Big Data Analytics
大数据分析

by

Oleh Tymchuk
讲师：奥勒·特姆恰克

Course Instructor

Oleh Tymchuk
奥勒·特姆恰克

- **Associate Professor of Taras Shevchenko National University of Kyiv 乌克兰国立基辅大学**
- **PhD in IT 信息技术博士**
- **Industry Experience: 从业经历：**
  Research Software Engineer, EPAM, Kyiv
  Backend Software Engineer, IMediaMAtch, Copenhagen
  研究软件工程师 亿磐 基辅
  后端软件工程师 IMediaMAtch 哥本哈根
- **Instructor & Mentor of Python Data Analytics courses:**
  IT Career Hub, Berlin
  PROG Academy, Kyiv
  Python数据分析课程讲师&导师
  IT职业中心 柏林
  编程学院 基辅

Course Structure 课程结构

Module 1. Introduction to Big Data 模块1. 大数据简介

What is Big Data? 什么是大数据？

- Exploring Careers in Big Data 大数据职业探索
- Data Sources 数据资源

Course Structure 课程结构

Module 1. Introduction to Big Data 模块1. 大数据简介

Module 2. Introduction to Python 模块2. Python简介

What is Python? 什么是Python？

- Python Interpreter
- Python解释器
- IDEs (Jupyter Notebook, Google Colab)
- 集成开发环境（Jupyter Notebook、Google Colab）
- Python practice
- Python练习

Course Structure 课程结构

Module 1. Introduction to Big Data 模块1. 大数据简介

Module 2. Introduction to Python 模块2. Python简介

Module 3. In-Memory Analytics with Pandas 模块3. 用Pandas进行内存分析

- Introduction to Pandas
- Data Cleaning and Preparation
- Exploratory Data Analysis (EDA)
- Chart Visualization
- Grouping and Aggregating Data
- ABC and XYZ Analysis

- Pandas简介
- 数据清洗与准备
- 探索性数据分析（EDA）
- 图表可视化
- 数据分组与聚合
- ABC和XYZ分析

Course Structure 课程结构

Module 1. Introduction to Big Data

模块1. 大数据简介

Module 2. Introduction to Python

模块2. Python简介

Module 3. In-Memory Analytics with Pandas

模块3. 使用Pandas进行内存分析

Module 4. Efficient In-Memory Analytics with Polars

模块4. 使用Polars进行高效内存分析

Course Structure 课程结构

Module 1. Introduction to Big Data

模块1. 大数据简介

Module 2. Introduction to Python

模块2. Python简介

Module 3. In-Memory Analytics with Pandas

模块3. 使用Pandas进行内存分析

Module 4. Efficient In-Memory Analytics with Polars

模块4. 使用Polars进行高效内存分析

Module 5. Big Data with Dask

模块5. 使用Dask处理大数据

Swiss UMEF
UNIVERSITY OF APPLIED SCIENCES INSTITUTE

Big Data Analytics
大数据分析

01: Introduction to Big Data
01：大数据简介

Instructor: Oleh Tymchuk
讲师：奥勒·特姆恰克

#01: Agenda 课程安排

1. What is Big Data? 什么是大数据？
2. Exploring Careers in Big Data 大数据职业探索
3. Data Sources 数据来源

What is Big Data?
什么是大数据？

Definition 定义

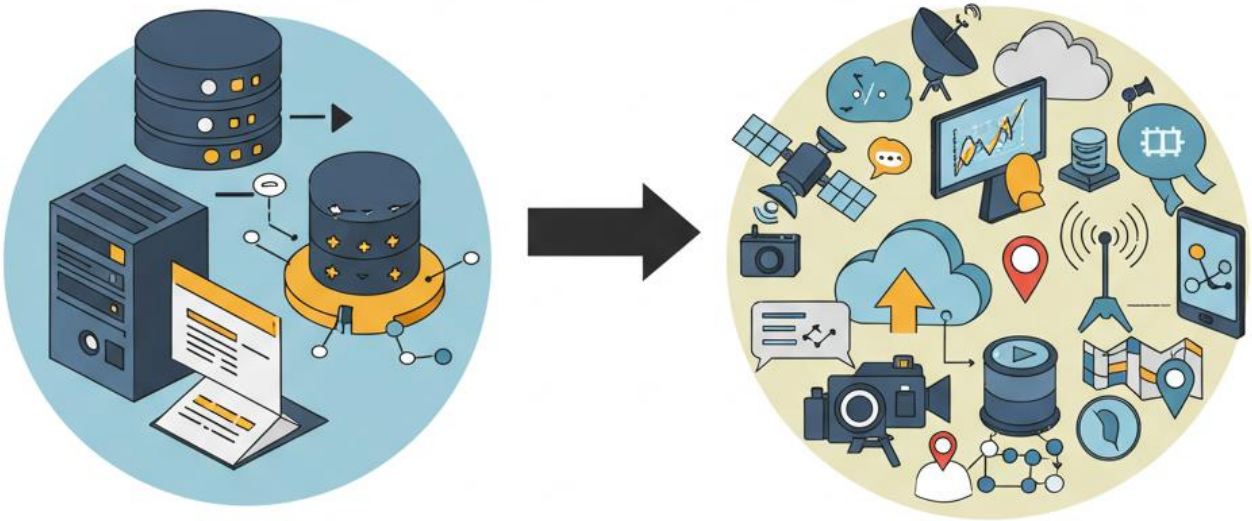
Big data refers to massive, complex data sets that traditional data management systems cannot handle.

