曹瑞升-5140219073 秦格华-5140219335

04/May/2017

Project2 实验报告

一. 数组平方和计算

Spark 基于 Ubuntu 的部署因为作业介绍中已经有了详细的介绍，这里我们仅通过图 1即可看出 Spark 已成功部署。

学号后三位 335 (335%11 + 20 = 25), 我们使用 $randint(0, 25)$, 运行截图参见图 1

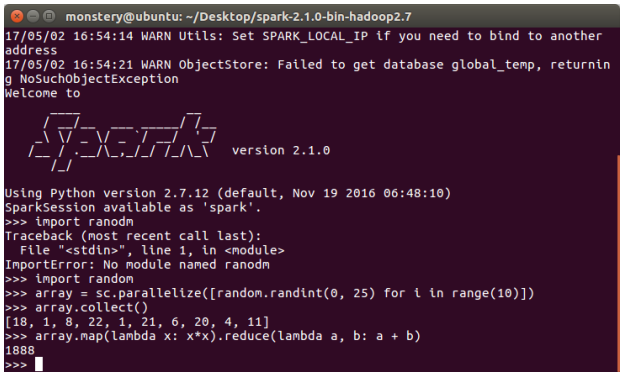

图 1: 数组平方和

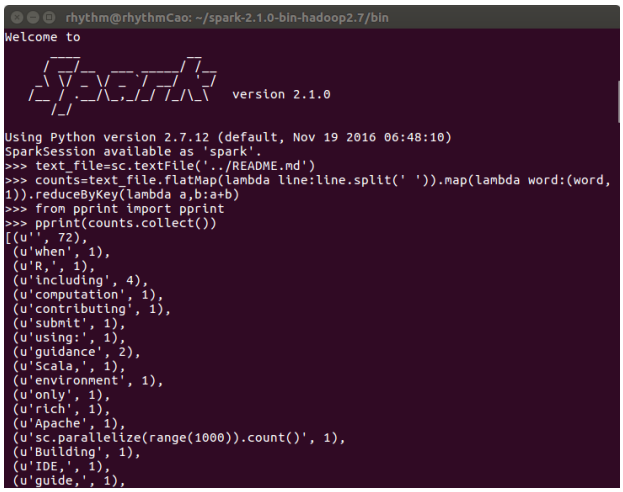

图 2: word count 统计

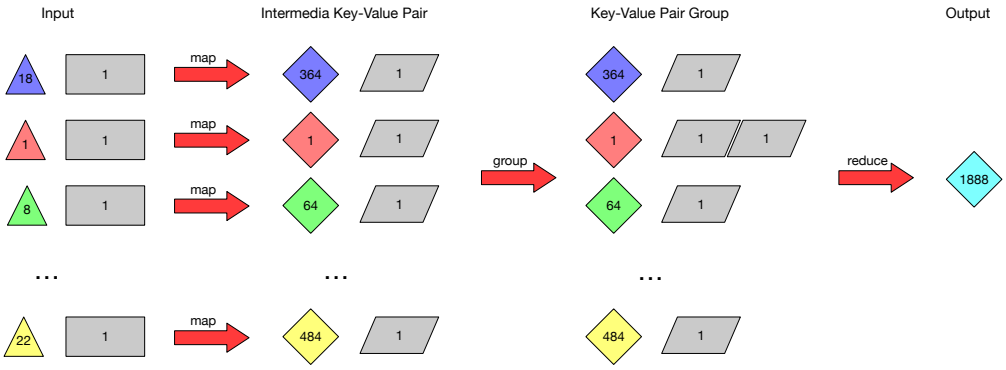

图 3: 数组平方和 MapReduce 示意图

二. word count 统计

运行截图参见图 2。

以 word count 计数为例, 输入文件 *README.md*, 由于该文件缺少代表性，我们以一个三行的文本为例. 任务分为 map 和 reduce 两个阶段:

