**天津工业大学**

**计算机科学与软件学院**

**实习报告**

课 程 名 称: 企业实训3

实习企业名称: 东方标准

实 训 题 目: 网站视频管理及视频

推荐系统架构的实现

专 业 班 级: 软件Z1802

项目分组情况：

|  |  |  |
| --- | --- | --- |
|  | **学号** | **姓名** |
| **组长** | 1811650424 | 许光皓 |
| **组员1** | 1811650536 | 杨晓庆 |
| **组员2** | 1811650423 | 顼古月 |
| **组员3** | 1811650631 | 刘欣妍 |
| **组员4** | 1811650634 | 石晓霏 |
| **组员5** | 1811650524 | 杨昊霖 |

指导教师签名： 刘洋 2021年 03月 17 日

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **实训时间** | | | **2020/10/24-2021/3/19** | |
| **实训地点** | **教学B511** | | **实训学时** | **180** |
| **实训目的** | | 1. **本实训要求学生掌握大数据环境下的推荐系统的架构搭建和实现，要求学生对不同 行业里大数据推荐系统的业务逻辑有所了解。** 2. **大数据环境配置能力上有所提高，能够熟练完成大数据环境配置。**   **3、能够理解在大数据背景下的推荐系统架构的的主流技术** | | |
| **实训内容** | | **本实训项目采用离线模式推荐架构 整个过程分成以下几个步骤：**   1. **数据来源** 2. **数据采集** 3. **数据汇集** 4. **数据预处理** 5. **数据仓库技术** 6. **ETL** 7. **推荐引擎**   **8、可视化显示** | | |
| **实训分工安排** | | **组长1人：项目组长负责整体项目的组织与安排工作，包括时间 控制，资源分配调整，任务的划分与分配以及项目部分代码的编写。在校外实习阶段负责协调组内工作交互，把控项目进度。**  **组员4-5人：技术骨干应能按照项目制定的标准进行开发与测试， 完成系统核心代码的编写与整体项目的测试，最终实现完整的系统。项目组成员应按照组长制定的计划进行相关模块的 设计与编码，协助组长及技术骨干完成整体项目的编 码与测试。** | | |
| **实训要求** | | **根据实训课题，以组为单位完成项目所规定的相关功能，进行项目答辩，提交实训相关材料。** | | |
| **实训主要设施设备** | | **PC机 台/人** | | |
| **实训成果** | | **完成网站视频管理及视频推荐系统架构的实现，完成英文文章单词统计功能** | | |

**1 引言**

1.1系统概述

本项目利用视频网站的数据，进行数据预处理和数据清洗，使用推荐引擎实现视频智能推荐功能的实现。

对于英文文章进行数据处理，实现统计英文文章单词的功能。

最后对视频数据，英文文章单词统计，视频推荐进行可视化展示。

1.2 需求分析

如今对于视频的应用量越来越大，尤其是对于短视频和电视剧电影等不同视频的数目也越来越多，对于数据库的要求也越来越高，需要进行改善管理。而对于不同的视频，人们往往总是面临着选择，从而会浪费许多时间。而不同人的喜好也大不相同，所以针对不同人群和不同个体的不同喜好进行视频推荐。

对视频进行统计与推荐，能够采用更直观的方式向用户呈现最贴近用户喜爱度的视频，有利于提升用户使用体验，从而使用户在众多视频网站中更愿意选择本视频网站。

视频统计紧跟实时热度与历史口碑。既可以为用户提供最新的影视咨询与影视资源，还可以为用户提供经久不衰的优秀作品。可以在与其他视频网站的竞争中，更加吸引用户。

视频推荐基于用户自身的历史记录和所喜爱的影视作品，为用户推荐同类型，同导演或同演员的作品，努力贴近用户的实际需求，预测用户可能会喜欢的作品。视频推荐可以帮助用户在观看完一部作品后尽快找到下一步喜爱的作品，有利于保持用户的长期使用。

1.3 运行环境

硬件环境：磁盘容量 大于等于 500G 2 个 4 核/8 核 CPU 16-64GB

内存软件环境：Python3、Xshell6、Xftp6、HDFS、MapReduce、CentOS7、MySQL5、Eclipse、 Tomcat7、JDK7、Maven、ActiveMQ，IE浏览器

