Generate PySpark running scripts in Pycharm

1. In Scratches and Consoles/Scratches create scratch. py as below. You might modify based on your own new parameter values:

```
code_bucket="lazard-test-client-master"
commons_branch="development"
dpl_branch="bugfix/persisting-agrent-xref"
prefix="stg_2120_"
bucket_loc="s3://lazard-emr-test-data/MDM/SPRINT5/SV_AGR1/"
match_crit = 'SV'
opt = ""#--executor-memory 20g --driver-memory 20g"
ss = f"/usr/bin/spark-submit --num-executors 8 {opt} --jars s3://lazard-
test-client-master/code/jars/postgresql-42.2.5.jar --master yarn --conf
spark.pyspark.python=/usr/bin/python3.6 --py-files s3://{code_bucket}
/code/commons/{commons_branch}/edm-commons.zip,s3://{code_bucket}/code
/dpl/{dpl_branch}/edm-dpl.zip s3://{code_bucket}/code/dpl/{dpl_branch}
/mdm_driver.py -s {bucket_loc} -m false -p {prefix} -e -x {match_crit}"
print("#### MDM ####")
print(ss)
code_bucket="lazard-test-client-master"
commons_branch="master"
dpl_branch="feature/sv-updated-at-2.0"
#conf_loc="s3://lazard-emr-test-data/MDM/1118_sfsvsf_test
/TEST_salesvision_firm_PROD_profile_09_25_2019.gz_1571063968662"
# conf_loc="s3://lazard-emr-test-data/MDM/1118_sfsvsf_test
/TEST_salesvision_office_PROD_profile_09_25_2019.gz_1571063973192"
# conf_loc="s3://lazard-emr-test-data/MDM/1118_sfsvsf_test
/TEST_salsvesvision_person_PROD_profile_09_25_2019.gz_1571063979621"
conf_loc="s3://lazard-emr-test-data/MDM/SPRINT5/SV_AGR2
/CLTMSTRT_ADX_LQE_Asset_File_2020-01-24_PROD_PROD.CSV.gz_1580402743076
s3://lazard-emr-test-data/MDM/SPRINT5/SV_AGR2
/CLTMSTRT_ADX_LQE_Holding_Files_2020-01-24_PROD_PROD.CSV.
gz_1580402762026 s3://lazard-emr-test-data/MDM/SPRINT5/SV_AGR2
/CLTMSTRT_ADX_LQE_Trade_Files_2019-10-18_0000000000107.CSV.
gz_1571674280263"
# conf_loc="s3://lazard-emr-test-data/MDM/SPRINT5/AMG1/amg_data_10-16-
2019.csv_1571241791546 s3://lazard-emr-test-data/MDM/SPRINT5/AMG1
/ldw_batch_data_range-01-01-2019-10-15-2019.csv_1571242077474"
ss= f"/usr/bin/spark-submit --num-executors 2 --jars s3://lazard-test-
client-master/code/jars/postgresql-42.2.5.jar --master yarn --conf
spark.pyspark.python=/usr/bin/python3.6 --py-files s3://{code_bucket}
/code/commons/{commons_branch}/edm-commons.zip,s3://{code_bucket}/code
/dpl/{dpl_branch}/edm-dpl.zip s3://{code_bucket}/code/dpl/{dpl_branch}
/etl_driver.py -c {conf_loc} -m false -e"
print("#### ETL ####")
print(ss)
code_bucket="lazard-test-client-master"
commons_branch= "development"
dpl_branch="feature/sv-export"
message= "sv_fop_export"
```

```
ss = f"/usr/bin/spark-submit --jars s3://lazard-test-client-master/code
/jars/postgresql-42.2.5.jar --master yarn --conf spark.pyspark.python=
/usr/bin/python3.6 --py-files s3://{code_bucket}/code/commons/
{commons_branch}/edm-commons.zip,s3://{code_bucket}/code/dpl/
{dpl_branch}/edm-dpl.zip s3://{code_bucket}/code/dpl/{dpl_branch}
/dbsync_driver.py -s {message}"
print("#### DB ####")
print(ss)
code_bucket="lazard-test-client-master"
commons_branch= "development"
dpl_branch= "development"
message="sf_entity_api"
ss = f"/usr/bin/spark-submit --jars s3://lazard-test-client-master/code
/jars/postgresql-42.2.5.jar --master yarn --conf spark.pyspark.python=
/usr/bin/python3.6 --py-files s3://{code_bucket}/code/commons/
{commons_branch}/edm-commons.zip,s3://{code_bucket}/code/dpl/
{dpl_branch}/edm-dpl.zip s3://{code_bucket}/code/dpl/{dpl_branch}
/dbsync_driver.py -s {message}"
print("#### DB ####")
print(ss)
code_bucket="lazard-test-client-master"
commons_branch="development"
dpl_branch="release/sprint3_1025"
ss = f"/usr/bin/pyspark --num-executors 12 --jars s3://lazard-test-
client-master/code/jars/postgresql-42.2.5.jar --conf spark.pyspark.
