



华北电力大学

NORTH CHINA ELECTRIC POWER UNIVERSITY

第2章 大数据采集与处理

目 录

2.1

数据采集方法

2.2

数据清洗

2.3

数据存储

2.4

数据处理技术



PART 01

2.1

数据采集方法

2.1 数据采集

数据采集是指从各种数据源获取信息的过程，它是大数据分析的第一步。有效的数据采集包括确定数据需求、选择合适的数据源、使用技术手段获取数据，并确保收集到的数据质量符合分析要求。这个过程可能涉及到从网站、社交媒体、企业内部数据库、传感器等多种来源收集结构化和非结构化数据。数据采集的目标是为后续的数据分析、处理和解释提供准确和全面的数据基础。随着技术的发展，数据采集方法和工具也在不断进步，使得从大量和多样化的数据源中快速、高效地收集数据成为可能。



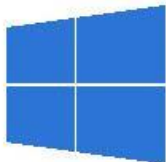
**一手数据
的获取：
网络爬虫**

**二手数据
的获取：
公开数据**

后羿采集器最新版本2.1.19正式发布! 点击查看版本更新日志!

下载后羿采集器最新版

一键安装, 免费采集网站数据, 无需编程免配置



↓ Windows

Windows 7, 8, 10



↓ Linux

Debian, Ubuntu, Centos, Fedora



↓ Mac

macOS 10.9+

后羿采集器一款**真免费**的采集软件，目前免费版本支持功能如下：

- ※ 智能模式：智能识别列表和分页，一键采集
- ※ 流程图模式：可视化操作，可以模拟人为操作
- ※ 采集任务：100个任务，支持多任务同时运行，无数量限制
- ※ 采集网址：无数量限制，支持手动输入，从文件导入，批量生成
- ※ 采集内容：无数量限制
- ※ 下载图片：无数量限制
- ※ 导出数据：导出数据到本地（无数量限制），导出格式：Excel、Txt、Csv、Html
- ※ 发布到数据库：无数量限制，支持发布到本地和云端服务器，支持类型：MySQL、PgSQL、SqlServer、MongoDB
- ※ 数据处理：字段合并，文本替换，提取数字、提取邮箱，去除字符、正则替换等
- ※ 筛选功能：根据条件组合对采集字段进行筛选
- ※ 预登录采集：采集需要登录才能查看内容的网址



主页



采集教程...



http://www.houyicaiji.com/?type=list&cat_id=148



预登录



电脑浏览器

· 如何设置预执行操作

· 如何采集需要登录才能查看的网页

· 编辑任务时遇到验证码怎么处理

· 如何在编辑任务时切换代理

· 切换浏览器模式有什么作用

· 如何设置页面类型

· 如何设置分页

· 如何对采集字段进行配置

· 如何进行数据筛选

· 如何设置数据筛选条件

· 如何设置采集范围

【智能模式】如何设置采集范围

本教程为大家介绍如何在智能模式中设置页面采集范围

2019-10-25 13:56:09

【智能模式】如何设置预执行操作

本教程为大家介绍如何设置预执行操作

2019-10-25 13:40:38

【智能模式】智能模式任务编辑界面介绍

本教程为大家介绍如何在智能模式的任务编辑页面进行任务设置

2019-10-12 15:06:20

【智能模式】如何设置分页

本教程主要给大家介绍在智能模式中如何设置分页。

2019-09-03 21:33:10

【智能模式】【流程图模式】编辑任务时遇到验证码怎么处理

本教程为大家介绍编辑任务时遇到验证码要怎么办

2019-08-23 16:28:31

【智能模式】【流程图模式】如何在编辑任务时切换代理

本教程为大家介绍如何在编辑任务时切换代理

2019-08-16 15:15:33

1

2

3

>

>>

到

4

页

GO

收起

页面类型

自动识别

分页设置

自动识别分页(成功)

