**河南科技学院2021~2022学年第二学期**

**《云计算》设计报告**

**题目：基于hadoop的电影可视化系统**

|  |  |
| --- | --- |
| **学 号：** | **20202214230 20202214227**  **20202214219 20202214222** |
|  |  |
| **姓 名：** | XXX XXX  黄启帆 XXX |
|  |  |
| **专 业：** | 人工智能 |
|  |  |
| **学 院：** | 人工智能学院 |

**摘 要**

**随着互联网与移动互联网迅速蒲剧，网上电影娱乐信息数量相当庞大，人们对获取感兴趣的电影娱乐信息的需求越来越大，对电影口碑的要求也逐渐升高。如何给用户可靠的电影推荐结果，并保证用户能满意推荐系统推荐的电影是系统需要满足的一个重要条件。本文电影评分预测系统使用hadoop与centos7搭建的虚拟机系统，mapreducer处理网络爬虫获取的豆瓣热播电影，并对电影进行排序以及可视化。本系统最主要是对未上映的电影进行评分预测，给予消费者更多的权益，使消费者获得更好的观影感受，商家也可以根据消费者喜好合理安排电影场次。该系统设计的预测未上映电影评分与上映后的评分进行比较，分数比较接近，电影评分预测系统设计成功。**

**关键词: 电影，Hadoop,爬虫**

目 录

[1 绪论 1](#_Toc123656719)

[1.1目的和意义 1](#_Toc123656720)

[1.2研究现状 1](#_Toc123656721)

[1.3设计目标 2](#_Toc123656722)

[2 项目分析 2](#_Toc123656723)

[2.1 需求分析 2](#_Toc123656724)

[2.2 可行性分析 2](#_Toc123656725)

[2.2.1市场可行性 2](#_Toc123656726)

[2.2.2 技术可行性 3](#_Toc123656727)

[2.3受众分析 3](#_Toc123656728)

[2.4功能分析 3](#_Toc123656729)

[2.4.1电影信息查询： 3](#_Toc123656730)

[2.4.2用户偏好推荐 3](#_Toc123656731)

[2.4.3用户登录功能 3](#_Toc123656732)

[3 环境准备 3](#_Toc123656733)

[3.1 软件安装 3](#_Toc123656734)

[3.2 环境搭建 4](#_Toc123656735)

[3.2.1 mapreduce搭建 4](#_Toc123656736)

[3.2.2 hive搭建 5](#_Toc123656737)

[3.2.3 mapper-reducer编程搭建 6](#_Toc123656738)

[4 数据集 8](#_Toc123656739)

[5 工作原理 10](#_Toc123656740)

[5.1 执行过程 10](#_Toc123656741)

[5.1.1 数据采集(pachong.py)、预处理 10](#_Toc123656742)

[5.1.2 数据统计及可视化 11](#_Toc123656743)

[5.1.3 大数据分析 15](#_Toc123656744)

[5.1.4 Hive数仓分析 18](#_Toc123656745)

[5.1.5 影片类型与票房统计图 20](#_Toc123656746)

[5.1.6 导演与影片类型关系图 20](#_Toc123656747)

[5.1.7电影票房预测 21](#_Toc123656748)

[5.1.8 电影评分预测scorepredict.py 21](#_Toc123656749)

[5.2 基本原理 22](#_Toc123656750)

[6 设计实现 23](#_Toc123656751)

[7 结论 25](#_Toc123656752)

[参考文献 27](#_Toc123656753)

1 绪论

* 1. 目的和意义

如今，电影已经成为人们日常生活中最普遍的一种娱乐消遣方式，对电影的评头论足也成为了普通大众进入讨论公共社会话题的方便之门。电影评分是电影质量的一个直观反映，对未上映的电影评分进行预测是非常重要的[1]。电影预测评分因其便捷、高效，成为大部分观众评价电影的首选方式，也是电影口碑最直观的体现，评分高低直接影响到大众的观影取向和消费选择。与文字、视频相比，大数据搜集、统计的评分信息可以让我们对观众的观影体验，审美趣味、情感态度一目了然。比如烂番茄、豆瓣等公开评分。视频网站、电影门户网站等的后台会根据观众的评分、打星和消费行为来推断用户的审美趣味和消费需求，从而为洪湖提供个性化推荐和引导。还有一些猫眼电影、百度糯米、淘票票等平台，则利用大数据的整合与分析能力，将线上的口碑评价、信息分享与线下的消费、购买行为直接关联。

电影作为文化产业的中流砥柱，带来了巨大的经济效益和社会效益，而电影票房是衡量电影经济效益最重要的指标。中国电影市场正进入飞速发展的阶段，年全国电影票房收入接近亿元，中国电影市场己经成为全球票房收入增长的主引擎。由于电影具有高投入高风险的特性，提前预测电影的评分是确保电影发行回报、控制发行风险的手段，对于电影的投资决策具有重要参考价值。预测电影评分也可影响大众的观影取向和消费选择，为消费者选择高质量电影提供了便利。此外，这也是一种可以帮助消费者在是否观看此电影的决策过程中提供其他消费者的一种快速有效的方式[2]。

* 1. 研究现状

电影评分是衡量一部电影优劣的重要标准，对于投资商和观影者极具参考价值，因此电影评分的预测成为电影领域的热点[3]。目前，对电影评分的预测研究方法有多种，较为常见的方法有随机森林算法的电影评分预测研究和基于文本矢量特征的电影评分预测模型研究等。随机森林算法于2001年由 Leo Breiman 提出，它首先通过自助法（bootstrap）采样技术，从大小为 N 的原始训练样本集中随机有放回并重复抽取 k 个样本，生成新的训练样本集。其次根据自助样本集生成k个分类树组成随机森林，新数据的分类结果视分类树投票分数而定[4]。黄东晋等[5]提出基于文本矢量特征的预测电影评分模型首先通过收集搜狗的语料库，经过预处理过后用Word2Vcc模型对其进行训练，生成词向量模型。接着对电影文本进行预处理，如分词，去掉停止词等。然后，把预处理后的词导入词向量模型中，分别生成每个词对应的词向量。接着对每个词向量给予相应的权重，这个权重是通过对预处理后的文本用TF-IDF算法计算得到的，表示每个词的重要程度。最后，把给过权重后的词向量以及预处理过后的电影特征信息一起输入到LISTM模型中训练，最终得到预测的得分值。

* 1. 设计目标

该系统是面向广大电影爱好者的，包括学校和社会上的各个阶层的人士，所以在页面设计和布局上一定要贴近群众，符合大众的浏览习惯。系统旨在通过网络爬虫爬取豆瓣电影的热播电影，获取电影名称，关键字，类别，电影评分，电影影评以及影评点赞数等信息，生成xlxs变革文件或者csv表格文件，向用户提供其查询的电影的全面信息，并根据用户的偏好按照电影类别以及关键词向用户推送相关电影的热门影评资讯。与其它电影评分预测系统不同的是，本系统通过对相同类型影片和出演演员受用户的喜爱程度等对即将上映的电影进行评分预测，此功能对于用户可以提前知道电影的评分，根据自己的喜好选择是否去购买电影片去观看影片；对于商家可以按照用户的喜好提前安排票价和场次，获取更多的盈利。该系统适用于用户和商家，是个双赢的电影评分预测系统。

2 项目分析

**2.1 需求分析**

由市场使用情况来看，豆瓣网的影音社区是行业中做的比较好的将影视评分做成几乎成了行业标准。从影评的角度来说，豆瓣网的影评内容发布门槛较低，任何人都可以在豆瓣上发表自己的短评或者长评，形成百花齐放但良莠不齐的影评内容。而我们做这个项目的初衷即将将豆瓣评论中的关键字，点赞数，评论及数量等信息呈现在用户面前，给用户该电影的直观评价。

**2.2可行性分析**

要成功地实现一个项目，首先应该进行功能上的需求分析，这样才能令设计出的项目满足用户的各项功能需求。可行性分析也称为可行性研究，是在项目调查的基础上，针对新项目的开发是否具备必要性和可能性，对新项目的开发从技术、经济、社会的方面进行分析和研究，以避免投资失误，保证新项目的开发成功。

