**实验序号及名称：实验 二 统计某电商网站买家收藏商品数量**

实验时间∶ 2019 年 4 月 15 日

1. **实验目的和要求**
2. 编写mapreduce代码统计每位用户收藏商品的数量。
3. 掌握eclipse如何运行mapreduce程序。
4. 进一步巩固对mapreduce的了解
5. **实验任务**

现有某电商网站用户对商品的收藏数据，记录了用户收藏的商品id以及收藏日期，名为buyer\_favorite1。buyer\_favorite1包含：买家id，商品id，收藏日期这三个字段，数据以“\t”分割，样本数据及格式如下：

1. 买家id   商品id    收藏日期
2. 10181   1000481   2010-04-04 16:54:31
3. 20001   1001597   2010-04-07 15:07:52
4. 20001   1001560   2010-04-07 15:08:27
5. 20042   1001368   2010-04-08 08:20:30
6. 20067   1002061   2010-04-08 16:45:33
7. 20056   1003289   2010-04-12 10:50:55
8. 20056   1003290   2010-04-12 11:57:35
9. 20056   1003292   2010-04-12 12:05:29
10. 20054   1002420   2010-04-14 15:24:12
11. 20055   1001679   2010-04-14 19:46:04
12. 20054   1010675   2010-04-14 15:23:53
13. 20054   1002429   2010-04-14 17:52:45
14. 20076   1002427   2010-04-14 19:35:39
15. 20054   1003326   2010-04-20 12:54:44
16. 20056   1002420   2010-04-15 11:24:49
17. 20064   1002422   2010-04-15 11:35:54
18. 20056   1003066   2010-04-15 11:43:01
19. 20056   1003055   2010-04-15 11:43:06
20. 20056   1010183   2010-04-15 11:45:24
21. 20056   1002422   2010-04-15 11:45:49
22. 20056   1003100   2010-04-15 11:45:54
23. 20056   1003094   2010-04-15 11:45:57
24. 20056   1003064   2010-04-15 11:46:04
25. 20056   1010178   2010-04-15 16:15:20
26. 20076   1003101   2010-04-15 16:37:27
27. 20076   1003103   2010-04-15 16:37:05
28. 20076   1003100   2010-04-15 16:37:18
29. 20076   1003066   2010-04-15 16:37:31
30. 20054   1003103   2010-04-15 16:40:14
31. 20054   1003100   2010-04-15 16:40:16

