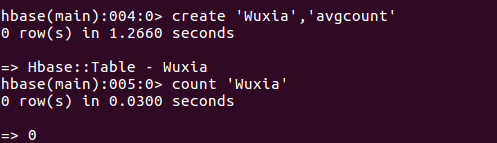
HBase、Hive的安装与使用

# 实验过程及结果

## HBase

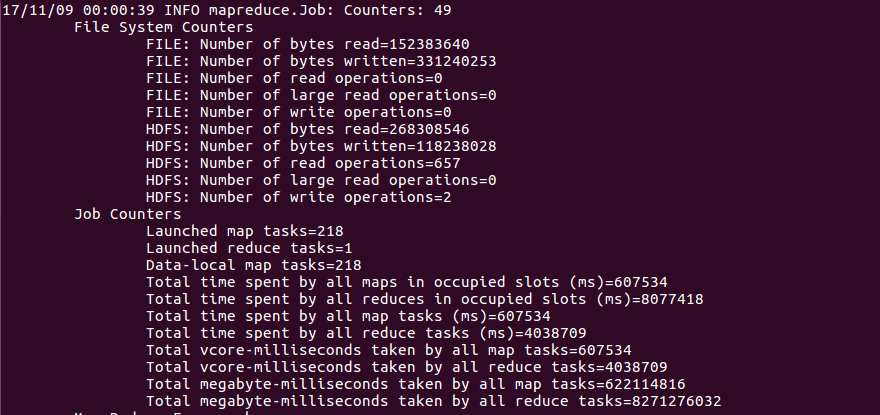
安装HBase，使用HBase创建表：Wuxia