**2 概要设计**

2.1 设计思路

后端：读取视频数据，推荐结果数据，英文文章单词统计结果数据。

前端：视频数据可视化展示，英文文章单词统计结果展示，视频推荐可视化展示。

数据库：Hive操作，英文文章单词统计操作，视频推荐操作。

2.2 功能设计分工

配置Hadoop及相应Hive版本：全员

数据清洗：杨晓庆，顼古月，杨昊霖

数据统计：顼古月，刘欣妍，石晓霏

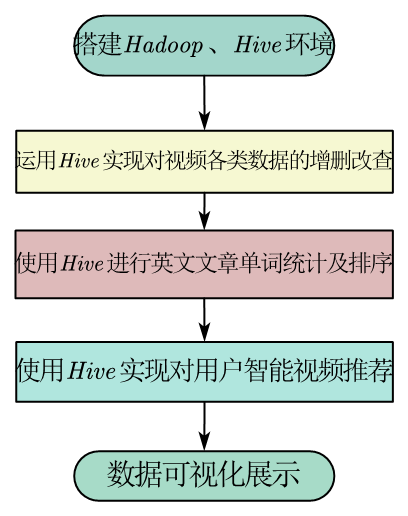
英文文章单词统计：刘欣妍，顼古月

智能推荐操作：许光皓，杨昊霖

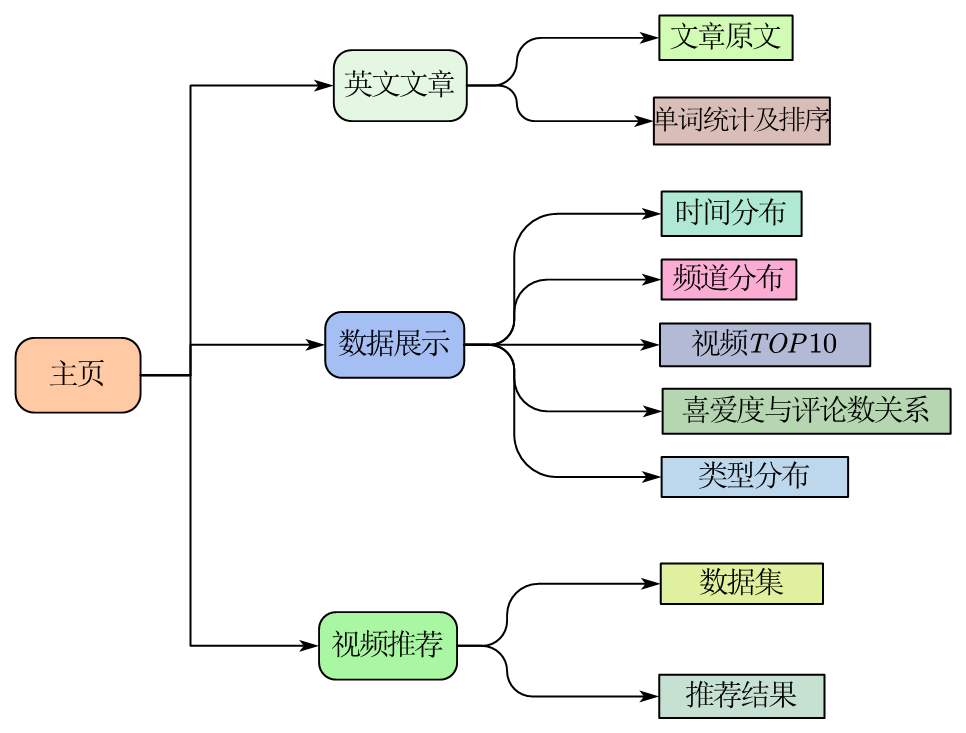
可视化展示：石晓霏，许光皓，杨晓庆

项目整合及展示：许光皓

2.3 模块结构图



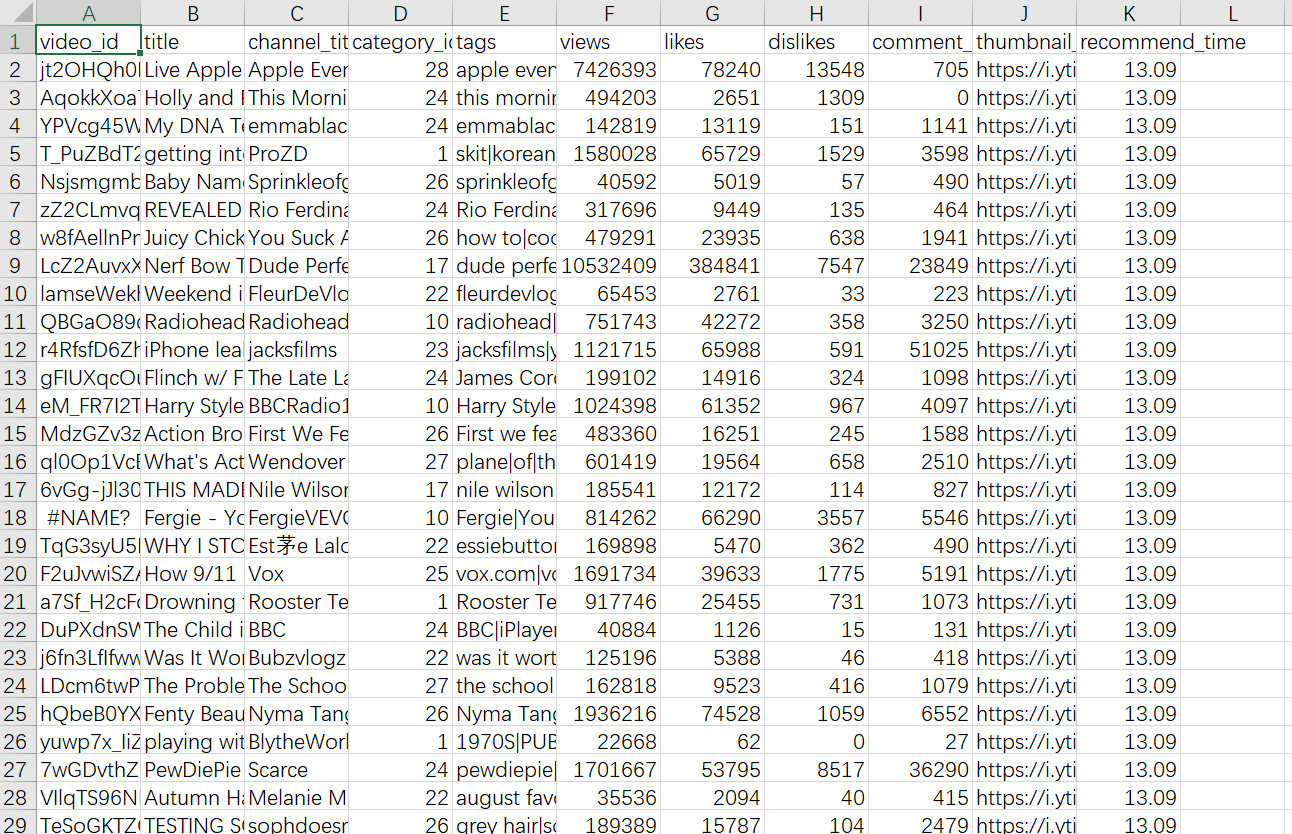
2.4 程序流程图