要求编写MapReduce程序，统计每个买家收藏商品数量

1. **实验准备方案**

包括以下内容：（硬件类实验：实验原理、实验线路、设计方案等）

（软件类实验：所采用的系统、组件、工具、核心方法、框架或流程图、程序清单等）

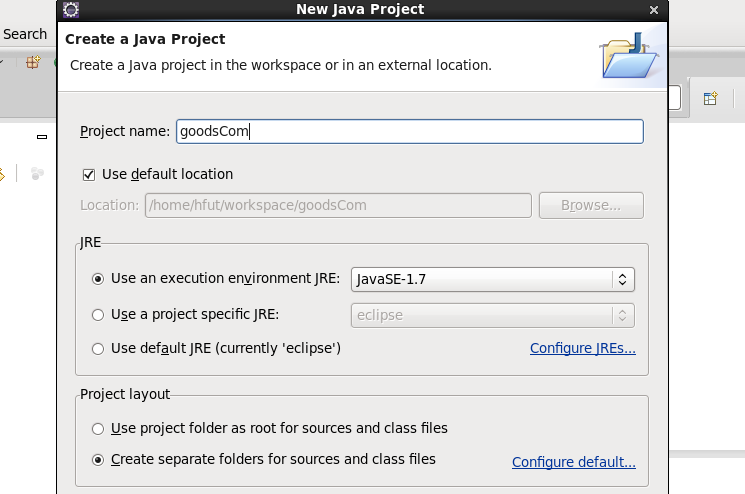
实验采用的系统：centos 6.5、hadoop2.5.2

工具：VMware WorkStation

1. **实验用仪器、设备**

实验仪器：计算机一台

1. **实验内容与步骤（过程及数据记录）：**
2. 新建工程



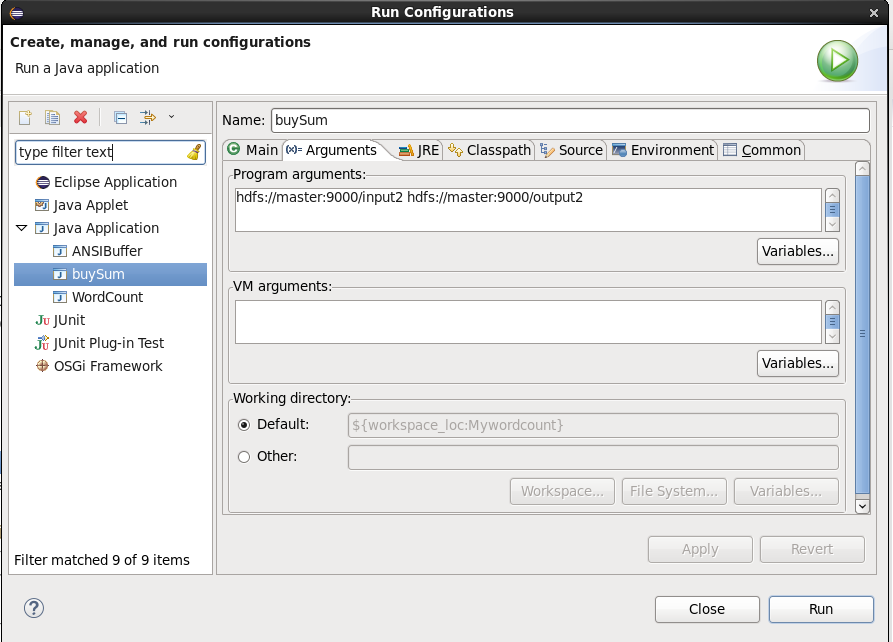
图一

1. 待处理数据结构

表一 原始数据结构

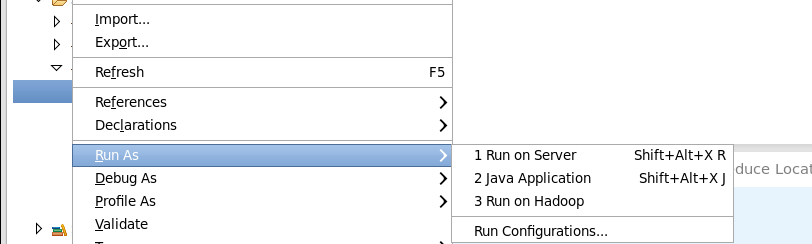
|  |  |  |
| --- | --- | --- |
| 买家ID | 商品ID | 收藏时间 |
| 10181 | 1000481 | 2010-04-04 16:54:31 |
| 20001 | 1001597 | 2010-04-07 15:07:52 |
| 20001 | 1001560 | 2010-04-07 15:08:27 |
| 20042 | 1001368 | 2010-04-08 08:20:30 |
| 20067 | 1002061 | 2010-04-08 16:45:33 |
| 20056 | 1003289 | 2010-04-12 10:50:55 |
| 20056 | 1003290 | 2010-04-12 11:57:35 |
| 20056 | 1003292 | 2010-04-12 12:05:29 |

1. 设置输入输出路径

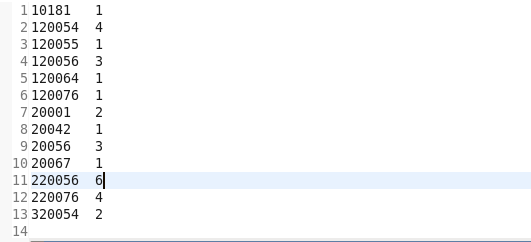


图二 设置输入输出路径

1. Run->run on Hadoop:



1. 实验结果



图三 统计结果

1. 处理数据流程

在map函数中将每一个key、value对的value进行分割，分隔符为’\t’,然后提取提取分割后的列表中的第一项（即用户ID），然后发送一个key、value对。在reduce函数中将相同key值的value相加，即可得到用户收藏的商品总数。