Semester Project Part 1 (Volume and Velocity)

Big Data and Semi-/Unstructured Data

Marc Schaaf

HS2025

1 Task

The project work for the first half of the lecture is subdevided into two tasks:

1.0.1 Task 1.A: Towards Volume

a) For an artificial company define a *small* normalized (3rd NF/ redunancy free) relational schema (not more then 5 tables) suitable for sample operational workloads.
b) Define two analytical questions that your imaginary company needs answered.
c) Define a mapping of the normalized database schema to a de-normalized form suitable for the analysis. The later one can be based on MongoDB or a relational DBMS like MySQL.
d) Implement the needed analysis queries either with MongoDB or a relational DBMS.

1.0.2 Task 1.B: Velocity

a) Define a *small* set of event types that could be generated in your scenario which would be relevant to the analytical questions from Task 1.A.
b) Implement a basic event generator that simulates the creation of the defined events and publishes them to Apache Kafka (sample code is available on Moodle).
c) Implement a basic stream processing for your events based on Apache Kafka Streams.

Note 1: The following two tasks closely follow the structure of the guided exercises which can thus be used as a foundation / inspiration for your project work.
Note 2: Even though the discussed principles are for the processing of large amounts of data, please use only very small datasets as we have only limited capacity on the demo systems.

2 Requirements

- The analytical processing must be non-trivial. E.g. a simple count of the number of rows will not do.
  - Must contain some form of aggregation based on a grouping
  - Should contain a join like operation
- A description of the analytics scenario is mandatory
  - What do you plan to achieve and what would the potential value for a business

  - Aspects:
    * Stakeholders: Who are the Stakeholders and what are their analytical/information needs?
    * Triggers: in which context(s) will stakeholders need to access the processing results?
    * Questions: what analytical needs / questions do the stakeholders have in the described context(s)? Please formulate these needs in the form of questions!
    * Expected result: which type of output will satisfy the stakeholders' information needs? Outline the results by providing mock results in the desired form.

3 Group Work

The project is to be conducted as a project group. All participants of the group have to contribute to all parts / need to fully understand all parts of the project. Participation will be verified through questions during the final presentation.

4 Initial Meeting

- Initial Meeting: Present/outline your analytics process idea to get feedback of its suitability for the project. Include details on
  - What is the scenario
  - What kind of data and how it will be obtained
  - What kind of processing technology is to be used
- Further coaching is of course available on demand during the lecture.

5 Presentation:

- Result Presentation: Present your finished project
  - Scenario / Data / Technology
  - Running analytics task
  - Results of the analytics and discussion of the findings
  - Short reflection: central challenges

6 Submission of the Results

- Deadline: See semester plan on Moodle
- Submission of the code / documentation via Moodle or Switch Drive if too large for Moodle (link is available on Moodle)