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：谢佳、房敏

目录

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（15分）

Hadoop环境配置与基本操作

掌握Hadoop分布式环境的配置方法；理解Mapreduce作业的原理和操作方法。

实验二（15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；编码实现一个经典的数据挖掘算法。

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：
  bigdata2021fall@163.com

- **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 3.2.2 & Java 1.8

一、小组合作搭建Hadoop分布式环境

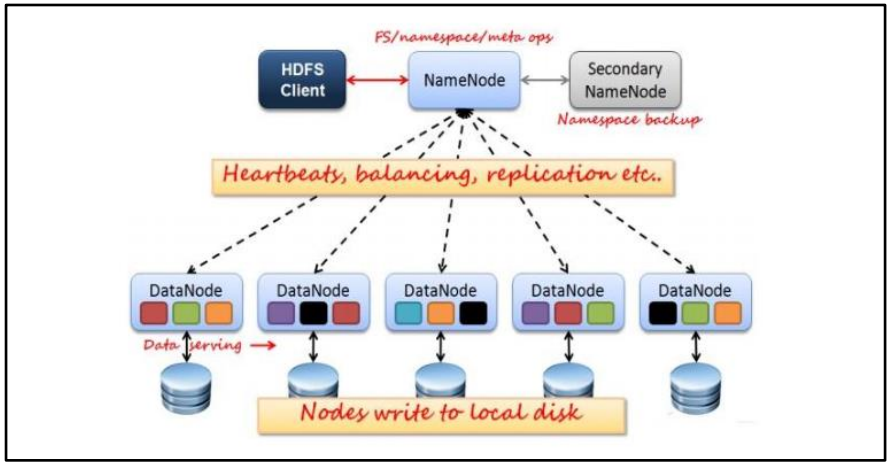
二、配置Hadoop客户端，上传预处理后的实验数据

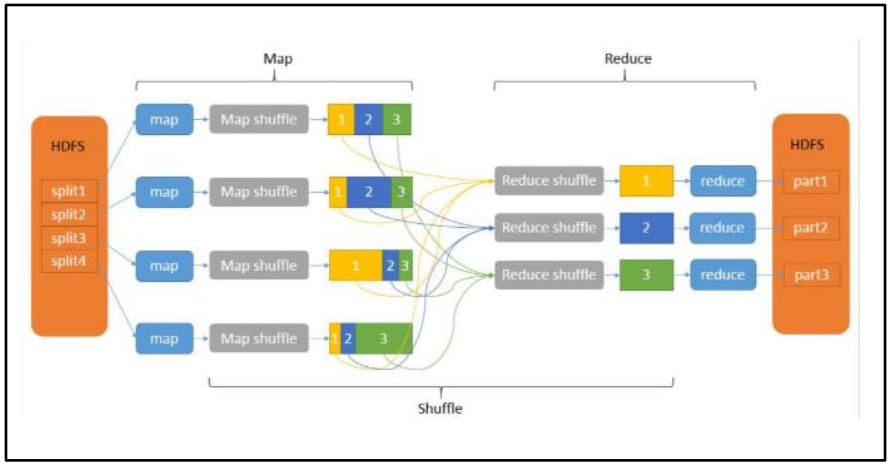
三、使用MapReduce实现WordCount程序任务

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

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分布式环境的搭建

主要包括以下几个步骤

集群机器配置

配置ssh无密码登录

安装JDK和Hadoop

配置Hadoop集群

启动hadoop集群

Hadoop分布式环境的搭建

集群机器配置

① 基础配置：apt更新、查看防火墙、修改主机名
② 创建用户：新增用户hitsz，添加sudo权限
③ 配置Hosts：域名和IP地址映射

计算机名	IP地址	自定义域名
hitsz-master	10.248.5.xx	master
hitsz-slave1	10.248.5.xx	slave1
hitsz-slave2	10.248.5.xx	Slave2

注意：执行后续步骤时务必注意hitsz用户和root 用户的切换！！！

Hadoop分布式环境的搭建

配置ssh无密码登录

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。

注意:

①每个节点需为hitsz用户配置到所有节点（包括自身）的免密登录。

②Master节点还需为root用户配置到所有节点（包括自身）的免密登录。

Hadoop分布式环境的搭建

安装JDK和Hadoop

① 解压 jdk 和 hadoop 安装包至 /opt/module 目录

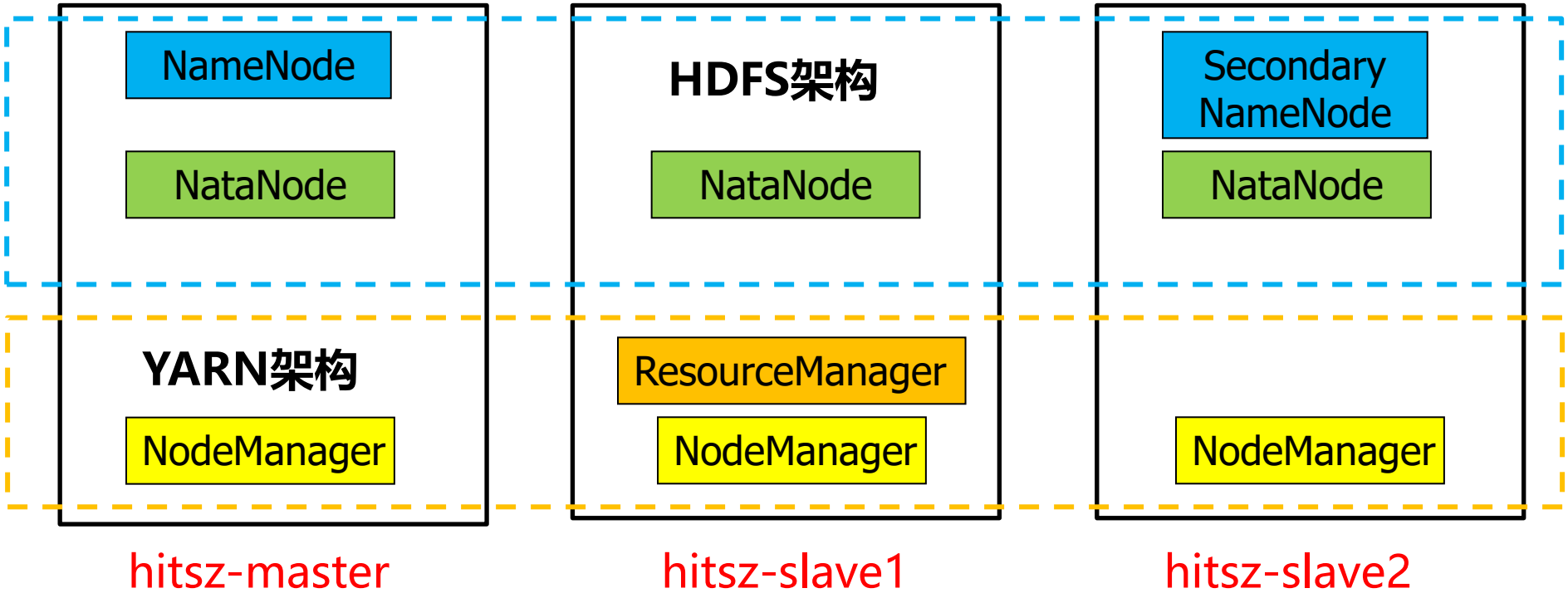
② 添加 java 和 hadoop环境变量 至 my_env.sh

③ 使添加的环境变量生效

注意：上述操作只在hitsz-master上进行

Hadoop分布式环境的搭建

配置Hadoop集群



Hadoop分布式环境的搭建

配置Hadoop集群

① 修改核心核心配置文件：core-site.xml

② 修改 HDFS 配置文件：hdfs-site.xml

③ 修改 Yarn 配置文件：yarn-site.xml

④ 修改 MapReduce 配置文件：mapred-site.xml

⑤ 修改Worker文件，配置DataNode节点

⑥ 使用xsync脚本同步/opt/module目录及my_env.sh至其他节点

注意：上述操作只在hitsz-master上进行

Hadoop分布式环境的搭建

启动Hadoop集群

①格式化NameNode（首次启动执行）

②启动hdfs（NameNode）

③启动yarn（ResourceManager）

④启动历史记录服务器

⑤使用jps命令查看

⑥使用WebUI界面查看

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

①配置hadoop客户端

②上传预处理后的实验文本数据至HDFS的input目录（hadoop fs）

③为MapReduce 应用程序添加集群运行配置

④在客户端执行WordCount程序

⑤查看output目录的统计结果

大数据导论实验

同学们，请开始实验吧！