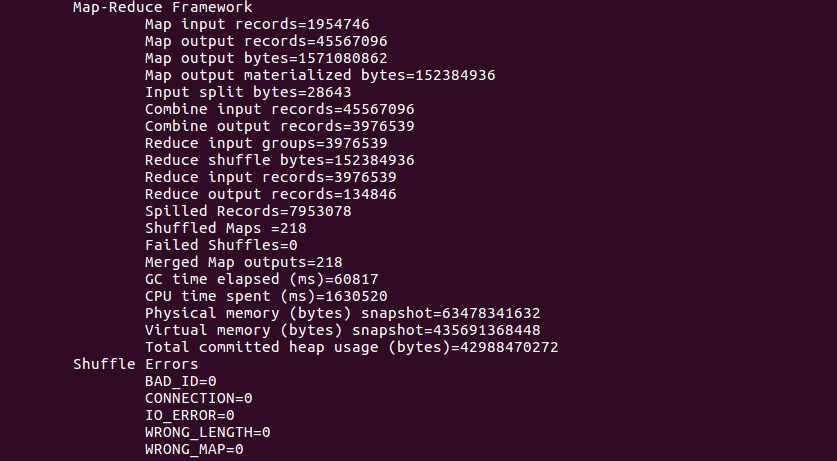


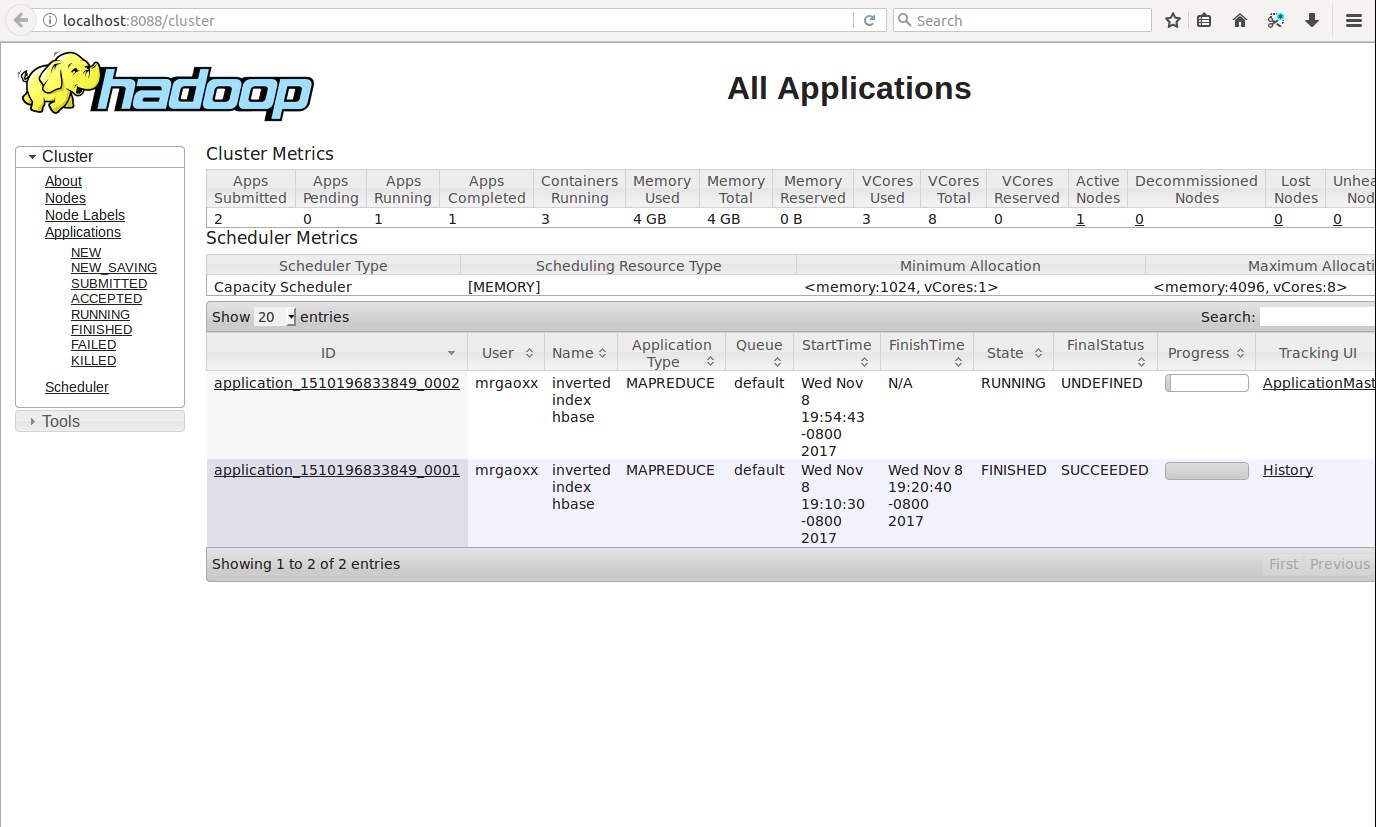
修改倒排索引的reduce程序，将每个词语及其对应的平均出现次数写入 Wuxia表中，运行修改后的倒排索引程序：

