哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

大数据分析综合案例

李春山

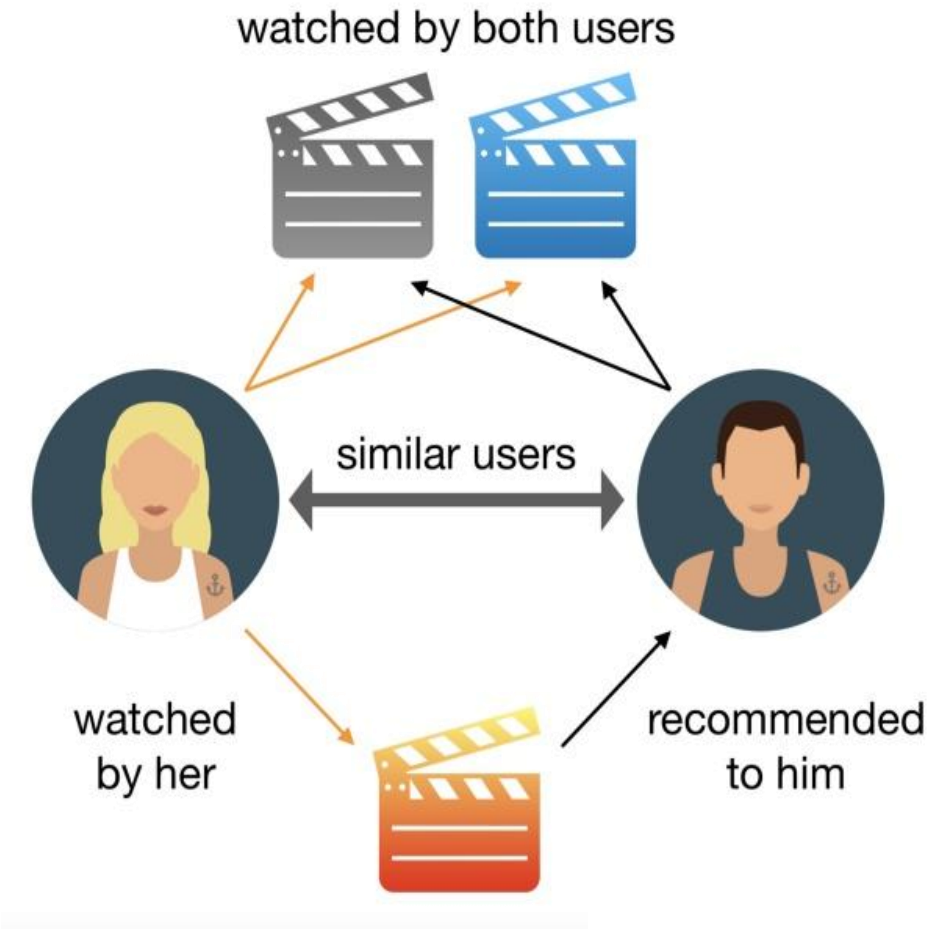
主要内容

- **案例任务**
- **系统设计**
- **技术选择和系统实现**

哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

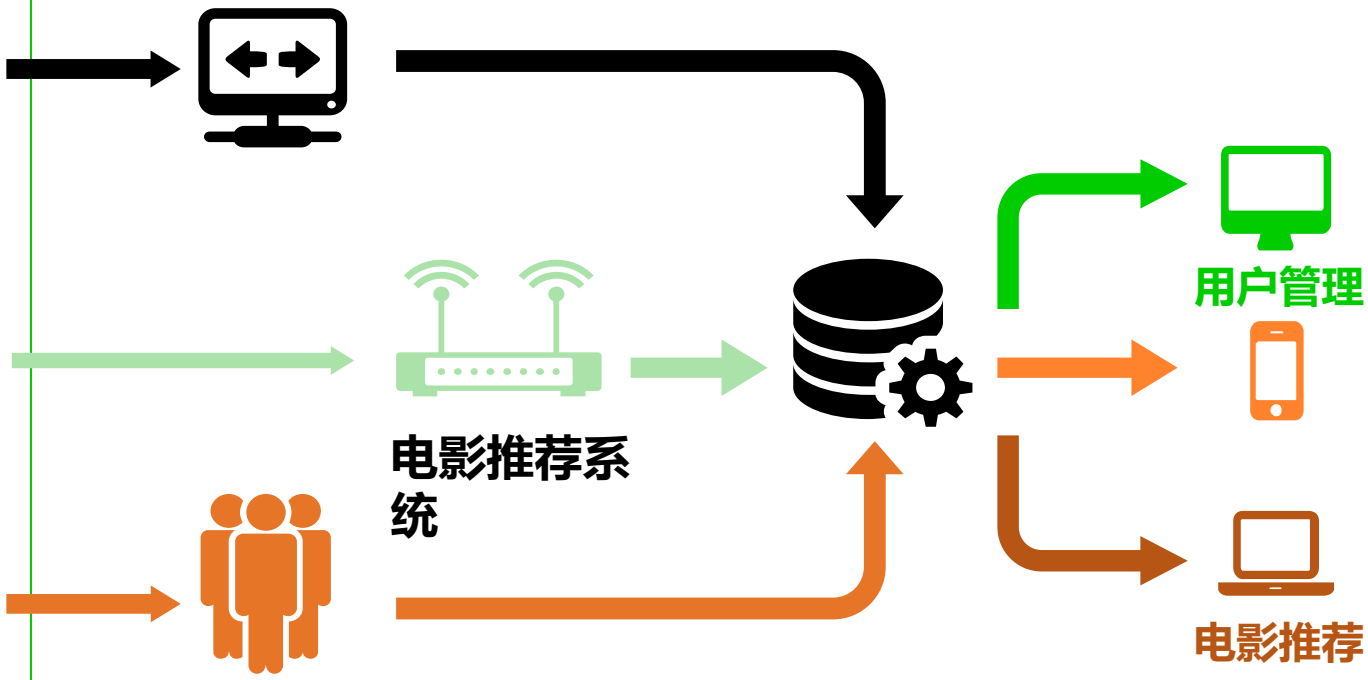
案例任务