**2.2.1市场可行性**

市场的可行性是项目完成后是否能够在实际中很好的推广。本项目提升了观影感受，拉近了电影和观影人的距离，操作简单、方便。

**2.2.2 技术可行性**

技术上的可行性分析主要分析技术条件能否顺利完成开发工作，硬、软件能否满足开发者的需要等。该项目采用了Browser/Server模式进行开发。.Browser/Server体系结构紧密的结合了Intemnet/ Intranet技术，是技术发展的大势所趋，它把网站带入了一个崭新的发展时代。数据库服务器SQL数据库，它能够处理大量数据，同时保持数据的完整性并提供许多高级管理功能。它的灵活性、安全性和易用性为数据库编程提供了良好的条件。因此，项目的开发平台已成熟可行。

**2.3受众分析**

该项目是面向广大电影爱好者的，包括学校和社会上的各个阶层的人士，所以在页面设计和布局上贴近群众，符合大众的浏览习惯。其次就是项目的压力测试，因为项目面对人员较广，具有不确定性和不稳定性，故考虑到了流里急剧增大对此项目造成的打开页面变慢等诸多问题。

**2.4功能分析**

项目将用户喜好的电影类型与豆瓣影评系统结合，满足用户的需求。该电影数据分析项目包含以下三个功能模块：

**2.4.1电影信息查询：**

用户通过查询电影名称后向用户展示电影的评分关键字与影评以及影评热度等信息。

**2.4.2用户偏好推荐**

用户注册后进入个人定制页界面，选择偏好影片类型与关键词。系统根据用户偏好在首页向用户推荐相关电影资讯影评等。

**2.4.3用户登录功能**

用户通过输入注册的账号密码进行登入操作或通过手机验证码登入系统，完成对目标电影评论信息的查询。

3 环境准备

3.1 软件安装

windows10：Pycharm+Python3+ mysql5.7+VMware16.1.2

虚拟机Centos7：Centos7+java1.8.0+hadoop2.7.7+ hive-2.3.9

3.2 环境搭建

下面主要以虚拟机centos7基于java1.8.0、hadoop2.7.7搭建的伪分布式环境为基础，进行的后续环境搭建展开介绍：

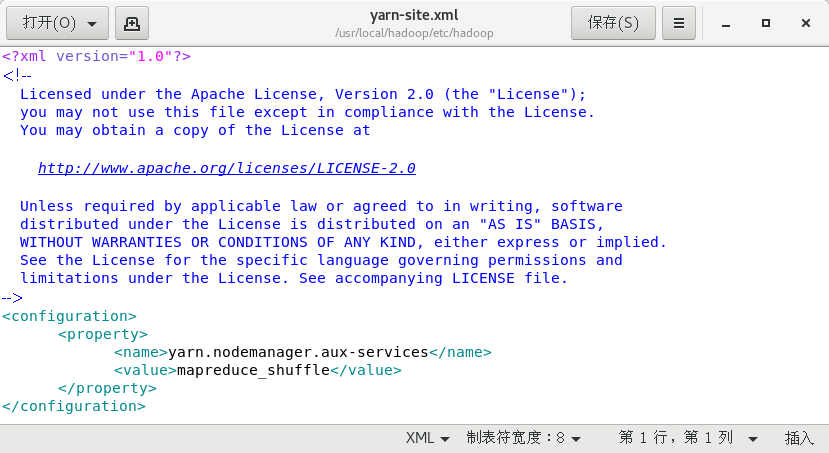
3.2.1 mapreduce搭建

编写mapred-site.xml内容为图3-1所示



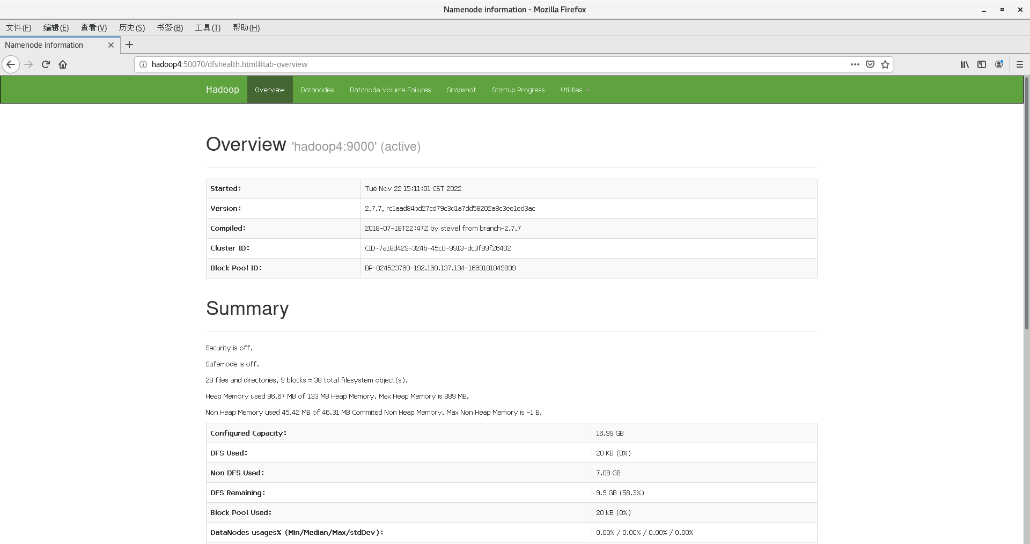
**图3-1 mapred-site.xml**

编写同目录下yarn-site.xml文件内容为图3-2所示

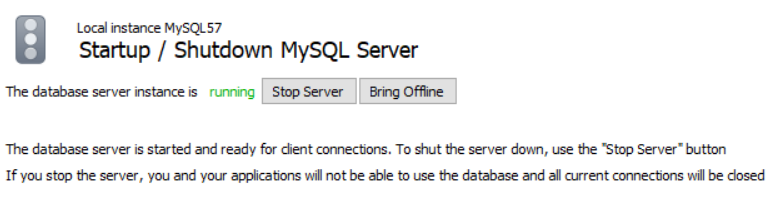


**图3-2 修改mapred-site.xml**

格式化namenode节点即可把配置文件的更改更新到hadoop[6]中。保证本机mysql5.7处于运行状态，启动hadoop集群，确保本地端web可以打开。



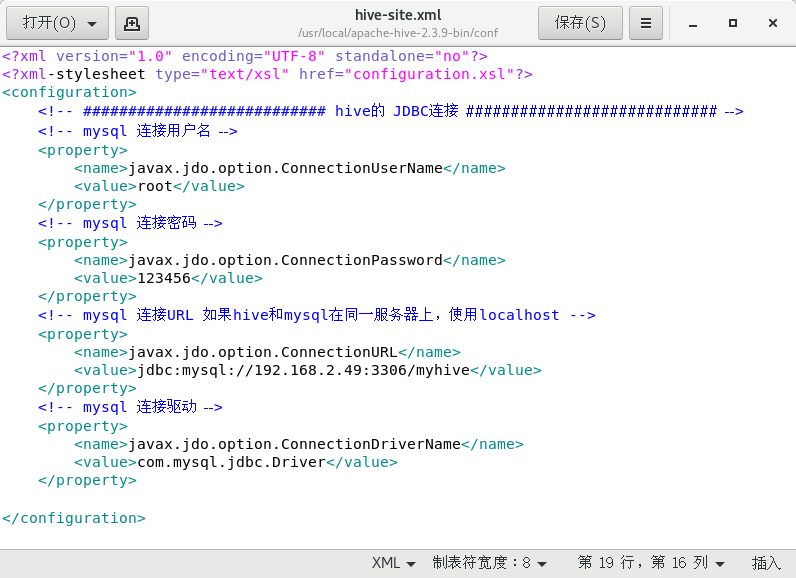
**图3-3 启动集群**



**图3-4 启动mysql数据库**

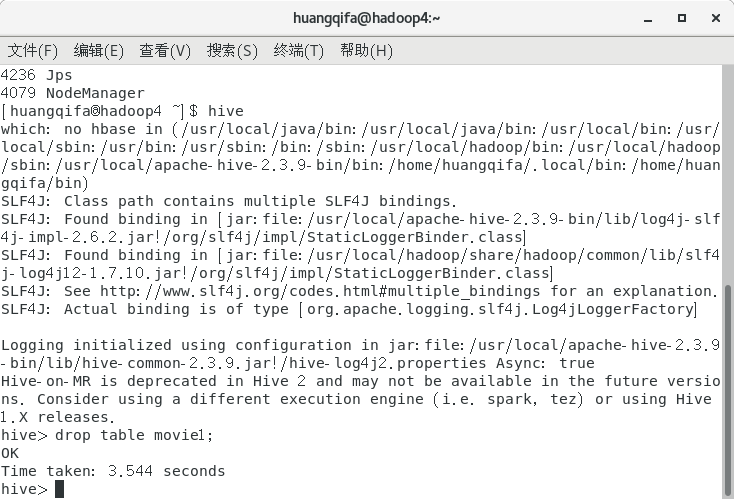
3.2.2 hive搭建

下载hive-2.3.9安装包并解压到/usr/local目录下，在hive[7]安装的conf目录下修改hive-site.xml文件为如下内容



**图3-5 修改hive-site.xml**

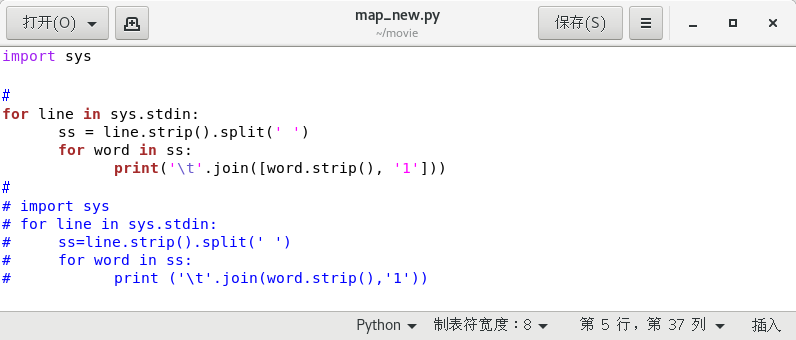
