Project 1. Data Architecture & Modeling

Introduction

The purpose of this project is to practice **core data engineering skills**: combining datasets, designing a **data architecture** to support ingestion and quality, and creating a **dimensional model (star schema)** for analytics. The project links business needs to technical implementation, helping you understand how data architecture and modelling enable analysis. The project does *not* involve a physical implementation of the data architecture and model – yet. This will be handled in the next project.

Requirements

1. Business Brief
Your project must start with a short business brief (max 1 page) including:

- Objective: one-sentence project goal
- Stakeholders: who will use the results (e.g., finance team, policymakers)
- Key Metrics (KPIs): 2–3 relevant measures
- Business Questions: at least 5 questions answerable with your model

2. Datasets

- Use at least two distinct datasets, each with ≥ 1000 rows and ≥ 8 columns.
- Sources:
    - Open data portals (e.g., data.gov, Eurostat, Estonian open data)
    - Kaggle datasets
    - APIs (e.g., Spotify, Twitter/X, WHO, World Bank)
- Themes could include: finance, public health, mobility, e-commerce, sports, environment, etc.

3. Tooling
The project does not involve implementation, but to the best of your knowledge, indicate which tools that we will cover in the course you would use in which stage of the data engineering lifecycle.

You can find the tools listed under practice sessions here:
https://courses.cs.ut.ee/2025/dataeng/Main/Lectures

4. Data Architecture
Provide a diagram showing:

- Data flow (source → ingestion → storage → warehouse → reporting)

- Ingestion method (API pull, file load, streaming, etc.)
- Update frequency (hourly, daily, weekly, monthly?)
- At least one example data quality check (e.g. null check, uniqueness check)

5. Data Model

- Design a star schema with:
    - ≥ 1 fact table (state the grain clearly)
    - ≥ 3 dimension tables
- For each dimension, justify your choice of Slowly Changing Dimension (SCD) type:
    - Type 1, Type 2, or Static (with reasoning)

6. Data Dictionary

- Provide table & column descriptions, including data types.

7. Demo Queries

- Write SQL queries answering the business questions from your brief.

Deliverables

- Report.pdf (main submission)
    - Include each group members' roles and contribution (% per group member)
    - LLM disclosure (links to any AI/LLM chats used for this project)
- GitHub link with README and any relevant DDL, DML or sample data
    - Can be pseudocode SQL for creating tables in star schema, answering business questions

---

Grading (15 points)

- Business brief & questions – 2p
- Datasets & tooling choice – 2p
- Data architecture – 3p
- Data model, grain & SCD justification – 4p
- Data dictionary – 2p
- Demo queries – 2p

Other

You can use datasets from your company or other courses (if allowed there). Ideally, the datasets should have relatively high update frequency (e.g, daily/weekly or even near real-time), and high granularity (APIs, sensor data. Open data is often pre-aggregated, which can make it difficult for you to fulfil some of the requirements of this and next projects).

It is OK to use synthetic data for one of the datasets, but you will need to ground this on research and make it as real-life as possible.

If you have any general clarifying questions (relevant for everyone), please ask them in the Moodle Q&A forum.

If you have any specific questions (relevant for your group), please send an e-mail Kristo.raun@ut.ee

Timeline

Datetime	Event
2025-09-15	Project 1 published
2025-09-28 23:59	Indicate any questions that you may have
2025-10-05 23:59	Project 1 submission deadline

The project will be followed by peer grading, where you will assess the work of 2 other groups.