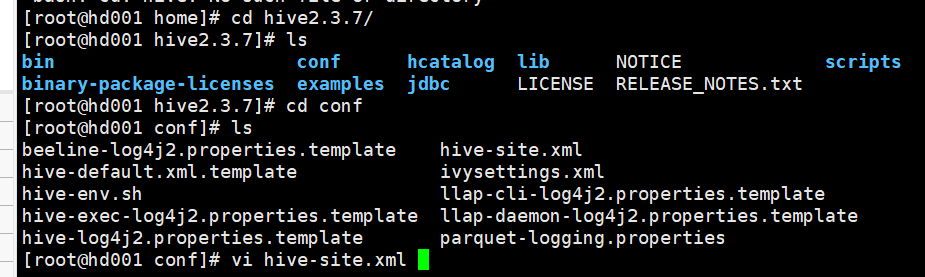
2.5 个人模块功能介绍

本人负责数据搜集整理，hive的jdbc连接操作，以及数据的分析统计

数据集 来源：<https://www.kaggle.com/datasnaek/youtube-new#USvideos.csv>

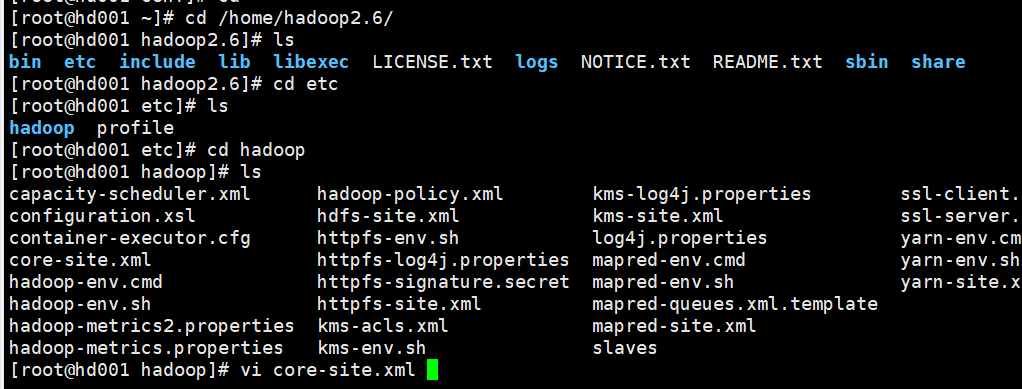


hive的jdbc操作过程及截图：

进入hive-site.xml配置文件中

添加如下代码



进入core-site.xml配置文件中

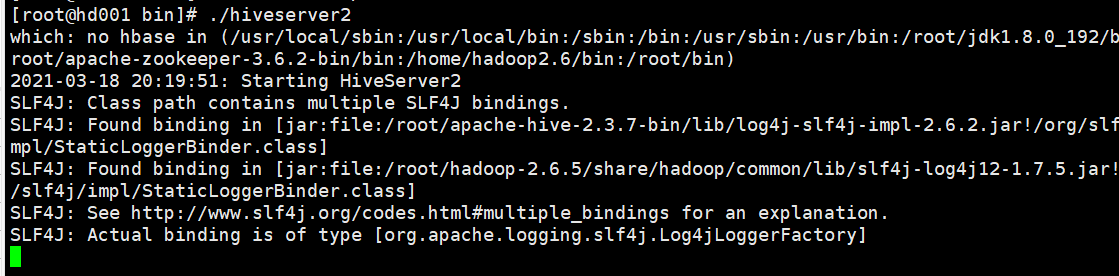
添加如下代码



将此配置文件发送到另外两台机器中去

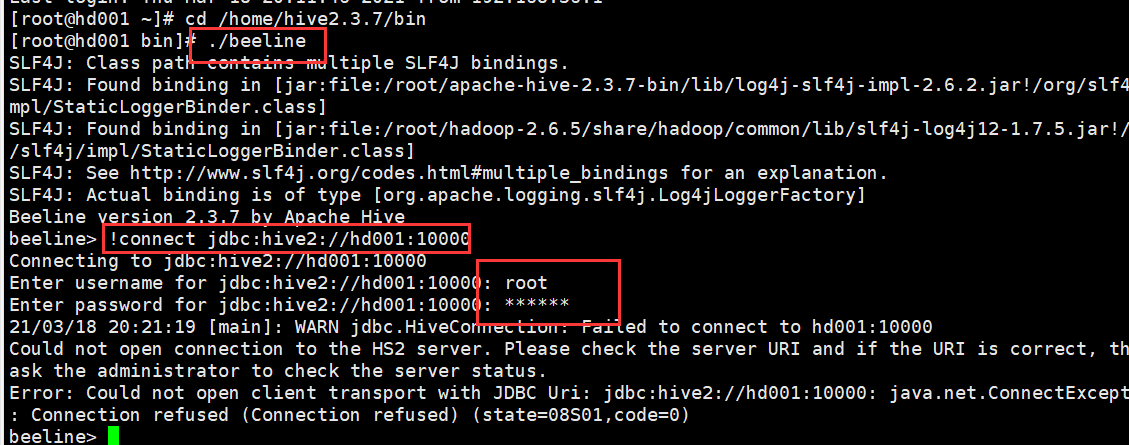
之后启动hadoop

