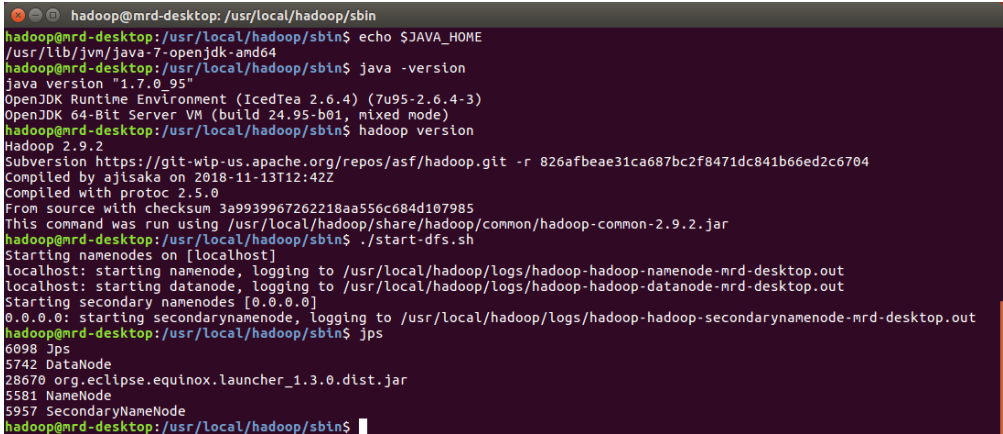


计算机科学与技术学院_大数据管理与分析_课程实验报告

实验题目: 安装单机 Hadoop 系统与 WordCount 程序实验		学号: 201605130116
日期: 2019. 3. 10	班级: 2016 级泰山学堂	姓名: 杜洪超
Email: 1503345074@qq.com		
实验目的: 安装和熟悉单机 Hadoop 系统, 并运行简单的 WordCount 程序, 加深对 Hadoop MapReduce 程序开发的理解, 熟悉实验环境, 为后续实验打下基础。		
实验软件和硬件环境: 软件环境: 系统: Ubuntu16.04 LTS 64 位 软件: openjdk-7-jre, openjdk-7-jdk, java1.7.0_95 Hadoop 2.9.2 Eclipse, ssh 硬件环境: CPU: Intel® Core™ i5-6260U CPU @ 1.80GHz × 4 磁盘: 121.8 GB 内存: 7.7 GiB		
实验原理和方法: 1. 安装和配置环境; 配置 java 和 Hadoop 环境, 安装 ssh 和 eclipse 用于远程登陆和编写程序; 2. 运行 Hadoop, 编写 WordCount 程序导出为 jar 包, 或利用 Hadoop 自带程序, 收集测试数据, 运行并观察结果。		
实验步骤: (不要求罗列完整源代码) 1. 安装与配置环境 安装好 java 和 Hadoop 环境, 测试如下图:  <pre>hadoop@mrd-desktop: /usr/local/hadoop/sbin\$ echo \$JAVA_HOME /usr/lib/jvm/java-7-openjdk-amd64 hadoop@mrd-desktop: /usr/local/hadoop/sbin\$ java -version java version "1.7.0_95" OpenJDK Runtime Environment (IcedTea 2.6.4) (7u95-2.6.4-3) OpenJDK 64-Bit Server VM (build 24.95-b01, mixed mode) hadoop@mrd-desktop: /usr/local/hadoop/sbin\$ hadoop version Hadoop 2.9.2 Subversion https://git-wip-us.apache.org/repos/asf/hadoop.git -r 826afbae31ca687bc2f8471dc841b66ed2c6704 Compiled by ajsaka on 2018-11-13T12:42Z Compiled with protoc 2.5.0 From source with checksum 3a9939967262218aa556c684d107985 This command was run using /usr/local/hadoop/share/hadoop/common/hadoop-common-2.9.2.jar hadoop@mrd-desktop: /usr/local/hadoop/sbin\$./start-dfs.sh Starting namenodes on [localhost] localhost: starting namenode, logging to /usr/local/hadoop/logs/hadoop-hadoop-namenode-mrd-desktop.out localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-hadoop-datanode-mrd-desktop.out Starting secondary namenodes [0.0.0.0] 0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-hadoop-secondarynamenode-mrd-desktop.out hadoop@mrd-desktop: /usr/local/hadoop/sbin\$ jps 6098 Jps 5742 DataNode 28670 org.eclipse.equinox.launcher_1.3.0.dist.jar 5581 NameNode 5957 SecondaryNameNode hadoop@mrd-desktop: /usr/local/hadoop/sbin\$</pre>		
Web 界面如下:		