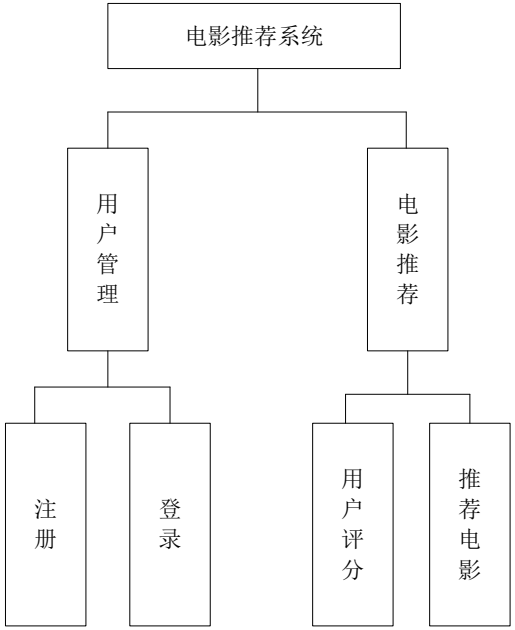
案例任务



电影推荐系统可**根据用户的喜好**，向用户推荐可能感兴趣的电影

案例任务





案例任务

为你推荐

换一批

超人3

金酸梅奖作品

x战警2

异能勇士大乱斗

蜘蛛侠2

蜘蛛侠也有成长的烦恼

谁偷了我的DNA

学渣屌丝基因突变获...

绿巨人2

最帅绿巨人诺顿

宇宙追缉令

李连杰大闹平行世界

蝙蝠侠：侠影之谜

蝙蝠侠的诞生

黑衣人2

黑衣人重生踏上回归...

超能联盟

逗比四人组意外获得...

黑客帝国3：矩阵...

