# /Users/jacqueline/Desktop/WechatIMG38.png

# 基于Hadoop的豆瓣电影数据分析平台

|  |  |  |
| --- | --- | --- |
| **姓名** | 骆振鹏 | 喻敬峰 |
| **学号** | 19040124 |  |
| **专业** | 软件工程 | |
| **班级** | 19软工 | |
| **日期** | 2021-12-21 | |

目录

1. **引言**

1.1编写目的

1.2参考资料

1.3使用的工具

1. **平台设计**

2.1设计概述

2.2设计框架图

2.3 运行环境

1. **数据处理流程**

3.1数据集下载

3.2 上传数据到hadoop

3.3 利用程序进行数据分析

3.4 分析结果存入数据库

1. **项目实战**

4.1 数据的下载与上传

4.2 hadoop与Python连接

4.3 运行程序得到可视化结果

4.4 电影推荐

1. **源码下载**

**一、引言**

**1.1 编写目的**

本文档旨在介绍应用如何设计，包括程序描述、流程逻辑、算法等，同时介绍项目完成情况。

**1.2 参考资料**

[1]Python如何连接mysql数据库 https://www.py.cn/jishu/jichu/12706.html

[2]Movielens数据集详细介绍 https://blog.csdn.net/zwqhehe/article/details/75912003

[3]Python3调用Hadoop的API https://www.cnblogs.com/sss4/p/10443497.html

[4]Collaborative Filtering(协同过滤算法详解) https://www.cnblogs.com/ECJTUACM-873284962/p/8729010.html

**1.3 使用的工具**

windows 10

hadoop 2.8.3

pycharm

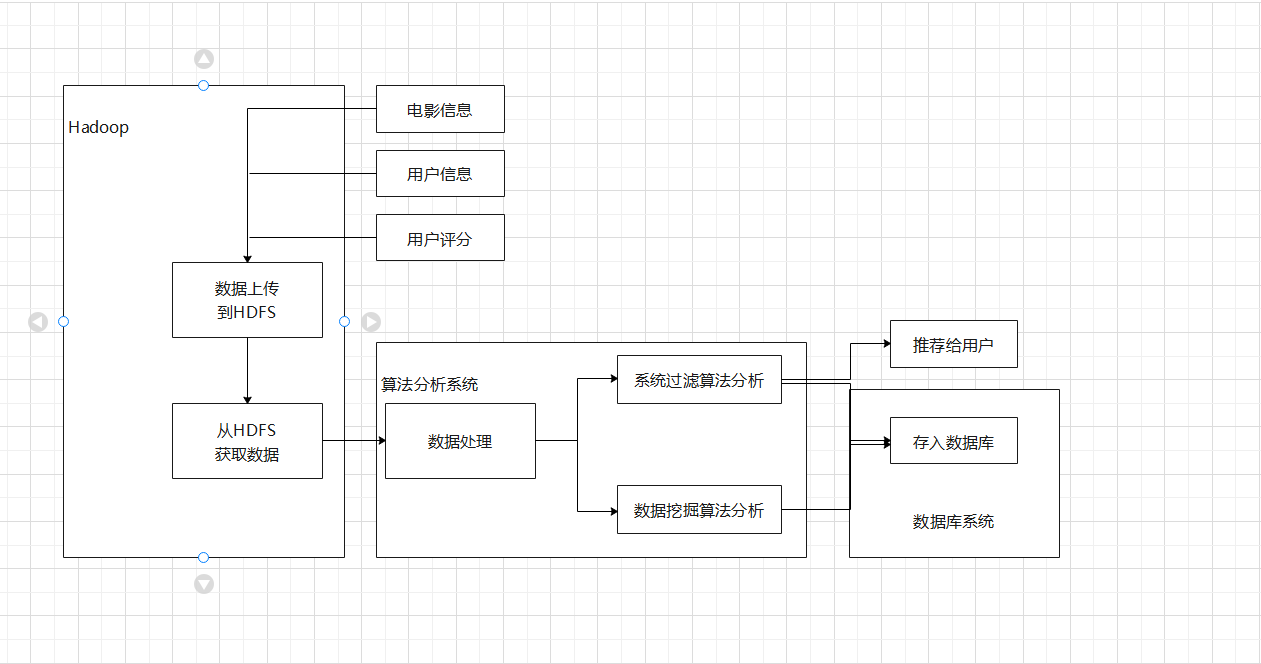
MySQL

**二、平台设计**

**2.1 设计概述**

为了使平台的设计更加完善，提高平台的可维护性和可拓展性。我们采用三种系统相结合的方式进行总的系统层次设计。分为Hadoop为主的分布式系统，python为主的算法分析系统和MySQL为主的数据库存储系统。三种系统之间相互独立，仅采用特定的程序接口进行调用，可以最大程度的增强推荐系统的稳健性。