When properly collected, managed and analyzed, big data can help organizations discover new insights and make better business decisions.

大数据是指传统数据管理系统无法处理的海量、复杂的数据集。

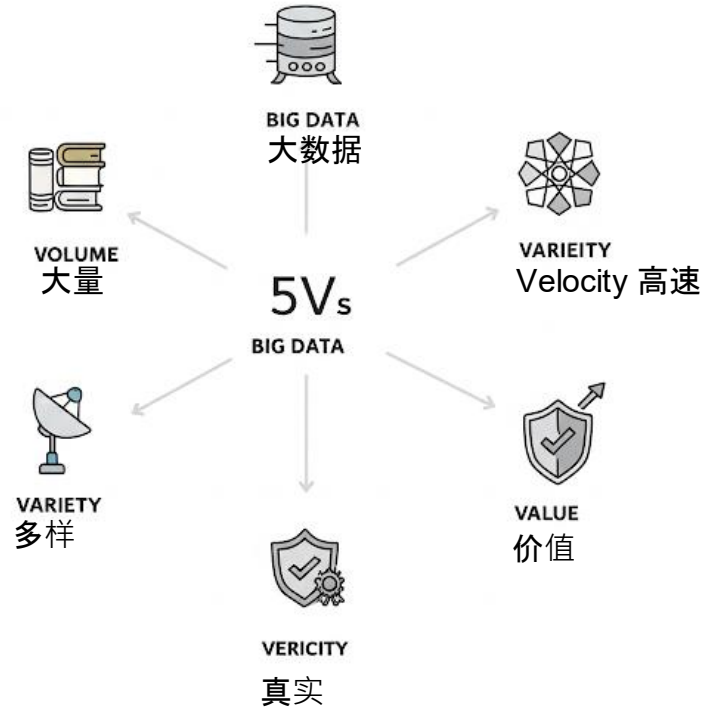
当正确收集、管理和分析时，大数据可以帮助组织获得**新的**见解，**做出更好的商**业决策。

Historical evolution 历史演变



Data never sleeps 数据永不眠

The 5Vs of Big Data 大数据的5V特征



Real-world examples | business, science, government
实际案例 | 商业、科学、政府

Healthcare: analyzing millions of patient records to detect disease outbreaks early
医疗保健：分析数百万患者记录，早期检测疾病爆发

Retail: Amazon tracks your clicks and recommends products in real time
零售：亚马逊跟踪**用户的点**击行为，并实时推荐产品

Government: cities use sensor data to manage traffic, pollution, and public safety
政府：**城市利用**传感器数据管理交通、污染和公共安全

Science: CERN generates petabytes of data from particle collisions in the Large Hadron Collider
科学：**欧洲核子研究中心**（CERN）**在大型**强子对撞机中从粒子碰撞中产生PB级数据

Real-world examples | business, science, government
实际案例 | 商业、科学、政府

Healthcare: analyzing millions of patient records to detect disease outbreaks early
医疗保健：分析数百万患者记录，早期检测疾病爆发

Retail: Amazon tracks your clicks and recommends products in real time
零售：亚马逊跟踪**用户的点**击行为，并实时推荐产品

Government: cities use sensor data to manage traffic, pollution, and public safety
政府：**城市利用**传感器数据管理交通、污染和公共安全

Science: CERN generates petabytes of data from particle collisions in the Large Hadron Collider
科学：**欧洲核子研究中心**（CERN）**在大型**强子对撞机中从粒子碰撞中产生PB级数据