设置采集范围

数据筛选

清空所有

深入采集

添加字段

	标题	标题链接	list-post-excerpt	list-post-date
1	【智能模式】【流程图模式】如何配置采集任务	http://www.houyicaiji.com/?type=post&pid=7955	本教程为大家介绍如何配置采集任务	2019-10-29 16:28:47
2	【智能模式】【流程图模式】任务运行界面介绍	http://www.houyicaiji.com/?type=post&pid=7809	本教程为大家介绍任务运行界面	2019-10-25 14:04:00
3	【智能模式】【流程图模式】如何设置数据去重	http://www.houyicaiji.com/?type=post&pid=7807	本教程为大家介绍了如何设置数据去重	2019-10-25 14:03:31
4	【智能模式】【流程图模式】如何设置防屏蔽	http://www.houyicaiji.com/?type=post&pid=7805	本教程为大家介绍如何设置防屏蔽功能	2019-10-25 14:01:24
5	【智能模式】如何设置采集范围	http://www.houyicaiji.com/?type=post&pid=7803	本教程为大家介绍如何在智能模式中设置页面采集...	2019-10-25 13:56:09

1. 采集教程_智能模式_全面掌握后...

开始采集

保存





主页

采集教程...



http://www.houyicaiji.com/?type=list&cat_id=148



预登录



电脑浏览器

功能点介绍

操作提示

已识别到列表，包含10个同类元素，您可以选择以下操作：

- 修改列表识别结果
- 提取列表中的数据
- 依次点击全部同类元素
- 提取该元素中的数据
- 点击一次该元素
- 循环点击该元素
- 移动鼠标到该元素

取消选择

编辑任务时遇到验证码怎么处理

如何在编辑任务时切换代理

功能点介绍 > 智能模式

【智能模式】【流程图模式】如何配置采集任务

2019-10-29 16:28:47

本教程为大家介绍如何配置采集任务

【智能模式】【流程图模式】任务运行界面介绍

2019-10-25 14:04:00

本教程为大家介绍任务运行界面

【智能模式】【流程图模式】如何设置数据去重

2019-10-25 14:03:31

本教程为大家介绍了如何设置数据去重

【智能模式】【流程图模式】如何设置防屏蔽

2019-10-25 14:01:24

本教程为大家介绍如何设置防屏蔽功能

【智能模式】如何设置采集范围

2019-10-25 13:56:09

本教程为大家介绍如何在智能模式中设置页面采集范围

【智能模式】如何设置预执行操作

2019-10-25 13:40:38

本教程为大家介绍如何设置预执行操作

【智能模式】智能模式任务编辑界面介绍

2019-10-12 15:06:21

收起

点击

提取数据

定时等待

滚动页面



点击



选择点击的元素

- ☐ 循环点击循环组件中的分页按钮
- ☐ 依次点击循环组件中列表内的元素
- ☒ 手动點選元素

/html/body//h2[contains(c

点击方式

- ☒ 单击
- ☐ 双击

1. 采集教程_智能模式_全面掌握后...

开始采集

保存



选择导出方式

导出到文件

Excel

CSV

TXT

HTML

导出到数据库

MySQL

SQLServer

PostgreSQL

MongoDB

神箭手数据源

Excel设置

*保存地址:

C:\Users\79775\Desktop\导出测试\2019-12-13-13-31-42-7

浏览..

*导出类型:

Excel(*.xlsx)

手动导出设置

☐ 导出所有未标记为已导出的数据☐ 导出数据页勾选的数据☒ 导出范围: 从第 条 到第 条

自动导出设置

☐ 启用自动导出 (个人旗舰版及以上专属) ☐ 打印自动导出错误日志 C:\Program Files (x86)\后羿采集器\ho ...

