实训课程报告

项目名称：大内影视库大数据分析项目

学生姓名：

所在学院：人工智能与大数据学部

专

学

业：软件工程

号：210111316 210111319 210111324

210111315 210111320 210111318

210111321

指导教师：彭金渝

朱方灏钧 黄乙鑫 张庆 周慧

李浇 杨阳 罗培铭

2022年5月

卡通画

中度可信度描述已自动生成

1

宜宾学院人工智能与大数据学部

目录

[第一章 项目需求 4](#_Toc104049083)

[第二章 开发环境 5](#_Toc104049084)

[1 Linux操作系统Centos 7发行版 5](#_Toc104049085)

[2 MySQL Server 8.0 5](#_Toc104049086)

[3 Hadoop 3.1.0 5](#_Toc104049087)

[4 Hive 3.1.2 5](#_Toc104049088)

[第三章 项目任务分解 6](#_Toc104049089)

[1 数据上传 6](#_Toc104049090)

[2 数据清洗 6](#_Toc104049091)

[3 建表、关联数据 6](#_Toc104049092)

[4 数据分析 8](#_Toc104049093)

[第四章 项目实现思路与设计 9](#_Toc104049094)

[1 用户上传的视频数量分类 9](#_Toc104049095)

[2 每个用户的朋友数分类 9](#_Toc104049096)

[3 按上传者的视频日期进行分组 9](#_Toc104049097)

[6 按视频长度length进行分组查询 10](#_Toc104049098)

[7 查询最多 10](#_Toc104049099)

[第五章 项目代码与效果展示 11](#_Toc104049100)

[1 用户上传的视频数量分类 11](#_Toc104049101)

[4 按视频类别category分组查询每个类别的视频总数量 17](#_Toc104049102)

[5 按视频评分rate进行分组查询 19](#_Toc104049103)

[7 查询最多 23](#_Toc104049104)

[第六章 实训总结. 26](#_Toc104049105)

# 分工明细

1. 朱方灏钧 组长

分配任务 编写SQL,可视化

1. 周慧 组员

思考数据分析

1. 张庆 组员

思考数据分析

1. 李浇 组员

查询数据

1. 杨阳 组员

查询数据

1. 黄乙鑫 组员

PPT制作，文案

1. 罗培铭 组员

PPT制作，文案

# 项目需求

电影公司制作或引进一部新电影需要推向市场时，要想获得成功，通常要了解电影市场的趋势。观众目前喜好的电影类型，电影的发行情况，改编电影和原创电影的收益等情况都能通过流量、评分、评论数与观看数有更直观的认识。

当前各行业发展流行趋势很快 ，如果不能正确认识到市场走向，观众喜好，不能提供观众满意的作品就无法在决策中作出更正确的判断。

为了让电影行业相关的公司通过数据分析的结果能作出更正确的判断，从而获得更大的收益，并能更好地运营。因此，我们可以利用大数据技术，通过对电影的评分、流量、观众互动比例的分析，结合电影类型、收益、发行情况推测电影的发展趋势。

通过对大量电影用户数据的分析，我们基本可以了解电影行业的种种信息，从而为更好的运营提供有力的数据支持。

# 开发环境

## Linux操作系统Centos 7发行版

Centos是一款Linux操作系统发行版，主要用于服务器运维和开发，Linux的高稳定性和对技术人员的高友好性，占用资源小是我们选择它作为大数据运行环境的主要原因。Linux也是服务器运行的主要环境。

## MySQL Server 8.0

MySQL Server是一款数据库服务端软件，主要用于存放数据，对于大数据所需要存储的大量数据文件，而且它支持Hadoop和Hive，所以我们首选MySQL作为后台数据库运行。

## Hadoop 3.1.0

大数据服务端软件我们使用Hadoop 3.1.0，Hadoop是大数据技术的首要技术支持软件，要开发大数据相关项目，离不开Hadoop。

## Hive 3.1.2

要开发大数据相关项目，除了Hadoop后端以外，我们还需要用一个客户端进行操作和连接，在这里，我们最常使用的就是Hive，用于与MySQL和Hadoop进行沟通，并提供了前台数据相关操作相关接口。