其中‘root’为主机mysql账户，‘123456’为账户密码，‘192.168.2.49’为主机ip，主机端通过命令行创建myhive数据库，并保持数据库正常运行，之后下载mysql驱动jar包，上传jar包到hive的安装的lib目录下，然后centos终端执行./schematool -initSchema -dbType mysql初始化元数据库，最后将hive的环境变量添加到/etc/profile文件，执行source /etc/profile更新环境变量使更改生效，至此hive安装成功。



**图3-6 本地测试hive安装**

3.2.3 mapper-reducer编程搭建

编写map\_new.py、red\_new.py统计词频文件内容如下

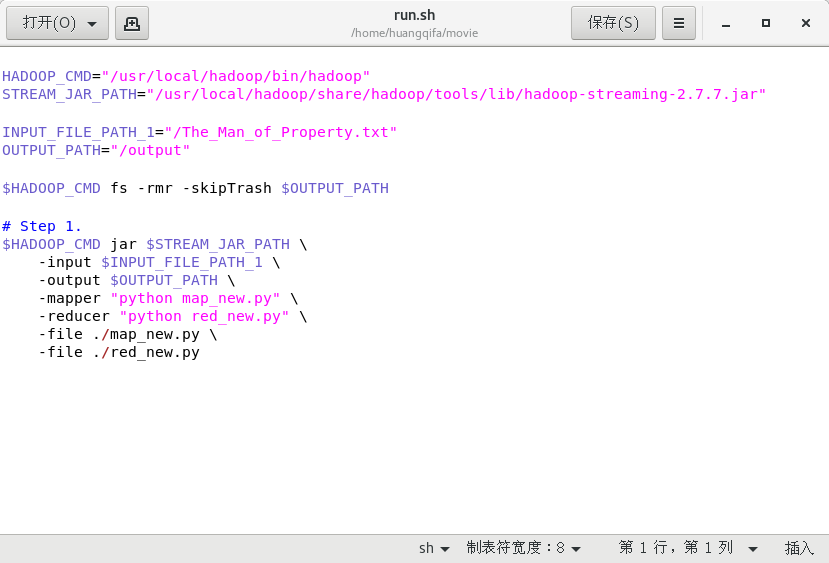


**图3-7 map\_new.py**



**图3-8 red\_new.py**

启动hadoop将The\_Man\_of\_Property.txt上传到hdfs中，编写run.sh将多个命令融合到该文件，一起执行，调用hadoop可执行程序、map\_new.py、red\_new.py文件与jar例程，处理电影字幕文件The\_Man\_of\_Property.txt，程序内容如下，之后赋予文件可执行权限

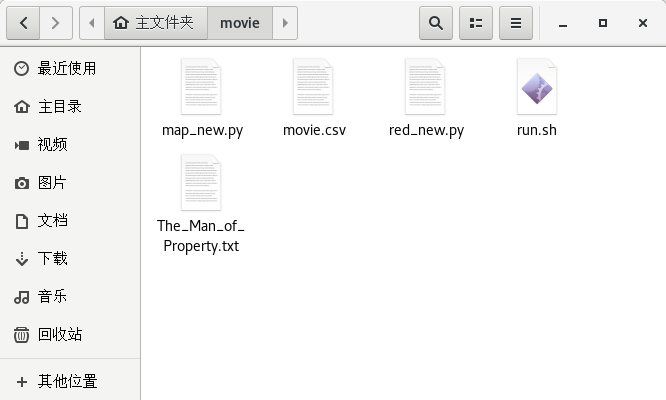


**图3-9 run.sh**

其中HADOOP\_CMD是hadoop可执行文件在系统中的路径，STREAM\_JAR\_PATH表示jar历程文件路径，INPUT\_FILE\_PATH\_1表示输入文件The\_Man\_of\_Property.txt在hdfs中的路径，OUTPUT\_PATH指定输出文件在hdfs中的位置，之后删除输出文件夹内容，如果为空则跳过，最后执行jar命令，进入mapreduce阶段。

hdfs是一个分布式文件系统，使用hdfs作为数据源。文件在hdfs上是以block块为单位进行存储，将block划分成只包含位置信息的数据分片Split。 数据块分好后，MapReduce[8]程序将这些分片以key-value的形式读取出来，进入map阶段， Map函数将电影字幕中的单词以<k1,1>形式输出，k1表示单词。接下来进入reduce阶段，reduce函数对map阶段的输出进行统计，输出<k2,v>,k2表示单词，’v’表示单词数量。之后将统计结果上传到hdfs中。

最终文件结构如图3-10所示



**图3-10 Centos文件结构**

4 数据集

4.1 采集数据

通过编写爬虫程序采集豆瓣电影top250电影信息，采集电影名称、电影简介、电影评分、其他信息、电影连接等字段，抓取电影票房总收入排名情况（取前20），删除冗余和空值字，利用Python的PyMysql库连接本地Mysql数据库并导入movies表（前提安装好mysql数据库），可以将数据保存到本地，从而进行数据可视化展示，也可将数据导入到大数据的Hive数仓工具中，用于大数据分析。

（1）排序:阿拉伯数字格式，例：1，2，3......

（2）影片名称:主要为中文电影，格式为汉语中文。

（3）类型:分为动作、动画、科幻、喜剧、剧情五类。

（4）总票房：以万为单位。

（5）场均人次：以实际为准。

（6）上映日期：格式为20xx/x/xx。

采集数据展示如图4-1所示



**图4-1采集数据展示**

4.2 影片短评信息，以《囧妈》为例。

（1）点赞数:以实际为准。

（2）评级：分为力荐、推荐、还行、较差、很差五类。

（3）用户名：有中文和英文两种格式。

（4）用户链接：用户在豆瓣网的链接。

（5）时间：格式为20xx-xx-xx.

（6）评论：以读者主观为准。

影片短评信息展示如图4-2所示



**图4-2影片短评信息展示**

4.3电影数据

（1）排序:阿拉伯数字格式，例：1，2，3......

