

## 实验3-3 spark编程

### 0. scala安装

ubuntu: apt-get install scala 为了后面配置scala的环境变量，这里先找出scala的位置 通过which scala 可以查看到java的执行路径（不同于安装路径）

/usr/bin/scala

执行 ls -lrt /usr/bin/scala

执行 ls -lrt /etc/alternatives/scala，其显示的/usr/share/scala-2.11就是我们需要查询的scala安装目录

```
root@h0:/home/Hadoop/spark-2.4.4/conf# ls -lrt /usr/bin/scala
lrwxrwxrwx 1 root root 23 Dec  3 00:13 /usr/bin/scala -> /etc/alternatives/scala
root@h0:/home/Hadoop/spark-2.4.4/conf# ls -lrt /etc/alternatives/scala
lrwxrwxrwx 1 root root 31 Dec  3 00:13 /etc/alternatives/scala -> /usr/share/scala-2.11/bin/scala
```

### 1. spark安装

版本&环境: Ubuntu 18 + java 1.8 +hadoop-3.2.1，使用集群模式，两个节点h0和h1，h0作为master，仅有h1作为worker

spark使用hadoop的resourcemanager分配资源

下载，解压，配置相关文件见<https://www.jianshu.com/p/a4a0e7e4e4b7>

其中spark-env.sh配置见下图

```
#!/usr/bin/env bash
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export SCALA_HOME=/usr/share/scala-2.11
export HADOOP_CONF_DIR=/home/Hadoop/hadoop-3.2.1/etc/hadoop
export HADOOP_HOME=/home/Hadoop/hadoop-3.2.1
#export HADOOP_HDFS_HOME=/usr/local/hadoop
export SPARK_HOME=/home/Hadoop/spark-2.4.4
export SPARK_MASTER_IP=h0
#export SPARK_MASTER_PORT=7077
export SPARK_MASTER_HOST=h0
#export SPARK_WORKER_PORT=8901
#export SPARK_WORKER_INSTANCES=1
export SPARK_MASTER_WEBUI_PORT=8080
```

start-all.sh

```
h0 root@h0:/home/Hadoop/spark-2.4.4# jps
1248 Jps
355 NameNode
1190 Master
824 ResourceManager
574 SecondaryNameNode

h1 root@h1:/home/Hadoop/spark-2.4.4# jps
24384 NodeManager
24274 DataNode
24582 Jps
23958 MainGenericRunner
24535 Worker
```

运行样例程序

```
root@h0:/home/Hadoop/spark-2.4.4# bin/run-example SparkPi 2>&1 | gre
p "Pi is"
Pi is roughly 3.1386956934784673
```

webUI



Spark Master at spark://h0:7077

URL: spark://h0:7077

Alive Workers: 1

Cores in use: 1 Total, 0 Used

Memory in use: 1024.0 MB Total, 0.0 B Used

Applications: 0 Running, 0 Completed

Drivers: 0 Running, 0 Completed

Status: ALIVE

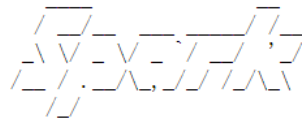
## 2. spark编程实践

### 2.1 pyspark配置及使用

实验环境：阿里云服务器，docker spark集群

- 安装anaconda，并配置jupyter见教程<https://www.jianshu.com/p/670486953d9e>（因为是在docker中配置，还需要在创建容器的时候配置端口映射，在阿里云安全组中开放对应的端口，关闭防火墙）
- 配置jupyter的自动补全<https://www.jianshu.com/p/0ab80f63af8a>
- 运行jupyter，并在本机浏览器访问：xx.xx.xx.x:端口号（注意使用chrome浏览器，默认edge不行）
- 配置pyspark<https://blog.csdn.net/dxyna/article/details/79772343>