在hive的bin目录下启动./hiveserver2

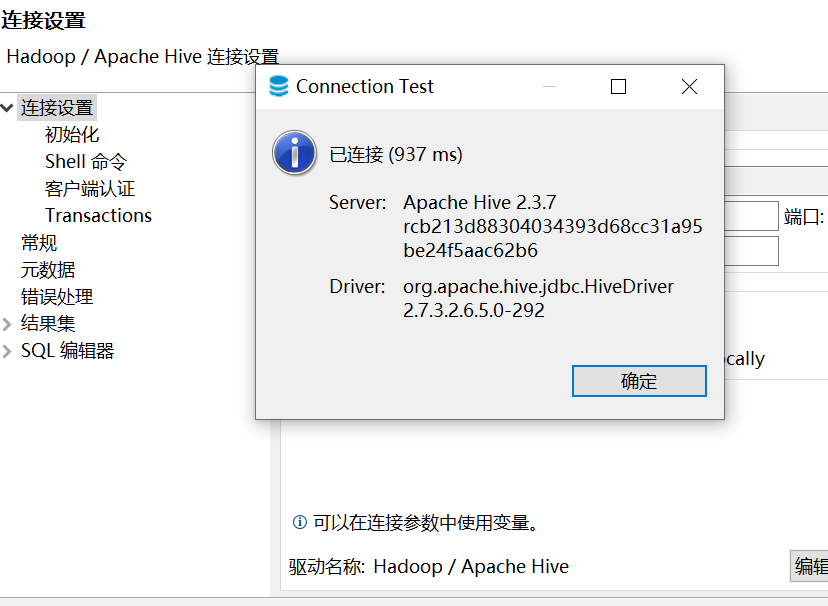


在另一个终端中启动./beeline

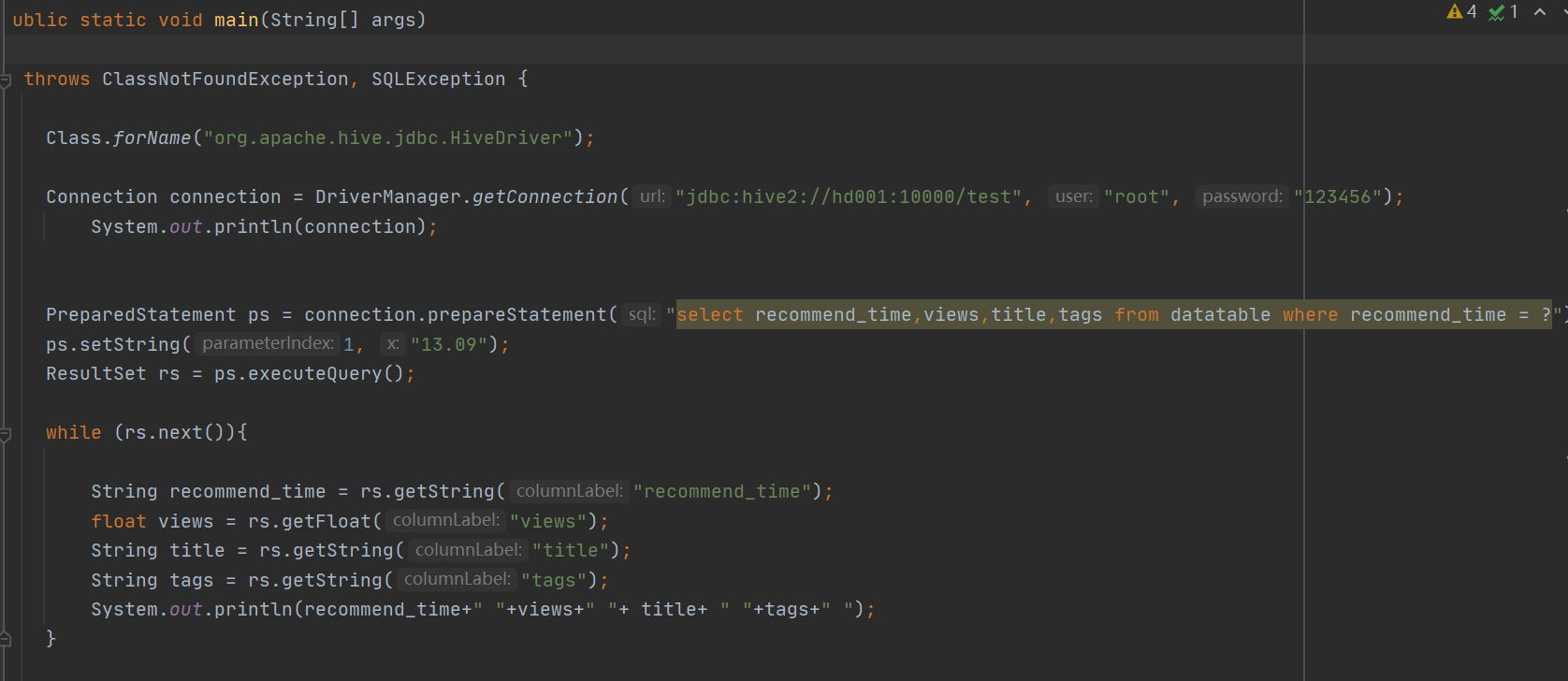
输入!connect jdbc:hive2://hd001:10000

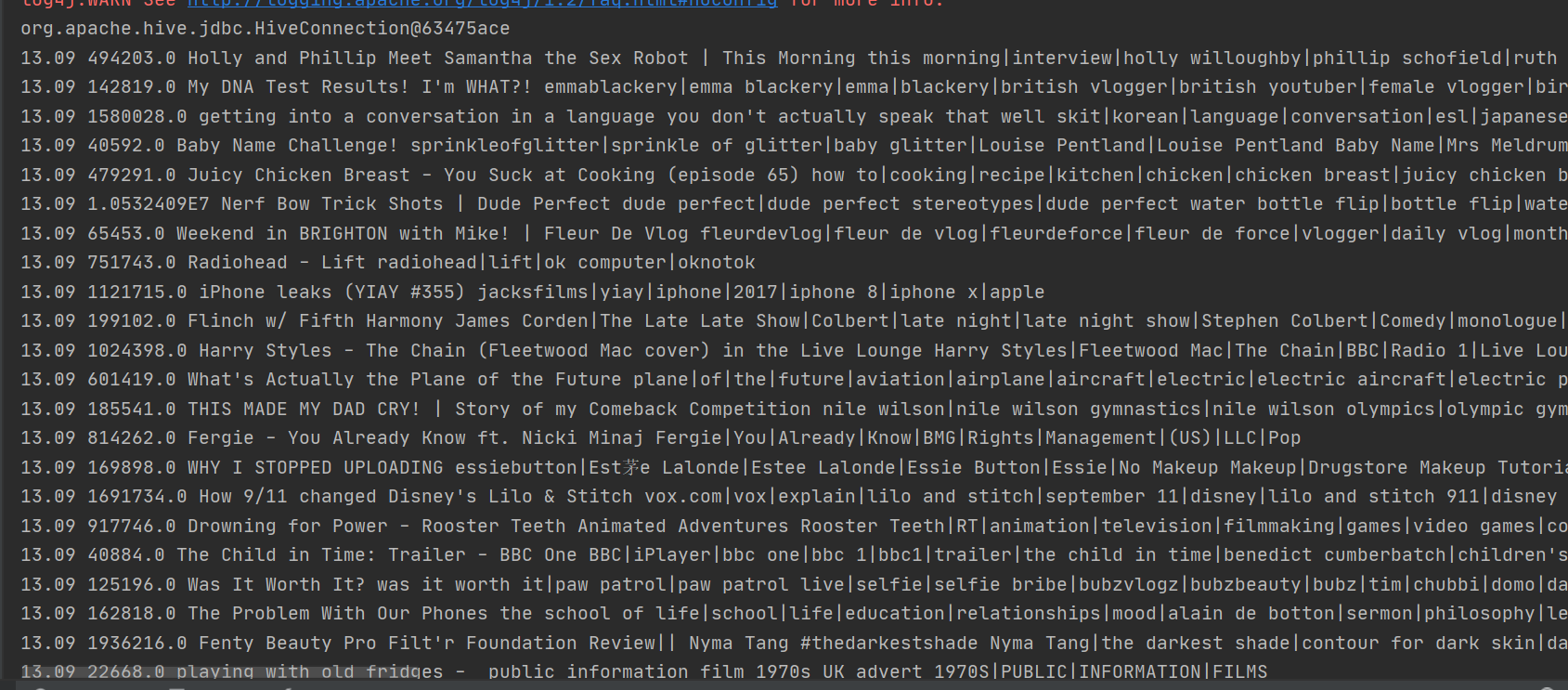


使用dbever进行hive连接测试



Idea中的代码及截图如下



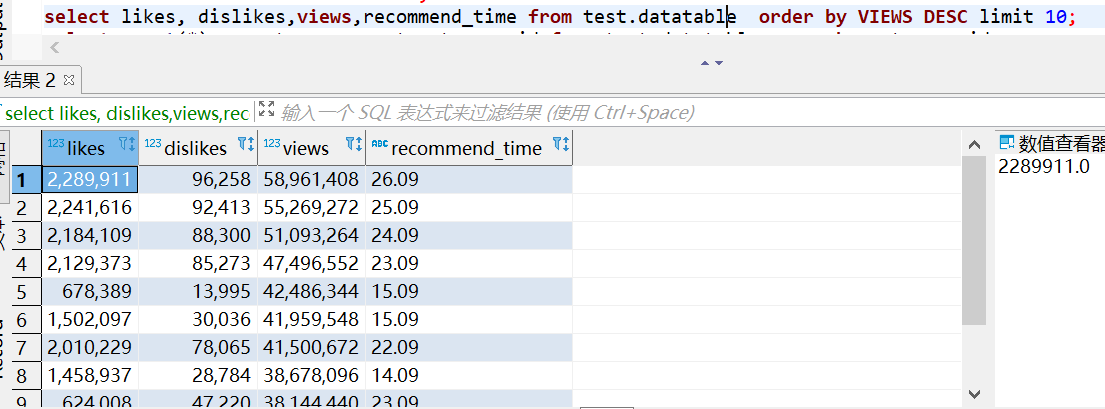


至此，完成了jdcb操作

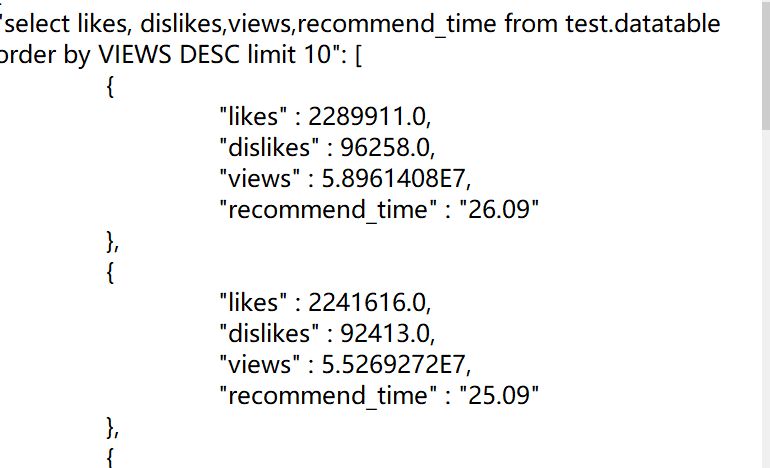
数据分析操作：

使用dbeaver工具，对hive仓库中的数据进行相关查询

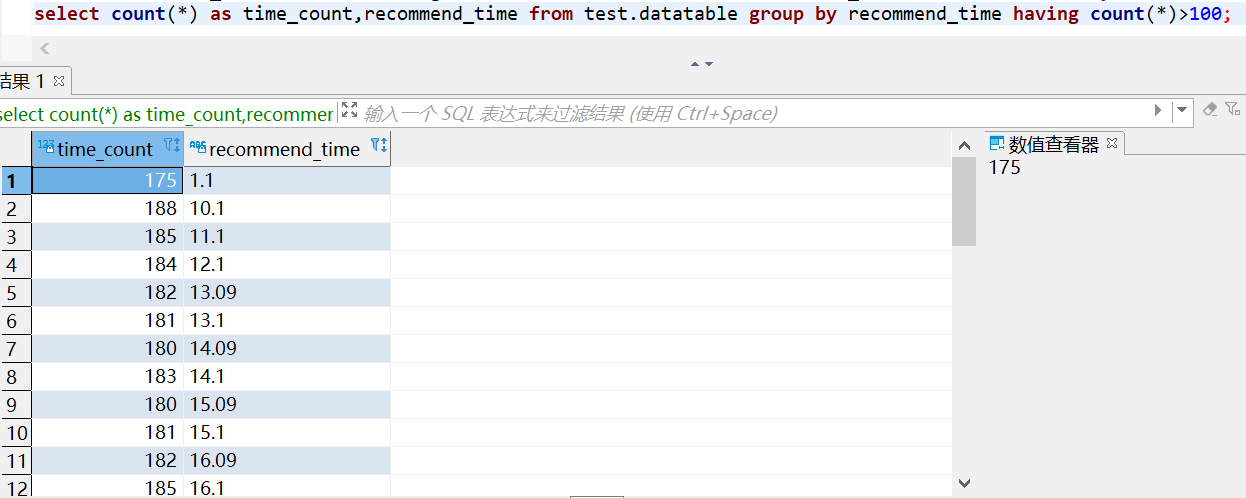
