大数据导论实验

实验一
Hadoop环境配置与基本操作

主讲教师：叶允明

实验教师：房敏、谢佳

目录

- 实验课程总体介绍
- 实验一任务
- 预备知识
- Hadoop的安装与使用
- WordCount 程序任务

本学期实验总体安排

实验课程共4个学时，2个实验项目，总成绩为30分。

实验一（10分）

Hadoop环境配置与基本操作

掌握大数据存储与检索的开源软件工具，并能完成给定的大数据存储与检索任务。

实验二（20分）

数据理解、数据预处理及决策树的应用

通过应用案例实践数据预处理方法；

编码实现一个经典数据挖掘算法。

实验作业提交

- **截止时间**
  请实验课后一周内（晚12：00）提交实验作业至指定邮箱：
  657253554@qq.com (1-4班)
  853669786@qq.com (5-7班)

- **提交内容**
  实验一：实验报告+词频统计结果
  实验二：实验报告+工程文件
  请使用实验报告模板，内容需包含实验目的、实验内容、实验过程、实验结果与分析。

- **命名要求**
  文件夹、邮件标题及实验报告命名规则：
  学号_姓名_实验编号

实验一任务

实验环境：
Ubuntu 16.04 & Hadoop 2.7.2

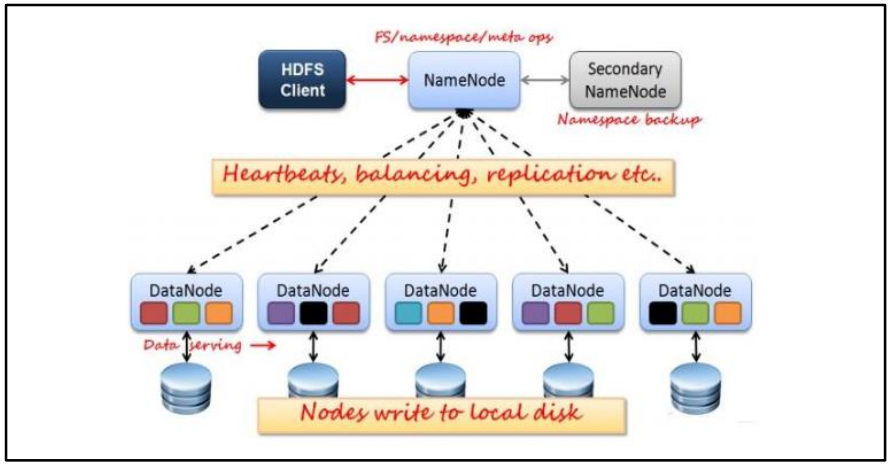
一、独立搭建Hadoop单机环境和伪分布式环境

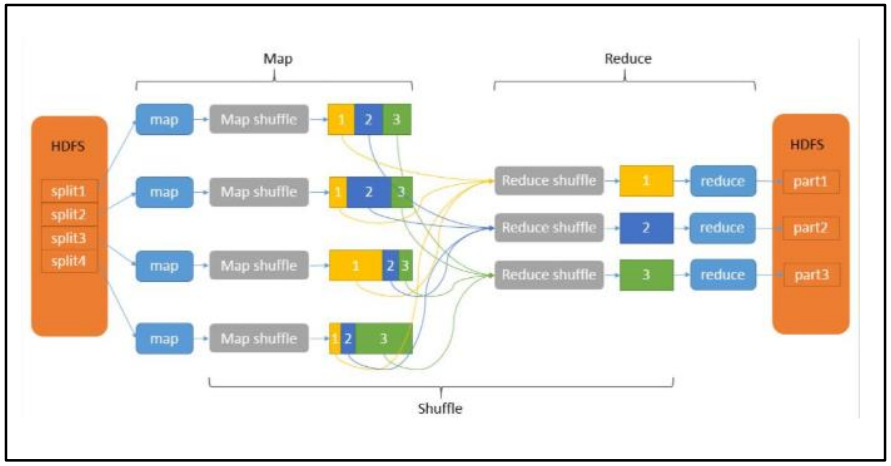
二、使用MapReduce实现WordCount程序任务

预备知识

Hadoop是一个由Apache基金会所开发的分布式系统基础架构。用户可以在不了解分布式底层细节的情况下，开发分布式程序。

Hadoop两大核心：HDFS + MapReduce





预备知识

Linux操作系统

实验环境：Ubuntu系统（双系统安装）

```
uname –a
```

shell

命令解析器，类似于DOS下的command。它接收用户命令，然后调用相应的应用程序。

sudo

一种权限管理机制，管理员可以授权给一些普通用户去执行一些需要root权限执行的操作。

vi/vim

linux命令行下的最著名的文本编辑器。

- 命令模式
- 编辑模式
- 末行模式

图形化编辑器 **Gedit**

预备知识

Hadoop安装方式

单机环境

Hadoop 默认模式为非分布式模式（本地模式），无需进行其他配置即可运行。