- Map 阶段
  并行读取文件每一行的输入, 对读取的单词进行 Map 操作, 每个单词都以 <key, value> 的形式生成, 参见图 5.

  flatMap(lambda line:line.split(‘ ‘)).map(lambda word: (word, 1))

- Reduce 阶段
  对 map 的结果进行排序, 合并, 最后得出词频, 参见图 6.

  reduceByKey(lambda a, b: a+b)

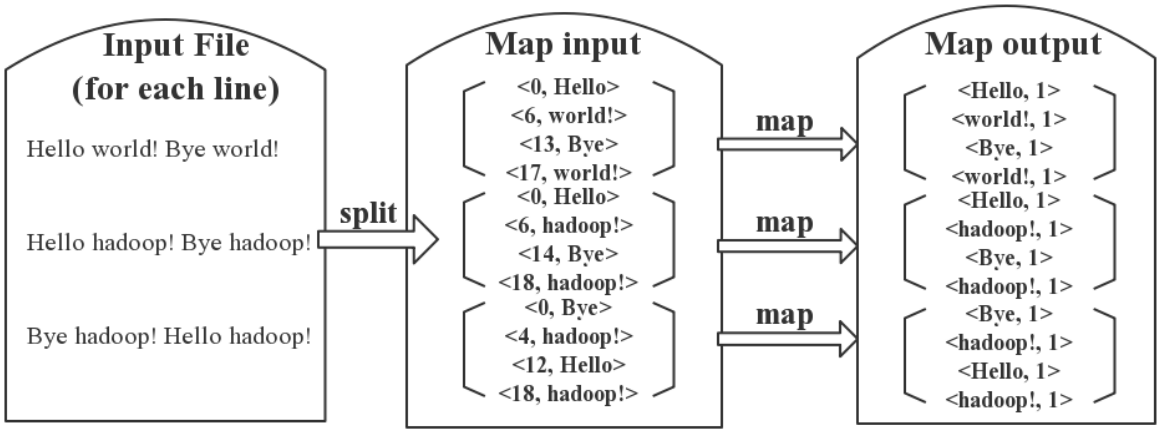

图 5: Map 阶段

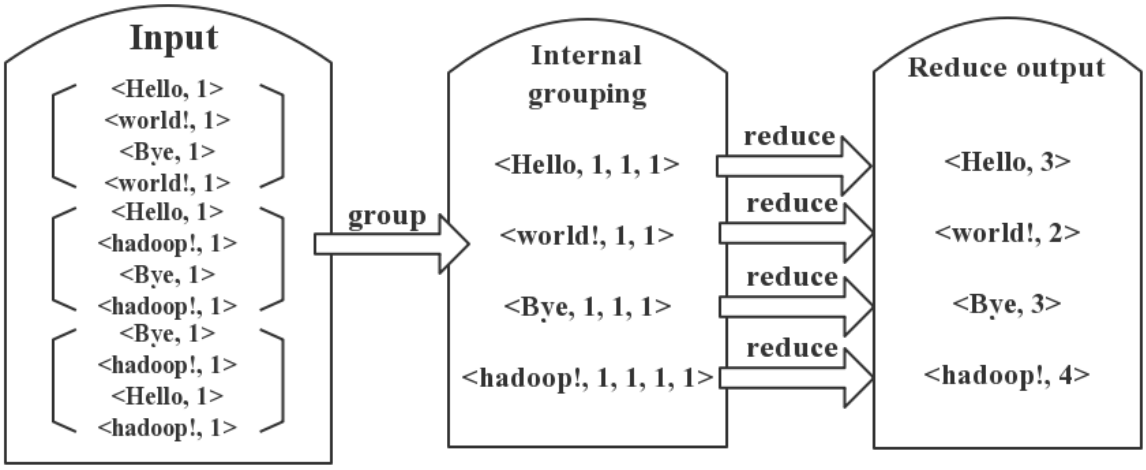

图 6: Reduce 阶段