喜欢与不喜欢人数占比查询 对应数据可视化页面中间的图



导出的json文件



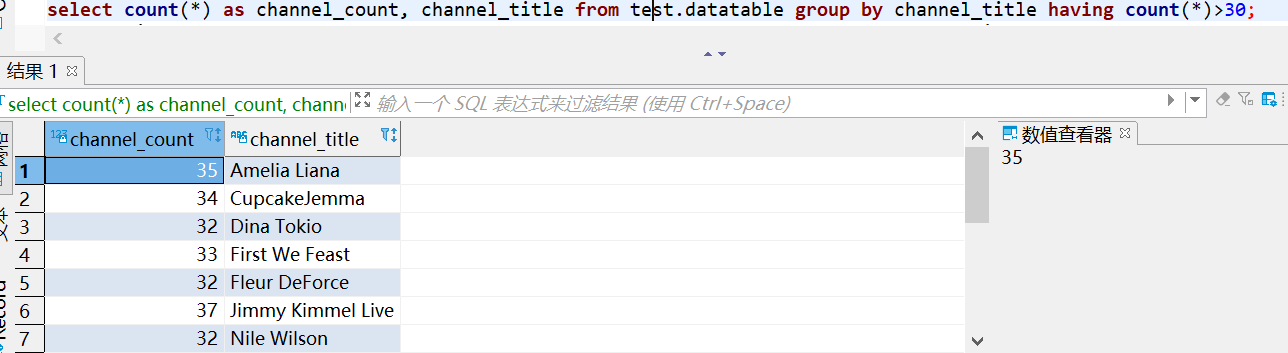
各日期视频数量查询 对应数据可视化页面左上的图



导出的json文件



各频道视频的数量占比对应数据可视化页面左下的图

导出的json文件如下



Js部分代码截图





**3 详细设计**

3.1 前端功能

前端将实现数据可视化展示，英文文章统计结果展示，视频推荐可视化展示。主要分为三个页面，每个页面选取不同角度对相关数据进行可视化展示。

涉及技术包括：HTML5、CSS3、Js、jQuery、Ajax。

3.2 后端功能

后台将实现读取视频数据，视频推荐数据，英文文章单词统计结果数据。

涉及技术包括：JAVA。

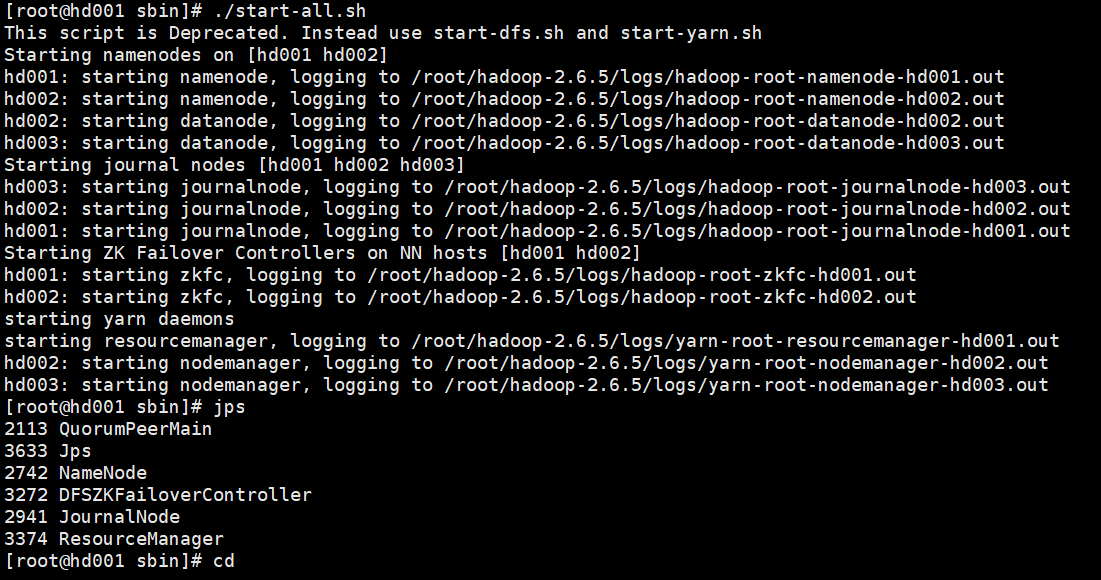
