实验1：单机Hadoop系统与程序开发工具

实验内容与要求

任务1：单机Hadoop系统与WordCount程序

- 每人在自己本地电脑上正确安装和运行伪分布式Hadoop系统。
- 安装完成后，自己寻找一组英文网页数据，在本机上运行Hadoop系统自带的WordCount可执行程序文件，并产生输出结果。

任务2：基本开发工具的安装使用

- 在本机上安装Maven、Git，创建自己的Github帐号
- 在任务分配Excel文件中找到自己的Task ID，然后查看对应ID的任务要求：https://github.com/PasaLab/MR-Course-Assignments/blob/spring-2022/issue_list.md
- 将fluid-cloudnative/fluid（https://github.com/fluid-cloudnative/fluid）fork到自己的Github仓库，clone到本地，建立新的分支完成给定任务；
- 完成任务后提交commit，并push到自己Github帐号远程仓库相应分支，最后创建并提交pull request至fluid-cloudnative/fluid；
- 及时处理PR页面中他人提出的修改意见，本地修改后push到自己Github帐号远程仓库即可（不需要重新创建PR），并等待最终merge。

实验1：单机Hadoop系统与程序开发工具

实验内容与要求

- 任务1要求书写一个实验报告
  - 实验报告的内容其中包括：
    - 系统安装运行情况
    - 实验数据说明（下载的什么网页数据，多少个HTML或text文件）
    - 程序运行后在Hadoop Web作业状态查看界面上的作业运行状态屏幕拷贝
    - 实验输出结果开头部分的屏幕拷贝
    - 实验体会
  - 实验报告文件命名规则： MPLab1-学号-姓名.docx
  - 实验报告提交至：本科教学支撑平台 http://cslabcms.nju.edu.cn/
  - 实验报告提交截止日期：3月31日（包含当天）
- 任务2不要求书写实验报告，以PR的提交与merge为完成依据
  - 助教将根据开源社区实际情况分批下发任务，任务下发及截止时间另行通知
  - 请所有同学在实验前仔细阅读实验注意事项，实验过程中请务必严格遵守开源社区规范，提交的PR标题和内容写实际的修改描述，不要写作业/Assignment/学号等不相关信息。助教会根据PR自行判断属于哪位同学提交的，如发现有同学临近DDL未提交PR，也会私信提醒一次。具体注意事项见下文。

实验注意事项

- 每人只能完成自己的任务，若完成他人的任务则不计分；
- 注意pull request标题及描述格式：

注意: Fluid项目的PR提交时格式如下：：

```
I. Describe what this PR does
II. Does this pull request fix one issue?
III. List the added test cases (unit test/integration test) if any, please explain if no tests are needed.
IV. Describe how to verify it
V. Special notes for reviews
```

这里只需要填写**I. Describe what this PR does**即可，例如：

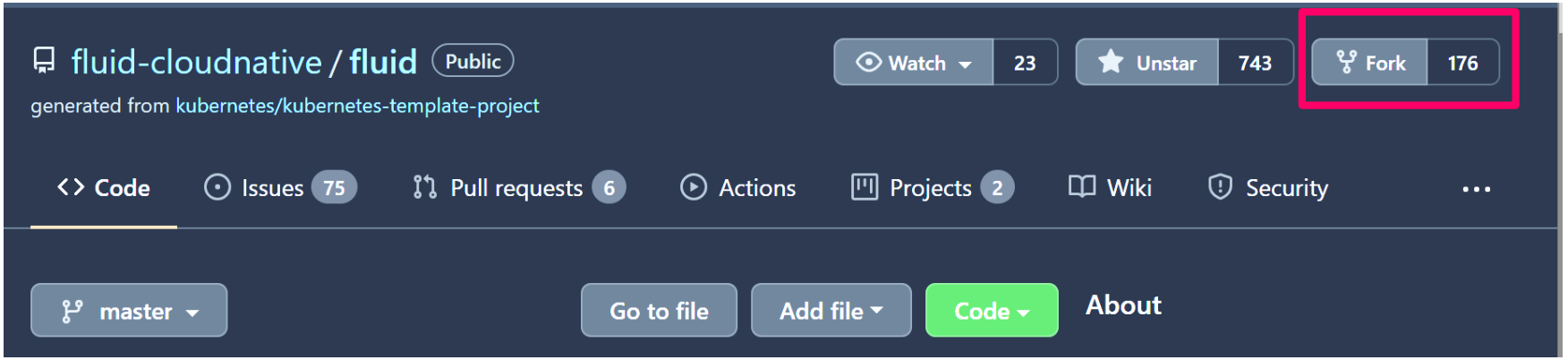
```
This pr is to add + 功能描述
```

实验注意事项

- 每人只能完成自己的任务，若完成他人的任务则不计分；
- 注意pull request标题及描述格式；
- 时常关注pull request页面，并及时处理相应评论；
- 提示：PR的处理通过需要时间，请合理规划实验进度，以免错过Deadline。

