Swiss UMEF

UNIVERSITY OF APPLIED SCIENCES INSTITUTE

Big Data Analytics

by

Oleh Tymchuk

Course Instructor

Oleh Tymchuk

- **Associate Professor of Taras Shevchenko National University of Kyiv**
- **PhD in IT**
- **Industry Experience:**
  Research Software Engineer, EPAM, Kyiv
  Backend Software Engineer, IMediaMAtch, Copenhagen
- **Instructor & Mentor of Python Data Analytics courses:**
  IT Career Hub, Berlin
  PROG Academy, Kyiv

Course Structure

Module 1. Introduction to Big Data

- What is Big Data?
- Exploring Careers in Big Data
- Data Sources

Course Structure

Module 1. Introduction to Big Data

Module 2. Introduction to Python

- What is Python?
- Python Interpreter
- IDEs (Jupyter Notebook, Google Colab)
- Python practice

Course Structure

Module 1. Introduction to Big Data

Module 2. Introduction to Python

Module 3. In-Memory Analytics with Pandas

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

Course Structure

Module 1. Introduction to Big Data

Module 2. Introduction to Python

Module 3. In-Memory Analytics with Pandas

Module 4. Efficient In-Memory Analytics with Polars

Course Structure

Module 1. Introduction to Big Data

Module 2. Introduction to Python

Module 3. In-Memory Analytics with Pandas

Module 4. Efficient In-Memory Analytics with Polars

Module 5. Big Data with Dask

Swiss UMEF

UNIVERSITY OF APPLIED SCIENCES INSTITUTE

Big Data Analytics

# 01: Introduction to Big Data

Instructor: Oleh Tymchuk

#01: Agenda

1. What is Big Data?
2. Exploring Careers in Big Data
3. Data Sources

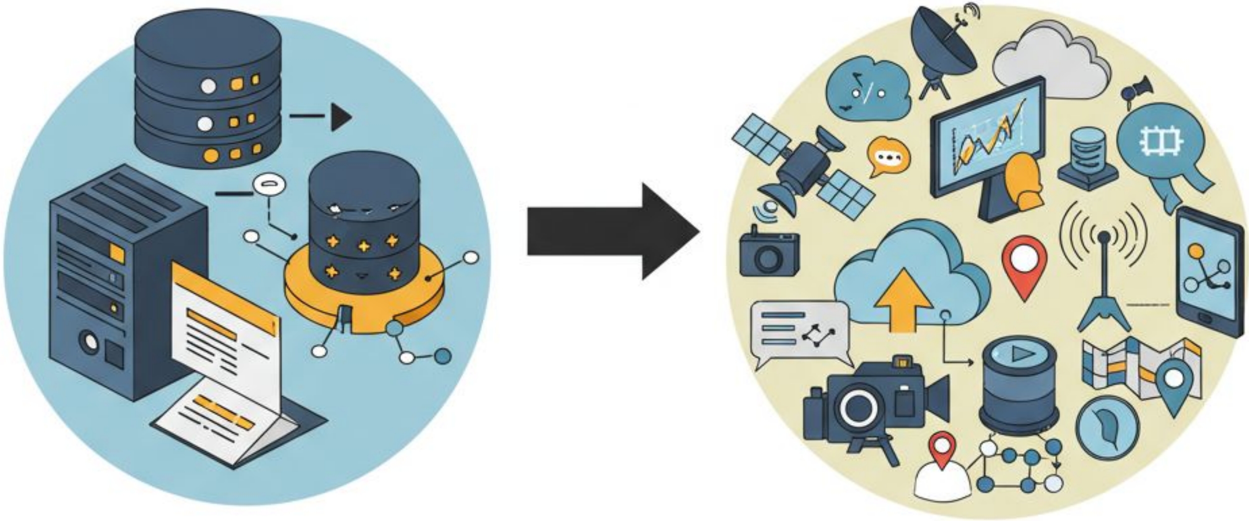
What is Big Data?