伪分布式环境

Hadoop在单个节点运行，该节点既是NameNode也是DataNode，读取HDFS 中的文件。

分布式环境

Hadoop在多个节点构成的集群环境上运行，读取 HDFS 中的文件。

Hadoop的安装与使用

主要包括以下几个步骤

创建Hadoop用户

SSH登录权限设置

安装Java环境

单机安装配置

伪分布式安装配置

Hadoop的安装与使用

创建Hadoop用户

打开终端窗口：

① 创建hadoop用户并使用bash作为shell

② 设置hadoop 用户密码

③ 为hadoop 用户增加管理员权限

```
sudo useradd -m hadoop -s /bin/bash
```

注意：创建hadoop用户后务必切换至该用户再执行下面的步骤！！！

Hadoop的安装与使用

SSH登录权限设置

SSH 为 Secure Shell 的缩写，是建立在应用层和传输层基础上的安全协议。

Hadoop名称节点（NameNode）需要启动集群中所有机器的Hadoop守护进程，这个过程需要通过SSH登录来实现。因此，为了能够顺利登录每台机器，需要将所有机器配置为名称节点可以无密码登录它们。

```
ssh-keygen -t rsa
```

```
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
```

Hadoop的安装与使用

安装Java环境

① 更新软件包列表

```
sudo apt-get update
```

② 在Ubuntu中直接通过命令安装 OpenJDK 8

```
sudo apt-get install openjdk-8-jre openjdk-8-jdk -y
```

③ 配置 JAVA_HOME 环境变量

```
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
```

Hadoop的安装与使用

单机环境安装配置

① 解压安装包 hadoop-2.7.2.tar.gz 到 /usr/local 文件夹下并重命名

```
sudo tar -xzvf hadoop-2.7.2.tar.gz -C /usr/local

sudo mv  /usr/local/hadoop-2.7.2  /usr/local/Hadoop

sudo chown –R hadoop:hadoop /usr/local/Hadoop
```

