TECHNICAL DATA ANALYSIS

General notes, assessement criteria. Introduction to ETL. Tools.

LST ILS

Call me

Patryk Hubar-Kołodziejczyk

Duty hours: Tuesdays, 9:30-11 (email me first)
Contact: p.hubar@uw.edu.pl
Other resources:

- https://www.wdib.uw.edu.pl/pracownicy/biogramy-i-dyzury/90-pracownicy/specjalisci-naukowo-techniczni/3106-mgr-patryk-hubar
- https://twitter.com/patt_hub
- https://orcid.org/0000-0001-5582-2042

Assessement criteria

Attendance

- dozwolone 3 nieusprawiedliwione nieobecności

In-class group assignments

- After each class
- 40%

Final assigment

- Individual
- 60%

Assessement criteria

Grading system:

- 0-50% - **2**
- 51-65% - **3**
- 66-79% - **4**
- 80%-95% - **5**
- 96%-100% - **5!**

Organisational and introductory classes

COURSE PLAN

- preparation of the development environment, installation of packages
- ETL approach
  - Command-line interface
  - Bash scripting
  - Regular expressions,
  - Introduction to database management

Questions?

The ETL Process Explained

Extract

Retrieves and verifies data from various sources

Transform

Processes and organizes extracted data so it is usable

Load

Moves transformed data to a data repository

EXTRACT, TRANSFORM, LOAD

In computing, extract, transform, load (ETL) is a three-phase process where data is extracted from an input source, transformed (including cleaning), and loaded into an output data container [wikipedia].

ETL is a process in data engineering that gathers data from various sources, transforms it into the required format, and loads it into a target system (e.g., data warehouse).

Essential for data integration, analytics, and business intelligence.

EXTRACT, TRANSFORM, LOAD

- **Goal:** Gather data from heterogeneous sources (databases, files, APIs, etc.)
- **Key Tasks:**
  - Connect to data sources
  - Select relevant data
  - Ensure data quality and consistency
- **Challenges:**
  - Data format variations
  - Incomplete or inconsistent data

EXTRACT, **TRANSFORM**, LOAD

- **Goal**: Prepare data for the target system
- **Key Tasks:**
  - Data cleaning: Remove duplicates, handle missing values
  - Data aggregation: Summarize or group data
  - Data mapping: Align data from different sources
  - Normalization: Ensure consistency in formats (e.g., dates, units)
  - Data enrichment: Add calculated fields or new attributes
- **Challenges:**
  - Complex transformation logic
  - High computational costs for large data volumes

EXTRACT, TRANSFORM, **LOAD**

- **Goal:** Load transformed data into the target system (e.g., database, data warehouse).
- **Key Tasks:**
  - Full load: Load entire datasets.
  - Incremental load: Load only new or updated data.
- **Challenges:**
  - Optimizing load times.
  - Managing system resources during large-scale operations.

Questions?

Task one

Please set up individual accounts on Github, complete the form below and follow https://github.com/patthub

https://docs.google.com/spreadsheets/d/1KZKt-Tu4C5ejvnQUv5hbG_jqMQdfExnNHR9eCjBMaP4/edit?usp=sharing

PROGRAMMING ENVIRONMENT

- Command-line console/terminal
- IDE
  - PyCharm
  - VS Code
  - Spyder
- Text editors
  - Sublime Text
- Notebooks
  - Jupyter Notebooks
  - **Google Colabolatory (Colab)**

GOOGLE COLABOLATORY (COLAB)

https://colab.research.google.com