Definition

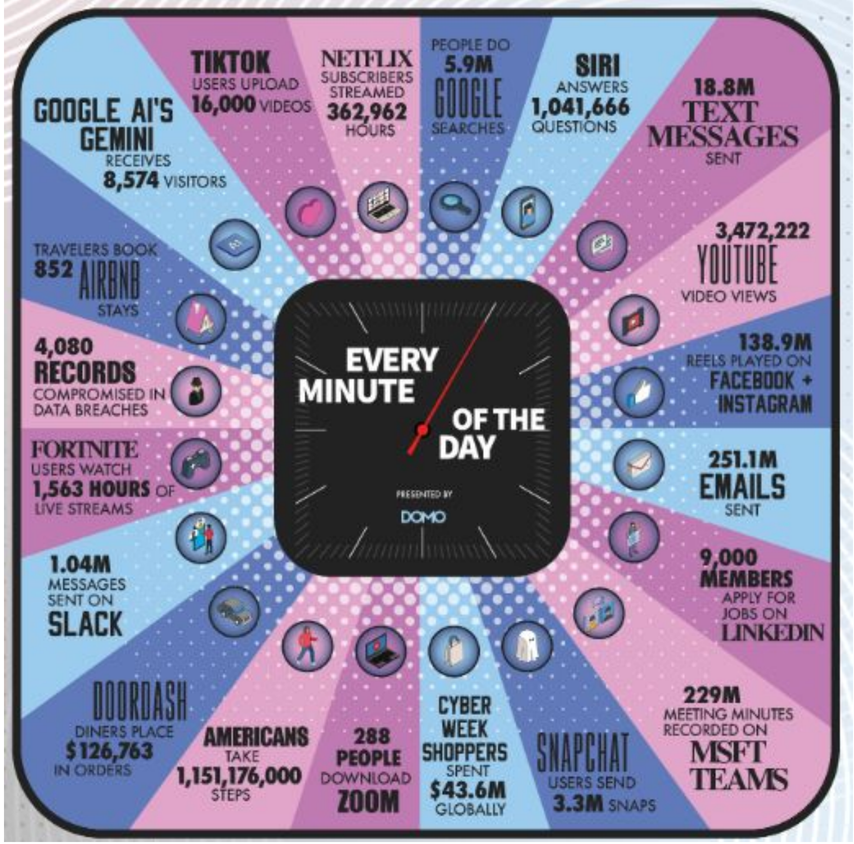
Big data refers to massive, complex data sets that traditional data management systems cannot handle.

