Spark on Yarn 之Python环境定制

1 动机

不同用户的应用程序使用的python版本及需要的python依赖可能存在不同，若每次缺少依赖都请op去update所有节点，对于大集群的维护来说相对比较困难，而且走变更流程等还影响spark 用户的工作效率。

为解决上述问题，我们探索一种用户可自定义Python环境的方法，用户可依据自己的需要，定制自己的python 并做简单指定即可。

2 方法

2.1 构建Python

```
cd /root/
#wget https://www.python.org/ftp/python/3.8.2/Python-3.8.2.tgz
yum -y install gcc openssl-devel bzip2-devel expat-devel gdbm-devel readline-devel sqlite-devel
tar -zxvf Python-3.8.2.tgz
cd Python-3.8.2/
./configure --prefix=/home/bf_hpt/Python-3.8.2
make
make altinstall
```

2.2 安装依赖

使用Python的包管理工具pip安装依赖,若无pip请自行安装。

```
/home/bf_hpt/Python-3.8.2/bin/pip3.8 install urllib3
```

2.3 打包

```
//进入python根目录， 注意打包路径决定下面的spark配制方式。11
cd /home/bf_hpt/Python-3.8.2
tar -zcf python-3.8.2.tgz *
```

3. Spark 配制

因客户机存在公用的可能，且每个应用程序的需求不同，为了降低不同用户之间的影响，我们推荐在提交命令中配制的作法。

3.1 上传python-3.8.2至Spark Executor工作目录

```
sudo -u hdfs hadoop fs -put python-3.8.2.tgz  /system/
```

4 应用提交示例

在cdh5种测试成功

```
spark2-submit \
--master yarn \
--deploy-mode cluster \
--queue xy_yarn_pool.development \
--num-executors 1 \
--executor-memory 1G \
--conf "spark.yarn.dist.archives=hdfs:///system/python-3.8.2.tgz#pyenv" \
--conf "spark.yarn.appMasterEnv.PYSPARK_PYTHON=pyenv/bin/python3.8" \
--conf "spark.yarn.appMasterEnv.PYSPARK_DRIVER_PYTHON=pyenv/bin/python3.8" \
hdfs:///scripts/pi.py
```

在livy中提交：

```
curl -X POST --data '{
"name": "gtm-ins2-ld-lc-fcsting",
"file": "/user/g2m-df/GoToMarket/application/app.py",
"pyFiles": ["/user/g2m-df/GoToMarket/application/foo.zip"],
"jars": ["/user/g2m-df/GoToMarket/environment/jars/mysql-connector-java-5.7.jar"],
"files": ["/user/g2m-df/GoToMarket/environment/app-setting-files/default_model_settings.json","/user/g2m-df/GoToMarket/environment/app-setting-files/mysql_info.cfg"],
"driverCores": 4,
"driverMemory": "10g",
"executorMemory": "8g",
```

```
"executorCores": 4,
"numExecutors": 10,
"queue": "default",
"conf": {
"spark.master": "yarn",
"spark.submit.deployMode":"cluster",
"spark.dynamicAllocation.enabled":"true",
"spark.dynamicAllocation.cachedExecutorIdleTimeout":"60",
"spark.yarn.dist.archives":"/user/g2m-
df/GoToMarket/environment/python/loreal_env.zip#pyenv",
"spark.yarn.appMasterEnv.PYSPARK_PYTHON":"pyenv/loreal_env/bin/python",
"spark.yarn.appMasterEnv.PYSPARK_DRIVER_PYTHON":"pyenv/loreal_env/bin/python"

},
"args": ["{\"taskId\":3456538467372}"]
}' -H "Content-Type: application/json" 10.162.64.103:8998/batches
```

5 缺点分析

上述方案虽可实现用户自定义python环境，但执行过程中每个Executor从HDFS下载一次
python环境，增加RPC等开销，在开启动态资源伸缩功能时，下载次数会更多……