# 项目任务分解

## 数据上传

在hdfs根目录中创建tv总目录，在进入其中分别创建user、video目录，将数据分别上传至目录中

## 数据清洗

hadoop jar Teduvideo-1.0-SNAPSHOT.jar cn.tedu.Teduvideo.etl.TeduvideoETLDriver /tv/video /tv/etl-video

## 建表、关联数据

创建原始数据表：tv\_ori，tv\_user\_ori，

创建最终表：tv\_orc，tv\_user\_orc

（1）创建原始数据表tv\_ori

create table tv\_ori(

videoId string,

uploader string,

age int,

category array<string>,

length int,

views int,

rate float,

ratings int,

comments int,

relatedId array<string>)

row format delimited fields terminated by "\t"

collection items terminated by "&"

stored as textfile;

（2）创建原始数据表: tv\_user\_ori

create table tv\_user\_ori(

uploader string,

videos int,

friends int)

row format delimited

fields terminated by "\t"

stored as textfile;

创建orc存储格式带snappy压缩的表：

（1）tv\_orc

create table tv\_orc(

videoId string,

uploader string,

age int,

category array<string>,

length int,

views int,

rate float,

ratings int,

comments int,

relatedId array<string>)

stored as orc

tblproperties("orc.compress"="SNAPPY");

（2）tv\_user\_orc

create table tv\_user\_orc(

uploader string,

videos int,

friends int)

row format delimited

fields terminated by "\t"

stored as orc

tblproperties("orc.compress"="SNAPPY");

（3）向ori表插入数据

load data inpath "/tv/etl-video" into table tv\_ori;

load data inpath "/tv/user" into table tv\_user\_ori;

（4）向orc表插入数据

insert into table tv\_orc select \* from tv\_ori;

insert into table tv\_user\_orc select \* from tv\_user\_ori;

## 数据分析

略

# 项目实现思路与设计

## 用户上传的视频数量分类

每个用户上传视频数 0-10（普通用户）

每个用户上传视频数 11-30（活跃用户）

每个用户上传视频数 31-50（视频小达人）

每个用户上传视频数 51-80（主播博主）

每个用户上传视频数 81-100

每个用户上传视频数 大于100

## 每个用户的朋友数分类

0-100的朋友数

101-200的朋友数

201-500的朋友数

501-1000的朋友数

1000以上的朋友数

## 按上传者的视频日期进行分组

（a.一年内 b.两年内 c.>两年以上）查询每组uploader视频上传者的总数

* 1. **按视频类别category分组查询每个类别的视频总数量**
  2. **按视频评分rate进行分组查询**

每个评分区间（a.0-3.00 b.3.00-4.00 c.4.00-5.00）不同视频类别category在这个评分区间的视频总量

## 按视频长度length进行分组查询

每个长度区间（a.<100 b.100<=length<200 c.200<=length<300 d.300<=length<400 e.400<length<=500 f.>500)的视频总数

## 查询最多

A. 查询出每个类别category中观看次数views最多的视频videoId 视频唯一id（String）以及视频长度length、views观看次数（Int）、视频评分（Double）rate、Ratings流量（Int）、评论数（Int）conments

B 查询出每个类别category中评分rate最高的视频videoId 视频唯一id（String）以及视频长度length、views观看次数（Int）、视频评分（Double）rate、Ratings流量（Int）、评论数（Int）conments

C. 查询出每个类别category中评论数conments最多的视频videoId 视频唯一id（String）以及视频长度length、views观看次数（Int）、视频评分（Double）rate、Ratings流量（Int）、评论数（Int）conments

# 项目代码与效果展示

## 用户上传的视频数量分类













|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 视频数量 | 0-10 | 11-30 | 31-50 | 51-80 | 81-100 | >100 |
| 用户人数 | 1264697 | 590112 | 142909 | 71264 | 20581 | 49166 |

通过图表结构可以得出大部分用户都是普通用户上传视频数量比较少，部分用户会尝试投稿，只有少部分用户会坚持上传视频。

* 1. **每个用户的朋友数分类**











|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 朋友数量 | 0-100 | 101-200 | 201-500 | 501-1000 | >1000 |
| 用户人数 | 2080284 | 31133 | 18566 | 5526 | 3600 |