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

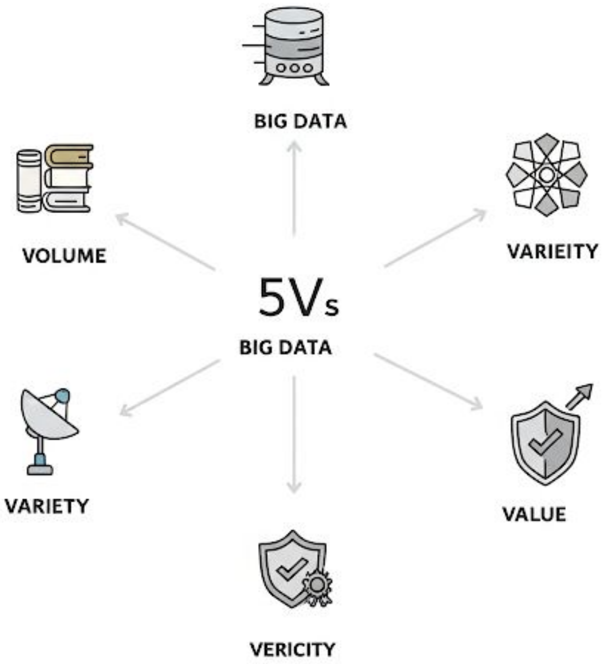
Historical evolution



Data never sleeps


EVERY MINUTE OF THE DAY
PRESENTED BY
DOMO
GOOGLE AI'S GEMINI RECEIVES 8,574 VISITORS
TIKTOK USERS UPLOAD 16,000 VIDEOS
NETFLIX SUBSCRIBERS STREAMED 362,962 HOURS
PEOPLE DO 5.9M GOOGLE SEARCHES
SIRI ANSWERS 1,041,666 QUESTIONS
18.8M TEXT MESSAGES SENT
3,472,222 YOUTUBE VIDEO VIEWS
138.9M REELS PLAYED ON FACEBOOK + INSTAGRAM
251.1M EMAILS SENT
9,000 MEMBERS APPLY FOR JOBS ON LINKEDIN
229M MEETING MINUTES RECORDED ON MSFT TEAMS
SNAPCHAT USERS SEND 3.3M SNAPS
CYBER WEEK SHOPPERS SPENT $43.6M GLOBALLY
288 PEOPLE DOWNLOAD ZOOM
AMERICANS TAKE 1,151,176,000 STEPS
DOORDASH DINERS PLACE $126,763 IN ORDERS
1.04M MESSAGES SENT ON SLACK
FORTNITE USERS WATCH 1,563 HOURS OF LIVE STREAMS
4,080 RECORDS COMPROMISED IN DATA BREACHES
TRAVELERS BOOK 852 AIRBNB STAYS

The 5Vs of Big Data



Real-world examples | business, science, government

Healthcare: analyzing millions of patient records to detect disease outbreaks early

Retail: Amazon tracks your clicks and recommends products in real time

Government: cities use sensor data to manage traffic, pollution, and public safety

Science: CERN generates petabytes of data from particle collisions in the Large Hadron Collider

Real-world examples | business, science, government

Healthcare: analyzing millions of patient records to detect disease outbreaks early

Retail: Amazon tracks your clicks and recommends products in real time

Government: cities use sensor data to manage traffic, pollution, and public safety

Science: CERN generates petabytes of data from particle collisions in the Large Hadron Collider

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

Exploring Careers in Big Data

Roles

🧑‍💻 Data Scientist

- 🔍 Analyze complex data
- 🧠 Build ML models
- 📊 Discover insights
- 🤝 Collaborate with stakeholders

Roles

🧱 Data Analyst

- 📈 Perform data analysis
- 📂 Clean and transform
- 📉 Identify trends
- 📝 Create dashboards

Roles

🧱 Data Engineer

🧱 Build data pipelines

⚙️ Manage infrastructure

🧪 Test data flows

🚀 Optimize performance

Roles

🤖 Machine Learning Engineer

- 🧠 Design ML algorithms
- 🛠️ Implement models
- 📦 Deploy solutions
- 🔁 Retrain and improve

Roles

📊 Business Intelligence Analyst

- 📊 Analyze business data
- 📌 Track KPIs
- 🧾 Generate reports
- 🎯 Support decisions

Roles

🧩 Data Visualization Specialist

- 🖼️ Create visual dashboards
- 🎨 Design clear charts
- 📤 Present insights
- 👀 Highlight patterns

Roles

🏛️ Data Architect

- 🧩 Design data systems
- 🏛️ Define structure
- 🔗 Ensure integration
- 🛡️ Enforce standards

Typical analyst’s toolset

🗃️ Data Storage & Query

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

Typical analyst’s toolset

🧮 Data Processing

- Python
- Pandas
- NumPy
- PySpark

Typical analyst’s toolset

📊 Visualization & BI

- Tableau
- Power BI
- Matplotlib
- Seaborn
- Plotly

Typical analyst’s toolset

📦 Machine Learning & Analytics

- Scikit-learn
- TensorFlow

Typical analyst’s toolset

🛠️ Collaboration & Versioning

- Git
- Jupyter Notebooks, Google Colab

Why Python is useful for Big Data analytics

- **Versatile & Powerful**
  Works for data processing, analysis, visualization, ML

- **Rich Ecosystem**
  Pandas, NumPy, PySpark, Scikit-learn, TensorFlow

- **Great for ML & AI**
  Ready-made libraries for advanced analytics

- **Easy Integration**
  Connects with Hadoop, Spark, SQL, NoSQL, APIs

- **Readable & Beginner-Friendly**
  Clean syntax, large supportive community

Data Sources

Types of Big Data. Structured Data

- Highly organized & schema-based
- Stored in databases or spreadsheets
- Examples: CRM data, financial records, HR databases
- Easy to query (e.g., SQL), fast analysis

