建立hdfs目录： `hadoop fs -mkdir -p /workspace/3.rs_project/project2/meiduo_mall`

建立以下sh文件mysqlToHDFS.sh，可将mysql中数据库meiduo_mall里指定表全部导入hdfs中：

```
#!/bin/bash
array=(tb_goods tb_goods_category tb_goods_specification tb_sku tb_sku_specificat
ion tb_specification_option)

for table_name in ${array[@]};
do
    import sqoop \
        --connect jbdc:mysql://hadoop-master/meiduo_mall \
        --username root \
        --password password \  # mysql数据库的密码
        --table $table_name \  # 要读取数据库database中的表名
        --m 5 \ #并行化是启动多个map task实现的，-m(或--num-mappers)参数指定map t
ask数，默认是四个
        --target-dir /project2-meiduo-rs/meiduo_mall/$table_name #指的是HDFS中导
入表的存放目录(注意：是目录)
done
```

执行:

- chmod +x mysqlToHDFS.sh
- source mysqlToHDFS.sh

使用spark读取数据

注意默认情况下，导入数据，每列数据之间是用逗号分隔，因此可以如同csv文件一样进行读取

In [ ]:
```
import os
# 配置pyspark和spark driver运行时 使用的python解释器
JAVA_HOME = '/root/bigdata/jdk'
PYSPARK_PYTHON = '/miniconda2/envs/py365/bin/python'
# 当存在多个版本时，不指定很可能会导致出错
os.environ['PYSPARK_PYTHON'] = PYSPARK_PYTHON
os.environ['PYSPARK_DRIVER_PYTHON'] = PYSPARK_PYTHON
os.environ['JAVA_HOME'] = JAVA_HOME
# 配置spark信息
from pyspark import SparkConf
from pyspark.sql import SparkSession

SPARK_APP_NAME = "TransferMySQLToHDFS"
SPARK_URL = "yarn"

conf = SparkConf()    # 创建spark config对象
config = (
    ("spark.app.name", SPARK_APP_NAME),    # 设置启动的spark的app名称，没有提供，将随机
    ("spark.executor.memory", "2g"),    # 设置该app启动时占用的内存用量，默认1g
    ("spark.master", SPARK_URL),    # spark master的地址
    ("spark.executor.cores", "1"),    # 设置spark executor使用的CPU核心数
    ("spark.executor.instances", "1") # yarn时才会设置该项
)

conf.setAll(config)

# 利用config对象，创建spark session
spark = SparkSession.builder.config(conf=conf).enableHiveSupport().getOrCreate()
```

In [ ]:
```
# ======跳过该cell，延申学习：使用saprk7077端口运行的配置=====
'''
conf = SparkConf()    # 创建spark config对象
config = (
    ("spark.app.name", SPARK_APP_NAME),    # 设置启动的spark的app名称，没有提供，将随机
    ("spark.executor.memory", "2g"),    # 设置该app启动时占用的内存用量，默认1g，指一
    ("spark.master", SPARK_URL),    # spark master的地址
    ("spark.executor.cores", "2"),    # 设置spark executor使用的CPU核心数，指一台虚拟机
    # 以下三项配置，可以控制执行器数量
#     ("spark.dynamicAllocation.enabled", True),
#     ("spark.dynamicAllocation.initialExecutors", 1),    # 1个执行器
#     ("spark.shuffle.service.enabled", True)
#   ('spark.sql.pivotMaxValues', '99999'),  # 当需要pivot DF，且值很多时，需要修改，默
)
# 查看更详细配置及说明：https://spark.apache.org/docs/latest/configuration.html

conf.setAll(config)

# 利用config对象，创建spark session
spark = SparkSession.builder.config(conf=conf).getOrCreate()
'''
```

In [2]:
```
!hadoop fs -ls /
```

```
Found 12 items
drwxr-xr-x   - root supergroup          0 2020-11-06 10:24 /cate_count.csv
drwxr-xr-x   - root supergroup          0 2020-12-12 19:23 /checkPoint
drwxr-xr-x   - root supergroup          0 2020-12-12 18:39 /data
drwxr-xr-x   - root supergroup          0 2020-12-11 11:11 /hbase
drwxr-xr-x   - root supergroup          0 2020-11-11 21:33 /headlines
-rw-r--r--   1 root supergroup       4358 2020-12-11 16:52 /iris.csv
drwxr-xr-x   - root supergroup          0 2020-11-09 10:18 /meiduo_mall
drwxr-xr-x   - root supergroup          0 2020-12-12 21:12 /models
drwxr-xr-x   - root supergroup          0 2020-10-30 12:35 /output
-rw-r--r--   1 root supergroup         84 2020-10-30 12:35 /test.txt
drwx------   - root supergroup          0 2020-10-30 19:26 /tmp
drwxr-xr-x   - root supergroup          0 2020-11-09 14:17 /user
```

In [4]:
```
ret = spark.read.csv('/meiduo_mall/tb_goods')
ret
```

Out[4]:
```
DataFrame[_c0: string, _c1: string, _c2: string, _c3: string, _c4: string, _c5: string, _c6: string, _c7: string, _c8: string, _c9: string, _c10: string, _c11: string, _c12: string]
```

In [5]:
```
ret.show()
ret.select('_c3','_c10').show(100)
```

```
+-----+-------------------+-------------------+-------------------+---+---+---+---+---+---+----+----+----+
|  _c0|                _c1|                _c2|                _c3|_c4|_c5|_c6|_c7|_c8|_c9|_c10|_c11|_c12|
+-----+-------------------+-------------------+-------------------+---+---+---+---+---+---+----+----+----+
|65233|2018-11-01 13:53:...|2018-11-01 13:53:...|万胶鼎 (整箱装多至72卷)透明胶...|  0|  0|  1|  5| 49|189|null|null|null|
|65234|2018-11-01 13:53:...|2018-11-01 13:53:...|3M 思高1468 钛金属不粘剪刀...|  0|  0|  1|  5| 49|189|null|null|null|
|65235|2018-11-01 13:53:...|2018-11-01 13:53:...|信发（TRNFA）农行定制 办公纸...|  0|  0|  1|  5| 49|189|null|null|null|
|65236|2018-11-01 13:53:...|2018-11-01 13:53:...|斯图 （sitoo)布基胶带 地毯...|  0|  0|  1|  5| 49|189|null|null|null|
|65237|2018-11-01 13:53:...|2018-11-01 13:53:...|齐心(COMIX) 省力按键式重型...|  0|  0|  1|  5| 49|189|null|null|null|
|65238|2018-11-01 13:53:...|2018-11-01 13:53:...|红绿蓝 无线抢答器S100型抢答器...|  0|  0|  1|  5| 49|189|null|null|null|
|65239|2018-11-01 13:53:...|2018-11-01 13:53:...|✅订书机 大号重型省力加厚订书机 ...|  0|  0|  1|  5| 49|189|null|null|null|
```