3.3 数据库

数据库实现使用基于 Hadoop 之上的 Hive 将规整数据映射成表，在 Hive 进行数据查询写 sql 导出结果分析出推荐数据写入到推荐结果中。

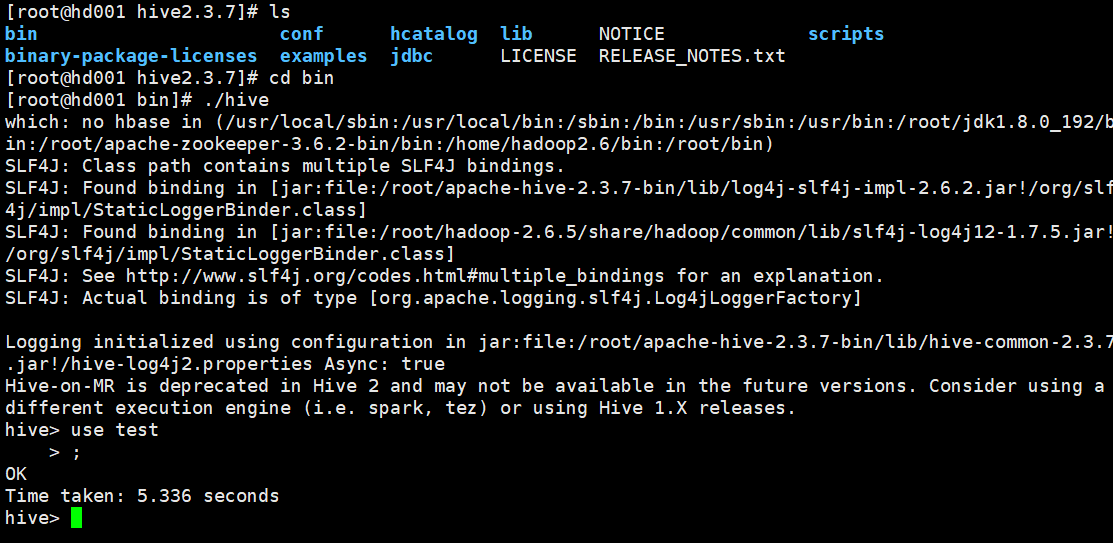
涉及技术包括：Linux，Hadoop,Hive,Hbase

**4 实现方案**

**4.1 配置环境及启动**

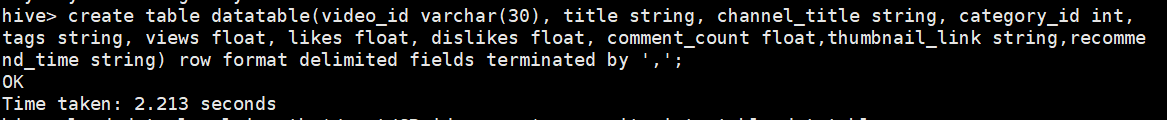


Hadoop配置启动

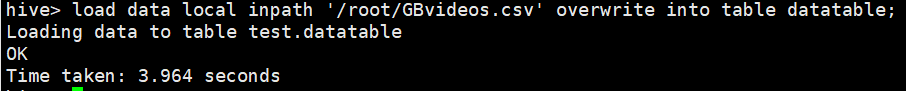


Hive配置启动

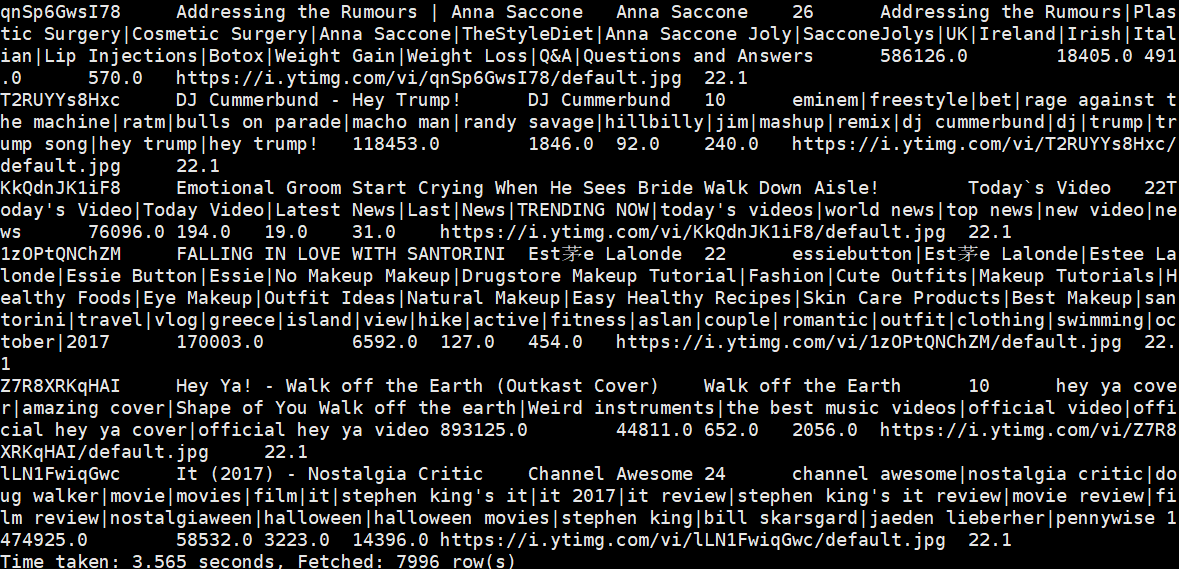
**4.2 使用Hive实现对视频各类数据的增删改查**



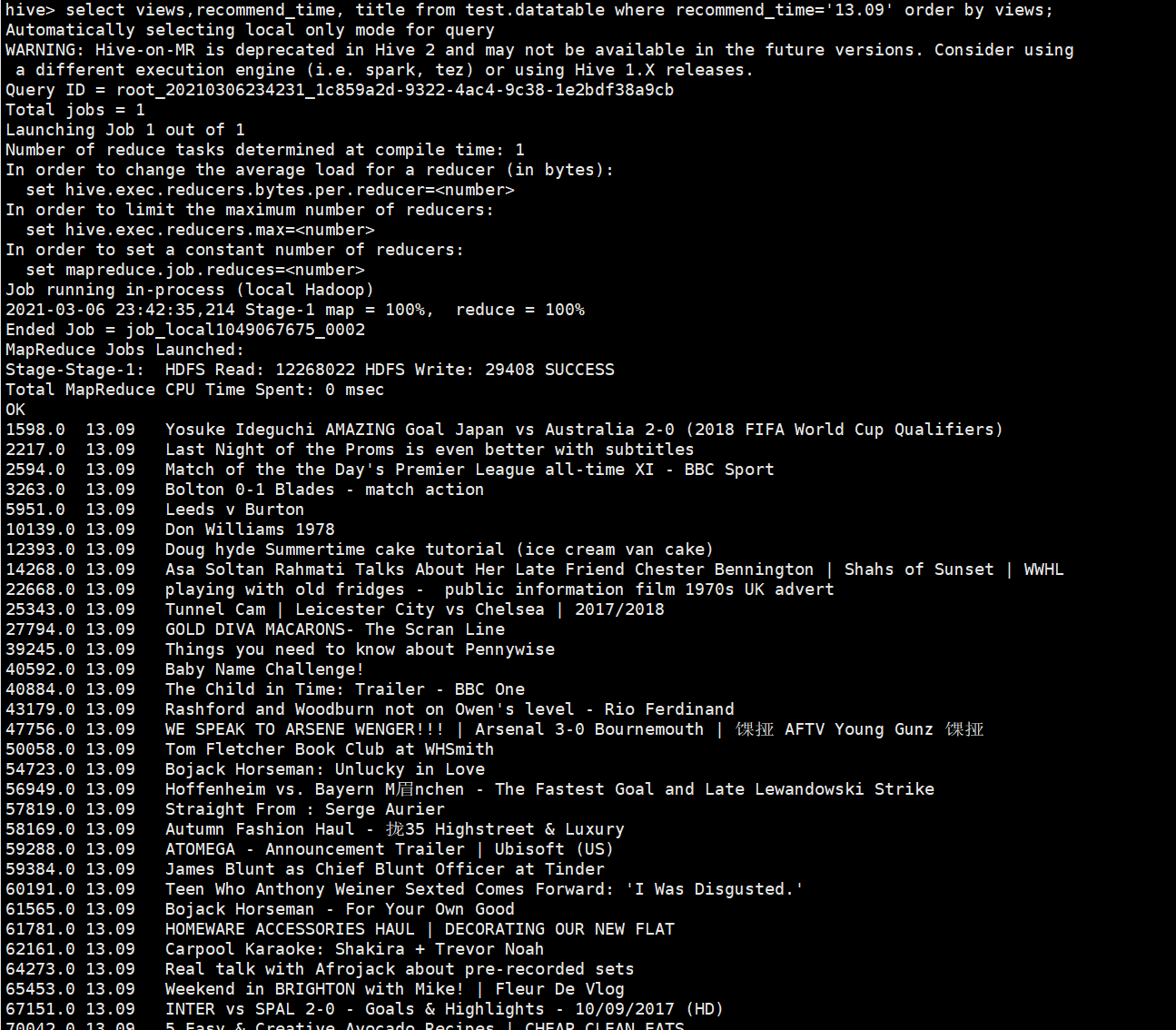