Welcome to



version 2.4.4

In [ ]: 1

## 2.2 编程实践

- ```
root@hadoop:/home/hadoop/share-files/spark_local_output# cat out/*
('天津市', [(('1288', 130), ('656', 115), ('1213', 89), ('692', 88), ('662', 82), ('1142', 80), ('389', 75), ('737', 74), ('177', 71), ('664', 67))])
('天津市', [(('1288', 120), ('656', 116), ('662', 83), ('177', 80), ('692', 76), ('1213', 72), ('389', 72), ('1142', 71), ('1401', 67), ('737', 64))])
('山西', [(('656', 123), ('1288', 119), ('1401', 83), ('692', 79), ('177', 78), ('1213', 77), ('664', 73), ('1142', 73), ('420', 72), ('737', 70))])
('上海市', [(('656', 125), ('692', 108), ('1288', 104), ('1142', 83), ('177', 81), ('1213', 76), ('1401', 75), ('662', 75), ('737', 73), ('389', 72))])
('香港', [(('1288', 118), ('656', 107), ('177', 95), ('1142', 82), ('737', 78), ('662', 78), ('692', 68), ('1401', 67), ('1611', 66), ('1553', 65))])
('云南', [(('656', 124), ('1288', 119), ('692', 111), ('1213', 84), ('389', 79), ('662', 77), ('737', 72), ('1401', 70), ('1142', 69), ('420', 65))])
('青海', [(('656', 122), ('1288', 121), ('1142', 94), ('1401', 92), ('737', 91), ('692', 80), ('1213', 79), ('662', 74), ('177', 74), ('389', 72))])
('陕西', [(('1288', 138), ('656', 119), ('692', 89), ('1213', 86), ('177', 83), ('1142', 81), ('662', 80), ('389', 75), ('737', 74), ('1401', 73))])
('安徽', [(('656', 141), ('1288', 120), ('692', 101), ('737', 85), ('1213', 84), ('1401', 80), ('662', 77), ('664', 68), ('420', 67), ('1142', 66))])
('贵州', [(('1288', 126), ('656', 110), ('692', 85), ('1213', 79), ('737', 77), ('1142', 77), ('1553', 76), ('389', 76), ('662', 75), ('1401', 72))])
('四川', [(('656', 128), ('1288', 105), ('692', 101), ('737', 94), ('662', 80), ('420', 77), ('1401', 73), ('1553', 71), ('177', 69), ('389', 67))])
('澳门', [(('656', 129), ('1288', 117), ('1213', 93), ('692', 87), ('662', 86), ('177', 80), ('1142', 78), ('737', 76), ('1401', 70), ('1408', 69))])
('江西', [(('656', 128), ('1288', 126), ('692', 106), ('737', 90), ('177', 90), ('662', 80), ('1213', 78), ('1142', 78), ('1401', 70), ('389', 70))])
('山东', [(('656', 121), ('1288', 99), ('692', 97), ('662', 95), ('389', 90), ('1142', 87), ('177', 86), ('737', 85), ('1401', 83), ('420', 74))])
('湖南', [(('656', 135), ('1213', 96), ('737', 95), ('1288', 92), ('692', 92), ('177', 88), ('1142', 85), ('662', 75), ('389', 67), ('1401', 64))])
('台湾', [(('656', 109), ('692', 98), ('1288', 89), ('1213', 88), ('662', 80), ('389', 79), ('1438', 77), ('1142', 75), ('177', 75), ('1401', 73))])
('广西', [(('656', 127), ('1288', 113), ('692', 90), ('737', 84), ('1401', 82), ('389', 82), ('662', 82), ('1142', 71), ('1213', 66), ('1611', 60))])
('广东', [(('656', 132), ('1288', 118), ('692', 95), ('1401', 87), ('1213', 77), ('420', 77), ('1142', 77), ('737', 75), ('389', 71), ('662', 70))])
('河南', [(('656', 145), ('1288', 126), ('1401', 89), ('692', 86), ('737', 85), ('177', 75), ('1213', 74), ('389', 70), ('662', 68), ('898', 68))])