|  |
| --- |
| hadoop jar mapreduce-lab.jar hbase.InvertedIndex input output |

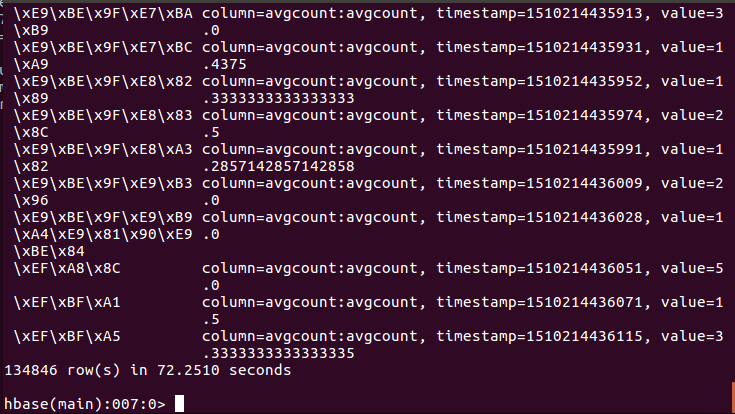
Job的运行结果如下图所示：



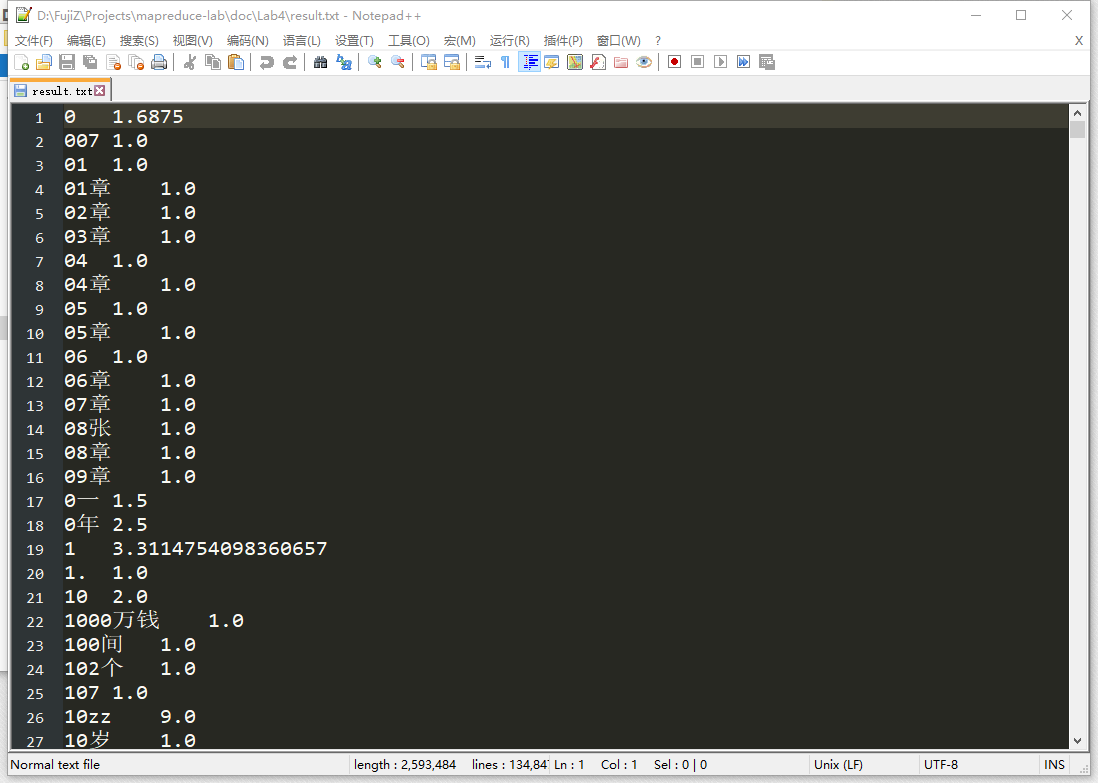




运行结束后通过scan命令遍历HBase中的Wuxia表，结果如下图所示，其中单词显示为16进制编码：



编写Java程序从HBase中读取计数信息，并将其保存到本地文件result.txt中。部分结果如下图所示：



## Hive

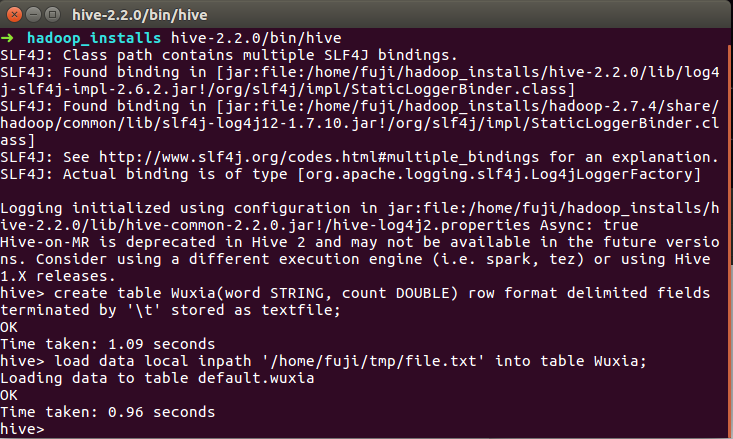
安装Hive，使用Hive Shell命令创建表：

|  |
| --- |
| **create** **table** Wuxia(word STRING, **count** DOUBLE) **row** format delimited fields terminated **by** '\t' stored **as** textfile; |