标记设置

☐ 导出后标记数据为已导出状态

导出文件规则

导出文件时, 遇到文件名相同则按照以下方式进行处理:

☒ 文件名添加时间前缀, 确保不重复☐ 新文件覆盖旧文件☐ 在旧文件中追加内容 (个人旗舰版及以上专属)

导出

清空数据

st-post-date

-10-29 16:28:47

-10-25 14:04:00

-10-25 14:03:31

-10-25 14:01:24

-10-25 13:56:09

-10-25 13:40:38

-10-12 15:06:20

-09-03 21:33:10

-08-23 16:28:31

-08-16 15:15:33

-08-08 20:06:31

-08-02 11:38:06

-04-23 20:22:15

-04-23 18:39:34

-04-23 16:59:45

-03-28 13:19:00

-02-19 20:05:02

-01-14 17:01:35

-12-26 19:57:35

-12-19 17:28:51





主页



采集教程...



http://www.houyicaiji.com/?type=list&cat_id=148



预登录



电脑浏览器

功能点介绍

智能模式

· 第一个采集

· 基本操作流

· 如何创建智

· 智能模式任

· 如何修改网

· 如何批量生

· 如何设置预

· 如何采集需

· 编辑任务时

· 处理

页面类型

自动识别

标题

1

【智能模式】【流程图

2

【智能模式】【流程图

3

【智能模式】【流程图模式】如何设置数据去重

4

【智能模式】【流程图模式】如何设置防屏蔽

5

【智能模式】如何设置采集范围

功能点介绍 > 智能模式

启动设置

定时启动

智能策略

自动导出

文件下载

加速引擎

数据去重

开发者设置

定时启动

☐ 循环采集 (个人专业版及以上专属)

间隔时间:

10分钟

单次运行时长:

不限制

☐ 定时启动 (个人专业版及以上专属)

启动频率:

☒ 一次☐ 每天☐ 每周

启动日期:

☒ 今天☐ 2019-12-20

启动时间:

☒ 现在☐ 00:00

停止时间:

☒ 采集完成☐ 00:00

启动



清空所有



深入采集



添加字段

list-post-date

2019-10-29 16:28:47

2019-10-25 14:04:00

2019-10-25 14:03:31

2019-10-25 14:01:24

2019-10-25 13:56:09

开始采集

保存

1. 采集教程_智能模式_全面掌握后...

针对不同基础的用户，它支持两种不同的采集模式，可以采集99%的网页。

1、智能采集模式：

该模式操作极其简单，只需要输入网址就能智能识别网页中的内容，无需配置任何采集规则就能够完成数据的采集。

欢迎使用后羿采集器

请输入要采集的网址（建议列表或表格页）

智能采集

👉 [新手入门必看](#)

🗺️ 流程图模式

可视化操作流程，根据提示在网页中点选内容即可生成采集规则，可以模拟任何人为操作

开始采集

🧠 智能模式

基于人工智能算法，输入网址即可自动识别网页内容和分页，无需配置采集规则，一键采集

开始采集

HOTSALE
促销活动!

买套餐！送定制！

立即咨询

我的任务

📄 + 🔍

> 示例

✓ 默认分组



创建流程图模式

可视化操作流程，根据显示在网页中点选内容即可生成采集规则，可以模拟任何人为操作



创建智能模式

基于人工智能算法，输入网址即可自动识别网页内容和分页，无需配置采集规则，一键采集

任务分组

默认分组

任务名称

使用网页标题

网址导入

手动输入

文件导入

批量生成

文件路径

选择本地文件

...

GBK

网址预览

文件导入网址没有数量限制
支持文件格式：txt、xlsx、csv

立即创建

🔊 购买个人旗舰版套餐，只需 8 元!

立即购买 x

 VIP专属客服

软件版本: 3.4.10

2、流程图采集模式:

完全符合人工浏览网页的思维方式，用户只需要打开被采集的网站，根据软件给出的提示，用鼠标点击几下就能自动生成复杂的数据采集规则；

欢迎使用后羿采集器

请输入要采集的网址（建议列表或表格页）

智能采集

[🔗 新手入门必看](#)