**2.2 设计框架图**

****

**2.3 运行环境**

HAdoop：在win10系统安装Hadoop直接运行。

Python代码：在pycharm中运行

数据库：mysql-8.0

**三、数据处理流程**

**3.1 数据集下载**

地址：<https://grouplens.org/datasets/movielens/100k/>

**3.2 上传数据到Hadoop**

打开cmd，启动Hadoop，下载的数据上传。

**3.3 利用程序进行数据分析**

主要采用了协调过滤算法分析和数据挖掘算法分析，分析结果都会存入数据库。

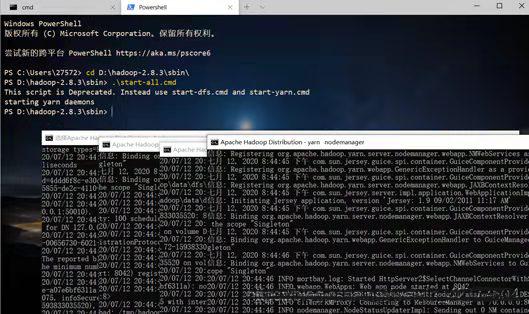
**3.4分析结果存入数据库**

Main函数里面有数据库的调用，存储简单。

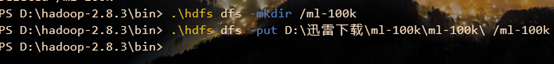
**四、项目实战**

**4.1数据的下载与上传**

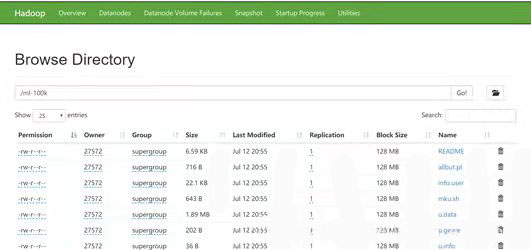
1. 从数据集下载地址下载数据
2. 启动Hadoop



1. 将ml-100k数据集上传到Hadoop下的ml-100k文件夹（该文件夹已提前建好）



上传结果：



**4.2 Hadoop与Python连接**

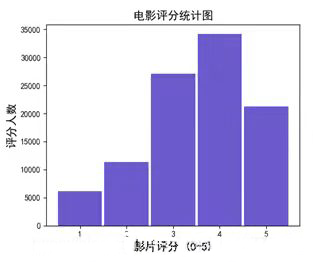
利用python中的hdfs包，用如下代码实现hadoop与python相连接

from hdfs import \*

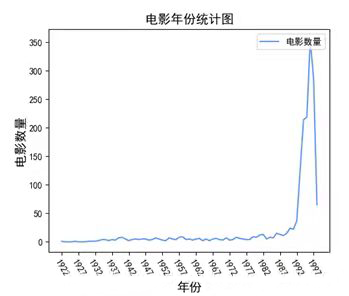
client = Client("http://localhost:50070")