从本地文件file.txt中导入数据至Wuxia：

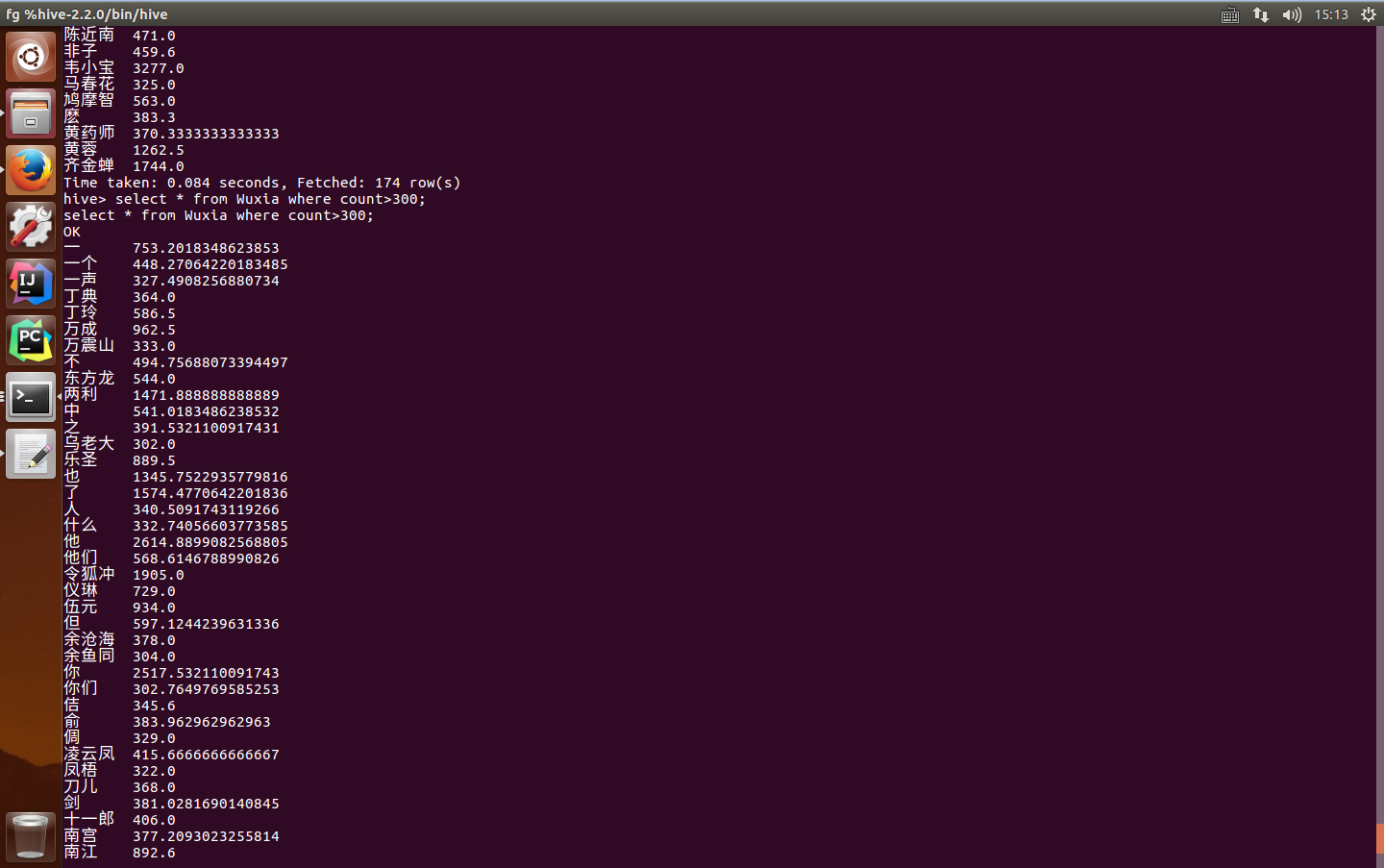
|  |
| --- |
| **load** **data** **local** inpath '/path/to/result.txt' **into** **table** Wuxia; |

运行结果如下图所示：



使用Hive查询出现次数大于300的词语：

|  |
| --- |
| **select** \* **from** Wuxia **where** **count**>300; |



查询前100个出现次数最多的词：

|  |
| --- |
| **select** \* **from** Wuxia sort **by** **count** **desc** **limit** 100; |

运行结果如下图所示：



# 实验体会

本以为HBase、Hive等环境的安装不会很费力，但安装过程中发现会遇到各种各样的问题，然后就需要不断地查找资料，大部分还是各种配置出现问题导致。当然，在安装配置问题没有解决的时候会非常烦躁，最后发现环境安装下来耗时要比编码还要久。想要熟练运用hadoop还是首先需要了解所涉及的系统的基本原理，如Hbase的非结构化的数据库，尤其是存储查询都很快，使得实验运行时间非常短。不得不佩服一下设计分布式大数据处理的工程师们。实验中我们还发现，书中有部分代码已不再适用，这也提醒我们系统在不断更新，我们也需要不断学习来跟上其步伐。