设置环境变量

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

jps 登录密码：9fe7oOdxJDwB3HfR

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

/home/mart-cis/ligang/submit-cis.sh ./submit-cis.sh xxx.py

#!/bin/sh

spark-submit \

--master yarn-client \

--conf spark.executor.instances=120 \

--conf spark.executor.memory=16g \

--conf spark.executor.cores=4 \

--conf spark.driver.cores=8 \

--conf spark.driver.memory=16g \

--conf spark.yarn.am.cores=8 \

--conf spark.yarn.am.memory=16g \

--conf spark.yarn.queue=root.bdp\_jmart\_cis\_union.bdp\_jmart\_cis\_offline\_spark \

--conf spark.default.parallelism=500 \

--conf spark.dynamicAllocation.enabled=false \

--conf spark.network.timeout=600s \

--conf spark.rpc.askTimeout=1200s \

--conf spark.rpc.lookupTimeout=600s \

--conf spark.rpc.numRetries=10 \

--conf spark.shuffle.io.maxRetries=10 \

--conf spark.serializer=org.apache.spark.serializer.KryoSerializer \

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\_m03\_self\_item\_sku\_da

======  订单明细表  =======  
gdm\_m04\_ord\_det\_sum

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

======  定价开放POP商品匹配表  ======

dev.dev\_pop\_vender\_sku\_match

======  品牌开放自营商品匹配表  =======

dev.dev\_self\_vender\_sku\_match

===== POP sku层级销售额数据 ====仅包含有销售数据的sku gmv，粒度天。包含销量，价格

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

====自营sku层级销售额数据 ===包含没有销售数据的sku gmv, 粒度天。含brand\_code, pv，uv和销量

dev.self\_sku\_pv

======  pop商家按三级分类汇总gmv =======

dev.op\_pop\_cid3\_gmv\_da

======  pop商家按品牌分类汇总gmv =======

dev.op\_pop\_brand\_gmv\_da