('河北', [(('1288', 123), ('656', 122), ('692', 95), ('662', 94), ('737', 83), ('1213', 76), ('1142', 76), ('1401', 68), ('1553', 60), ('420', 58))])
('重庆市', [(('1288', 117), ('656', 107), ('1213', 94), ('177', 90), ('692', 88), ('662', 83), ('737', 81), ('1401', 76), ('389', 66), ('1553', 65))])
('甘肃', [(('1288', 120), ('656', 104), ('692', 102), ('177', 99), ('737', 86), ('1213', 81), ('662', 80), ('1553', 72), ('1142', 69), ('389', 69))])
('北京市', [(('656', 131), ('692', 111), ('1288', 101), ('177', 90), ('1213', 85), ('1142', 80), ('737', 80), ('1438', 77), ('662', 76), ('1401', 75))])
('宁夏', [(('656', 130), ('1288', 121), ('692', 113), ('737', 89), ('662', 89), ('177', 85), ('1213', 79), ('1438', 75), ('1401', 74), ('1142', 73))])
('海南', [(('1288', 121), ('656', 107), ('177', 85), ('1213', 82), ('1401', 81), ('389', 78), ('692', 77), ('662', 75), ('1553', 74), ('1438', 72))])
('辽宁', [(('1288', 121), ('656', 118), ('177', 96), ('692', 89), ('1401', 82), ('662', 82), ('1213', 80), ('1438', 75), ('1142', 73), ('737', 66))])
('新疆', [(('1288', 105), ('656', 104), ('737', 90), ('662', 78), ('1213', 77), ('177', 71), ('692', 70), ('1401', 70), ('1438', 66), ('389', 64))])
('内蒙古', [(('1288', 118), ('656', 105), ('662', 92), ('692', 86), ('177', 81), ('737', 76), ('1142', 73), ('1401', 72), ('389', 70), ('420', 67))])
('湖北', [(('656', 121), ('1288', 112), ('1213', 89), ('692', 87), ('737', 79), ('662', 77), ('1401', 75), ('1142', 70), ('177', 68), ('389', 65))])
('浙江', [(('1288', 129), ('656', 112), ('692', 95), ('177', 92), ('662', 89), ('737', 89), ('664', 69), ('1142', 69), ('1213', 66), ('1401', 65))])
('江苏', [(('656', 131), ('1288', 131), ('662', 101), ('692', 100), ('1213', 89), ('737', 86), ('177', 78), ('1438', 68), ('1401', 67), ('1142', 66))])
('吉林', [(('656', 129), ('1288', 119), ('177', 89), ('692', 89), ('389', 82), ('737', 78), ('662', 75), ('1438', 73), ('1213', 69), ('664', 64))])
('西藏', [(('656', 116), ('1288', 100), ('662', 85), ('389', 82), ('177', 82), ('692', 80), ('1142', 76), ('737', 75), ('1213', 74), ('1438', 65))])
('黑龙江', [(('656', 129), ('1288', 118), ('177', 91), ('1401', 90), ('1213', 90), ('692', 84), ('662', 82), ('1142', 74), ('737', 71), ('389', 70))])
```

```
data=data.map(lambda x: list(x.split(',')))
```

```
data=data.filter(lambda x: x[7]=='2')
print(data.first())
data=data.map(lambda x: (x[10],x[2]))
print(data.first())
data=data.groupByKey()# 返回[(key,pyspark.resultiterable),(,)]
data=data.mapValues(list)
#print(data.first())
def cat_10 (x):
    dct={}
    for key in x:
        dct[key]=dct.get(key,0)+1
    lst=sorted(dct.items(),key=lambda y:y[1],reverse=True)
    return lst[:10]
data=data.mapValues(cat_10)
print(data.first())
data.saveAsTextFile('file:///home/Hadoop/share-files/spark_local_output/out2')
#这里结果保存在本地，也可以选择saveAsHadoopData等保存在hdfs上
```