通过图表结构可以得出几乎所有用户的好友数都不会超过100，只有极少数用户的好友数会在100以上，说明大部分用户都是个人用户，没有粉丝需求。

* 1. **按上传者的视频日期进行分组**





|  |  |  |  |
| --- | --- | --- | --- |
| 视频日期 | 一年以内 | 两年以内 | 两年以上 |
| 用户人数 | 21338 | 689075 | 27094 |

通过图表结构可以得出两年前上传视频的用户并不多，九成上传者都在近两年上传视频，尤其是去年涨幅非常大，今年的涨幅应该会比去年更大。

## 按视频类别category分组查询每个类别的视频总数量



|  |  |
| --- | --- |
| 种类 | 数量 |
| "Autos","Vehicles" | 14284 |
| "Comedy" | 87818 |
| "Entertainment" | 127674 |
| "Film","Animation" | 73293 |
| "Gadgets","Games" | 59817 |
| "Howto","DIY" | 18257 |
| "Music" | 179049 |
| "News","Politics" | 35925 |
| "People","Blogs" | 48890 |
| "Pets","Animals" | 10496 |
| "Sports" | 67329 |
| "Travel","Places" | 14675 |
| "UNA" | 6062 |

通过图表结构可以得出音乐类视频数量是最多的，可以增加此类视频的推送；而政治、自然类相对要少很多，体现用户对此类视频缺少兴趣。

## 按视频评分rate进行分组查询







|  |  |  |  |
| --- | --- | --- | --- |
| 评分 | 0-2 | 3-4 | 5 |
| Autos Vehicles | 4668 | 2960 | 6656 |
| Comedy | 24833 | 15105 | 47880 |
| Entertainment | 24862 | 17395 | 85417 |
| Film Animation | 9924 | 7951 | 55418 |
| Gadgets Games | 11851 | 11237 | 36729 |
| Howto DIY | 4874 | 3510 | 9873 |
| Music | 29117 | 20700 | 129232 |
| News Politics | 8803 | 7196 | 19926 |
| People Blogs | 17707 | 7590 | 23593 |
| Pets Animals | 3843 | 1918 | 4735 |
| Sports | 16945 | 10471 | 39913 |
| Travel Places | 6186 | 2265 | 6224 |
| UNA | 868 | 1370 | 3824 |

通过图表结构可以得出娱乐向视频更容易获得高评分，也意味着娱乐向视频更受到用户喜爱，非娱乐向视频则缺少关注度，评分数量远不如娱乐向视频。

* 1. **按视频长度length进行分组查询**









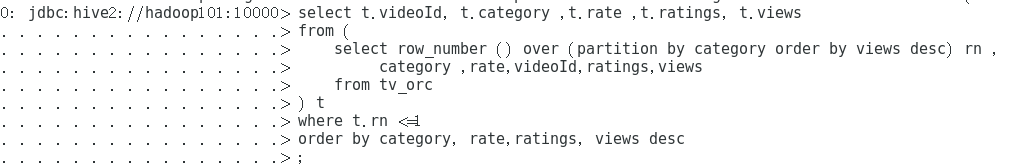


|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 视频长度 | 0-100 | 100-200 | 200-300 | 300-400 | 400-500 | >500 |
| 视频数量 | 211365 | 174635 | 188238 | 71528 | 38623 | 59180 |

通过图表结构可以得出当今用户更偏爱时长更短的短视频，而时长偏长的视频则相对缺少吸引力。

## 查询最多

A. 查询出每个类别category中观看次数views最多的视频videoId 视频唯一id（String）以及视频长度length、views观看次数（Int）、视频评分（Double）rate、Ratings流量（Int）、评论数（Int）conments



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 视频ID | 种类 | 评分 | 流量 | 访问量 |
| RjrEQaG5jPM | ["Autos","Vehicles"] | 3.89 | 5034 | 2803140 |
| dMH0bHeiRNg | ["Comedy"] | 4.68 | 87520 | 42513417 |
| sdUUx5FdySs | ["Film","Animation"] | 4.78 | 42417 | 5840839 |
| pFlcqWQVVuU | ["Gadgets","Games"] | 4.68 | 8784 | 3651600 |
| hut3VRL5XRE | ["Howto","DIY"] | 4.5 | 7097 | 2684989 |
| QjA5faZF1A8 | ["Music"] | 4.84 | 120506 | 15256922 |
| hr23tpWX8lM | ["News","Politics"] | 4.35 | 513 | 4706030 |
| 2GWPOPSXGY | ["Pets","Animals"] | 4.58 | 12086 | 3660009 |
| Ugrlzm7fySE | ["Sports"] | 4.32 | 8541 | 2867888 |
| bNF\_P281Uu4 | ["Travel","Places"] | 4.81 | 29152 | 5231539 |
| aRNzWyD7C9o | ["UNA"] | 3.75 | 36815 | 8825788 |

