设置环境变量

setx PATH "%PATH%;%HADOOP\_HOME%\bin;%SPARK\_HOME%\bin"

启动spark

172.21.9.17： pyspark --master yarn-client --conf spark.executor.memory=4g --conf spark.executor.instances=200 --conf spark.default.parallelism=500 --conf spark.yarn.am.cores=8 --conf spark.yarn.am.memory=16g --conf spark.pyspark.python=python2.7 --conf spark.serializer=org.apache.spark.serializer.KryoSerializer --conf spark.driver.memory=16g --conf spark.yarn.queue=root.bdp\_jmart\_cis\_union.bdp\_jmart\_cis\_offline\_spark --conf spark.sql.shuffle.partitions=500

172.19.142.148：  pyspark --total-executor-cores 100  --conf spark.executor.memory=2g

pyspark  --master spark://172.19.142.130:7077 \  
         --deploy-mode client \  
         --total-executor-cores 40  \  
         --conf spark.sql.shuffle.partitions=600 \  
         --conf spark.default.parallelism=600 \  
         --driver-memory 2g \  
         --executor-memory 10G

需要用到的表

=========  SKU信息  ======  
gdm\_m03\_item\_sku\_da gdm.gdm\_m03\_self\_item\_sku\_da

======  订单  =======  
gdm\_m04\_ord\_det\_sum

======  流量数据  =======  
gdm\_m14\_online\_log\_item\_d

======  POP商品匹配数据  =======

dev.dev\_pop\_vender\_sku\_match

======  POP sku销售数据 =======

app.app\_op\_pop\_top\_N\_sku\_gmv\_di

======  自营sku销售额数据 ======= 包含没有销售数据的sku,粒度每天

dev.self\_sku\_pv