127.0.0.1:50070/dfshealth.html#tab=overview

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Overview 'localhost:9000' (active)

Started:	Sun Mar 10 03:36:03 +0800 2019
Version:	2.9.2, r826afbeae31ca687bc2f8471dc841b66ed2c6704
Compiled:	Tue Nov 13 20:42:00 +0800 2018 by ajsaka from branch-2.9.2
Cluster ID:	CID-1d6ce16f-9db2-4b0f-8443-a371722c2ec7
Block Pool ID:	BP-613047386-127.0.1.1-1551870566866

Summary

Security is off.
 Safemode is off.
 10 files and directories, 4 blocks = 14 total filesystem object(s).
 Heap Memory used 73.77 MB of 222.5 MB Heap Memory. Max Heap Memory is 889 MB.
 Non Heap Memory used 34.84 MB of 36.44 MB Committed Non Heap Memory. Max Non Heap Memory is 214 MB.

Configured Capacity:	112.92 GB
DFS Used:	80 KB (0%)
Non DFS Used:	53.62 GB
DFS Remaining:	53.54 GB (47.41%)
Block Pool Used:	80 KB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	1 (Decommissioned: 0, In Maintenance: 0)

2. 测试数据

测试数据使用了山东大学官网及相关链接的网页内容文件，一共爬取了 154 个网页文件，共 2.73M。

3. 编写程序，导出 jar 包；将数据和 jar 包复制到 HDFS 后，利用 jar 命令下发作业，执行情况如下：

```
hadoop@mrd-desktop: ~/hadoop
hadoop@mrd-desktop:~/hadoop$ hdfs dfs -ls /usr
Found 3 items
-rw-r--r-- 1 hadoop supergroup      10740 2019-03-10 00:27 /usr/WordCount.jar
drwxr-xr-x - hadoop supergroup         0 2019-03-10 03:50 /usr/test-in
drwxr-xr-x - hadoop supergroup         0 2019-03-10 03:57 /usr/test-out
hadoop@mrd-desktop:~/hadoop$
```

MapReduce Application application_1552220604043_0001

Active Jobs

Job ID	Name	State	Map Progress	Maps Total	Maps Completed	Reduce Progress	Reduces Total	Reduces Completed
job_1552220604043_0001	word count	RUNNING	<div></div>	154	52	<div></div>	1	0

Showing 1 to 1 of 1 entries

```

hadoop@mr-ds-tp: ~/hadoop
hadoop@mr-ds-tp:~/hadoop$ hadoop jar WordCount.jar /usr/data-in /usr/data-out
19/03/10 20:38:00 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
19/03/10 20:38:00 INFO input.FileInputFormat: Total input files to process : 154
19/03/10 20:38:00 INFO mapreduce.JobSubmitter: number of splits:154
19/03/10 20:38:00 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. I
nstead, use yarn.system-metrics-publisher.enabled
19/03/10 20:38:00 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1552220604043_0001
19/03/10 20:38:01 INFO impl.YarnClientImpl: Submitted application application_1552220604043_0001
19/03/10 20:38:01 INFO mapreduce.Job: The url to track the job: http://mr-ds-tp:8088/proxy/application_1552220604043_
0001/
19/03/10 20:38:01 INFO mapreduce.Job: Running job: job_1552220604043_0001
19/03/10 20:38:08 INFO mapreduce.Job: Job job_1552220604043_0001 running in uber mode : false
19/03/10 20:38:08 INFO mapreduce.Job: map 0% reduce 0%
19/03/10 20:38:24 INFO mapreduce.Job: map 4% reduce 0%
19/03/10 20:38:36 INFO mapreduce.Job: map 5% reduce 0%
19/03/10 20:38:37 INFO mapreduce.Job: map 8% reduce 0%
19/03/10 20:38:47 INFO mapreduce.Job: map 12% reduce 0%
19/03/10 20:38:58 INFO mapreduce.Job: map 16% reduce 0%
19/03/10 20:39:09 INFO mapreduce.Job: map 19% reduce 0%
19/03/10 20:39:10 INFO mapreduce.Job: map 20% reduce 0%
19/03/10 20:39:20 INFO mapreduce.Job: map 21% reduce 0%
19/03/10 20:39:21 INFO mapreduce.Job: map 23% reduce 0%
19/03/10 20:39:29 INFO mapreduce.Job: map 24% reduce 0%
19/03/10 20:39:30 INFO mapreduce.Job: map 27% reduce 0%
19/03/10 20:39:32 INFO mapreduce.Job: map 27% reduce 9%
19/03/10 20:39:38 INFO mapreduce.Job: map 28% reduce 9%