- 统计各省的双十一前十热门销售产品（购买最多前10的产品）-- 和MapReduce作业对比结果

```
root@hdt:/home/Hadoop/share-files/spark_local_output# cat out2/*
('天津市', [(('1859899', 8), ('783997', 7), ('317673', 6), ('191499', 6), ('1180222', 6), ('493761', 5), ('698879', 5), ('864805', 5), ('823766', 5), ('3081', 5))],
('福建省', [(('783997', 10), ('67897', 8), ('713695', 7), ('108215', 7), ('836876', 6), ('823766', 6), ('1059899', 5), ('201485', 5), ('179830', 5), ('141675', 5))],
('山西省', [(('191499', 9), ('559647', 7), ('353560', 7), ('713695', 7), ('221663', 7), ('936203', 6), ('1059899', 6), ('539608', 5), ('349999', 5), ('654894', 5))],
('上海市', [(('191499', 12), ('353560', 10), ('1059899', 6), ('713695', 6), ('514725', 6), ('944554', 5), ('67897', 5), ('1838146', 5), ('213297', 5), ('1844140', 5))],
('云南省', [(('191499', 10), ('1059899', 7), ('48664', 5), ('655904', 5), ('349999', 5), ('1010145', 5), ('514725', 4), ('147751', 4), ('413046', 4), ('181387', 4))],
('香港', [(('936203', 8), ('118347', 6), ('107407', 6), ('1059899', 5), ('191499', 5), ('713695', 5), ('276750', 5), ('926069', 5), ('89953', 5), ('856368', 4))],
('青海省', [(('353560', 8), ('944554', 7), ('800025', 6), ('191499', 6), ('81360', 6), ('221663', 6), ('317073', 5), ('1075577', 5), ('713695', 5), ('316514', 5))],
('陕西省', [(('191499', 9), ('353560', 9), ('514725', 8), ('936203', 7), ('107407', 7), ('221663', 7), ('181387', 6), ('28895', 6), ('1091980', 6), ('1059899', 5))],
('安徽省', [(('353560', 6), ('676215', 6), ('823766', 6), ('186456', 6), ('636863', 5), ('108215', 5), ('1059899', 5), ('801860', 5), ('514725', 5), ('554408', 5))],
('贵州省', [(('783997', 6), ('936203', 6), ('179830', 6), ('353560', 5), ('28895', 5), ('713695', 5), ('343432', 5), ('191499', 5), ('823766', 5), ('1039919', 4))],
('四川', [(('514725', 10), ('191499', 7), ('783997', 6), ('1059899', 6), ('221663', 6), ('209821', 6), ('181387', 5), ('328160', 5), ('803999', 5), ('28186', 5))],
('澳门', [(('353560', 7), ('1059899', 6), ('936203', 6), ('349999', 5), ('783997', 5), ('191499', 5), ('952198', 5), ('3081', 5), ('825218', 4), ('713695', 4))],
('江西', [(('191499', 11), ('349999', 6), ('107407', 6), ('698879', 6), ('676215', 5), ('181387', 5), ('783997', 5), ('713695', 5), ('514725', 5), ('229233', 5))],
('山东', [(('1059899', 7), ('713695', 7), ('823766', 7), ('221663', 6), ('191499', 6), ('783997', 5), ('186456', 5), ('28186', 5), ('981145', 5), ('81901', 5))],
('湖南', [(('191499', 9), ('81901', 7), ('992911', 6), ('1039919', 5), ('107407', 5), ('67897', 5), ('181387', 5), ('655904', 5), ('823766', 5), ('353560', 5))],
('台湾', [(('191499', 8), ('353560', 7), ('349999', 7), ('1824557', 6), ('315345', 6), ('713695', 5), ('441588', 5), ('944554', 5), ('221663', 5), ('823766', 5))],
('广西', [(('221663', 10), ('936203', 8), ('353560', 7), ('889995', 7), ('191499', 6), ('671759', 5), ('773802', 5), ('783997', 5), ('676215', 5), ('49881', 4))],