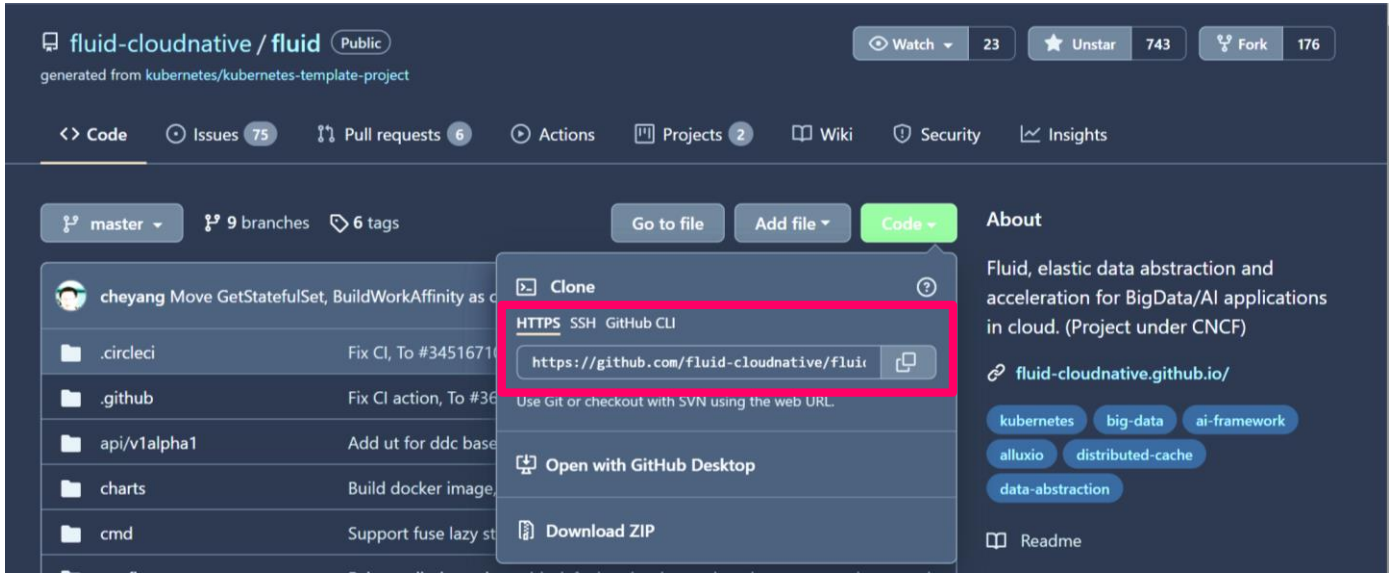
相关软件安装以及fork Fluid仓库

- 在本机安装Git；
- 创建自己的Github账号；
- 将Fluid项目（https://github.com/fluid-cloudnative/fluid）fork到自己Github远程仓库
- 本地配置git，命令行下运行以下命令
  - $git config --global user.name "your_name"
  - $git config --global user.email "your_github_email"



Clone代码到本地、添加远程源并fetch最新代码

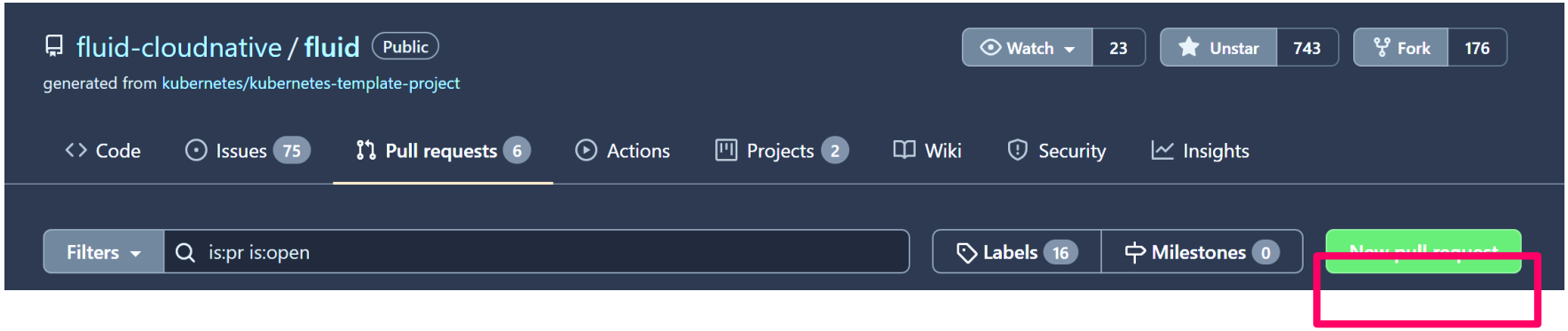
- git clone https://github.com/someone/fluid.git
- 进入项目根目录：cd fluid
- git checkout master
- git remote add upstream https://github.com/fluid-cloudnative/fluid.git
- git fetch upstream
- git merge upstream/master



新建分支并完成指定任务

- git checkout -b <my-branch-name>
- 修改相应文件
- 查看当前Git修改状态 git status
- git add <修改的文件> 或者 git add .
- git commit -m "description （简单描述你的修改）" -s
- git push origin <my-branch-name>:<my-branch-name>

注意，push到自己Github远程仓库就行了，不需要再次创建PR，相应的PR会检测到这次更改



实验1：单机Hadoop系统与程序开发工具

交互式统一大数据编程计算平台

- 交互式统一大数据编程计算平台是PASA实验室开发的在线实验平台，大数据处理课程的后续试验均需要在该平台上完成。
- 该平台集成了MapReduce、Spark并行计算平台；支持HDFS分布式存储和Alluxio内存存储。
- 实验时，通过向平台的编程工作空间上传JAR包，并在线输入JAR包运行指令，即可使用平台的集群资源运行大数据处理程序。
- 平台地址： http://114.212.190.95:8082/

后续实验安排

- 自第二次实验开始，实验均在交互式统一大数据编程计算平台上完成，以小组形式完成后续实验及最后的课程设计。
- 可自由组合实验小组，每组2-4人

后续实验安排

- 近日助教将根据分组情况为各小组分配账号，请各位同学按照个人实际情况填写分组问卷
- 问卷地址：https://wj.qq.com/s2/9754506/b731/
- 也可扫二维码填写问卷：