🗺️ 流程图模式

可视化操作流程，根据提示在网页中点选内容即可生成采集规则，可以模拟任何人为操作

开始采集

🤖 智能模式

基于人工智能算法，输入网址即可自动识别网页内容和分页，无需配置采集规则，一键采集

开始采集

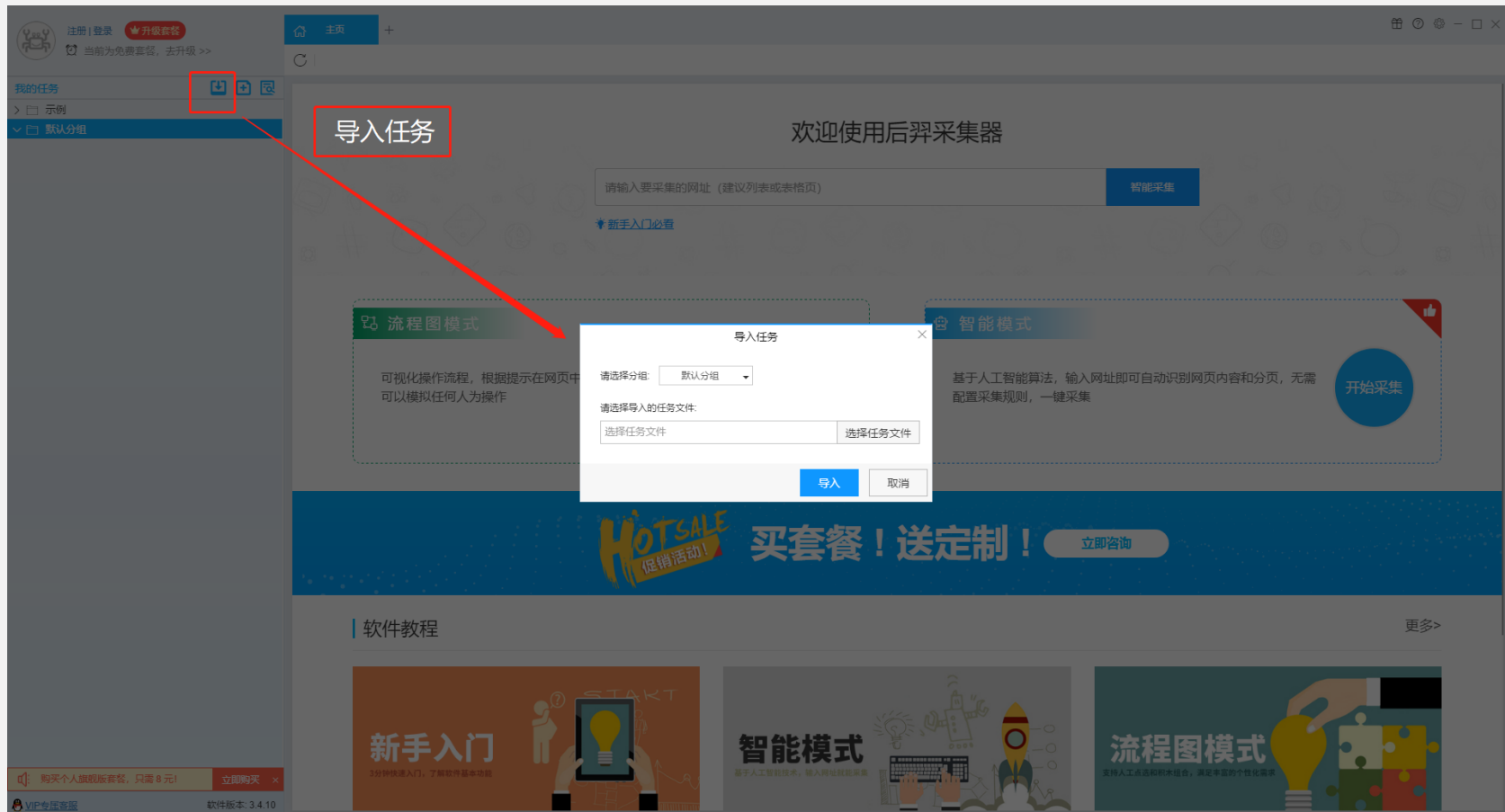


买套餐！送定制！

立即咨询

如何导入和导出采集任务

1、导入采集任务



2、导出采集

注册 | 登录

升级套餐

当前为免费套餐, 去升级 >>

我的任务

> 示例

默认分组

采集教程_功能点介绍_全面掌握后羿采集器...