（2）电影名称:格式为汉语中文。

（3）评分：整型、最多一位小数的浮点型。

（4）评论数：以实际为准。

（5）类型:分为动作、动画、科幻、喜剧、剧情五类。可多选

（6）地区：该影片上映的地区。

（7）上映年份：以影片实际上映年份为准。

（8）总分：以影片实际获得总分为准。

电影数据展示如图4-3所示



**图4-3电影数据展示**

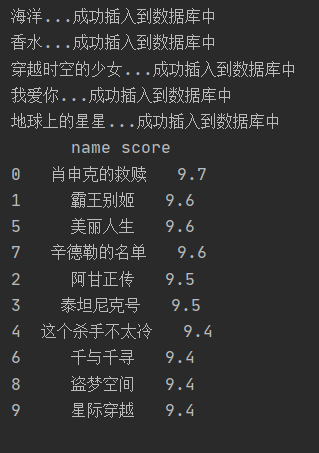
5 工作原理

5.1 执行过程

5.1.1 数据采集(pachong.py)、预处理

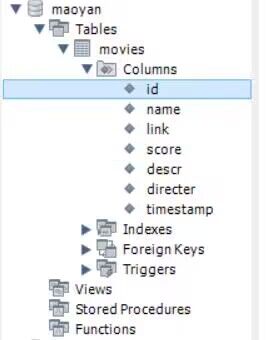
采集豆瓣电影top250电影信息，采集电影名称、电影简介、电影评分、其他信息、电影连接等字段，抓取电影票房总收入排名情况（取前20），删除冗余和空值字，利用Python的PyMysql库连接本地Mysql数据库并导入movies表，可以将数据保存到本地，从而进行数据可视化展示，也可将数据导入到大数据的Hive数仓工具中，用于大数据分析。

采集数据展示：



**图5-1 采集数据展示**

maoyan数据库movie表结构：



**图5-2 maoyan数据库movie表结构**

5.1.2 数据统计及可视化

数据可视化能使数据更加直观，更有利于分析，可以说可视化技术是数据分析与挖掘最重要的内容。Matplotlib作为基于Python语言的开源项目，旨在为Python提供一个数据绘图包，实现专业丰富的绘图功能。

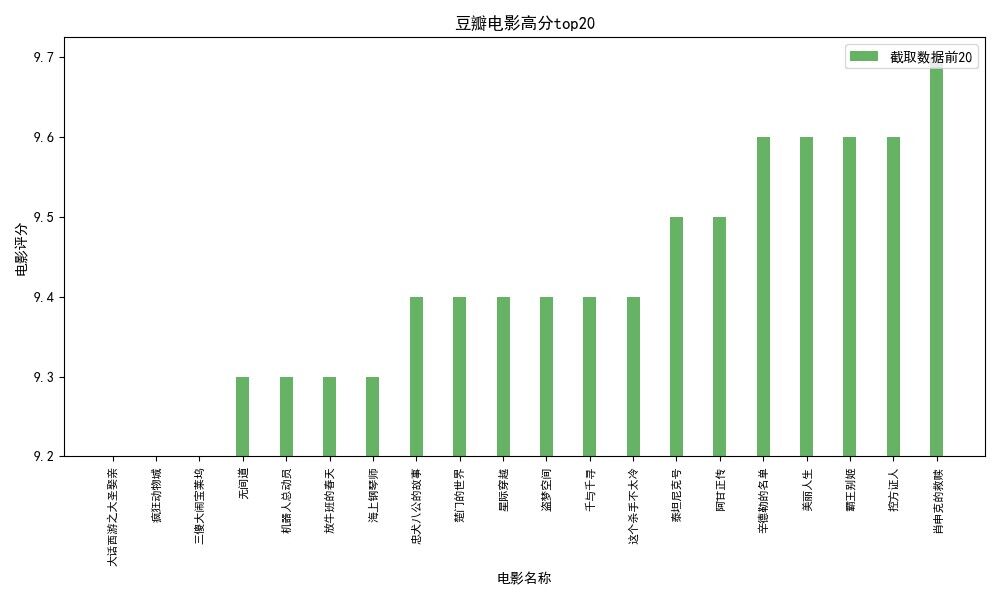
1. 电影票房排名：

利用Python中Matplotlib绘图库及Pandas中的pd.readcsv()方法读取Excel电影数据文件，读取每列数据进行，设置参数将结果绘制成折线图。

1. 电影评分排名douanscore.py

利用Python的Request、Beautifulsoup库进行爬虫，模拟请求获取网页数据，结合正则表达式匹配提取数据，并将豆瓣电影top250电影数据存储到Mysql数据库中,通过数据库语句使用order by实现电影top评分统计。

利用数据库语句统计movies进行电影评分top20，并将结果通过Python的Matplotlib库进行数据可视化，绘图结果如下：



**图5-3 绘图结果**

movie表结构：



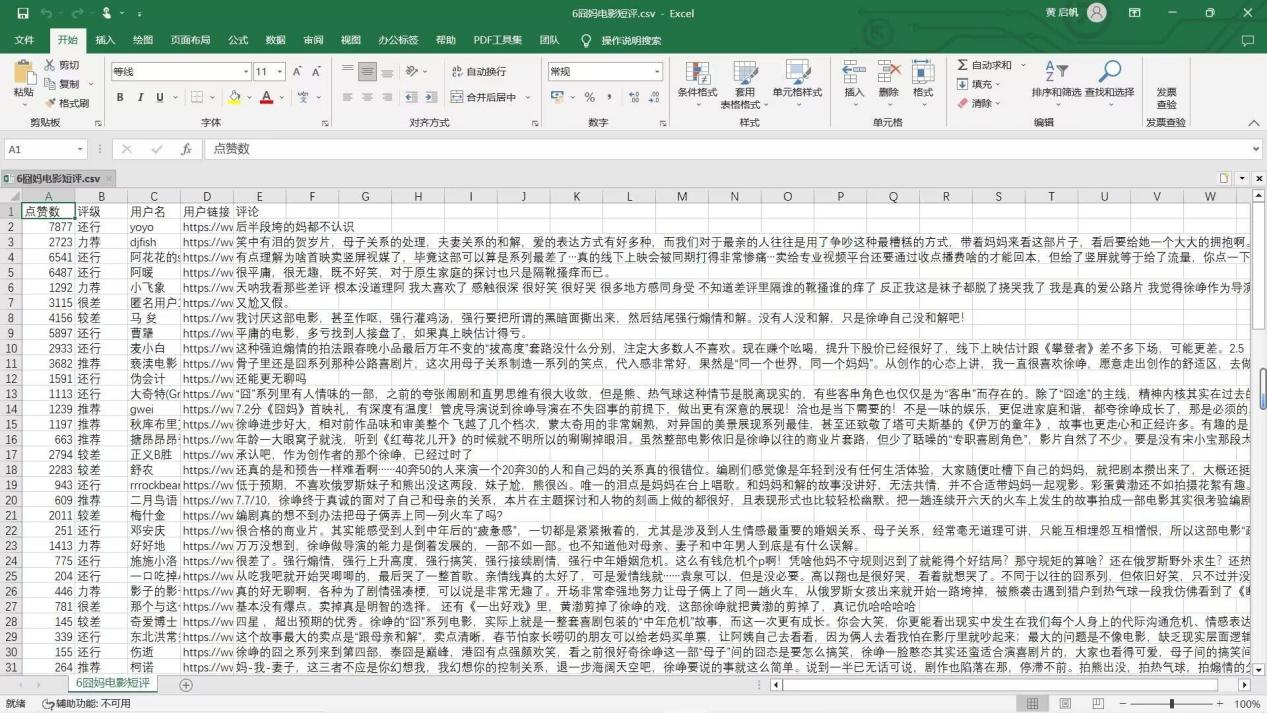
**图5-4 movie表结构**

1. Echarts最近上映电影

Echarts 主要用于数据可视化展示，是一个开源的JavaScript库，兼容现有绝大部分浏览器。在Python中，Echarts被包装成数据可视化工具库Pyecharts。它提供直观、丰富、可个性化定制的数据可视化图表，包括常规的折线图、柱状图、散点图、饼图等。

1. 影片《囧妈》短评信息

大年初一电影《囧妈》网络首映映，截止目前其豆瓣电影评分6.0分，通过电影《囧妈》的豆瓣热门短评进行案例分析，以八爪鱼软件为数据采集工具进行数据爬虫，采集字段有用户名、评级、点赞数和评论内容等信息，利用正则表达式匹配字段标签，根据豆瓣电影提供的评级星数系统显示力荐、推荐、还行、较差、很差等五个评级，满分为五星