黑客系列最终篇

惊变28天

人心比丧失还恐怖

绿灯侠

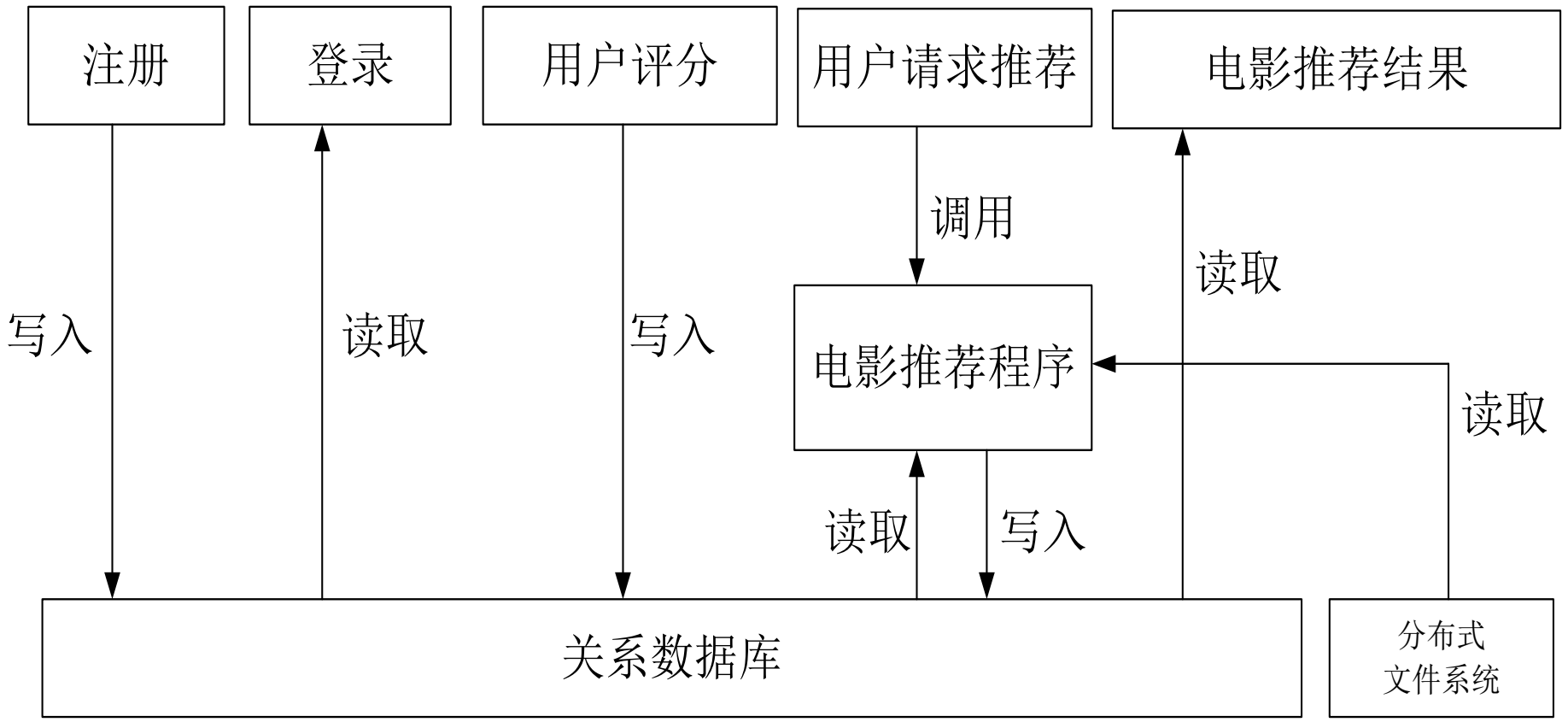
花花公子成大英雄

哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

系统设计

系统设计

系统设计

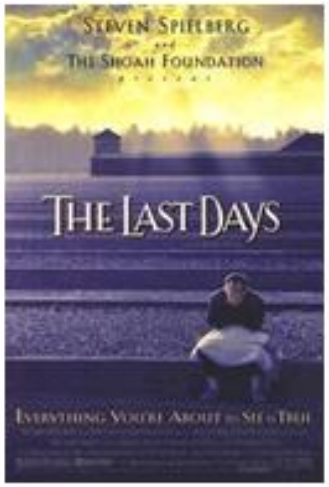


网站、电影推荐程序和数据库三个部分之间的关系

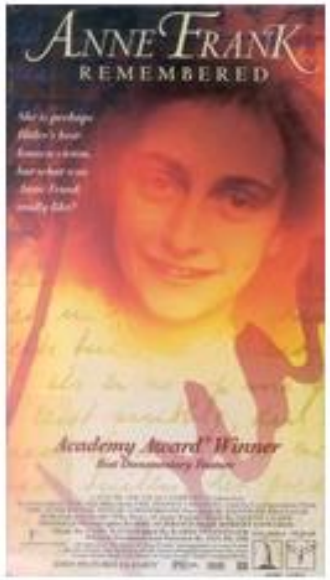
系统设计

亲爱的用户:hadoop,猜你喜欢电影:

Last Days, The (1998)



Anne Frank Remembered (1995)



McCabe & Mrs. Miller (1971)

Smashing Time (1967)

Man of the Century (1999)