数据条数: 0

2019-12-19 14:...

开启加速引擎, 享最高10倍加速

启动任务...

编辑任务

查看数据

修改名称...

复制任务...

导出任务...

修改分组...

删除

欢迎使用后羿采集器

请输入要采集的网址 (建议列表或表格页)

智能采集

新手入门必看

流程图模式

可视化操作流程, 根据提示在网页中点选内容即可生成采集规则, 可以模拟任何人为操作

开始采集

智能模式

基于人工智能算法, 输入网址即可自动识别网页内容和分页, 无需配置采集规则, 一键采集

开始采集

HOTSALE

促销活动!

买套餐! 送定制!

立即咨询

软件教程

更多>

新手入门

3分钟快速入门, 了解软件基本功能

智能模式

基于人工智能技术, 输入网址就能采集

流程图模式

支持人工点选和脚本结合, 满足丰富的个性化需求

购买个人旗舰版套餐, 只需 8 元!

立即购买

VIP专属客服

软件版本: 3.4.10




PART 02

2.2

数据清洗



CONTENTS

1. 缺失值处理
 2. 数据一致性处理
 3. 异常值处理
 4. 数据规范化
 5. 数据质量验证
- 



01

缺失值处理

缺失值处理方法：

缺失数据的处理方式包括：删除含有缺失值的记录、填补缺失值或使用模型预测缺失值。

参考资料：.....

缺失值处理方法

删除含有缺失值的记录:

对于缺失数据严重的记录，可以选择删除整个记录以确保数据质量。但需谨慎处理，避免数据丢失过多。

填补缺失值:

填补方法包括使用平均值、中位数、众数等进行简单填补，也可以使用基于其他变量的预测模型进行填补。





02

数据一致性处理



数据一致性处理

数据一致性处理方法：

包括数据转换、统一单位和格式，或者手动校正数据错误。

实际案例分享：

数据一致性处理对业务的影响及改进措施。

数据一致性处理方法

数据录入错误的纠正：

根据业务逻辑对数据进行修正，确保数据的一致性和准确性。

数据格式统一：

对于日期、地址等数据进行格式化处理，使其符合统一标准。





03

异常值处理

异常值处理

01

异常值识别与处理：

需要根据具体情况决定是删除异常值、进行变换处理还是保留异常值。

02

实用技巧分享：

使用统计学方法识别和处理异常值的技巧。

异常值识别与处理



异常值的影响:

异常值可能对数据分析结果产生重大影响，需要谨慎处理。



处理策略:

根据异常值的性质，采取合适的处理策略，如删除、转换或保留。



04

数据规范化

数据规范化



数据规范化方法：

对数据进行规范化处理，包括统一日期格式、地址格式、文本的大小写等。



工具推荐：

数据规范化常用工具及其特点比较。

数据规范化方法

日期格式规范化:

统一日期格式，便于数据分析和报告呈现。

文本大小写统一:

统一文本数据的大小写格式，提高数据的一致性和可比性。





05

数据质量验证



数据质量验证

通过数据质量检查来验证数据的准确性、完整性和一致性。

数据质量验证方法

数据统计量对比:

对比清洗前后的关键统计量，评估数据清洗效果。

数据可视化重审:

重新进行数据可视化分析，验证数据的一致性和趋势是否符合预期。



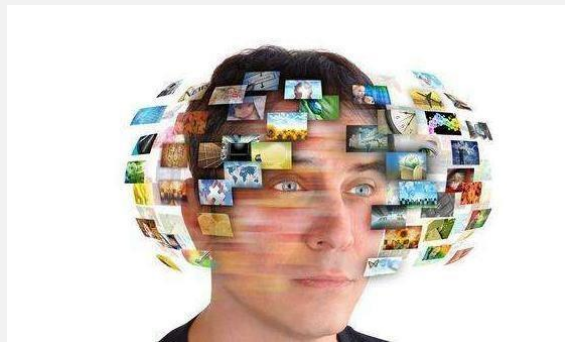
PART 03

2.3

数据存储

1.3 大数据的结构类型

数据存储涉及将数字信息保存在各种形式的媒介上，以便于未来的检索和使用。随着技术的进步，数据存储已经从传统的物理硬盘和磁带演变到了复杂的数据库系统和云存储解决方案。这些技术不仅提高了存储效率和数据访问速度，还增加了数据的安全性和可靠性。在大数据时代，有效的数据存储方案是支撑数据分析、业务智能和信息管理的基础，对于企业和组织来说至关重要。



CONTENTS

- 数据库技术
- 数据存储技术
- 数据存储策略
- 云存储和计算服务





01

数据库技术



数据库技术

关系型数据库
(RDBMS)

非关系型数据
库 (NoSQL)

新型数据库技
术



关系型数据库 (RDBMS)

- 特点：基于固定架构，使用SQL进行操作，强调ACID属性（原子性、一致性、隔离性、持久性）。
- 应用场景：适合事务性强、结构固定的传统业务系统，如财务、人力资源管理系统。银行系统和金融机构（如：摩根大通、花旗银行）广泛使用关系型数据库管理客户信息、交易记录和账户信息，确保数据的一致性和完整性。
- 示例：MySQL, Oracle, SQL Server。



- 示例：MySQL, Oracle, SQL Server。

MySQL是一个广泛使用的关系型数据库管理系统，它基于SQL（结构化查询语言）进行数据库管理，为多种应用提供了数据库服务。

特点:

- 开源性
- 可靠性
- 可移植性
- 稳定性
- 高性能
- 可扩展性
- 易于使用
- 社区支持



非关系型数据库 (NoSQL)

- - 特点：灵活的数据模型，高扩展性和可用性，不一定遵循ACID属性。
- - 应用场景：适合大规模数据集的存储和访问，如社交网络、大数据分析和实时应用。社交媒体（如：Twitter、Facebook）

社交媒体平台使用NoSQL数据库来处理大量的非结构化数据，如用户帖子、评论和社交关系图。NoSQL数据库支持快速数据读写和水平扩展，适合社交媒体的高数据吞吐需求。

- - 示例：MongoDB（文档型），Cassandra（宽列存储），Redis（键值存储）。



新型数据库技术

- - 时序数据库：专门用于处理时间序列数据（如股票价格、传感器数据），优化时间序列数据的存储和查询。
- - 图数据库：优化了图形结构数据的存储和查询，适合处理复杂的关系网络，如社交网络分析、推荐系统。例如：推荐系统（如：LinkedIn）
领英使用图数据库来管理复杂的职业网络，使得推荐算法能够有效地发现潜在的职业联系和机会。





02

数据存储技术



数据存储技术

分布式文件系统 (如HDFS)

- 特点：能够跨多个服务器存储大量数据，提供高吞吐量的数据访问，增强了数据的可靠性和可扩展性。
- 应用场景：大数据处理和分析，如Hadoop和Spark生态系统。例如：互联网搜索Google使用分布式文件系统来存储和处理Web页面索引，支持其搜索引擎高效地处理大规模数据集。

数据仓库

- 特点：集成多个数据源的数据，支持复杂的查询和分析操作，优化了读操作的性能。
- 应用场景：商业智能、报表生成、数据分析。例如：亚马逊使用数据仓库技术来分析顾客购买行为，优化库存管理，并提供个性化的购物推荐。

数据湖

- 特点：存储原始格式的数据，无需预先定义模式，支持非结构化和半结构化数据。
- 应用场景：数据科学和探索性数据分析，允许用户灵活地访问和分析数据。例如：流媒体服务Netflix构建了一个庞大的数据湖，用于存储各种日志和事件数据，支持其复杂的数据分析需求，如用户观看偏好分析和内容推荐。