B 查询出每个类别category中评分rate最高的视频videoId 视频唯一id（String）以及视频长度length、views观看次数（Int）、视频评分（Double）rate、Ratings流量（Int）、评论数（Int）conments

文本

描述已自动生成

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 视频id | 类别 | 评分 | 流量 | 观看次数 |
| fkVPlSy3rSM | ["Autos","Vehicles"] | 5.00 | 4 | 971 |
| dse-nvEuMRM | ["Comedy"] | 5.00 | 1 | 209 |
| o2x1tc1O\_7s | ["Entertainment"] | 5.00 | 5 | 3361 |
| QH8OX76mjtI | ["Film","Animation"] | 5.00 | 1 | 69 |
| oM-M2rYzJjE | ["Gadgets","Games"] | 5.00 | 12 | 1426 |
| lPCQuiAi4yg | ["Howto","DIY"] | 5.00 | 2 | 1097 |
| JP9knK-DU5M | ["Music"] | 5.00 | 1 | 346 |
| CgTWKPkUSuw | ["News","Politics"] | 5.00 | 1 | 361 |
| 92A\_7UT3be0 | ["People","Blogs"] | 5.00 | 2 | 34 |
| d1slF6n0U00 | ["Pets","Animals"] | 5.00 | 1 | 16 |
| ed32C6OmImo | ["Sports"] | 5.00 | 1 | 30 |
| 0YHz0fvaf1U | ["Travel","Places"] | 5.00 | 3 | 2509 |
| RSW-lmwHgDc | ["UNA"] | 5.00 | 4 | 1959 |

C. 查询出每个类别category中评论数conments最多的视频videoId 视频唯一id（String）以及视频长度length、views观看次数（Int）、视频评分（Double）rate、Ratings流量（Int）、评论数（Int）conments

文本

描述已自动生成

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| id | 种类 | 评分 | 流量 | 观看次数 |
| RjrEQaG5jPM | ["Autos","Vehicles"] | 3.89 | 5034 | 2803140 |
| 0XxI-hvPRRA | ["Comedy"] | 4.49 | 80710 | 20282464 |
| vr3x\_RRJdd4 | ["Entertainment"] | 4.87 | 39079 | 10786529 |
| 6B26asyGKDo | ["Film","Animation"] | 4.58 | 31792 | 5147533 |
| Hfmx4sZNhXU | ["Gadgets","Games"] | 4.08 | 7882 | 1183046 |
| i7QVbJnSPQE | ["Howto","DIY"] | 4.04 | 1699 | 268068 |
| QjA5faZF1A8 | ["Music"] | 4.84 | 120506 | 15256922 |
| k8x14cLGh5o | ["News","Politics"] | 3.65 | 12377 | 1431638 |
| D2kJZOfq7zk | ["People","Blogs"] | 4.82 | 42162 | 11184051 |
| OeNggIGSKH8 | ["Pets","Animals"] | 4.36 | 5921 | 2457750 |
| Ugrlzm7fySE | ["Sports"] | 4.32 | 8541 | 2867888 |
| bNF\_P281Uu4 | ["Travel","Places"] | 4.81 | 29152 | 5231539 |
| aRNzWyD7C9o | ["UNA"] | 3.75 | 36815 | 8825788 |

# 实训总结.

实训，就是把我们在学校所学的理论知识，运用到客观实际中去，了解学到的理论知识应该怎样使用。只学不干，对于学习的知识水平就没有清楚的自我认识。理论就应与操作相结合，不仅能学习、熟悉所学知识，还能对自己的工作能力有个初步的自我了解，及时发现、查漏补缺。另一方面，实践也是在为以后找工作打基础。

通过这周的实训，学习了之前没有接触过的大数据知识。对于不熟悉的知识要学会从实践中学习，从学习中实践。在之后还要不断实践，在实践中提高熟练度，不断从各方面武装自我，才能在竞争中突出自我，表现自我