University of Basel

Prof. Dr. Heiko Schuldt
Sein Coray, M. Sc.
Silvan Heller, M. Sc.
Loris Sauter, M. Sc.
Alexander Stiemer, M. Sc.

# Databases Autumn 2019

## Data Analysis Project

In the data analysis project, the concepts from the database lecture are applied to an end-to-end scenario. In this hands-on exercise, you are required to perform the most significant steps in integrating multiple data sources to one unified data store which is later on used for analysis purposes. The project is divided in multiple steps:

**Part 0: Project idea.** Select multiple data sources and pose analysis goals.

**Part 1: Schema integration.** Analyze your data sources and design an integrated schema.

**Part 2: Data integration.** Integrate your data sources into a unified and normalized database.

**Part 3: Analysis and visualization.** Use the newly integrated data for performing analyses and creating visualizations.

# Organization

### Schedule

| Date | Phase |
|---|---|
| Thu., 26.09.2019 | Introduction to the project |
| Sun., 06.10.2019 | Hand-in of project ideas |
| 22.10., 23.10. & 24.10.2019 | Plenary presentation of project ideas |
| Sun., 03.11.2019 | Hand-in of integrated schema |
| Tue., 05.11. & Thu., 07.11.2019 | Presentation of integrated schema |
| Sun., 08.12.2019 | Hand-in of data integration methods |
| Tue., 10.12. & Thu., 12.12.2019 | Presentation of data integration methods |
| Sun., 12.01.2020 | Final hand-in |
| Tue., 28.01.2020 | Plenary presentation of full project |

**Groups** The project **must** be solved in groups of **2 students**.

**Assessment** The project is assessed with a grade. The lecture's final grade is the weighted sum of $\frac{1}{3}$ project grade and $\frac{2}{3}$ exam grade.

**Infrastructure** Deliverables are to be hand-in via ADAM[1]. The groups are given a Gitlab-repository. It is expected that any files related to the project are at any time accessible by the provided repository. Furthermore, it is to be expected that the assistants have access to the databases used. Please note, that you must not use the exercise database for this project, as these might be erased after the exercise. There are no provided databases servers.

# Details

## P0: Project Idea

**Task:** Select at least 3 interesting data sources with different formats (e. g., JSON files, a relational database dump and a CSV), sizes and from different sources, etc. We may force you to revise your project proposal if the combination of diversity and size in your data sources is not sufficient.

You may find interesting data on the following websites:

- `https://opendata.swiss`
- `https://opendata.swisscom.com`
- `https://data.sbb.ch/pages/home20/`
- `https://open-data.europa.eu`
- `https://www.data.gov`
- `https://github.com/BaselHack/BaselHack2017/tree/master/data`[2]
- `https://www.ukdataservice.ac.uk/get-data/other-providers/open-data`
- `https://www.reddit.com/r/datasets/`
- `https://www.kaggle.com`
- `https://www.cooldatasets.com/`

Come up with interesting analysis goals (i. e., research questions) which you want to be able to answer by combining the data sources. Make sure to pose your goals precisely as well as the expected data size after the integration.

**Example:** You may integrate crime data of the canton Basel City with restaurant reviews from a review website together with geographic information. Your analysis could focus on

[1] ADAM: `http://adam.unibas.ch`
[2] BaselHack: `https://www.baselhack.ch/`

whether restaurants in areas with a high criminality have lower review scores or whether restaurants which are closer to public transport have better scores.

**Hand-in:** Project proposal including details on languages and tools, database to be used, data sources and their size and analysis goals. Use the provided template to submit your proposal on ADAM.

To set up the GitLab access, each student needs to log in at least once into SciCore GitLab[3] in order to get their account created. To register your group and receive access to the GitLab repository send an email to `s.coray@unibas.ch` listing all group member's emails and GitLab usernames.

**Presentation:** Present your project idea (or the revision thereof), in particular your data sources inclusive size and your analysis goals in a $5-10$ minutes plenary presentation during the exercise slots / lecture.

## P1: Schema Integration

**Task:** Analyse your data sources and create entity relationship (ER) diagrams of the single sources representing their as-is state before integration. Create an integrated conceptual schema combining the various sources without dropping any data of the original sources. Derive a logical, relational schema from the conceptual schema.