② 配置 Hadoop 环境变量及 Hadoop 中的JAVA_HOME

③ 运行并完成 WordCount 程序任务

Hadoop的安装与使用

伪分布式环境安装配置

① 修改 core-site.xml配置文件

② 修改 hdfs-site.xml配置文件

③ 执行 NameNode 的格式化

```
/usr/local/Hadoop/bin/hdfs namenode -format
```

④ 开启 NameNode 和 DataNode 守护进程

```
/usr/local/Hadoop/sbin/start-dfs.sh
```

⑤ 运行并完成 WordCount 程序任务

WordCount程序任务

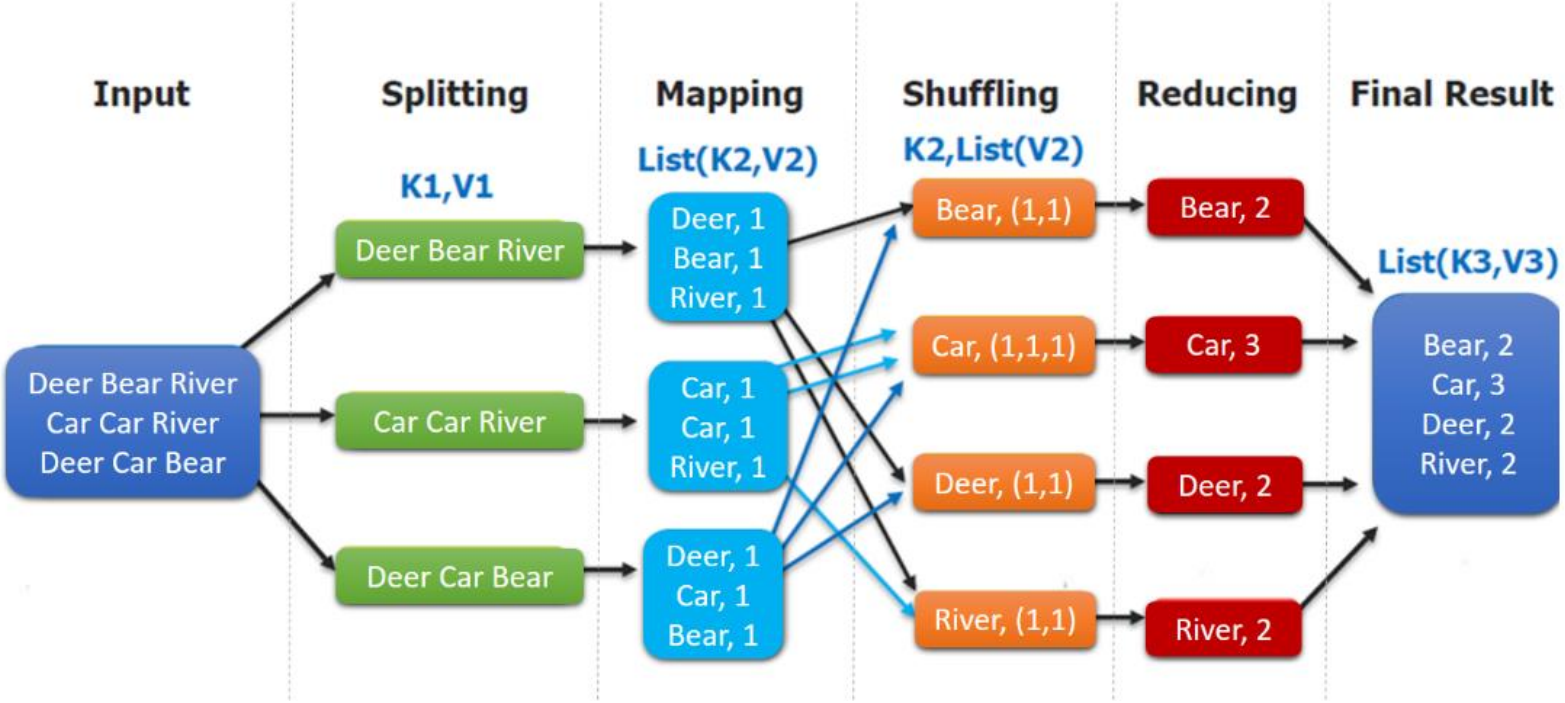
程序	WordCount
输入	一个包含大量单词的文本文件
输出	文件中每个单词及其出现次数（频数），并按照单词字母顺序排序，每个单词和其频数占一行，单词和频数之间有间隔

输入和输出示例

输入	输出
Hello World Hello Hadoop Hello MapReduce	Hadoop 1 Hello 3 MapReduce 1 World 1

WordCount程序任务

MapReduce 词频统计处理流程：



WordCount程序任务

单机环境:

```
hadoop jar
/usr/local/Hadoop/share/hadoop/mapreduce/hadoop-
mapreduce-examples-2.7.2.jar wordcount
/usr/test/hadoop/input /usr/test/hadoop/output
```

伪分布式环境:

```
hadoop jar
/usr/local/Hadoop/share/hadoop/mapreduce/hadoop-
mapreduce-examples-2.7.2.jar wordcount input
output
```

大数据导论实验

同学们，请开始实验吧！