Big Data Assignment 4: *Spark SQL, Dataframes and Datasets*

This assignment is to make you familiar with the Spark `Dataset` and related older `Dataframe` (now equivalent to `Dataset[Row]`) APIs, as well as the Spark SQL front-end with query optimizer built on top of these APIs.

In part I, we will be introducing dataframes together with using Apache's framework Sedona (formerly GeoSpark) that is built to process geospatial data. In part II, we will show you an example on how to use UDFs (user-defined functions). This is useful when Spark SQL's built-in functions do not suffice anymore and you'd like more freedom in processing your dataframes.

General instructions

Continue in Apache Zeppelin in the `big-data` container. The exercises can be found in the attachments on Brightspace. How to open the Zeppelin notebooks is written in previous assignment.

The deadline for this assignment is **April 23rd, 23:59**. Please hand in the Exercises notebook (`.zpln`) after filling it in.

After handing in the assignment, you can go to Brightspace Quizzes to test if you understood the assignment. Making the quiz is voluntary, but recommendable, as you can check your answers from the assignment there. If you have any questions (you find out your answer is incorrect and you do not know why), do not hesistate to come discuss it with us in the practicum (or send us an email / contact us in the matrix room, if going to campus is not an option for you).