**图5-5 短评信息**

1. 词图云cituyun.py

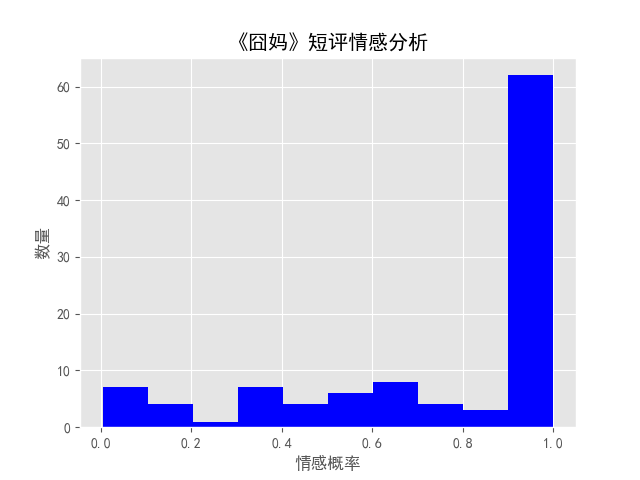
利用Python的jieba分词工具以及wordcloud库实现词云展示，截取电影《囧妈》评论一列，按照规定的停用词切割每行语句，实现分词功能。通过词图云展示可以直观地看出用户对电影的态度情况，数据展示结果如下：



**图5-6 词云**

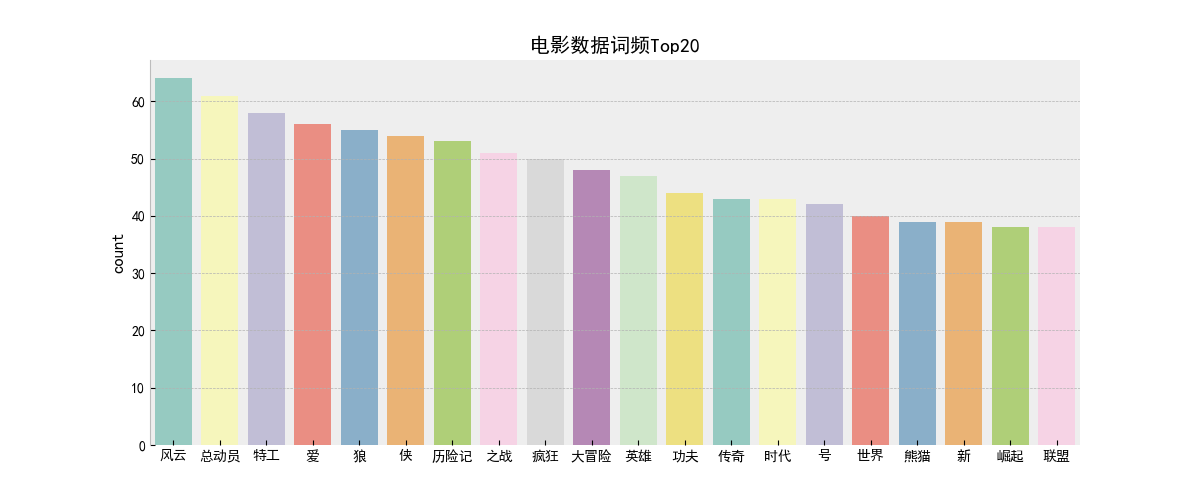
1. 情感分析emotion.py

运用Python机器学习中的情感分析库Snownlp和绘图库pyplot挖掘囧妈短评数据，做出情感分析并展示。利用Snownlp中s.sentiments方法计算情感分数，分数在0.5以上判断为是好评，可以看到电影的评论差距明显。情感分析效果图如下：



**图5-7 情感分析效果图**

1. python词频统计wordcount.py



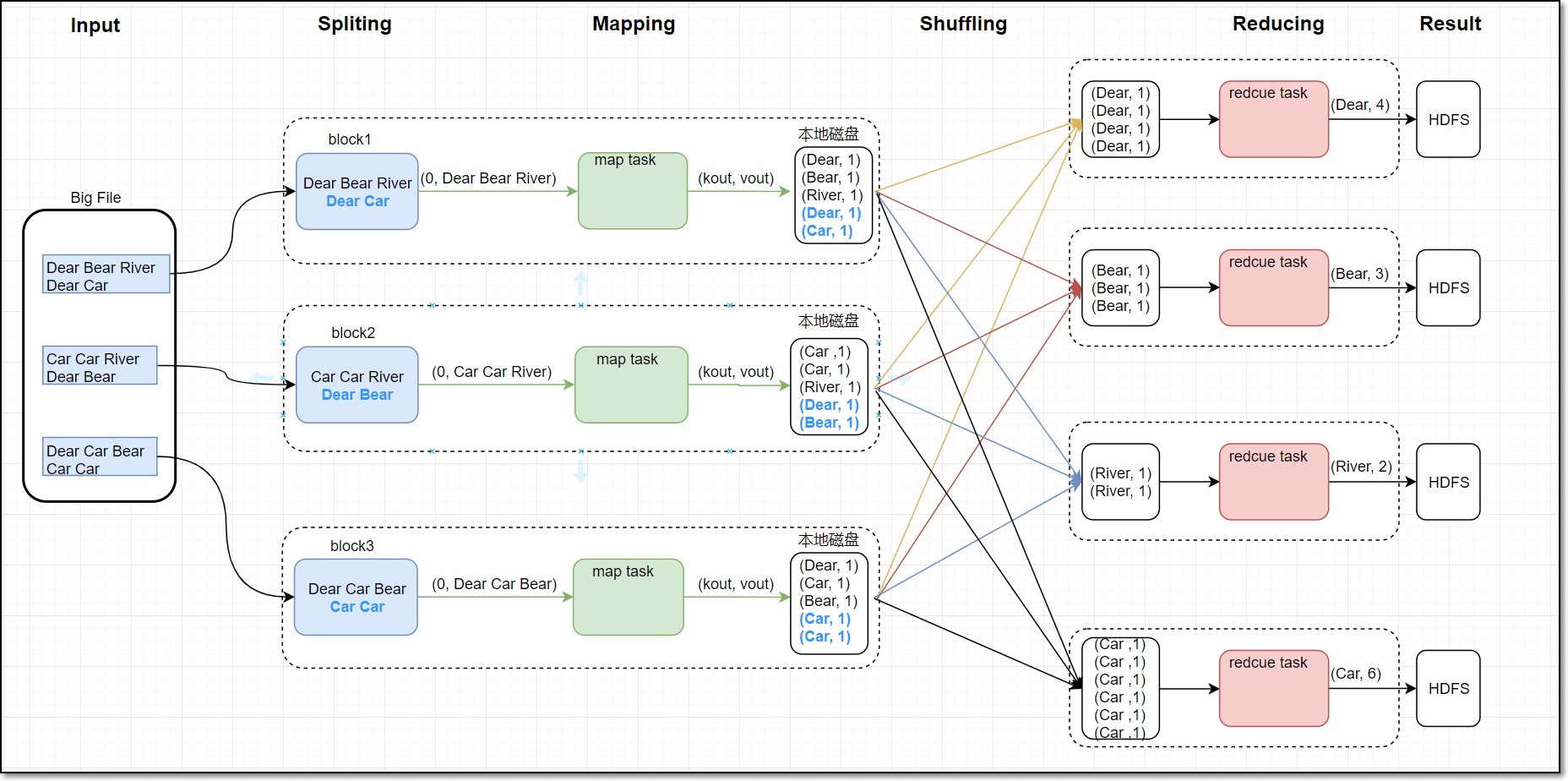
**图5-8 词频统计**

5.1.3 大数据分析

大数据处理最重要的环节就是数据分析，数据分析通常分为两种：批处理和流处理。批处理是对一段时间内海量的离线数据进行统一处理，对应的处理框架Mapreduce、Spark等；流处理则是针对动态实时的数据处理，即在接收数据的同时就对其进行处理，对应的处理框架有 Storm、Spark Streaming、Flink等。本文以离线计算为主，介绍电影数据分析。

