分布式系统第七次实验思路说明

17030140014 张笑天

SPARK

代码地址

参考文献： 子雨大数据之Spark入门教程(Python版)

基础设计

1. 学生成绩统计程序
    - 利用filter过滤选修

```
compulsoryScore = lineData.filter(lambda line: line[3] == "必修")
```

    - 将值作为一个(成绩, 1)的元组转换，从而可以同时加和成绩和课程数量。

```
scoreData = compulsoryScore.map(lambda line: (line[1], (int(line[4]), 1)))
scoreSum = scoreData.reduceByKey(lambda x, y: (x[0] + y[0], x[1] + y[1]))
```

    - 利用多个filter统计各个成绩段并且放在列表中，最后转换成RDD

```
rangeNumber = [
avgData.filter(lambda x: 90 <= x[1] <= 100).count(),
avgData.filter(lambda x: 80 <= x[1] < 90).count(),
avgData.filter(lambda x: 70 <= x[1] < 80).count(),
avgData.filter(lambda x: 60 <= x[1] < 70).count(),
avgData.filter(lambda x: 0 <= x[1] < 60).count()
]
sc.parallelize(rangeNumber).saveAsTextFile(sys.argv[3])
```

    - 使用的命令

```
python ./student_score.py file:///home/enigma/Documents/StudentScore.txt
file:///home/enigma/Desktop/SparkHomework/avg
file:///home/enigma/Desktop/SparkHomework/count
```

2. 祖孙关系

- 假设文件每一行对应父与子
- 利用flatMap方法同时加入父子关系和子亲关系

```
relationshipData = lineData.flatMap(lambda line: [(line[0], (line[1], "T")),
(line[1], (line[0], "F"))])=
```

- 利用键值对生成祖孙关系，使用传统写法

```
def valueProcess(v):
    child = []
    parent = []
    pair = []
    for i in v:
        if (i[1] == "T"):
            parent.append(i[0])
        else:
            child.append(i[0])
    for i in child:
        for j in parent:
            pair.append((i, j))

    return pair
groupRelationship =
relationshipData.groupByKey().mapValues(valueProcess).flatMap(lambda x:
x[1])
```

- 使用join算子的方法，更加优雅

```
childData = lineData.map(lambda x: (x[0], x[1]))
parentData = lineData.map(lambda x: (x[1], x[0]))
grandData = parentData.join(childData).map(lambda x: x[1])
```

- 使用的命令

```
python ./child_parent.py file:///home/enigma/Documents/ChildParent.txt
file:///home/enigma/Desktop/SparkHomework/CPOut
# 将产生的两个文本连接
cat ./CPOut/part-00001 >> ./CPOut/part-00000
```

问题与解决

1. 对于RDD的数据结构形式不熟练，与MapReduce混淆；认真阅读教材，并且打印数据解决。

```
data.foreach(print)
```

2. shell定位文件在HDFS中，实际上想使用Linux的文件系统；

   解决：file:/// + 绝对路径

3. 祖孙关系中祖孙的连接，还没有想到好的写法

结果

注意：pdf版本中对于文件链接会转化为绝对路径，可以使用md版本获取相对路径

平均分

人数

祖孙关系

心得

熟悉了Spark Python版本的写法，对于这类分布式数据的处理更为熟练了。