使用Hive实现数据表的创建、删除、修改操作



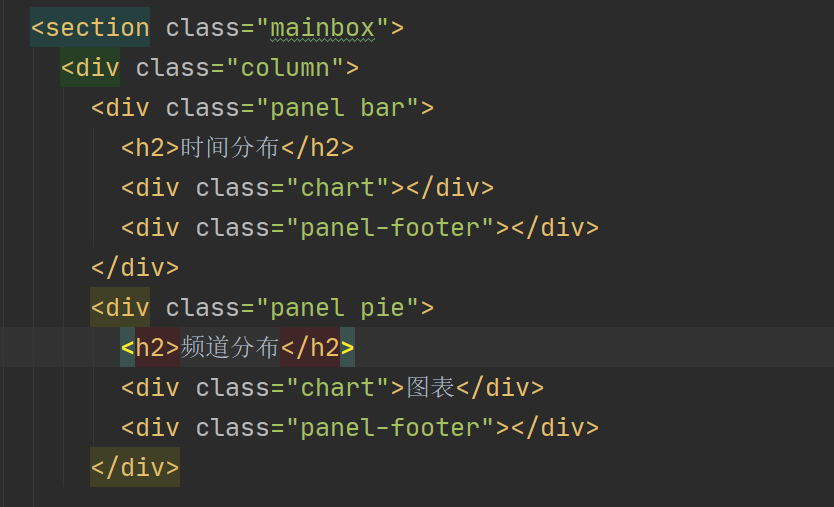
插入数据集中的数据



将数据集中的数据select查询



按照播放量降序排列查询某一天的相关数据



视频数据可视化代码



视频数据可视化代码



视频数据可视化代码

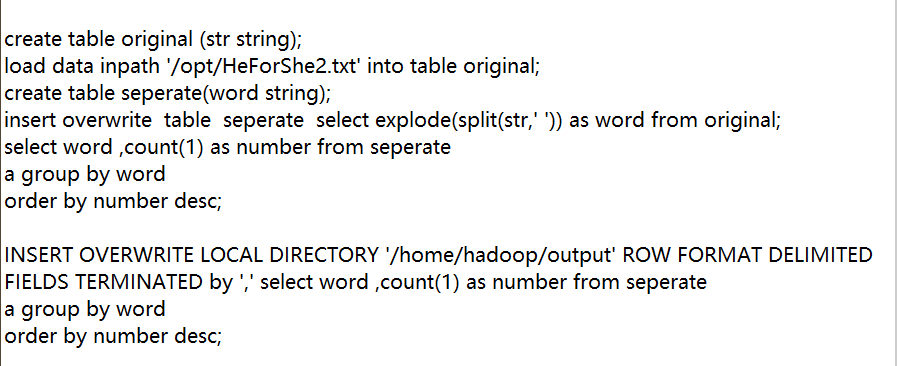


视频数据可视化代码

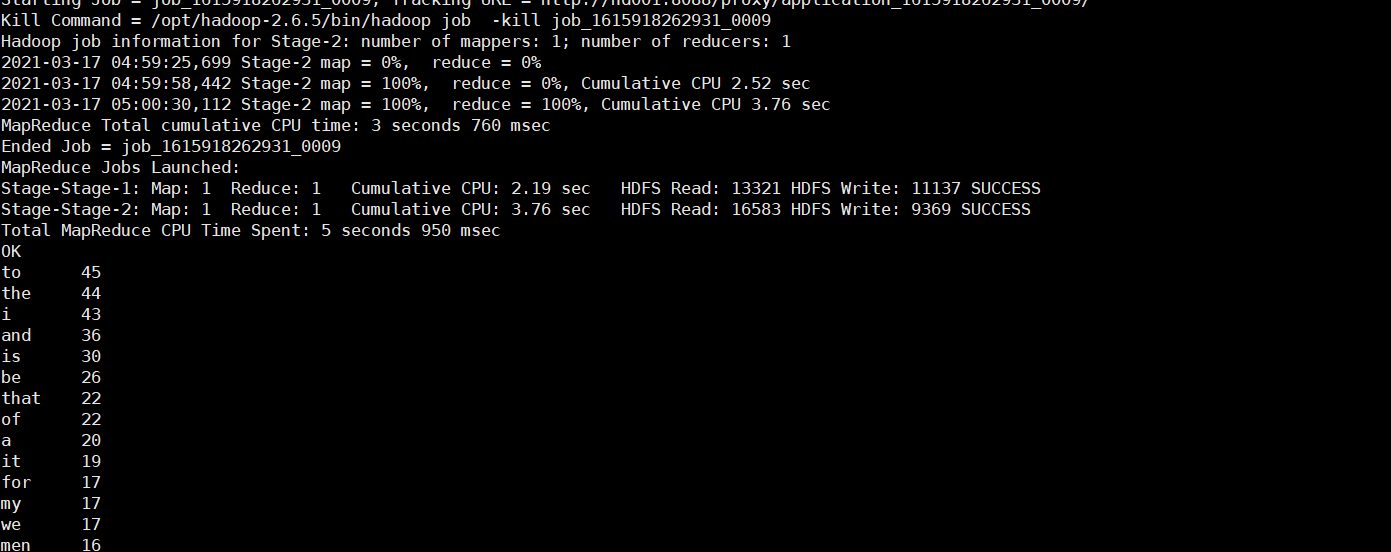


视频数据可视化展示

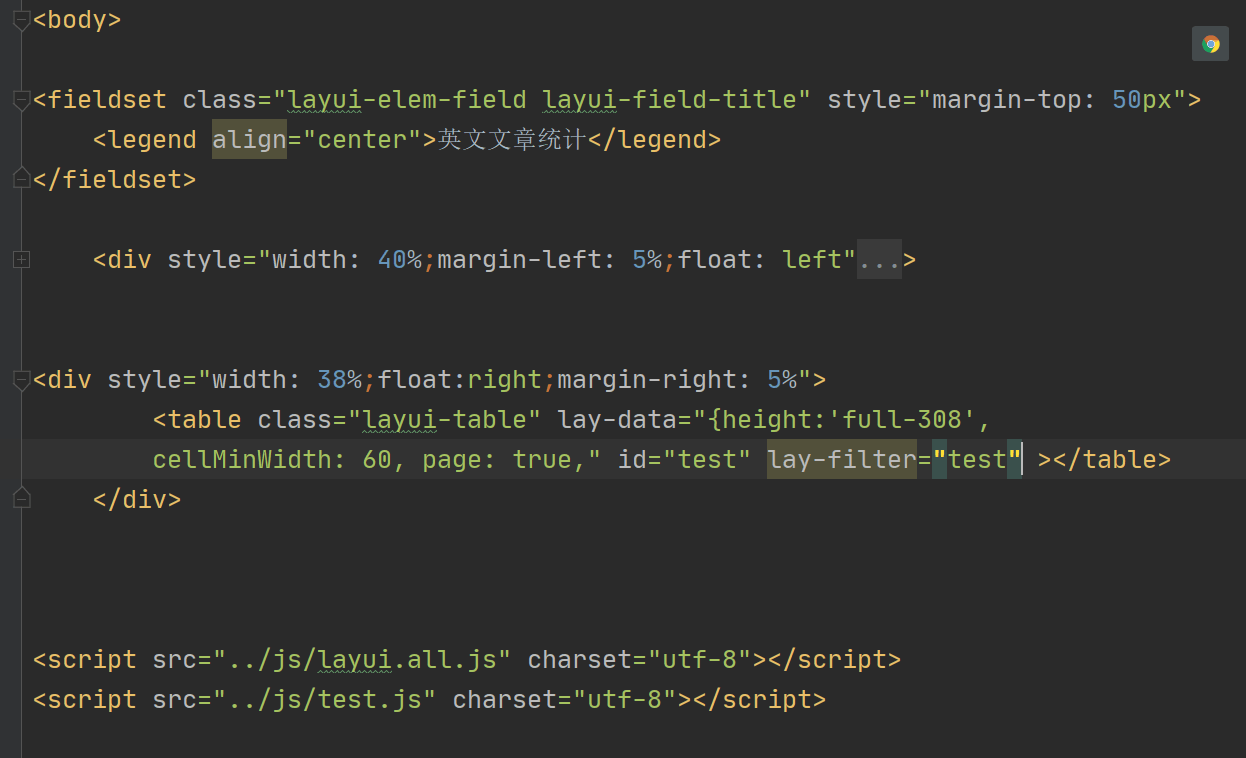
**4.3 使用Hive实现英文文章的单词统计**



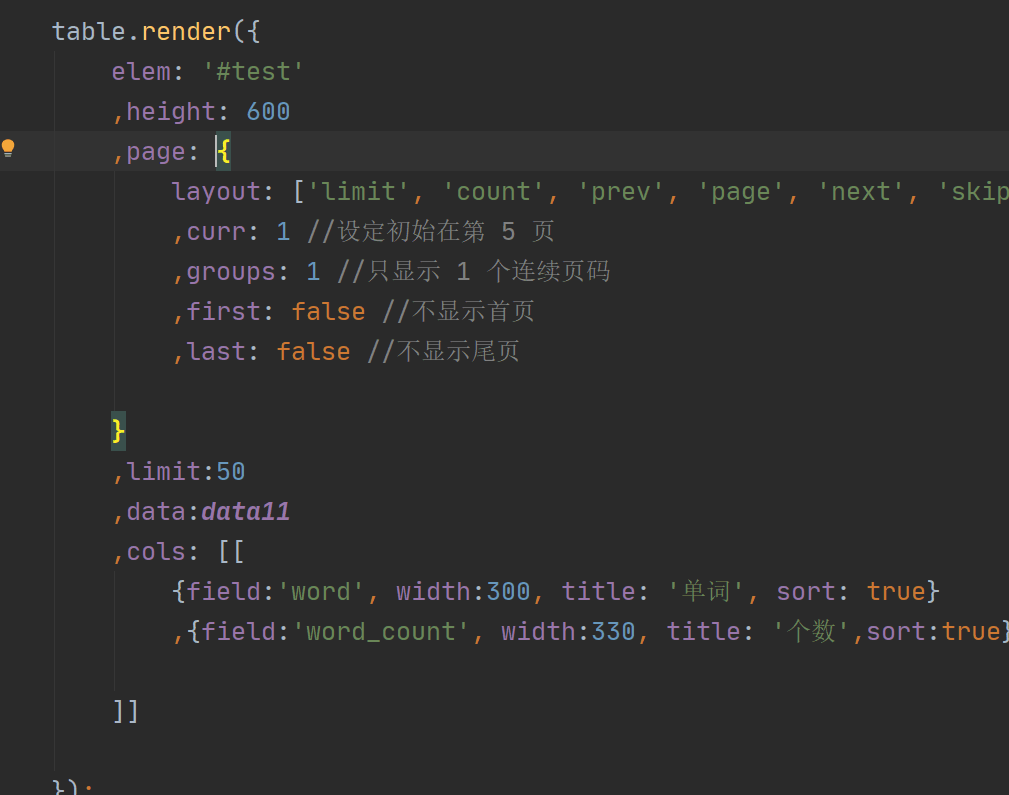
Hive代码整合



Hive结果截图