1. Mapreduce离线计算(mapreduce\_hive文件)

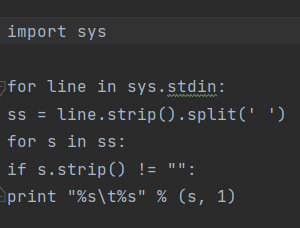
Mapreduce编程词频统计主要利用wordcount思想，通过按规定格式分割词句，实现单词统计词频。其统计数据为历史电影的上映信息，map阶段主要负责单词分割统计，map阶段把每个字符串映射成键、值对，按行将单词映射成（单词，1）形式，Shuffle过程会对map的结果进行分区排序，然后按照同一分区的输出合并在一起写入到磁盘中，最终得到一个分区有序的文件，最后reduce阶段会汇总统计出每个词对应个数，数据最终会存储在HDFS上。本文以电影词作为统计对象，实现单词统计词频功能。词频统计流程图如下：



**图5-9 词频统计流程图**

map阶段会把每个字符串映射成键、值对，按行将单词映射成（单词，1）形式输出，其中shuffle过程会对map的结果进行分区排序，然后按照同一分区的输出合并在一起写入到磁盘中，最终得到一个分区有序的文件。

Map阶段代码：

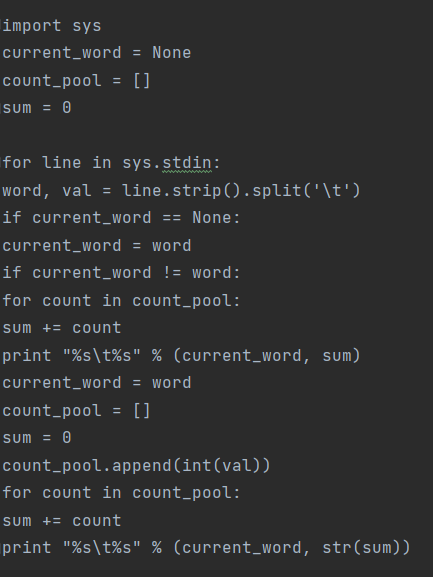


**图5-10 Map阶段代码**

1. reduce阶段

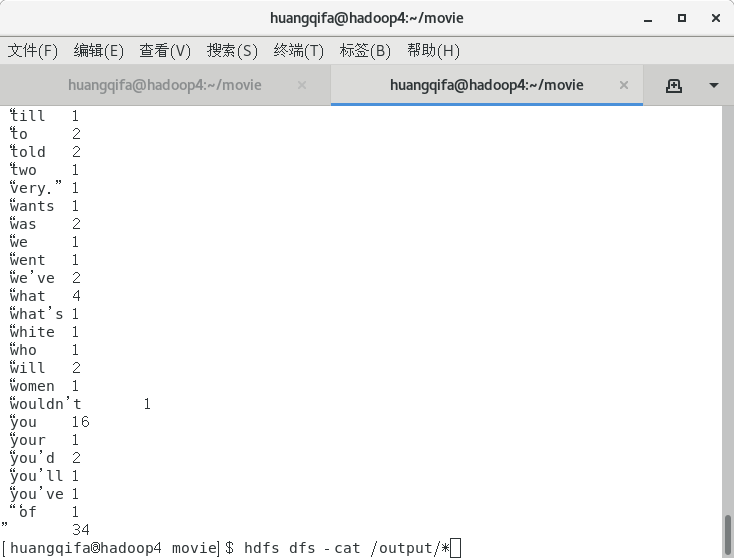
reduce阶段会汇总map阶段结果每个词对应个数，数据最终会存储在HDFS上。本文以《哈利波特》英文电影文本为例，实现词频统计功能。

Reduce阶段代码：



**图5-11 Reduce阶段代码**

利用Hadoop Streaming可以使用任何可运行程序或语言作为Map和Reduce的创建和执行MapReduce作业，通过编写shell脚本执行wordcount，需要进入mapreduce\_hive目录下运行执行命令sh run.sh 。

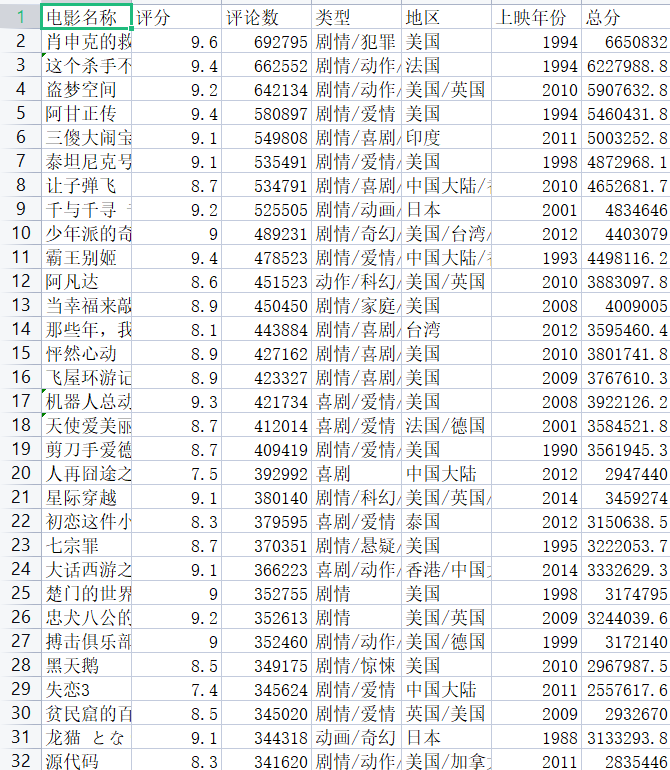


**图5-11 操作图**

5.1.4 Hive数仓分析

Hive是一个基于Hadoop的数据仓库工具,主要用于解决海量结构化日志的数据统计，可以将结构化的数据文件映射成一张表，通过类SQL语句的方式对表内数据进行查询、统计分析。利用Sqoop数据传输工具可以将Mysql数据库信息导入到Hive数仓。

运用Hive可以实现海量数据分析，并且支持自定义函数，省去MapReduce编程。本文针对历史豆瓣电影数据进行统计，数据经过清洗，删除空值、多余项，得到大约100000多条电影数据，数据文件对应movie.csv。

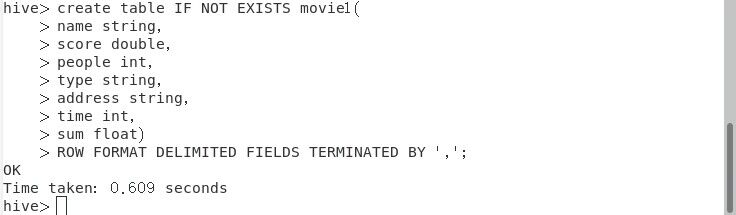


**图5-12 movie.csv文件截图**

1. 建表

Hive建表分为内部表和外部表两种。创建内部表，表内数据将会移动到数据仓库指向的路径，删除表时，数据会随之删除；而外部表在删除时，不会删除数据表原有信息，相对更加安全。

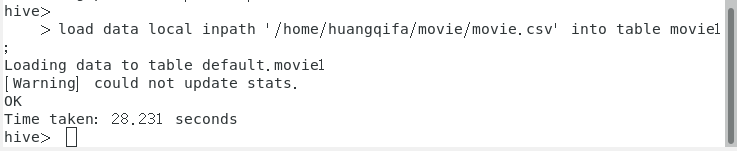
本文电影表包括电影名称、评分、评论人数、类型、上映年份，以及总分等字段，数据默认逗号分隔，其中总分=电影评分\*评论人数，数据创建命令如下：



**图5-13 数据创建命令**

1. 导入数据

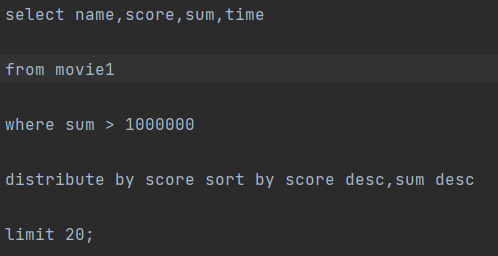
Hadoop支持各类型文件上传到HDFS，可以通过本地命令直接上传到Hive中，也可以利用Sqoop数据传输工具实现将Mysql数据库与Hive数据库互传。本地导入命令如下：



