Jupyter Lab Spark Setup w/ Miniconda3

Friday, July 2, 2021 11:52 AM

1. Delete your old conda3 installation in /s/anaconda/users/CDSID directory
   a. cd /s/anaconda/users/lguerra5
   b. >> ls
   c. >> rm -rf *

2. Backup and delete stuff in your home directory (~ or /u/CDSID)
   a. >>cp .bashrc backup_bashrc (this creates a backup of your .bashrc)
   b. >> ls -la (to see whats in the directory)
   c. Run this command to delete the files named below:

   >> rm rf .backup_miniconda2 .bash_history .beeline .cache .condarc .config .continuum .graphlab .ipynb_checkpoints .ipython . jupyter .local .python_history .scala_history .viminfo
   d. >> la -la (to verify that all the specified files are gone)
   e. >> rm -rf .bashrc (get rid of old bashrc)
   f. >>m -rf .bash_profile

3. Check how much space you have in your hom directory, make sure the output is less than 800M
   a. >>du -sh

4. Close this putty window
5. Open a new putty window
6. Install miniconda3 again:
   a. sh /s/anaconda/miniconda3/scripts/Miniconda3Py36.sh', this will take about 10 min.
   b. Every time you open a new putty window. You need to manually type 'miniconda3' then 'source activate py368nb' (you can replace py368nb with the python environment of your choice).

7. Install pyspark
   a. >>conda install pyspark
8. Copy these lines into your .bashrc:

```
#
# Added by Miniconda3Py36 script
alias miniconda3=/s/anaconda/miniconda3/bin/miniconda3

# HADOOP & SPARK
export DRIVER_MEMORY=4g
export EXECUTOR_MEMORY=8g
export EXECUTOR_CORES=2
export INIT_EXECUTORS=10
export MIN_NUM_EXECUTORS=$INIT_EXECUTORS
export EXECUTOR_OVERHEAD_MEMORY=4g
export DRIVER_OVERHEAD_MEMORY=4g
export MAX_NUM_EXECUTORS=2048
#export N_SHUFFLE_PARTITION=$((MAX_NUM_EXECUTORS*EXECUTOR_CORES))
export N_SHUFFLE_PARTITION=2048
export N_PARALLELISM=$N_SHUFFLE_PARTITION
export MAX_FAILURES=8
export KRYOSERIALIZER_BUFFER_MB=256

# My Aliases
alias lsh='ls -lh'
alias hls='hadoop fs -ls -h'
alias hdu='hadoop fs -du -h'
alias hmv='hadoop fs -mv'
alias hmvToLocal='hadoop fs -moveToLocal'
alias hmvFromLocal='hadoop fs -moveFromLocal'
alias hcp='hadoop fs -cp'
alias hcpToLocal='hadoop fs -copyToLocal'
alias hcpFromLocal='hadoop fs -copyFromLocal'
alias hmkdir='hadoop fs -mkdir'
```

9. Activate your .bashrc using this command: >>source .bashrc
10. Create a file called 'start_jupyter.sh' in your /u/cdsid folder with the following lines:

```
#!/bin/bash
export SPARK_HOME=/usr/hdp/current/spark2-client
export SPARK_MAJOR_VERSION=2
export HDP_VERSION='current'
export PYSPARK_PYTHON=`which python`
export PYSPARK_DRIVER_PYTHON=`which jupyter`
export PYSPARK_DRIVER_PYTHON_OPTS="${1:-lab} --NotebookApp.open_browser=False --NotebookApp.ip='*'"
export username=`whoami`
export app_name=${username}'_ps_'`date +%g%m%d-%H%M%S`
echo App Name: $app_name

  # --conf spark.default.parallelism=$N_PARALLELISM \
```

```
  # --jars /usr/hdp/current/hive_warehouse_connector/*.jar,/u/$username/myjars/*.jar \

pyspark --master yarn \
  --deploy-mode client \
  --verbose \
  --name $app_name \
  --conf spark.driver.maxResultSize=0 \
  --conf spark.dynamicAllocation.enabled=true \
  --conf spark.hadoop.metastore.catalog.default=hive \
  --conf spark.sql.execution.arrow.enabled=true \
  --conf spark.sql.execution.arrow.pyspark.enabled=true\
  --conf spark.sql.session.timeZone=UTC \
  --executor-memory $EXECUTOR_MEMORY \
  --executor-cores $EXECUTOR_CORES \
  --driver-memory $DRIVER_MEMORY \
  --jars /usr/hdp/current/hive_warehouse_connector/*.jar \
  --py-files /usr/hdp/current/hive_warehouse_connector/*.zip
```

11. Start jupyter lab like this: >>./start_jupyter.sh
12. Then look for the port number in the output (XXXX) and open a jupyter notebook with chrome (http://hpchdp2e.hpc.ford.com:XXXX)
13. Start a new notebook and use these lines to test spark connection:

```
from pyspark_llap.sql.session import HiveWarehouseSession
from pyspark.sql import SparkSession

spark = SparkSession.builder.enableHiveSupport().getOrCreate()
hive = HiveWarehouseSession.session(spark).build()

df = hive.executeQuery("select * from sakula5_test.hello_acid")
df.show()
```

14. FYI - link to track spark jobs:
    http://hpchdp2i4.hpc.ford.com:8088/cluster