Big Data Processing

Homework 2

要求

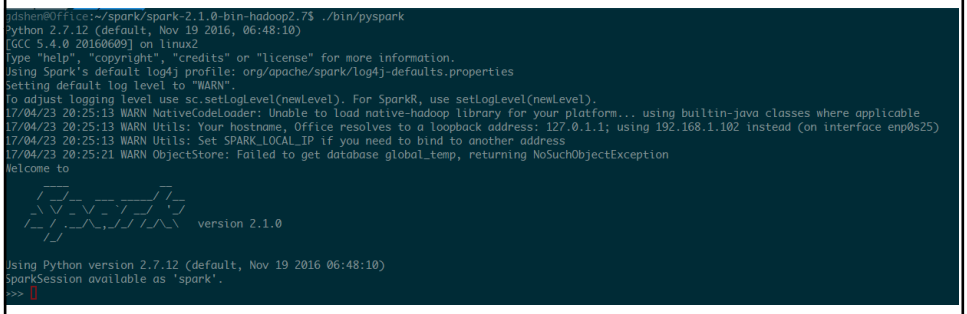
- 使用Spark完成数组求平方和 & word count
- 编写报告
- 两人组队（不组队也可，但不组队这件事自身不会带来任何加分）
- 本次作业的代码和预期的运行结果都以截图形式给出了，大家主要是进行验证性的工作，加深对于MapReduce的理解
- （后续的截图均是在Ubuntu 16.04下运行的结果）

Spark 部署简介

- Spark下载地址 http://mirrors.ustc.edu.cn/apache/spark/spark-2.1.0/spark-2.1.0-bin-hadoop2.7.tgz

- 安装Java环境 “sudo apt-get install openjdk-8-jdk”
- 下载Spark “wget http://mirrors.ustc.edu.cn/apache/spark/spark-2.1.0/spark-2.1.0-bin-hadoop2.7.tgz”
- 解压 “tar zxvf spark-2.1.0-bin-hadoop2.7.tgz”
- 进入spark文件夹下 “cd spark-2.1.0-bin-hadoop2.7”
- 启动spark自带的Python Shell “./bin/pyspark”

成功启动Spark pyshell截图



计算数组平方和

- 这一步存在随机化，所以不希望大家提交的报告中出现完全相同的数组（尽可能避免出现直接复制粘贴别人截图的报告）
- 大家可以将randint(0, 40)改为randint(0, 自己的学号后三位%11+20)

```
>>> import random
>>> array = sc.parallelize([random.randint(0, 40) for i in range(10)])
>>> array.collect()
[38, 22, 0, 16, 31, 0, 1, 11, 36, 3]
>>> array.map(lambda x: x*x).reduce(lambda a, b: a + b)
4572
```

对spark自带的README.md进行单词统计

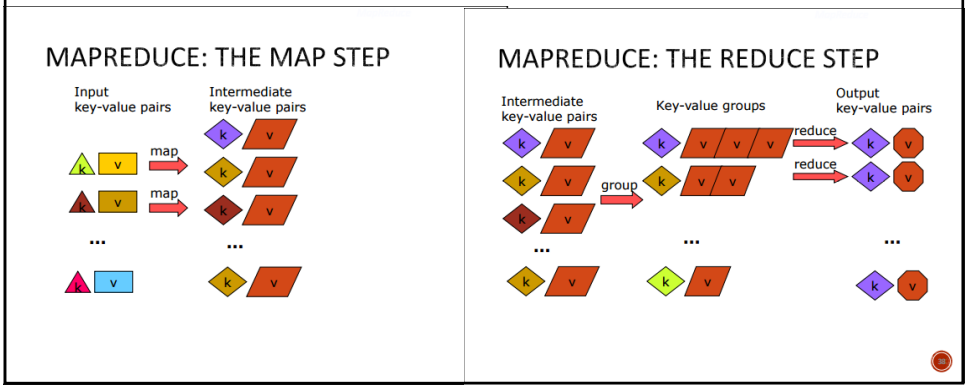
```
>>> text_file = sc.textFile('README.md')
>>> counts = text_file.flatMap(lambda line: line.split(' ')).map(lambda word: (word, 1)).reduceByKey(lambda a, b: a + b)
>>> from pprint import pprint
>>> pprint(counts.collect())
[(u'', 72),
 (u'when', 1),
 (u'R,', 1),
 (u'including', 4),
 (u'computation', 1),
 (u'contributing', 1),
 (u'submit', 1),
 (u'using:', 1),
 (u'guidance', 2),
 (u'Scala,', 1),
 (u'environment', 1),
 (u'only', 1),
 (u'rich', 1),
 (u'Apache', 1),
```

报告要求

- 使用Word，Pages, LaTeX或者markdown等编写都可以，但最后提交时转成PDF文件格式
- 报告的内容
  - 自己运行结果的截图
  - 从这两个实验中任意挑选一个，画出MapReduce执行的示意图。
- 报告的长度控制在两页以内

关于示意图

- 参考课件中的下述两页，并将对应的符号换成具体的值
- 不用画出所有的值，可以挑选一部分示意即可



提交

- 作业提交位置
  - ftp://public.sjtu.edu.cn username: shen_yao password: public
  - 提交到ftp中/upload/CS426/hw2/ 目录下
- 作业提交时间
  - ddl: 5月4号23:59:59
  - 晚交惩罚：每超时24小时，该次作业总分扣除20%成绩，不满24小时按照24小时计算，5月7日23:59:59之后提交的作业一概不接收。
  - 时间根据ftp服务器接收到文件的时间为准。
- 作业命名规则
  - 学号1_姓名1_学号2_姓名2_hw2.pdf (本次作业无需提交代码，有报告即可)

评分标准（满分10分）

- 成功部署Spark 2分
- 完成数组平方和计算 3分
- 完成word count 3分
- 示意图 2分
- （以上内容均需要在报告中进行截图体现）

- (分数的档次只会有0, 0.5, 1.0, 1.5, 2.0以此递推)

完成作业过程遇到任何问题，请发邮件到gdshen@sjtu.edu.cn