**图5-14 本地导入命令**

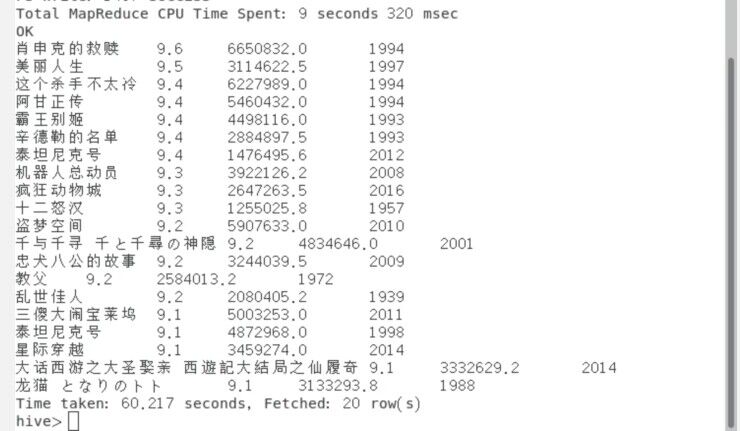
1. 统计分析

Hive底层基于Mapreduce执行，利用distribute by和sort by命令可以实现分组排序，统计总分（总分计算规则如上）在1000000分以上电影数据，并按照评分、总得分降序顺序排列，优化后命令如下：



