



大数据处理综合实验课程 (2025)

课程实验1-Hadoop系统安装、开源大数据系统实验工具链使用

南京大学 计算机学院



课程实验1：Hadoop系统安装、开源大数据系统实验工具链使用

- 实验目标：
 - 在本地安装 Hadoop, Git 开源大数据工具链；
 - 完成每个工具的简单示例操作，验证安装是否成功；
 - 熟悉开源大数据工具链的基本使用。
- 工具链说明：
 - Hadoop：分布式存储（HDFS）和计算（MapReduce）的基础框架。
 - Git：代码版本控制系统，用于跟踪、管理项目文件变更，支持多人协作开发。



课程实验1：Hadoop系统安装、开源大数据系统实验工具链使用

- 任务1：Hadoop 伪分布式环境安装与配置，HDFS基本操作；
- 任务2：执行Hadoop官方示例程序wordcount；
- 任务3：Git 安装与基本操作；
- 附加实验（可选）：执行更多的Hadoop官方示例程序。

实验环境：

- 操作系统：Linux；
- 软件环境：jdk-8(java 1.8.0_201), Hadoop-3.2.1；
- 资源需求：2核vCPU, 4G内存, 20G硬盘；



任务1：Hadoop 伪分布式环境安装与配置，HDFS基本操作

- 下载并解压Hadoop软件包，配置Hadoop核心配置文件
 - 配置core-site.xml, hdfs-site.xml, mapred-site.xml, yarn-site.xml。
- 格式化 HDFS 并启动 Hadoop相关进程
 - 在浏览器查看图形化的HDFS文件系统以及Yarn页面。
- 使用hdfs dfs / hadoop fs （这两个命令在大多数情况下等价）进行HDFS命令行基本操作。



YARN



任务2：执行Hadoop官方示例程序wordcount

1.运行wordcount示例程序，观察MapReduce任务执行过程

- ① 上传文本文件 example.txt 到HDFS
- ② 使用官方jar包执行wordcount任务
- ③ 观察MapReduce执行过程（同时可以在YARN Web页面查看执行情况）

执行结果示例：

```
master@MapReduce:~$ hdfs dfs -ls /test/output
Found 2 items
-rw-r--r--  3 master supergroup      0 2025-02-09 10:13 /test/output/_SUCCESS
-rw-r--r--  3 master supergroup    73 2025-02-09 10:13 /test/output/part-r-00000
master@MapReduce:~$ hdfs dfs -cat /test/output/part-r-00000
HBase      2
Hadoop     1
Hive       2
Spark      2
Zookeeper  1
hadoop     3
spark      1
zookeeper  2
master@MapReduce:~$ _
```



任务3：Git 安装与基本操作

- 在本机安装好Git，并完成Git的全局配置。
 - 在本机创建SSH公私钥对，将SSH公钥添加到个人Github账户上。
- 将fluid-cloudnative/fluid (<https://github.com/fluid-cloudnative/fluid>) 复刻（fork）到自己的Github仓库，克隆（clone）到本地，建立新的分支（branch）。
- 按照Git任务文档和任务分配文档完成指定任务。



任务3：Git 安装与基本操作

- 完成任务后提交（commit），并推送（push）到自己Github账户的远程仓库相应分支（branch）。
- 创建一个Issue到fluid-cloudnative/fluid，再创建一个合并请求（pull request）至fluid-cloudnative/fluid，等待pull request被接受。
- 及时处理PR页面中他人提出的修改意见，等待最终合并（merge）。



附加实验（可选）：执行更多的Hadoop官方示例程序

1. 本地运行 pi 示例程序，观察 MapReduce 任务执行过程（8个 map 任务，每个任务1,000,000个样本）；
2. 实验平台上运行 pi 示例程序，观察MapReduce任务执行过程（8个 map 任务，每个任务1,000,000个样本）；
3. 比较1和2的执行时间，说明你的发现。
4. 本地或实验平台上运行 grep 示例程序，找出给定的 example.txt 中“hadoop”这个单词出现的次数。



实验报告要求

实验报告的内容至少包括：

1. Hadoop 伪分布式环境安装与配置，HDFS基本操作
 1. HDFS, Yarn 启动截图（包括Web页面）
 2. HDFS 命令行基本操作执行截图
2. 执行Hadoop官方示例程序wordcount
 1. Wordcount 示例程序执行截图
3. Git 安装与基本操作（可后续单独提交）
 1. Git 安装成功截图
 2. 本地具体修改截图
 3. Issue提交、PR提交、PR被接受以及最终合并截图（注明提交的Issue和PR编号）
4. 附加实验（可选）
 1. 步骤1和步骤2的运行结果截图；
 2. 说明步骤1和步骤2在运行时间上的差异，尝试找出原因。
 3. 步骤4的运行结果截图。

实验报告的格式要求：

- 报告格式应为PDF；
- 代码及其他资源（如果有）打包为 zip；
- 报告/压缩包文件名：学号_姓名_BG_x，其中 x 表示第 x 次实验；
- 如果提交内容包含 Jar 包，应当附带 Jar 包的执行说明。
- **DDL**
 - 任务1+任务2：2025年3月18日
 - 任务3：本学期末