Big Data is **everywhere**, and it's growing faster than ever

大数据无处不在，并**以前所未有的速度增**长

Exploring Careers in Big Data
大数据职业探索

Roles 角色

👨💻 Data Scientist

数据科学家

分析复杂数据 🔍 Analyze complex data

构建机器学习模型 🧠 Build ML models

发现见解 📊 Discover insights

与利益相关者合作 🤝 Collaborate with stakeholders

Roles 角色

Data Analyst
数据分析师

进行数据分析 Perform data analysis

清洗和转换数据 Clean and transform

识别趋势 Identify trends

创建仪表板 Create dashboards

Roles

Data Engineer
数据工程师

构建数据管道 Build data pipelines

管理基础设施 Manage infrastructure

测试数据流 Test data flows

优化性能 Optimize performance

Roles 角色

Machine Learning Engineer
机器学习工程师

设计机器学习算法 — Design ML algorithms

实现模型 — Implement models

部署解决方案 — Deploy solutions

重新训练并改进 — Retrain and improve

Roles 角色

Business Intelligence Analyst

商业智能分析师

分析商业数据

追踪关键绩效指标（KPI）

生成报告

支持决策

- Analyze business data
- Track KPIs
- Generate reports
- Support decisions

Roles 角色

Data Visualization Specialist

数据可视化专家

创建可视化仪表板

设计清晰的图表

展示见解

突出显示模式

Create visual dashboards

Design clear charts

Present insights

Highlight patterns

Roles 角色

Data Architect

数据架构师

设计数据系统

定义结构

确保集成

强制标准

Design data systems

Define structure

Ensure integration

Enforce standards

Typical analyst's toolset 经典的分析师工具集

Data Storage & Query 数据存储&查询

- SQL (PostgreSQL, MySQL)
- NoSQL (MongoDB, Cassandra)
- Hadoop / HDF
- Amazon S3, Google BigQuery

Typical analyst's toolset 经典的分析师工具集

Data Processing 数据处理

- Python
- Pandas
- NumPy
- PySpark

Typical analyst's toolset 经典的分析师工具集

Visualization & BI 可视化与商业智能

- Tableau
- Power BI
- Matplotlib
- Seaborn
- Plotly

Typical analyst's toolset 经典的分析师工具集

Machine Learning & Analytics 机器学习&数据分析

- Scikit-learn
- TensorFlow

Typical analyst's toolset 经典的分析师工具集

🛠 Collaboration & Versioning 协作&版本控制

- Git
- Jupyter Notebooks, Google Colab

Why Python is useful for Big Data analytics
为什么Python对大数据分析很有用？

- **Versatile & Powerful 多功能&强大**
  Works for data processing, analysis, visualization, ML
  适用于数据处理、分析、可视化和机器学习
- **Rich Ecosystem丰富的生态系统**
  Pandas, NumPy, PySpark, Scikit-learn, TensorFlow

- **Great for ML & AI 适合机器学习&人工智能**
  Ready-made libraries for advanced analytics
  为高级分析提供现成库
- **Easy Integration 易于集成**
  Connects with Hadoop, Spark, SQL, NoSQL, APIs
  与Hadoop、Spark、SQL、NoSQL和API连接
- **Readable & Beginner-Friendly 易读且适合初学者**
  Clean syntax, large supportive community
  句法简洁，拥有庞大的支持社区

Data Sources
数据来源

Types of Big Data. Structured Data 大数据类型 结构化数据

- Highly organized & schema-based
- 高度组织化且基于模式

- Stored in databases or spreadsheets
- 存储在数据库或电子表格中

- Examples: CRM data, financial records, HR databases
- 示例：客户关系管理（CRM）数据、财务记录、人力资源数据库

- Easy to query (e.g., SQL), fast analysis
- 易于查询（例如SQL），分析速度快

Types of Big Data. Unstructured Data 大数据类型 结构化数据

- No predefined model, diverse formats
- 没有预定义模型，格式多样

- Examples: Text (emails, social media), multimedia (images, videos), IoT sensor data
- 示例：文本（电子邮件、社交媒体）、多媒体（图像、视频）、物联网（IoT）传感器数据

- Challenges: Requires NLP, ML, and advanced tools for analysis
- 挑战：需要自然语言处理（NLP）、机器学习（ML）和高级工具进行分析

Types of Big Data. Semi-Structured Data 大数据类型 半结构化数据

- Flexible structure
- 灵活的结构

- Examples: Web data, emails, NoSQL databases
- 示例：网络数据、电子邮件、NoSQL数据库

- Balance: Flexibility + easier analysis than unstructured data
- 平衡：灵活性 + 分析难度低于非结构化数据

Data Sources. Structured Data 数据来源 结构化数据

- Relational Databases (e.g., SQL)
  关系型数据库（例如SQL）

- Spreadsheets (Excel, Google Sheets)
  电子表格（Excel、谷歌表格）

- ERP & CRM Systems (Salesforce, SAP)
  企业资源规划（ERP）和客户关系管理（CRM）系统（Salesforce、SAP）