系统设计

用户信息表user

用户ID、用户名用户登录密码

电影信息表movieinfo

电影ID、电影名称、电影上映时间、电影导演主要演员、电影宣传海报、电影的平均评分

参与电影评分的人数、电影简介、电影类型

用户评分表personalratings

用户ID、电影ID、用户对电影的评分以及评分时间

电影推荐结果表recommendresult

用户ID、电影ID、电影评分、电影名称

系统设计

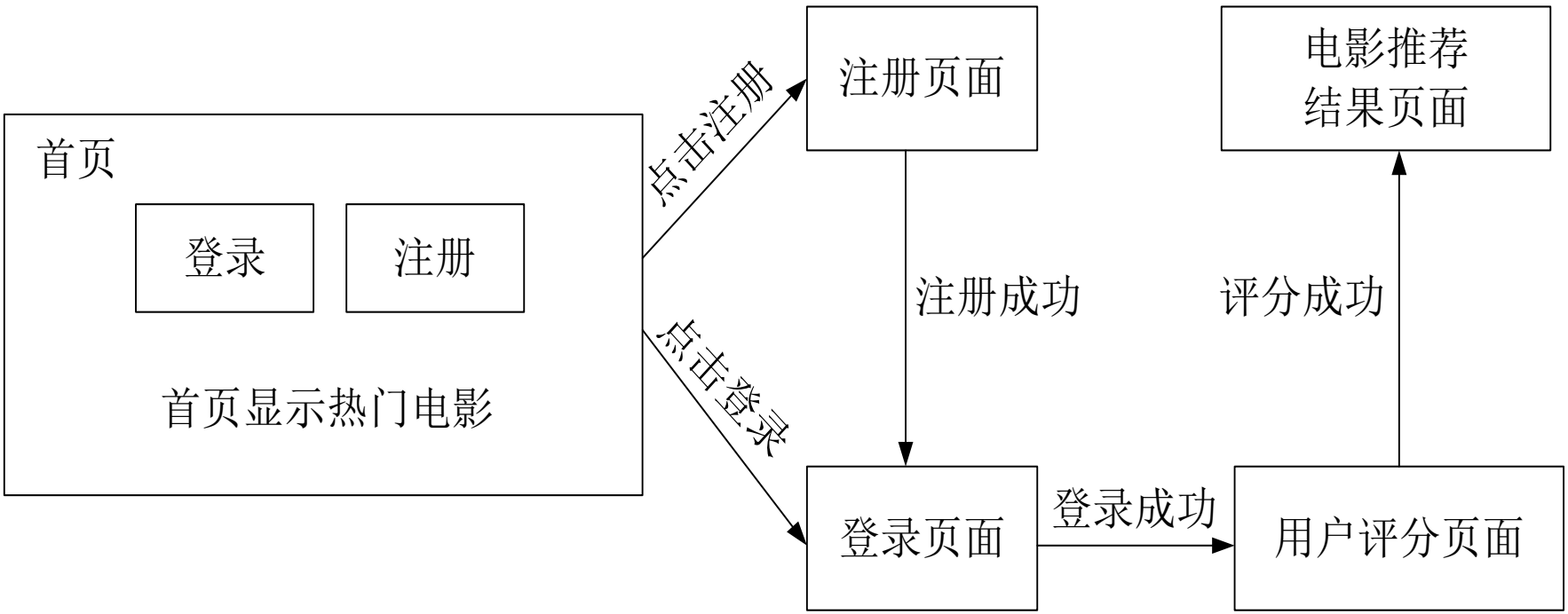


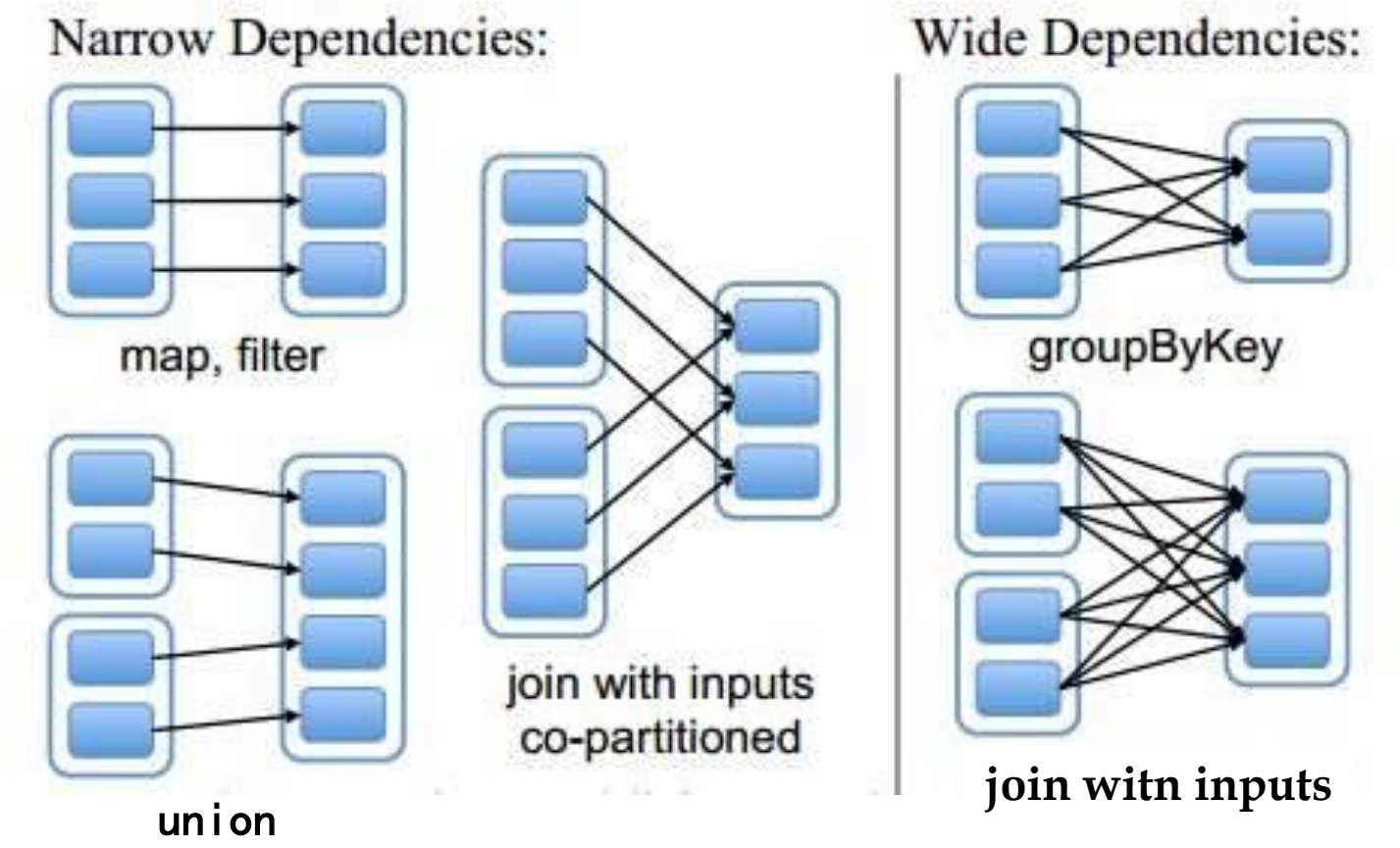
图 电影推荐系统的网页跳转示意图

系统设计

算法设计

电影推荐程序的设计是电影推荐系统设计的核心

算法设计



基于ALS矩阵分解的协同过滤算法

算法设计