('广东', [(('926069', 7), ('181387', 7), ('1059899', 6), ('713695', 6), ('1039919', 6), ('353560', 6), ('118347', 6), ('191499', 6), ('514725', 6), ('981145', 5))],
('河南', [(('191499', 12), ('1059899', 9), ('713695', 7), ('353560', 6), ('283850', 5), ('316514', 5), ('48664', 5), ('735931', 5), ('758374', 4), ('1044140', 4))],
('河北', [(('191499', 9), ('713695', 7), ('353560', 7), ('82431', 6), ('213297', 6), ('349999', 6), ('67897', 5), ('1059899', 4), ('698879', 4), ('102025', 4))],
('重庆市', [(('713695', 8), ('191499', 8), ('186456', 6), ('655904', 6), ('49881', 6), ('936203', 6), ('483836', 5), ('157763', 5), ('179830', 5), ('413846', 5))],
('甘肃', [(('353560', 11), ('107675', 7), ('181387', 6), ('813135', 6), ('28895', 5), ('514725', 5), ('28186', 5), ('191499', 5), ('1059899', 4), ('936203', 4))],
('北京市', [(('191499', 8), ('1059899', 8), ('514725', 6), ('944554', 6), ('698879', 5), ('492131', 4), ('965273', 4), ('982357', 4), ('353560', 4), ('89953', 4))],
('宁夏', [(('318890', 8), ('713695', 8), ('783881', 6), ('191499', 6), ('67897', 5), ('28895', 5), ('181387', 5), ('107407', 4), ('487805', 4), ('676215', 4))],
('海南', [(('1059899', 10), ('191499', 8), ('514725', 7), ('353560', 7), ('513855', 6), ('28895', 6), ('770668', 4), ('317073', 4), ('195478', 4), ('256896', 4))],
('辽宁', [(('698879', 8), ('936203', 8), ('783997', 7), ('655904', 7), ('886674', 6), ('514725', 5), ('823766', 5), ('107407', 5), ('713695', 5), ('846996', 5))],
('新疆', [(('1059899', 10), ('353560', 7), ('118347', 7), ('349999', 7), ('107407', 5), ('676215', 5), ('1186283', 5), ('315345', 5), ('928498', 4), ('713695', 4))],
('内蒙古', [(('191499', 8), ('353560', 8), ('770668', 6), ('1039919', 6), ('358797', 5), ('713695', 5), ('1059899', 5), ('226595', 5), ('878372', 4), ('343432', 4))],
('湖北', [(('81360', 7), ('191499', 6), ('107407', 6), ('353560', 5), ('147751', 5), ('655904', 5), ('823766', 5), ('142889', 4), ('573778', 4), ('779070', 4))],
('浙江', [(('1059899', 10), ('191499', 8), ('107407', 8), ('349999', 7), ('213297', 5), ('48664', 5), ('783997', 5), ('514725', 5), ('221663', 5), ('81360', 5))],
('江苏', [(('191499', 10), ('889995', 6), ('181387', 6), ('655904', 6), ('1039919', 6), ('110347', 5), ('272605', 5), ('944554', 5), ('203850', 5), ('843392', 4))],
('吉林', [(('1059899', 11), ('3081', 8), ('191499', 8), ('107407', 8), ('353560', 6), ('655904', 5), ('1039919', 5), ('822352', 5), ('15173', 4), ('784451', 4))],
('西藏', [(('191499', 11), ('353560', 8), ('1059899', 7), ('783997', 6), ('315345', 6), ('936203', 5), ('698879', 5), ('221663', 5), ('713695', 5), ('1180222', 4))],
('黑龙江', [(('191499', 10), ('353560', 7), ('823766', 7), ('783997', 5), ('1039919', 5), ('1073932', 5), ('722301', 5), ('713695', 4), ('281247', 4), ('272605', 4))])
```

