2020交通学院实训

目录

**[实训目标](#_Toc1004416167_WPSOffice_Level1)** **[2](#_Toc1004416167_WPSOffice_Level1)**

**[实训涉及技术范围](#_Toc1365551352_WPSOffice_Level1)** **[2](#_Toc1365551352_WPSOffice_Level1)**

**[日程安排](#_Toc432203616_WPSOffice_Level1)** **[3](#_Toc432203616_WPSOffice_Level1)**

**[选题与数据集介绍](#_Toc752851596_WPSOffice_Level1)** **[3](#_Toc752851596_WPSOffice_Level1)**

[题目设置](#_Toc1365551352_WPSOffice_Level2) [3](#_Toc1365551352_WPSOffice_Level2)

[选题要求](#_Toc432203616_WPSOffice_Level2) [4](#_Toc432203616_WPSOffice_Level2)

**[数据集简介](#_Toc1443373763_WPSOffice_Level1)** **[4](#_Toc1443373763_WPSOffice_Level1)**

[北京多站点空气质量数据集数据](#_Toc1437095813_WPSOffice_Level2) [4](#_Toc1437095813_WPSOffice_Level2)

[家电能耗预测数据集](#_Toc1055460125_WPSOffice_Level2) [5](#_Toc1055460125_WPSOffice_Level2)

[Adult](#_Toc1645529524_WPSOffice_Level2) [5](#_Toc1645529524_WPSOffice_Level2)

[BitcoinHeistRansomwareAddressDataset](#_Toc1775258556_WPSOffice_Level2) [5](#_Toc1775258556_WPSOffice_Level2)

[体育馆人数数据](#_Toc981950895_WPSOffice_Level2) [6](#_Toc981950895_WPSOffice_Level2)

[网上购物者购买意图数据集](#_Toc752851596_WPSOffice_Level2) [6](#_Toc752851596_WPSOffice_Level2)

[P2P信贷平台业务数据](#_Toc1443373763_WPSOffice_Level2) [6](#_Toc1443373763_WPSOffice_Level2)

[Give Me Some Credit](#_Toc766527907_WPSOffice_Level2) [6](#_Toc766527907_WPSOffice_Level2)

[Internet-companies-of-China](#_Toc1521221791_WPSOffice_Level2) [6](#_Toc1521221791_WPSOffice_Level2)

[B站《王者荣耀》相关高播放量视频数据](#_Toc963561613_WPSOffice_Level2) [7](#_Toc963561613_WPSOffice_Level2)

[B站《英雄联盟手游》相关高播放量视频数据](#_Toc2013356387_WPSOffice_Level2) [7](#_Toc2013356387_WPSOffice_Level2)

[MediumArticlesDataset](#_Toc1477302701_WPSOffice_Level2) [7](#_Toc1477302701_WPSOffice_Level2)

# 实训目标

根据当前学期的课程内容（大数据生态系统与大数据数据分析与挖掘技术）进行项目实训。

实训共两周，第一周周一开题，讲解实训基本内容，并提供几个数据集供学生选择，之后学生需要根据实训内容，分组选择实训所需数据集进行实际内容的编写。第一周周五中期检查，检查学生项目进展程度。第二周周五答辩。

对于实训中的每一天，上午根据案例进行技术讲解，同时学生同步编写有关代码。下午自行编码，根据上午讲解有关技术编写个人所选择的实际课题。

最终目标为使学生掌握大数据获取，大数据迁移，大数据存储，大数据分析与挖掘，大数据可视化有关基本技术，并使用有关技术编写完整的项目。

# 实训涉及技术范围

根据实际数据集的不同，学生从以下技术选择合理的方法完成最终结果：

（1）使用Flume获取数据 （2）使用sqoop进行数据迁移 （3）使用HDFS进行文件存储 （4）使用Hive构建大数据仓库 （5）使用Hbase构建大数据仓库 （6）使用MapReduce编写算法进行数据分析与挖掘 （7）使用hive进行数据分析 （8）使用hbase进行数据分析 （9）使用redis进行数据分析与挖掘结果的缓存 （10）使用echarts/d3/matplotlib进行数据可视化

# 日程安排

第一周周一：开题：大数据发展现状，大数据常用架构，数据集介绍与命题选择

第一周周二：如何快速构建大数据系统实操，基于云开发大数据系统实操，大数据获取与数据迁移实操

第一周周三：基于MapReduce统计分析

第一周周四：基于hive与hbase进行统计分析，结果缓存

第一周周五：中期检查与答疑

第二周周一：使用echarts/d3/matplotlib进行数据可视化（一）

第二周周二：使用echarts/d3/matplotlib进行数据可视化（二），实时数仓与离线数仓的构建

第二周周三：基于MapReduce的构建常用算法

第二周周四：扩展内容

第二周周五：答辩

注：实训内容会根据学生实际进度进行微调。

# 选题与数据集介绍

## 题目设置

一共提供了十二个数据集，其中带有时间标志的数据集有五个，一个数据集包含两个数据文件，文本数据集一个。还有三个数据集比较方便用于爬虫扩展，两个数据集官方提供了后续数据，可进行对比分析。

