ID2221 Data-Intensive Computing Proposal

Ahmad Al Khateeb
Kusumastuti Cahyaningrum
Aleksandra Burdakova

September 2025

1 Problem Description

• Problem Description: What problem will you be investigating?
The problem being investigated is:

> How to efficiently manage and analyze warehouse stock data in a batch data engineering environment?

Specifically, the project will focus on combining and analyzing transactional stock data from multiple geographically distributed warehouses. By using Apache Spark, we will ingest and process this raw data into a central database and derive meaningful insights, such as:
We will then build an interactive dashboard to display:

- Daily item trends.
- Warehouse-level stock balances.
- Receipts and issues trends over time.

Additionally, we plan to incorporate a machine learning component to project future sales trends using regression techniques.

2 Tools

In our project, we will utilize several key tools. Each tool serves a specific purpose as follows:

- MariaDB: for storing the aggregated transactional data.
- Spark: for processing and transforming raw data into meaningful insights.
- Kafka (optional): if needed for handling data streams, though not strictly necessary in our batch scenario.

- Redis: for caching intermediate results if needed during data analysis.
- Jupyter Notebook: for exploratory data analysis, visualization, and machine learning experimentation.
- matplotlib/plotly: for visualization.

3 Data

We will generate synthetic data using Python scripts to simulate stock transactions across multiple warehouses over a period of time. The generated data is inserted directly into the MariaDB table through SQL queries executed in Python. Transactional warehouse stock records, with the following attributes:

- STOCK_DATE: Date of stock entry.
- WAREHOUSE_ID: Unique ID of the warehouse (e.g., Stockholm, Gothenburg, Malmo).
- ITEM_NAME: Name of the inventory item (e.g., Tape Dispenser, Calculator, Notebook).
- OPENING_STOCK: Quantity available at the start of the day.
- RECEIPTS: Quantity received (inbound stock).
- ISSUES: Quantity issued (outbound stock).
- UNIT_VALUE: Price per unit of the item.

4 Methodology

We plan to use Apache Spark for data transformation and aggregation, followed by visualization through selected tools. Finally, the final products can be either:

- Transaction/Product Trend Dashboard
- Machine Learning to project future sales trend of items using Regression.