Spark集群使用说明

SPARK_VERISON 2.2.0
PYTHON_VERSION 2.7+
SPARK_URL: spark://10.190.2.112:7077
HDFS_URL: hdfs://10.190.2.112

例题：

在HDFS上放着一个消费者的消费历史记录的文件，写一个sparkjob，统计购买情况。

输入文件：

```
bin/hdfs dfs -cat /fu/UserPurchaseHistory.csv
John,iPhone Cover,9.99
John,Headphones,5.49
Jack,iPhone Cover,9.99
Jill,Samsung Galaxy Cover,8.95
Bob,iPad Cover,5.49
```

代码：

```
"""A simple Spark app in Python"""
from pyspark import SparkContext


sc = SparkContext("spark://10.190.2.112:7077", "First Spark App")
data = sc.textFile("hdfs://10.190.2.112/fu/UserPurchaseHistory.csv")\
   .map(lambda line: line.split(","))\
   .map(lambda record: (record[0], record[1], record[2]))
# let's count the number of purchases
numPurchases = data.count()
# let's count how many unique users made purchases
uniqueUsers = data.map(lambda record: record[0]).distinct().count()
# let's sum up our total revenue
totalRevenue = data.map(lambda record: float(record[2])).sum()
# let's find our most popular product
products = data.map(lambda record: (record[1], 1.0)).reduceByKey(lambda a, b: a + b).collect()
mostPopular = sorted(products, key=lambda x: x[1], reverse=True)[0]


# Finally, print everything out
```

```
print("Total purchases: %d" % numPurchases)
print("Unique users: %d" % uniqueUsers)
print("Total revenue: %2.2f" % totalRevenue)
print("Most popular product: %s with %d purchases" % (mostPopular[0], mostPopular[1]))

# stop the SparkContext
sc.stop()
```

提交命令：

本地提交：

```
bin/spark-submit --master spark://localhost:7077 ~/Downloads/pythonapp.py
```

远程提交：

```
bin/spark-submit --master spark://10.190.2.112:7077 ~/Downloads/pythonapp.py
```

运行结果（不含日志）：

```
Total purchases: 5
Unique users: 4
Total revenue: 39.91
Most popular product: iPhone Cover with 2 purchases
```

参考阅读

https://spark.apache.org/docs/2.2.0/submitting-applications.html