3.6.4                   py36_0
unicodecsv                0.14.1          py36ha668878_0
unixodbc                  2.3.6               h1bed415_0
urllib3                   1.22            py36hbe7ace6_0
wcwidth                   0.1.7           py36hdf4376a_0
webencodings              0.5.1           py36h800622e_1
werkzeug                  0.14.1                  py36_0
wheel                     0.31.1                  py36_0
widgetsnbextension        3.2.1                   py36_0
wrapt                     1.10.11         py36h28b7045_0
xlrd                      1.1.0           py36h1db9f0c_1
xlsxwriter                1.0.4                   py36_0
xlwt                      1.3.0           py36h7b00a1f_0
xz                        5.2.4               h14c3975_4
yaml                      0.1.7               had09818_2
zeromq                    4.2.5               h439df22_0
zict                      0.1.3           py36h3a3bf81_0
zlib                      1.2.11              ha838bed_2
```