03

数据存储策略



数据存储策略

冷热数据管理

根据数据访问频率采取不同的存储策略，优化存储成本和数据访问性能。

数据备份和恢复

确保数据的安全性和可靠性，防止数据丢失。

数据安全和隐私

采取加密、访问控制等措施保护数据不被未经授权访问。



04

云存储和计算服务



云存储和计算服务



特点：提供按需的存储和计算资源，用户无需管理底层硬件。

应用场景：适用于需要高度可扩展性和灵活性的应用，如大数据分析、机器学习项目。





PART 04

1.4

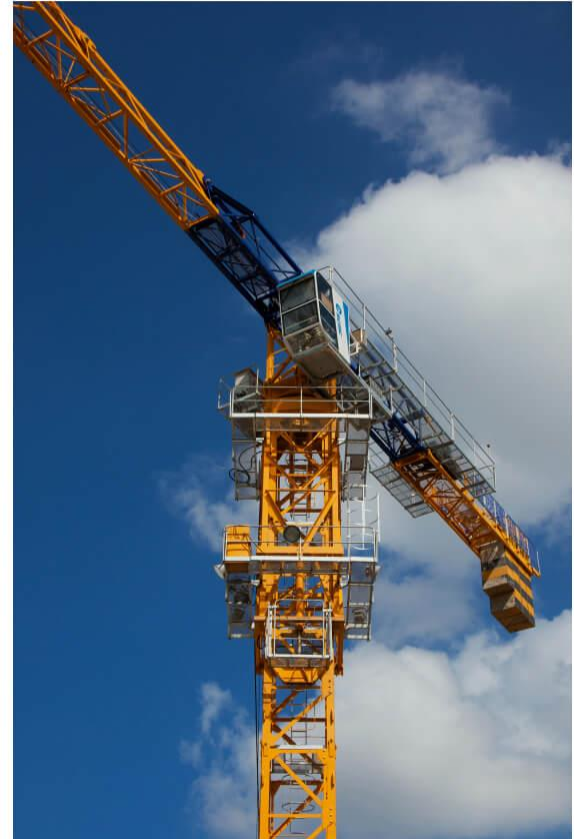
数据处理技术

1.4 数据处理技术

数据处理技术包括一系列方法和工具，旨在从原始数据中提取有用信息、清洗、转换和分析数据。这些技术支持从基本的数据清洗和预处理到复杂的数据挖掘、机器学习和深度学习等高级分析。随着技术的发展，数据处理变得更加自动化和智能化，使得个人和企业能够更有效地理解数据，做出基于数据的决策，并发现潜在的业务机会和洞察。

CONTENTS

- Excel
- Python





01

Excel





Excel

基础数据处理：

数据输入和格式化。

数据分析：

使用“分析工具”包进行统计分析。

数据可视化：

使用条件格式化和图形工具增强数据呈现。

自动化与高级功能：

利用Excel插件和外部工具扩展功能。

基础数据处理

使用公式和函数进行计算:

可以使用内置公式和函数执行各种计算，如求和、平均值等。

数据排序和筛选:

对数据进行升序、降序排列，并筛选出符合条件的数据。

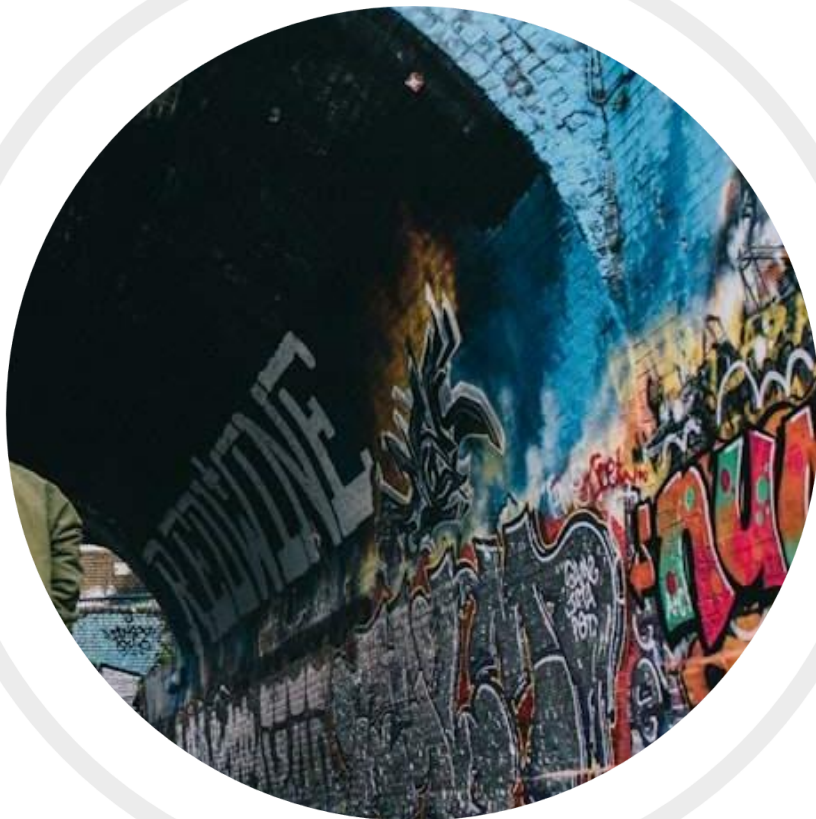
数据分析

条件格式化:

根据设定的条件对数据进行格式化，使关键数据更易于识别。

数据透视表和数据透视图:

通过数据透视表 and 透视图进行数据汇总和分析。



数据可视化

创建图表和图形:

制作各类图表，例如柱状图、折线图、饼图等，直观展示数据分析结果。

使用条件格式化和图形工具增强数据呈现:

通过颜色、图表等方式增强数据可视化效果。

自动化与高级功能

使用宏和VBA进行自动化操作:

编写宏和使用VBA进行自动化处理，提高工作效率。

利用Excel插件和外部工具扩展功能:

结合Excel插件和外部工具，拓展Excel功能，满足更多需求。



02

Python



Python

数据清洗与预处理：

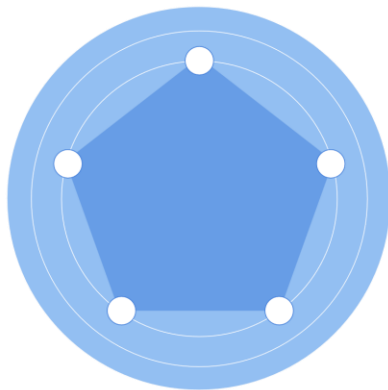
处理缺失值、删除重复项
、数据类型转换。

数据可视化：

Matplotlib和Seaborn

数据处理库：

Pandas和NumPy。



自动化与数据分析：

脚本编写和使用统计模型
、机器学习算法。

机器学习与深度学习：

Scikit-learn、TensorFlow、PyTorch

数据处理库



Pandas:

提供高效的数据结构（`DataFrame`）和数据分析工具，适用于大规模数据处理和分析。

NumPy:

用于处理大型多维数组和矩阵，支持高级数学函数，为科学计算提供了必要支持。

数据可视化

Matplotlib:

用于创建静态、动态和交互图形的库，可满足不同数据可视化需求。

Seaborn:

基于Matplotlib，用于制作更加美观的统计图表，增强数据可视化效果。



数据清洗与预处理

数据预处理：

标准化、归一化、独热编码、标签编码，
确保数据质量满足分析要求。



自动化与数据分析

脚本编写：

使用Python脚本自动化重复性数据处理任务，提高效率。

数据分析：

使用统计模型、机器学习算法进行数据探索和预测分析，发现数据之间的关联和趋势。



机器学习与深度学习

Scikit-learn:

提供简单高效的数据挖掘和数据分析工具，支持各种机器学习算法。

TensorFlow、PyTorch:

用于构建和训练深度学习模型，支持神经网络等深度学习应用。



THE END
THANKS