代码同上一题，只需把(x[10],x[2]) 替换为(x[10],x[1])即可

与mapreduce对比

```
in trust check: loc
上海市, 191499 12
上海市, 353560 10
上海市, 713695 6
上海市, 1059899 6
上海市, 514725 6
上海市, 213297 5
上海市, 926069 5
```

可以发现结果相同，但是用python 编写的spark程序简单非常非常多

- 查询双11那天浏览次数前十的品牌 -- 和Hive作业对比结果

## 1. 使用RDD编程

```
data=sc.textFile('file:///home/Hadoop/share-
files/million_user_log.csv')
data.cache()
data=data.map(lambda x: list(x.split(',')))
data=data.filter(lambda x: x[7]!='0')
data=data.map(lambda x:x[4])
lst=sorted(data.countByValue().items(), key= lambda x:
x[1],reverse=True)
print(lst[:10])
```

## 结果

```
[('1360', 49151), ('3738', 10130), ('82', 9719), ('1446', 9426), ('6215', 8568), ('1214', 8470), ('5376', 8282), ('2276', 7990), ('1662', 7808), ('8235', 7661)]
```

之前使用Hive的结果：

```
OK
1360      49151
3738      10130
82         9719
1446      9426
6215      8568
1214      8470
5376      8282
2276      7990
1662      7808
8235      7661
```

2. 此外还可以用dataframe进行查询（pyspark不支持Dataset，因为python本身不是一种类型安全的语言）

```
#使用Dataframe+sparkSQL
from pyspark.sql.types import *
#如果使用默认反射推断会将全部数据推断为String，并且淘宝数据没有header，这里用编程指定类型
schema = StructType([
    StructField("user_id", StringType()),
    StructField("item_id",StringType()),
    StructField("cat_id", StringType()),
    StructField("merchant_id", StringType()),
    StructField("brand_id", StringType()),
    StructField("month", StringType()),
    StructField("day", StringType()),
    StructField("action", StringType()),
    StructField("age_range", StringType()),
    StructField("gender", StringType()),
    StructField("province", StringType())])

data1=spark.read.csv('file:///home/Hadoop/share-
files/million_user_log.csv',header=False,schema=schema)
data1.createOrReplaceTempview('taobao')
data1.show(5)
```

| user_id | item_id | cat_id | merchant_id | brand_id | month | day | action | age_range | gender | province |
|---------|---------|--------|-------------|----------|-------|-----|--------|-----------|--------|----------|
| 328862  | 406349  | 1280   | 2700        | 5476     | 11    | 11  | 0      | 0         | 1      | 四川       |
| 328862  | 406349  | 1280   | 2700        | 5476     | 11    | 11  | 0      | 7         | 1      | 重庆市      |
| 328862  | 807126  | 1181   | 1963        | 6109     | 11    | 11  | 0      | 1         | 0      | 上海市      |
| 328862  | 406349  | 1280   | 2700        | 5476     | 11    | 11  | 2      | 6         | 0      | 台湾       |
| 328862  | 406349  | 1280   | 2700        | 5476     | 11    | 11  | 0      | 6         | 2      | 甘肃       |

only showing top 5 rows

```
data2=spark.sql('select brand_id, count ( action ) actions from taobao
where action = 0 group by brand_id order by actions desc limit 10')
data2.show()
```

| brand_id | actions |
|----------|---------|
| 1360     | 49151   |
| 3738     | 10130   |
| 82       | 9719    |
| 1446     | 9426    |
| 6215     | 8568    |
| 1214     | 8470    |
| 5376     | 8282    |
| 2276     | 7990    |
| 1662     | 7808    |
| 8235     | 7661    |

参考教程: <https://www.jianshu.com/p/cb0fec7a4f6d>

3. 此外还可以是加载hiveContext然后运行SQL语句进行查询

```
from pyspark.sql import HiveContext
hive_context=HiveContext(sc)
hive_context.sql('select brand_id, count ( action ) actions from tUser
where action = 0 group by brand_id order by actions desc limit
10').show()
```