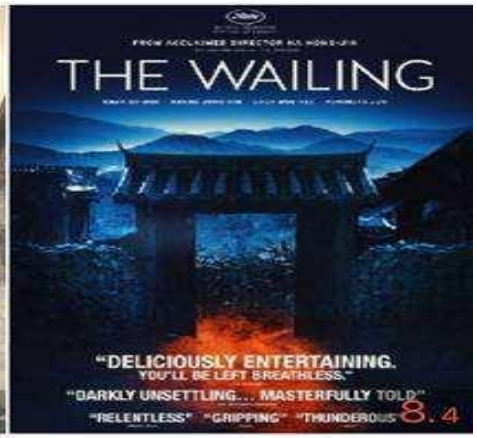
海边的曼彻斯特

萨利机长

谍影重重5

航海王之黄金城

神奇动物在哪里



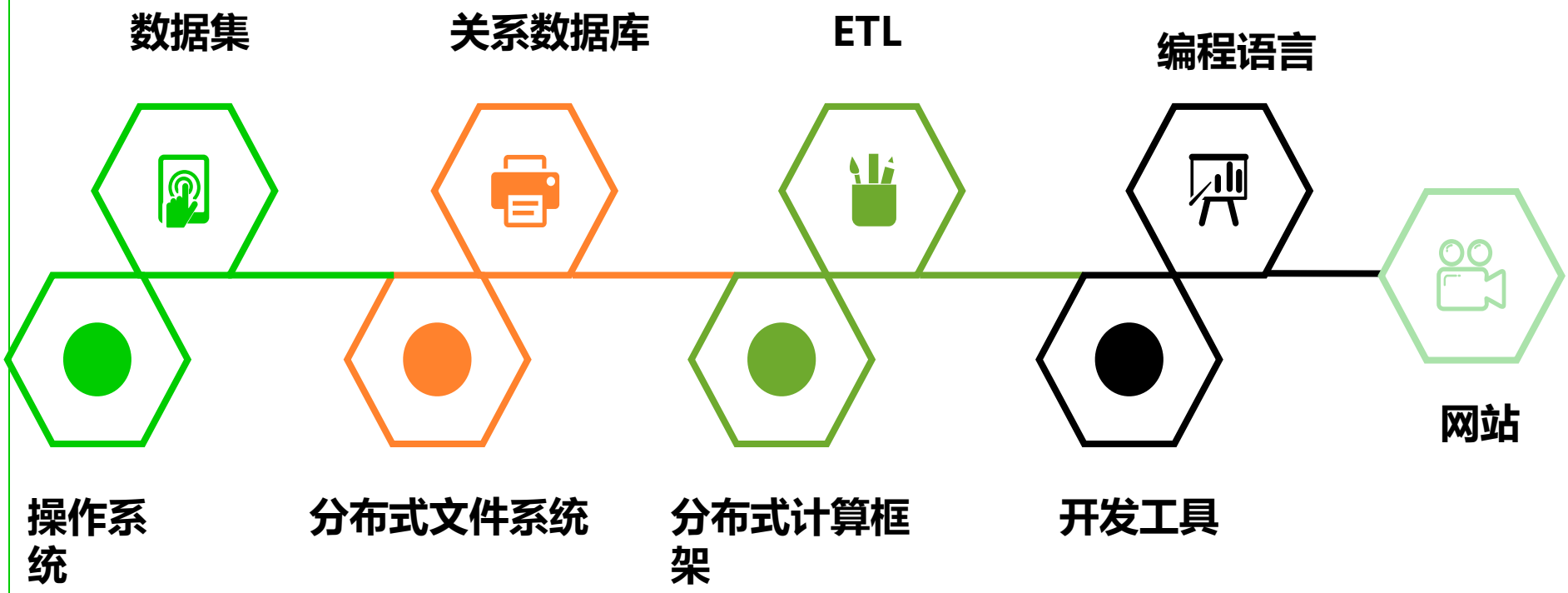
哭声

把评分高的电影推荐给该用户

哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

技术选择和系统实现

技术选择



技术选择

系统实现技术

项目	技术
数据集	Scrapy爬虫
操作系统	Linux系统，比如Ubuntu
关系数据库	MySQL
分布式文件系统	HDFS
ETL	Kettle
分布式计算框架	Spark MLlib
编程语言	Scala
开发工具	IntelliJ IDEA
网站	Node.js

系统实现

搭建环境

01