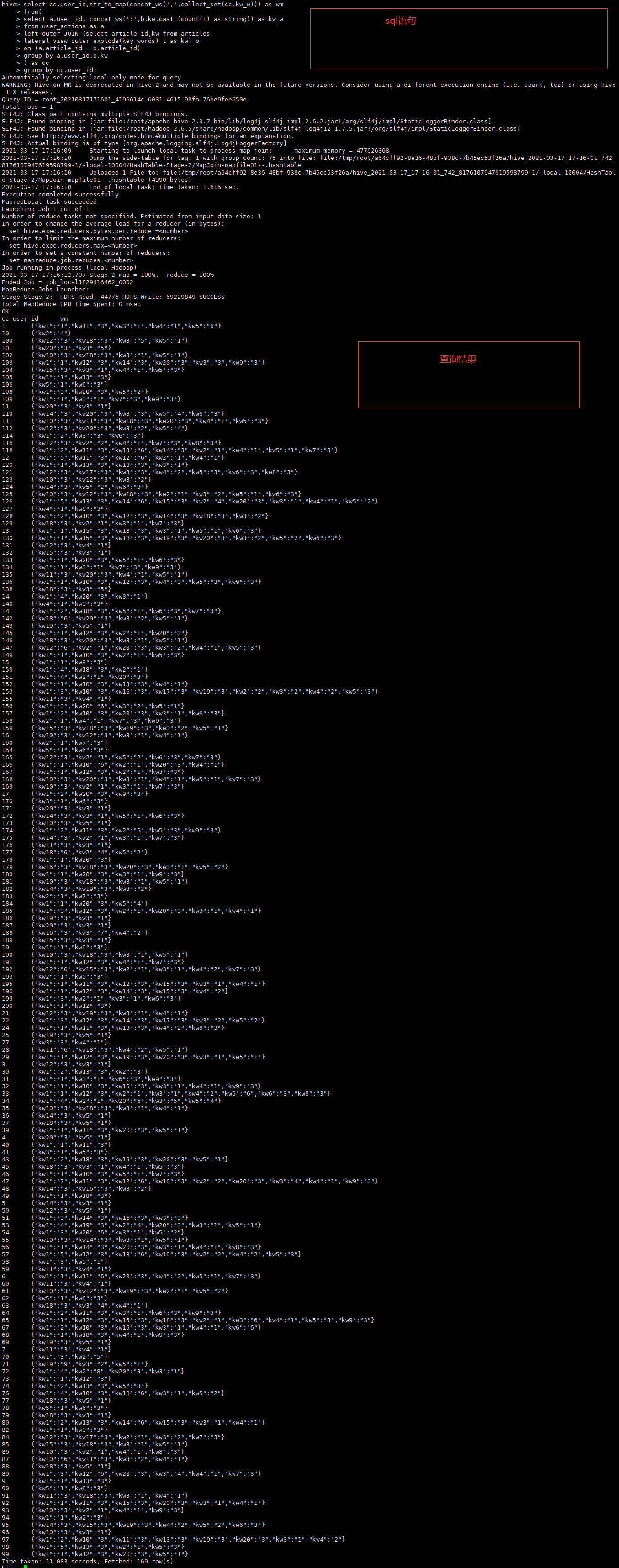
英文文章单词统计可视化代码



英文文章单词统计可视化代码

英文文章单词统计可视化展示

**4.4使用Hive实现对用户智能视频推荐**



视频推荐可视化展示



**5 测试**

开启服务器的情况下，使用浏览器作为测试，开启检查功能。如有错误，检查代码，设置断点逐步推查。

**心得与体会**

这次的大数据实训将要结束，在为期一个月的实训中，我从大数据小白到现在的了解并应用了部分大数据工具进行数据分析，收获了很多知识，对于环境的搭建，对于linux终端的命令操作，以及对于hive数据库的管理，对数据集中的数据进行分析，这都是未来成为一个数据分析师所必须要做且必须要熟练的工作。一遍遍的搭建环境，一次次的调试程序中，一个个问题进行百度，虽然有很多次都快要消磨耗尽我的耐心，但每次当问题解决，其中的成就感与满足感都是我进行下一步的动力。