Types of Big Data. Unstructured Data

- No predefined model, diverse formats
- Examples: Text (emails, social media), multimedia (images, videos), IoT sensor data
- Challenges: Requires NLP, ML, and advanced tools for analysis

Types of Big Data. Semi-Structured Data

- Flexible structure
- Examples: Web data, emails, NoSQL databases
- Balance: Flexibility + easier analysis than unstructured data

Data Sources. Structured Data

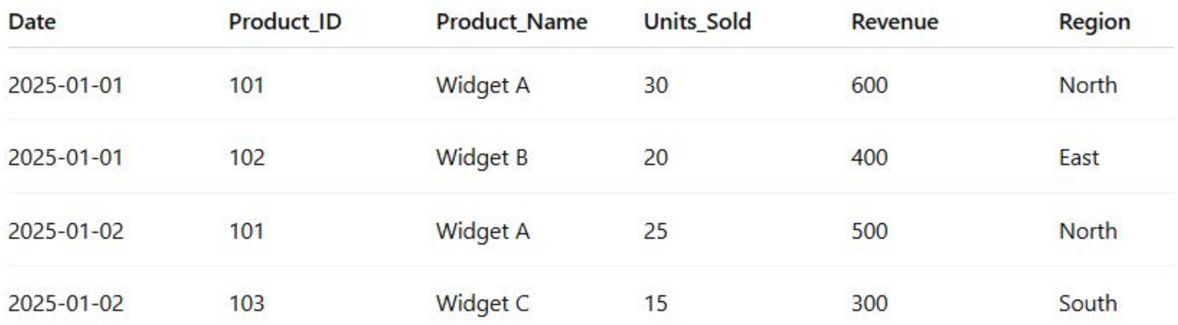
- 🗂️ Relational Databases (e.g., SQL)
- 📊 Spreadsheets (Excel, Google Sheets)
- 🧾 ERP & CRM Systems (Salesforce, SAP)
- 💳 Financial Transactions

Data Sources. Structured Data. Example

📈 Spreadsheets

A monthly sales report in Excel

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



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

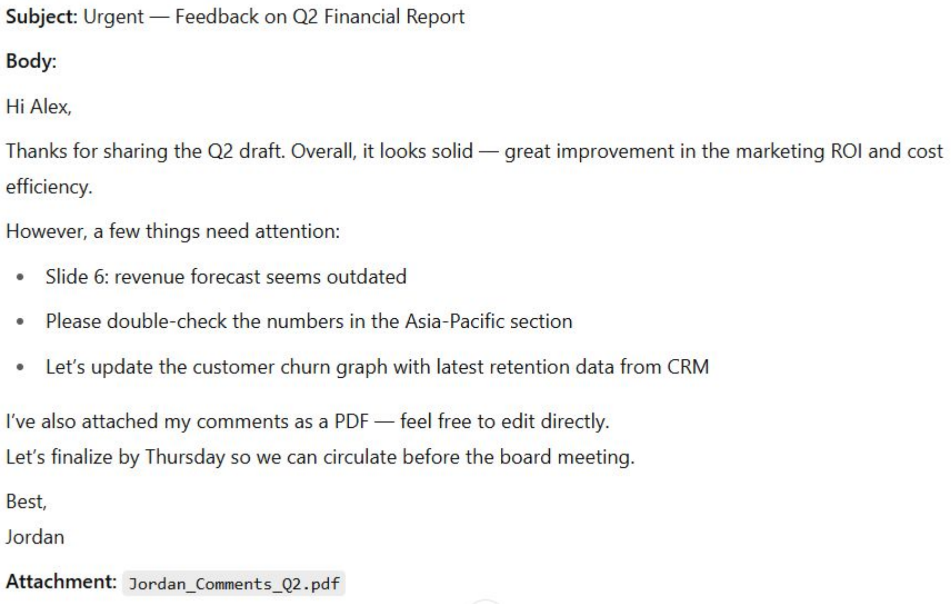
Data Sources. Unstructured Data

- 📱 Social Media (Twitter, Facebook)
- 📧 Emails & Docs (Office 365, Google Workspace)
- 📡 IoT Devices (sensors, cameras)
- 🎥 Streaming Platforms (e.g., YouTube, Netflix)

