Report

> PB20020480 王润泽
>
> Windows11，WSL2， Ubuntu20.04
>
> hadoop-1.0.4

Hadoop是一个开源的分布式计算框架，旨在处理大规模数据集的存储和处理。Hadoop的核心组件包括以下几个部分：

1. Hadoop分布式文件系统（HDFS）：HDFS是Hadoop的分布式文件系统，用于可靠地存储大规模数据集。它将数据划分为多个块，并复制到集群中的多个节点上，以实现容错性和数据可用性。
2. MapReduce编程模型：MapReduce是Hadoop的计算模型，用于并行处理大规模数据集。它将计算任务划分为两个阶段：Map和Reduce。Map阶段将输入数据划分为多个独立的部分，然后并行处理这些部分。Reduce阶段将Map阶段的输出合并并生成最终的结果。
3. YARN（Yet Another Resource Negotiator）：YARN是Hadoop的资源管理器，用于集群中的资源调度和作业管理。它允许多个应用程序在集群上共享资源，并提供了灵活的资源分配和任务调度机制。

本次实验在 Linux 系统上安装搭建 hadoop环境,并熟悉其常规操作

1 搭建环境

根据文件说明，配置SSH，安装Hadoop，并启动WordCount.java程序，

```
1. 安装jdk
2. 验证并安装ssh
3. 生成ssh秘钥对
4. 安装hadoop
5. 配置单机模式
6. 配置伪分布模式
7. 格式化HDFS
~/hadoop/hadoop-1.0.4/bin/hadoop namenode -format
8. 启动hadoop
~/hadoop/hadoop-1.0.4/bin/start-all.sh
9. 检测hadoop是否成功启动
10. 在HDFS中添加文件和目录
11. 在当前目录下新建一个WordCount.java文件
12. 编译WordCount.java
13. 运行Hadoop作业
14. 获得运行结果
15. 关闭hadoop集群
```

得到结果

```
a       1
b       1
bc      1
cd      1
dd      1
de      1
dhs     1
```

```
e       2
f       1
fg      1
g       1
gh      1
hdk     1
tt      2
```

2. 单词长度频率统计

修改 WordCount.java 以下内容，将映射从单词名称改成单词长度

```
public void map(Object key, Text value, Context context) throws IOException,
InterruptedException {
    StringTokenizer itr = new StringTokenizer(value.toString());
    while (itr.hasMoreTokens()) {
        word.set(itr.nextToken());
        Text truth = new Text(Integer.toString(word.getLength()));
        context.write(truth, one);
    }
}
```

重复 11-14步骤，得到以下结果

```
1       6
2       8
3       2
```

3. 总结

本次实验主要是根据手册配置Hadoop环境，熟悉了Hadoop的常用操作，并通过Hadoop运行程序得到结果