- Financial Transactions
  金融交易

Data Sources. Structured Data. Example
数据来源 结构化数据 举例

📈 Spreadsheets 电子表格

A monthly sales report in Excel Excel中的月度销售报告

Columns: Date, Product_ID, Product_Name, Units_Sold, Revenue, Region

列：日期、产品ID、产品名称、销售数量、收入、地区

Date	Product_ID	Product_Name	Units_Sold	Revenue	Region
2025-01-01	101	Widget A	30	600	North
2025-01-01	102	Widget B	20	400	East
2025-01-02	101	Widget A	25	500	North
2025-01-02	103	Widget C	15	300	South

Data Sources. Unstructured Data 数据来源：非结构化数据


Social Media (Twitter, Facebook)

社交媒体（推特、脸书）


Emails & Docs (Office 365, Google Workspace)

电子邮件和文档（Office 365、Google Workspace）


IoT Devices (sensors, cameras)

物联网设备（传感器、摄像头）


Streaming Platforms (e.g., YouTube, Netflix)

流媒体平台（例如：YouTube、网飞）

Data Sources. Unstructured Data. Example
数据来源：半结构化数据 举例

Emails & Documents 电子邮件&文档

Email bodies, attachments (PDFs, Word files), meeting notes, and collaborative docs
电子邮件正文、附件（PDF、Word文件）、会议记录和协作文档

Subject: Urgent — Feedback on Q2 Financial Report

Body:

Hi Alex,

Thanks for sharing the Q2 draft. Overall, it looks solid — great improvement in the marketing ROI and cost efficiency.

However, a few things need attention:

- Slide 6: revenue forecast seems outdated
- Please double-check the numbers in the Asia-Pacific section
- Let's update the customer churn graph with latest retention data from CRM

I've also attached my comments as a PDF — feel free to edit directly.
Let's finalize by Thursday so we can circulate before the board meeting.

Best,
Jordan

Attachment: `Jordan_Comments_Q2.pdf`

主题：紧急 — Q2财务报告反馈
正文：
嗨，Alex,
感谢您分享Q2草稿。总体来看，它看起来很扎实——ROI和成本效率有了很大的提升。
然而，有几件事需要注意：
第6页：收入预测似乎过时了；
请再次核对亚太地区的数字；
让我们用自CRM的最新保留数据更新客户流失图表。
我还附上了我的评论作为PDF文件，您可以直接编辑。
让我们在周四之前定稿，以便在董事会会议前分发。

此致，
Jordan
附件：Jordan_Comments_Q2.pdf

Data Sources. Semi-Structured Data数据来源 半结构化数据

- Web APIs (JSON, XML) 网络接口（JSON, XML）
- Email Metadata

  电子邮件元数据
- NoSQL Databases (MongoDB, Cassandra)

  NoSQL数据库 (MongoDB, Cassandra)
- Logs & Clickstreams

  日志和点击流

Data Sources. Semi-Structured Data. Example
数据来源 半结构化数据 举例

🌐 Web APIs (JSON, XML) 网络接口（JSON, XML）

API responses from public services (weather, stock prices, maps), typically in JSON or XML format
来自公共服务（天气、股票价格、地图）的API响应，通常以JSON或XML格式

```
{
  "location": "Kyiv",
  "temperature_celsius": 18.3,
  "humidity": 72,
  "forecast": [
    {"day": "Monday", "high": 21, "low": 13},
    {"day": "Tuesday", "high": 20, "low": 12}
  ]
}
```

Benefits of Big Data Analytics 大数据分析的优势

- **Real-Time Intelligence 实时智能**
  Instant insights for faster decisions 实时洞察，加快决策速度

- **Better Decisions 更佳决策**
  Trends, patterns, correlations revealed 揭示趋势、模式和相关性

- **Cost Savings 成本节约**
  Efficiency, waste reduction, forecasting 提高效率，减少浪费，进行预测

- **Customer Engagement 客户参与**
  Behavior insights, personalized marketing 行为洞察，个性化营销

- **Risk Management 风险管理**
  Early threat detection, predictive models 早期威胁检测，预测模型

Challenges of Big Data 大数据的挑战

- Data Quality & Management 数据质量与管理
  Keeping data clean and connected across fast, complex sources
  在快速、复杂的数据源中保持数据的清洁和连接
- Scalability 可扩展性
  Storing and processing growing volumes of data in real time
  实时存储和处理不断增长的数据量
- Privacy & Security 隐私与安全
  Protecting sensitive data and meeting regulatory requirements
  保护敏感数据，满足监管要求
- Integration Complexity 集成复杂性
  Combining structured, unstructured, and semi-structured data
  结合结构化、非结构化和半结构化数据
- Skilled Workforce Shortage 专业人才短缺
  Finding professionals with data science and engineering skills
  寻找具备数据科学和工程技能的专业人员

Useful Links 实用链接

IBM. What is big data? 什么是大数据？

IBM. What is big data analytics? 什么是大数据分析？

Google. What is big data? 什么是大数据?

DOMO. Data never sleeps 数据永不休眠