python=/usr/bin/python3.6 --py-files s3://{code_bucket}/code/commons/
{commons_branch}/edm-commons.zip,s3://{code_bucket}/code/dpl/
{dpl_branch}/edm-dpl.zip "
print("#### Pyspark ####")
print(ss)
code_bucket="lazard-test-client-master"
commons_branch="master"
dpl_branch="development"
ss = f"/usr/bin/spark-submit --num-executors 12 --jars s3://lazard-test-
client-master/code/jars/postgresql-42.2.5.jar --conf spark.pyspark.
python=/usr/bin/python3.6 --py-files s3://{code_bucket}/code/commons/
{commons_branch}/edm-commons.zip,s3://{code_bucket}/code/dpl/
{dpl_branch}/edm-dpl.zip run_all.py"
print("#### Spark Test ####")
print(ss)
```

2. Run it and get output like the below, which are PySpark commands you can run in EMR cluster

```
C:\Users\xiaop\AppData\Local\Continuum\anaconda3\python.exe C:/Users
/xiaop/.PyCharm2019.3/config/scratches/scratch.py
#### MDM ####
/usr/bin/spark-submit --num-executors 8  --jars s3://lazard-test-client-
master/code/jars/postgresql-42.2.5.jar --master yarn --conf spark.
```

```
pyspark.python=/usr/bin/python3.6 --py-files s3://lazard-test-client-
master/code/commons/development/edm-commons.zip,s3://lazard-test-client-
master/code/dpl/bugfix/persisting-agrent-xref/edm-dpl.zip s3://lazard-
test-client-master/code/dpl/bugfix/persisting-agrent-xref/mdm_driver.py
-s s3://lazard-emr-test-data/MDM/SPRINT5/SV_AGR1/ -m false -p stg_2120_
-e -x SV
#### ETL ####
/usr/bin/spark-submit --num-executors 2 --jars s3://lazard-test-client-
master/code/jars/postgresql-42.2.5.jar --master yarn --conf spark.
pyspark.python=/usr/bin/python3.6 --py-files s3://lazard-test-client-
master/code/commons/master/edm-commons.zip,s3://lazard-test-client-
master/code/dpl/feature/sv-updated-at-2.0/edm-dpl.zip s3://lazard-test-
client-master/code/dpl/feature/sv-updated-at-2.0/etl_driver.py -c
s3://lazard-emr-test-data/MDM/SPRINT5/SV_AGR2
/CLTMSTRT_ADX_LQE_Asset_File_2020-01-24_PROD_PROD.CSV.gz_1580402743076
s3://lazard-emr-test-data/MDM/SPRINT5/SV_AGR2
/CLTMSTRT_ADX_LQE_Holding_Files_2020-01-24_PROD_PROD.CSV.
gz_1580402762026 s3://lazard-emr-test-data/MDM/SPRINT5/SV_AGR2
/CLTMSTRT_ADX_LQE_Trade_Files_2019-10-18_0000000000107.CSV.
gz_1571674280263 -m false -e
#### DB ####
/usr/bin/spark-submit --jars s3://lazard-test-client-master/code/jars
/postgresql-42.2.5.jar --master yarn --conf spark.pyspark.python=/usr
/bin/python3.6 --py-files s3://lazard-test-client-master/code/commons
/development/edm-commons.zip,s3://lazard-test-client-master/code/dpl
/feature/sv-export/edm-dpl.zip s3://lazard-test-client-master/code/dpl
/feature/sv-export/dbsync_driver.py -s sv_fop_export
#### DB ####
/usr/bin/spark-submit --jars s3://lazard-test-client-master/code/jars
/postgresql-42.2.5.jar --master yarn --conf spark.pyspark.python=/usr
/bin/python3.6 --py-files s3://lazard-test-client-master/code/commons
/development/edm-commons.zip,s3://lazard-test-client-master/code/dpl
/development/edm-dpl.zip s3://lazard-test-client-master/code/dpl
/development/dbsync_driver.py -s sf_entity_api
#### Pyspark ####
/usr/bin/pyspark --num-executors 12 --jars s3://lazard-test-client-
master/code/jars/postgresql-42.2.5.jar --conf spark.pyspark.python=/usr
/bin/python3.6 --py-files s3://lazard-test-client-master/code/commons
/development/edm-commons.zip,s3://lazard-test-client-master/code/dpl
/release/sprint3_1025/edm-dpl.zip
#### Spark Test ####
/usr/bin/spark-submit --num-executors 12 --jars s3://lazard-test-client-
master/code/jars/postgresql-42.2.5.jar --conf spark.pyspark.python=/usr
/bin/python3.6 --py-files s3://lazard-test-client-master/code/commons
/master/edm-commons.zip,s3://lazard-test-client-master/code/dpl
/development/edm-dpl.zip run_all.py

Process finished with exit code 0
```