```

```

hadoop@mr-ds-tp: /usr/local/hadoop/sbin
hadoop@mr-ds-tp:/usr/local/hadoop/sbin$ hdfs dfs -ls /usr
Found 3 items
-rw-r--r-- 1 hadoop supergroup 10740 2019-03-10 20:36 /usr/WordCount.jar
drwxr-xr-x 0 2019-03-10 20:37 /usr/data-in
drwxr-xr-x 0 2019-03-10 20:42 /usr/data-out
hadoop@mr-ds-tp:/usr/local/hadoop/sbin$ hdfs dfs -ls /usr/data-out
Found 2 items
-rw-r--r-- 1 hadoop supergroup 0 2019-03-10 20:42 /usr/data-out/_SUCCESS
-rw-r--r-- 1 hadoop supergroup 1118632 2019-03-10 20:42 /usr/data-out/part-r-000000
hadoop@mr-ds-tp:/usr/local/hadoop/sbin$ hdfs dfs -tail /usr/data-out/part-r-000000
**山东大学陈子江教授团队的又一项全国多中心前瞻性随机对照临床研究，以原创研究文章形式在国际著名医学期刊《
柳叶刀》杂志（影响因子：53.254）发表，题为“体外受精冷冻单个囊胚移植与新鲜单囊胚移植的比较”（“Frozen
【本站讯】近日，材料科学与工程学院李辉教授课题组在金属纳米粒子氧化机理研究方面取得了重要进展，其研究成果“
Atomistic
【机械】举行青年教师教学科研交流会 2
【校友】李宇兵：勤奋为基石 1
【校奖得主】冯毫：医道传承，上下求索2019-01-25 1
【校奖得主】杜晓颖：心之所向，译者无疆2019-01-18 1
【物理】与实俱进实验室志愿服务活动启动 2
【综合事务】学校荣获计划生育工作先进单位称号 2
【综合事务】山东大学召开一校三地财务工作研讨会 2
【综合事务】山大召开资产与实验室工作联席会议 2
【药学】召开安全工作专题会议 2
hadoop@mr-ds-tp:/usr/local/hadoop/sbin$

```

结论分析与体会：

实验最终没有实现预期效果，出现把大段文字识别为单词的现象；但原因是 java 的字符串分解过于简单，不能做到有效分词，不过这与本次实验目的无关，可以通过使用其它更成熟的分词工具解决。

通过本次实验，我熟悉了 Hadoop 系统和 HDFS，对 MapReduce 编程有了更深入的理解，为以后的实验打下了良好的基础。

就实验过程中遇到和出现的问题，你是如何解决和处理的，自拟 1—3 道问答题：

1. Hadoop 安装时出现不能识别 java 类的错误；
解决办法：更换低版本的 Hadoop 或使用高版本的 java
2. 使用 jar 命令提交作业时不能识别 java 类；
解决办法：最终找出原因在于导出 jar 包时没有指定默认入口类，因此需要在命令中显示加上类的全称或提前指定；
3. 课件或书中提供的代码存在已被舍弃的用法警告；可使用其它合法语法
4. Datanode 节点出现无法启动现象；
把 tmp 目录中 data 和 name 下的 VERSION 文件中的集群 ID 统一即可。
5. Web 界面找不到 jobs 执行情况；安装 yarn 可解决。