安装Linux系统、JDK、关系型数据库MySQL、大数据软件Hadoop、大数据软件Spark、开发工具IntelliJ IDEA、ETL工具Kettle等

采集数据

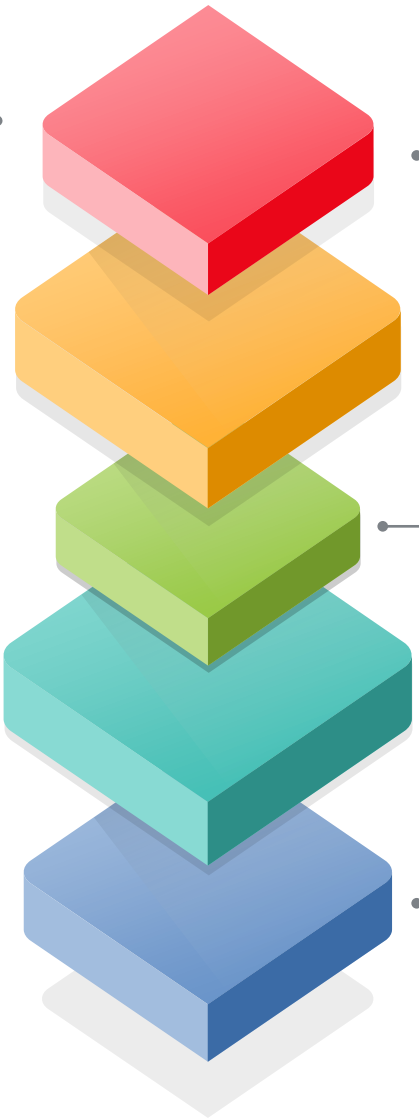
02

编写Scrapy爬虫从网络上获取电影评分数据

加载数据

03

使用ETL工具Kettle对数据进行清洗后加载到HDFS



存储和管理数据

04

使用HDFS和关系数据库MySQL对数据进行存储和管理

处理和分析

05

使用Scala语言和开发工具IntelliJ IDEA，编写Spark MLlib程序，根据HDFS中的大量数据进行模型训练

可视化

06

使用Node.js搭建网站，接受用户访问，并以可视化方式呈现电影推荐结果

系统实现

实现本案例

Linux操作系统、关系数据库、JDK基本知识、面向对象编程、Scala编程语言、网络爬虫、数据清洗、分布式文件系统、Spark、Spark SQL、Spark Mllib、JDBC、机器学习、数据挖掘、推荐系统、协同过滤算法、ASL算法、网页应用程序开发、HTML语言、数据可视化、系统设计等

专业技能的角度

Linux系统、JDK的安装、Hadoop的安装和基本使用方法、Spark的安装和基本使用方法、MySQL数据库的安装和基本使用方法、开发工具InteliJ IDEA的安装和使用方法、Scala程序开发方法、软件项目管理工具Maven的使用方法、ETL工具Kettle的安装和使用方法、Spark SQL程序的开发方法、ALS算法的使用方法、Spark MLlib程序开发方法、Node.js的安装和使用Node.js开发动态网页的方法等

哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

结束

2023年6月1日