

《并行计算》实验报告

天津大学

Hadoop 算法实现



学 院 计算机科学与技术
专 业 计算机科学与技术
年 级 2012 级
姓 名 王雨朦
学 号 3012216083

2015 年 5 月 10 日

1. 实验内容

1) 实验题目选择

✧ WordCount 应用

WordCount 是一个最简单的分布式应用实例，主要功能是统计输入目录中所有单词出现的总次数，如文本文件中有如下内容：

Hello world

则统计结果应为：

Hello 1

world 1

WordCount 可以使用多种方式实现，本次实验内容要求使用 Hadoop 或者 Spark 实现 WordCount 程序，并完成对应实验报告。

2) 实验环境选择及配置

✧ 实验环境：

Deepin 系统+ openjdk 1.7.0_79+ hadoop-2.6.0

✧ 配置过程

参考 ubuntu 下的 hadoop 配置:<http://www.powerxing.com/install-hadoop/>

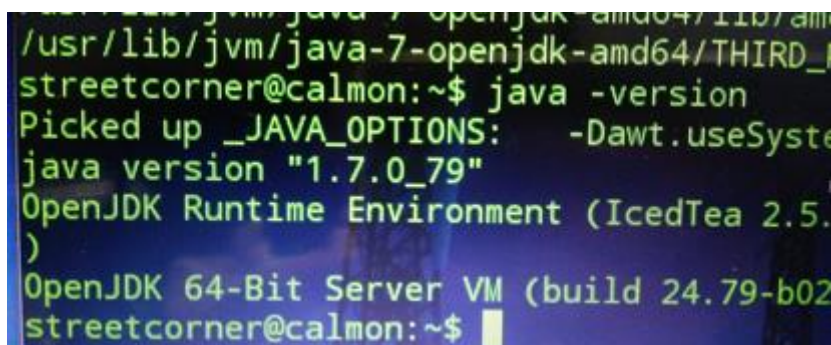
(1) 安装 Java 环境

直接通过命令安装 OpenJDK 7:

```
sudo apt-get install openjdk-7-jre openjdk-7-jdk
```

安装完成后检查: `java -version`

输出版本表示安装成功。



```
streetcorner@calmon:~$ java -version
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAConfig=false
java version "1.7.0_79"
OpenJDK Runtime Environment (IcedTea 2.5.4)
OpenJDK 64-Bit Server VM (build 24.79-b02)
streetcorner@calmon:~$
```

然后配置一下 JAVA_HOME 环境变量：
结果如下：

```
streetcorner@calmon:~$ vim ~/.bashrc
[2]+ 已停止                  vim ~/.bashrc
streetcorner@calmon:~$ vim ~/.bashrc
streetcorner@calmon:~$ source ~/.bashrc
streetcorner@calmon:~$ echo $JAVA_HOME
/usr/lib/jvm/java-7-openjdk-amd64
streetcorner@calmon:~$
```

(2) 安装 Hadoop 2

在链接 <http://mirror.bit.edu.cn/apache/hadoop/common/stable2/> 中下载，下载完成后解压经检查，下载可用：

```
streetcorner@calmon:~$ cd ../hadoop
streetcorner@calmon:~/hadoop$ ./bin/hadoop
Usage: hadoop [--config confdir] COMMAND
    where COMMAND is one of:
    fs                run a generic filesystem user client
    version            print the version
    jar <jar>         run a jar file
    checknative [-a|-h] check native hadoop and compression libraries availability
    distcp <srcurl> <desturl> copy file or directories recursively
    archive -archiveName NAME -p <parent path> <src>* <dest> create a hadoop archive
    classpath          prints the class path needed to get the
    credential         interact with credential providers
    daemonlog          Hadoop jar and the required libraries
    trace              get/set the log level for each daemon
    or                 view and modify Hadoop tracing settings
    CLASSNAME          run the class named CLASSNAME
    Most commands print help when invoked w/o parameters.
```

至此，实验环境安装完毕。

2. 设计实现

1) 实验模型

Hadoop 是 HDFS 和 MapReduce 为核心，为用户提供了系统底层细节透明的分布式基础架构。HDFS 在 MapReduce 任务处理过程中提供了文件操作和存储等支持，MapReduce 在 HDFS 的基础上实现了任务的分发、跟踪、执行等工作，并收集结果，二者相互作用，完成了 Hadoop 分布式集群的主要任务。

2) 实现方法

实验环境配置好后，用 hadoop 自带的 WordCount 检测了一下几个 XML 文件的内容。先将 xml 文件复制进 INPUT，再调用 WordCount 将结果输出到 OUTPUT，过程如下。结果部分截图在下一过程中。文件已附在附件中。

```
streetcorner@calmon:~/hadoop$ mkdir input
streetcorner@calmon:~/hadoop$ ls
bin  include  lib  LICENSE.txt  README.txt  share
etc  input    libexec  NOTICE.txt  sbin
streetcorner@calmon:~/hadoop$ cp ./etc/hadoop/*.xml input
streetcorner@calmon:~/hadoop$ ls
bin  include  lib  LICENSE.txt  README.txt  share
etc  input    libexec  NOTICE.txt  sbin
streetcorner@calmon:~/hadoop$ ./bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.6.0.jar wordcount input output
```

3. 实验结果

下图为部分结果，结果文件已于附件“OUTPUT”中。

```
深度终端
user.      2
user?     1
users     21
users,wheel".  18
uses      2
using     3
value     19
values    1
version   1
version="1.0"  5
version="1.0"?> 2
via       1
when      4
where     1
which     5
while     1
who       2
will      7
window    1
window,   1
with      27
within    1
without   1
work      1
writing,   8
you       9
streetcorner@calmon:~/hadoop$
```

4. 实验总结

学习使用新的技术或软件总是一件令人激动的事。HADOOP 便是如此。

通过配置环境和简单的 wordcount 的实现，体会到 HADOOP 的好用之处，安装在系统上就可以实现集群的效果，方便的实现了多个计算机一起运算数据。