本次选题为半开放选题：（1）数据集不限，除提供数据集外，也可使用其他数据集（2）所选技术栈进行一定的限制。

可选题目：（1）基于XXX数据集离线计算仓库的数据分析、挖掘与可视化系统

1. 基于XXX数据集的实时数仓数据分析、挖掘与可视化系统
2. XXX数据分析、挖掘，可视化数据报告

## 选题要求

1. 实时数仓需根据时间进行分时间段的有关数据分析与可视化。
2. 离线计算数仓则必选完成数据迁移与模型的定时更新。
3. 数据报告型课题，必须进行建模预测
4. 必须包含完整的数据获取（或数据导入），大数据分析与挖掘，结果缓存，数据可视化与展示完整的相关内容

# 数据集简介

## 北京多站点空气质量数据集数据

该每小时数据集考虑了北京多个站点的6种主要空气污染物和6种相关的气象变量，带有时序标志。可用于分类，回归或者时间序列分析。

## 家电能耗预测数据集

数据集为10分钟，持续约4.5个月。用ZigBee无线传感器网络监控房屋的温度和湿度条件。每个无线节点在3.3分钟左右传输温度和湿度条件。然后，将无线数据平均10分钟。能量数据每10分钟用m-bus能量计记录一次。最近的机场气象站（比利时基耶夫斯机场）的天气是从Reliable Prognosis（rp5.ru）的公共数据集中下载的，并使用日期和时间列与实验数据集合并在一起。数据集中包含两个随机变量，用于测试回归模型并过滤掉非预测性属性（参数）。 有关房屋，数据收集，R脚本和图形的更多信息，请参考该论文以及以下github存储库：[Web链接](https://github.com/LuisM78/Appliances-energy-prediction-data)

## Adult

美国收入调查数据，来自于UCI，提取是由Barry Becker从1994年人口普查数据库中进行的。使用以下条件提取了一组合理的干净记录：（（AAGE> 16）&&（AGI> 100）&&（AFNLWGT> 1）&&（HRSWK> 0））预测任务是确定一个人的年收入是否超过50K年。该数据集不包含时间，可通过美国人口普查数据获取其他数据进行对比分析，如果分析内容较少，需要进行建模预测。

## BitcoinHeistRansomwareAddressDataset

比特币数据集，该数据集较为复杂，需要进行一定的数据提取。同时包含时间序列。不过，该数据集已经进行了一定的清晰，也可使用该数据集的原始数据集进行本次命题。可以用于分类预测。

## 体育馆人数数据

体育馆人数数据集：某美国大学体育馆人数数据集，大约每隔十分钟统计一次，共26000人次的体育馆人员数量数据，除了人数以外还包括天气、学期等附加属性信息。

## 网上购物者购买意图数据集

在数据集的12,330个会话中，有84.5％（10,422）是不以购物结束的负面类样本，其余的（1908）是以购物结束的正面类样本。

## P2P信贷平台业务数据

所提供数据来自拍拍贷真实业务数据，从2015-01-01到2017-01-30的所有信用标的10%sample样本。数据集包含LC.csv（标的特征表数据）和LP.csv（标的还款计划和还款记录表）数据。 该数据集较为复杂，需要进行表连接，在hive或者hbase中可以进行更多分析，使用mapreduce较为复杂。可以进行分类预测。拍拍贷同时也在不断更新新数据集，可进行对比分析与预测。

## Give Me Some Credit

该数据集较为简单，可以基于所给的数据，预测借款人是否会预期，并构建信用卡评分模型。或者进行较为详细的分层分析。该数据集没有时间标志。

## Internet-companies-of-China

8W家中国互联网公司数据。可以作为可扩展数据集进行扩展，如通过爬虫获取这些企业的其他数据，进行进一步的分析。

## B站《王者荣耀》相关高播放量视频数据

b站《王者荣耀》相关视频，播放量最高的1000个视频数据 数据集 包含字段：标题 视频地址 图片 视频时长 播放量 发布时间 up主链接 up主。可以结合爬虫进一步获取新的数据。

## B站《英雄联盟手游》相关高播放量视频数据

b站《英雄联盟手游》相关视频，播放量最高的1000个视频数据（2020.11.1） 数据集 包含字段： 标题 标题\_链接 图片 视频时长 关键词1 播放量 上传时间 up主链接 up主ID。可以结合爬虫进一步获取新的数据。

## MediumArticlesDataset

来自7个不同出版物的2019年出版的6K +文章数据，该数据集比较复杂，需要牵扯到自然语言处理有关的内容。也可以使用中文数据集进行有关数据分析。