Data Sources. Unstructured Data. Example

📄 Emails & Documents

Email bodies, attachments (PDFs, Word files), meeting notes, and collaborative docs



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`

Data Sources. Semi-Structured Data

🌐 Web APIs (JSON, XML)

✉️ Email Metadata

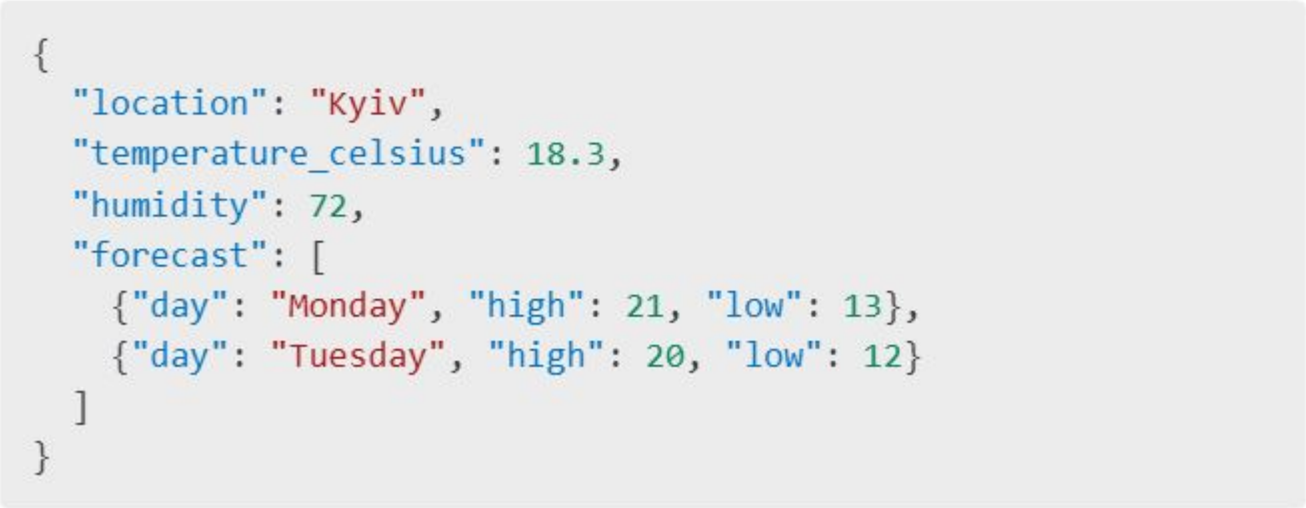
🗃️ NoSQL Databases (MongoDB, Cassandra)

📊 Logs & Clickstreams

Data Sources. Semi-Structured Data. Example

🌐 Web APIs (JSON, XML)

API responses from public services (weather, stock prices, maps), typically in JSON or XML format



```
{
  "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

- Real-Time Intelligence
  Instant insights for faster decisions
- Better Decisions
  Trends, patterns, correlations revealed
- Cost Savings
  Efficiency, waste reduction, forecasting
- Customer Engagement
  Behavior insights, personalized marketing
- Risk Management
  Early threat detection, predictive models

Challenges of Big Data

- Data Quality & Management

  Keeping data clean and connected across fast, complex sources

- Scalability

  Storing and processing growing volumes of data in real time

- Privacy & Security

  Protecting sensitive data and meeting regulatory requirements

- Integration Complexity

  Combining structured, unstructured, and semi-structured data

- Skilled Workforce Shortage

  Finding professionals with data science and engineering skills

Useful Links

IBM. What is big data?

IBM. What is big data analytics?

Google. What is big data?

DOMO. Data never sleeps