**图5-15 优化后命令**

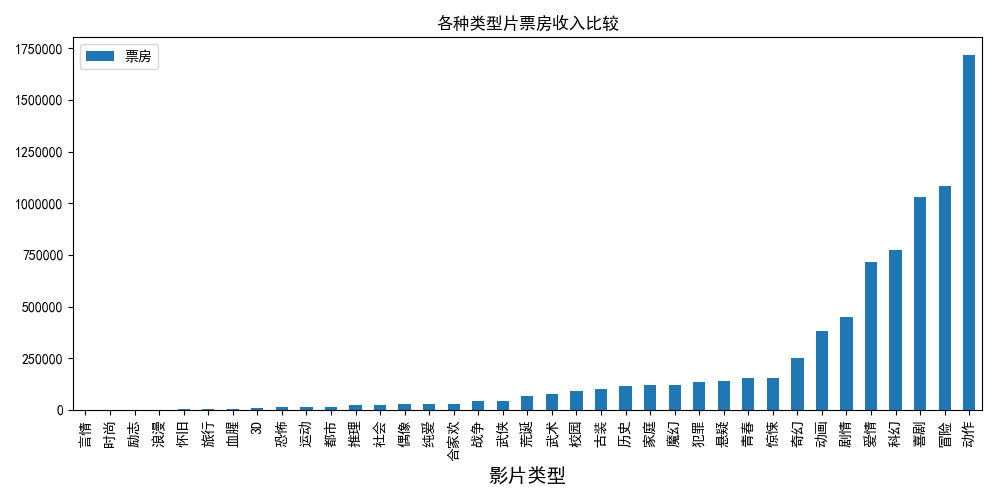
最终运行结果展示



**图5-16 运行结果**

5.1.5 影片类型与票房统计图

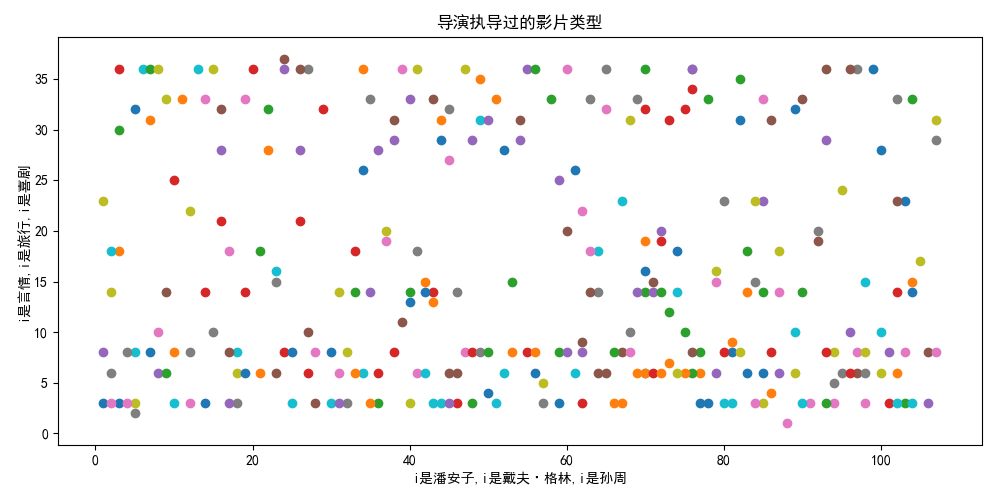
执行movietype.py



**图5-17 影片类型与票房统计图**

5.1.6 导演与影片类型关系图

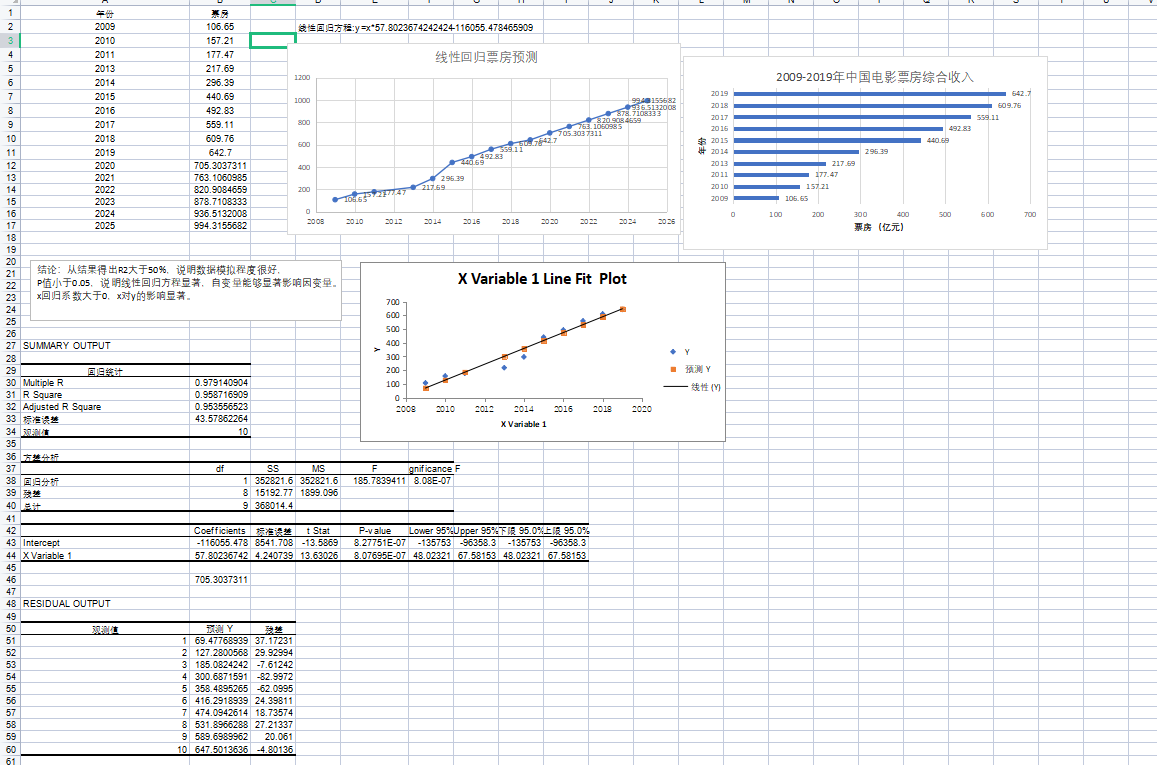
执行director.py



**图5-18 导演与影片类型关系图**

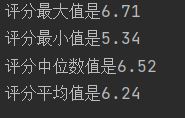
5.1.7电影票房预测

通过数据分析线性回归预测电影票房未来收入情况，并进行相关分析，得到电票票房预测.xls



**图5-19 电影票房预测**

5.1.8 电影评分预测scorepredict.py

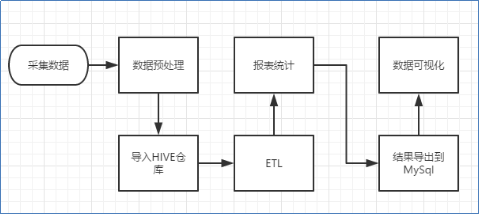


**图5-20 电影评分预测**

5.2 基本原理

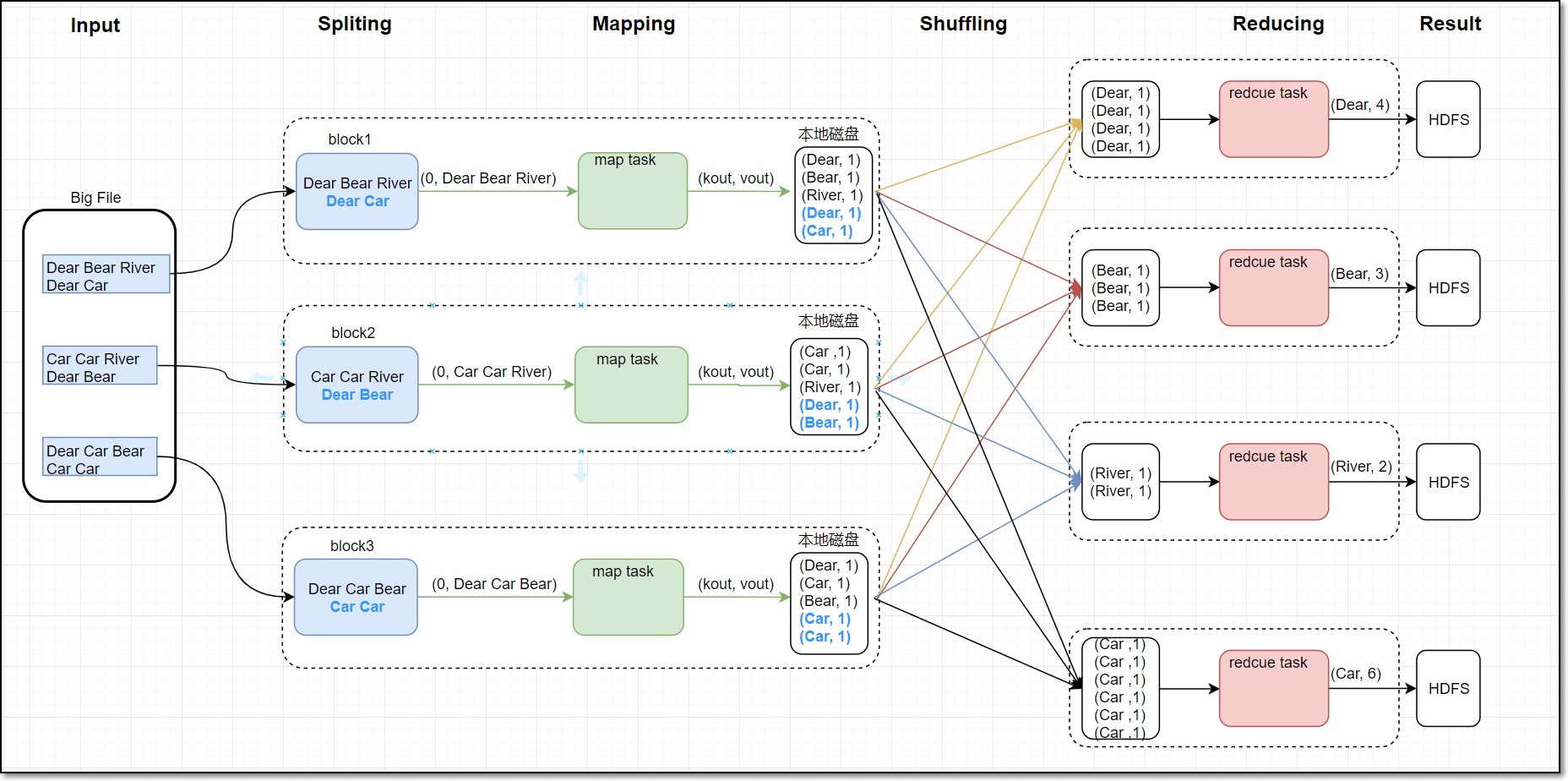
大数据处理最重要的环节就是数据分析，数据分析通常分为两种：批处理和流处理。批处理是对一段时间内海量的离线数据进行统一处理，对应处理框架Mapreduce、Spark等；流处理则是针对动态实时的数据处理，即在接收数据的同时就对其进行处理，对应的处理框架有 Storm、Spark Streaming、Flink等。

大数据分析流程图：



**图5-21 大数据分析流程图**

Mapreduce编程词频统计主要利用wordcount思想，通过按规定格式分割词句，实现单词统计词频。其统计数据为历史电影的上映信息，map阶段主要负责单词分割统计，map阶段把每个字符串映射成键、值对，按行将单词映射成（单词，1）形式，Shuffle过程会对map的结果进行分区排序，然后按照同一分区的输出合并在一起写入到磁盘中，最终得到一个分区有序的文件，最后reduce阶段会汇总统计出每个词对应个数，数据最终会存储在HDFS上。本文以电影词作为统计对象，实现单词统计词频功能。



**图5-22 单词统计词频流程图**

6 设计实现

本项目以电影数据为主题，以数据采集、处理、分析及数据可视化为项目流程，融合了包括python爬虫，Matplotlib绘图、Echarts数据可视化、结合mysql数据实现hive电影相关数据统计、Mapreduce词频统计、情感分析、词图云等多方面知识，实现了对电影评分的预测。使用该方法可实现百万级电影数据离线处理与计算。



**图6-1 整体结构**



**图6-2 豆瓣评分前20统计**



**图6-3 囧妈影评可视化**



**图6-4 电影词频统计**



**图6-5 Mapreduce词频分析**



**图6-6 hive统计高分电影**



**图6-7 电影评分预测**

7 结论

本项目以电影数据为主题，基于hadoop伪分布式搭建，结合hive数据仓库调用物理机mysql数据库实现电影相关数据统计、通过Mapreduce编程对hdfs文件系统的文件进行词频统计。使用python进行电影数据采集、处理、分析及数据可视化，融合了python爬虫，Matplotlib绘图、Echarts数据可视化、情感分析、词图云等多个功能，可实现百万级电影数据离线处理与计算。

这次小组作业我们巩固了前面学习的Hadoop环境搭建、mysql数据库相关知识，在此基础上对hive数据仓库、mapreduce工作原理进行学习。通过本次作业我们也学习到了一些linux系统的安装、基本使用、系统架构以及一些常见命令的使用等。通过小组学习，我们各有分工，相互协作，加深了对彼此的了解，学到新的知识。

参考文献

[1]刘林慧,王慧. 基于随机森林理论的电影评分预测[J]. 现代信息科技,2021,5(16):83-85,92. DOI:10.19850/j.cnki.2096-4706.2021.16.021.

[2]李香君,肖小玲. 基于机器学习的电影评分预测研究[J]. 电脑知识与技术,2021,17(27):109-111.

[3] 黄东晋, 耿晓云, 李娜,等. 基于混合特征的电影评分预测系统[J]. 计算机技术与发展, 2020, 30(12):6.

[4]谭家柱. 基于随机森林算法的IMDB电影评分预测研究[J]. 现代计算机,2021,27(30):24-31. DOI:10.3969/j.issn.1007-1423.2021.30.004.

[5]黄东晋,纪浩,耿晓云,等. 基于文本矢量特征的电影评分预测模型[J]. 现代电影技术,2019(3):44-50. DOI:10.3969/j.issn.1673-3215.2019.03.010.

[6]张黎平,段淑萍,俞占仓.基于Hadoop的大数据处理平台设计与实现[J].电子测试,2022,36(20):74-75+83.DOI:10.16520/j.cnki.1000-8519.2022.20.011.

[7]季斯皓. 基于Hive数据仓库的贵金属市场投资者交易行为管理系统的设计和实现[D].华东师范大学,2022.DOI:10.27149/d.cnki.ghdsu.2022.002081.

[8]耿俊豪. 基于MapReduce的分类算法研究[D].江西理工大学,2022.DOI:10.27176/d.cnki.gnfyc.2022.000449.