**Hand-in:** ER diagrams of the single sources representing their as-is state and of the integrated schema on ADAM.

**Presentation:** Present the schemas in a 5 minutes presentation to the assistants during the exercise slots.

## P2: Data Integration

**Task:** Integrate the selected data sources into a single relational database without dropping any data. Make sure to create all necessary index structures, clean and normalize the data to a unified format.

**Hand-in:** Push code to the provided git repository as well as guide to replay the integration step-by-step and provide the required SQL code to create the schema.

[3] SciCore Gitlab: `https://git.scicore.unibas.ch/`

**Presentation:** Present the methods applied for the data integration in a 5 minutes presentation to the assistants during the exercise slots.

## P3: Analysis and Visualisation

**Task:** Make use of analysis queries and visualizations to achieve your analysis goals you formulated in the first part of the project.

**Hand-in:** See final hand-in section.

**Presentation:** Present the full project in a 10 minutes plenary presentation describing your project and the methods used while having a strong focus on the analysis and the visualization.

## Final Hand-in

Your final hand-in should contain the following parts:

### Report

- Describe the problem setting, your analysis goals and the single sources in detail. Provide ER diagrams of the single data sets.
- Describe the steps occurring for integrating the schema and the data. Provide an ER diagram together with a logical schema of the integrated database. Detail possibly any cleaning, normalization, unification and transformation of the data. After all, the logical and physical schema should match.
- Describe the SQL queries required for your analysis. Give an overview of the visualization techniques used.
- Include screenshots / plots from your analysis in the report.
- Make sure to use the LaTeX template provided.
- The report has to be handed-in via ADAM.

### Code

- Any "source code" produced in the process (ad-hoc code, SQL queries, project files needed by external tools, etc.) has to be pushed to the provided git repository.

### Presentation

- Hand in the final plenary presentation via ADAM.

### Data

- Provide a (standard) SQL dump of your data via SWITCHfilesender[4] or give the assistants access to your DB.

## General requirements

- When specified two dates in the schedule, you are expected to be able to present (and to be present) on both dates. The order of presentations is announced by the assistants on the day of the presentation. Inform your assistants early in advance if you are unable to attend one of the two exercise slots.
- You are expected to use the provided git repository for your source code. Do not commit data sets! Any other document (except for the dump in the final hand-in) should be handed-in via ADAM.
- The solutions must be clearly readable. This refers in particular to diagrams describing conceptual schemas. Please do use a drawing tool (no pencil and paper) and please do adhere to standard ER diagram syntax as taught in the lecture.
- You are required to request the usage of any non-standard library and / or tool which is not already on the whitelist, published on ADAM. Non-standard refers to library-specifics, for instance you are not required to ask for permission to use the Java Standard Edition (but if you would like to use third party library *xy*). This also includes commercial tools mentioned in the next point.
- You are allowed to use commercial tools to analyse / transform the data and to make use of further external data sources for the purpose of cleaning the provided data, enriching your data, etc.

## Common Pitfalls

Do not try to solve this assignment close to last minute, and focus on the quality of your results. In particular, when working with large data sets, do not expect commands to finish almost instantly, although you can try to speed up your commands by using index structures in your database and applying batch updates.

Keep in mind the following hints:

- If possible test your code first on a smaller dataset and verify the changes, before running it on the full dataset.
- Use *prepared statements* to largely speed up your code.

```
String update = "UPDATE table SET sum = sum + ? WHERE name = ?";
PreparedStatement statement = con.prepareStatement(update);
statement.setInt(1, 10);
statement.setString(2, "Databases");
```

[4] SWITCHfilesender: `https://www.switch.ch/de/services/filesender/`

- Use *batch* operations to further speed up your code. Instead of

```
for (String query : queries) {
    statement.execute(query);
}
```

  it is better to write

```
for (String query : queries) {
    statement.addBatch(query);
}
statement.executeBatch();
```

- Temporarily remove foreign key constraints, as they need additional checking at insertion time. However, do not forget to enforce foreign constraints again after inserting the tuples.
- Think about methods for improving the performance of queries in relational databases.
- When working with Java, use the JVM option `-Xmx` to indicate how much memory Java is allowed to use (note that there is no space between the option and the memory indication, e. g., `-Xmx2g`).

# Contact

If there are any questions, feel free to ask the responsible assistant first:

`Sein Coray <s.coray@unibas.ch>`

Or eventually the other assistants:

`Silvan Heller <silvan.heller@unibas.ch>`

`Loris Sauter <loris.sauter@unibas.ch>`

`Alexander Stiemer <alexander.stiemer@unibas.ch>`