**4.3 运行程序得到可视化结果**

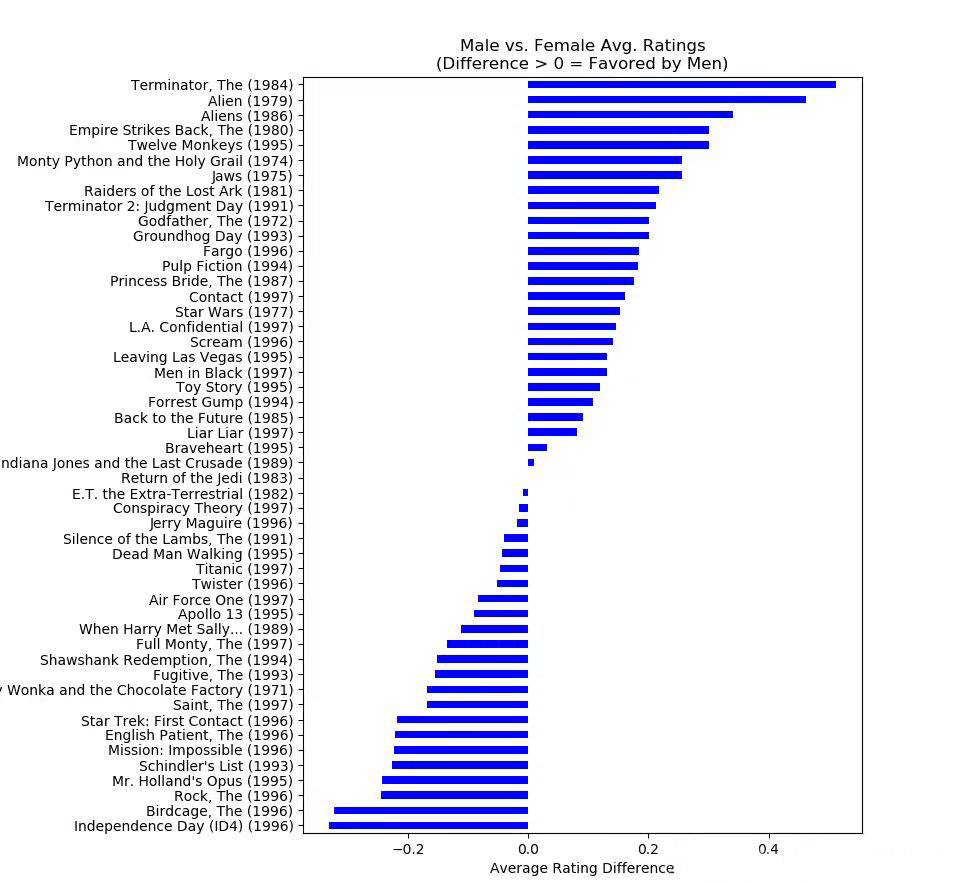
电影评分统计图



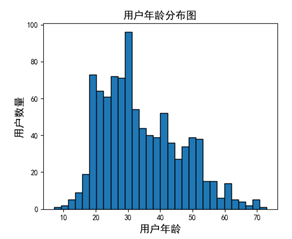
电影年份统计图



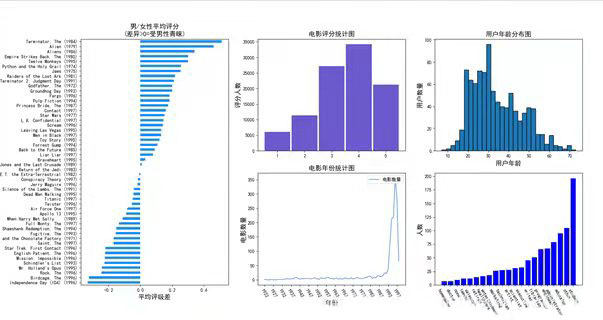
男女性平均评分



用户年龄分布图



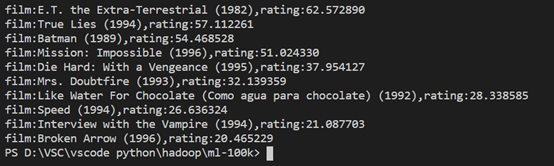
总览图



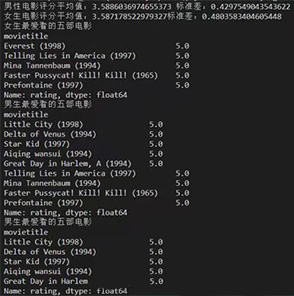
**4.4 电影推荐**

该部分，我们的算法是基于协同过滤算法实现的。

算法运行截图：



推荐结果：

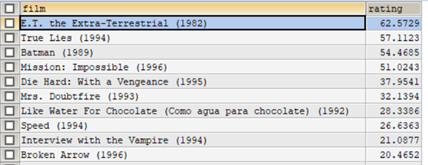


**4.5 将推荐结果存入数据库以便查阅**

实现代码：db = pymysql.connect(host='127.0.0.1', port=3306, user='root', passwd='lzp104613', db='test', charset='utf8')

cursor = db.cursor()

存储结果：



**五、源码下载**

下载地址：https://github.com/NormalLuo/big-data