

## 3. Übung Big Data II

Sommersemester 2024

### 1 Visualisierung mit D3

Bearbeiten Sie das (falls noch nicht geschehen) das Lab 4 unter <https://bna2022.github.io/BigData/docs/labs/lab4.html>.

### 2 SQL und Notebooks

In der Vorlesung haben wir grundlegende SQL Befehle wiederholt. Das folgende Notebook gibt ein Beispiel, wie man innerhalb Jupyter Notebooks Datenbankumgebungen nutzen kann. Die verwendete Datenbank DuckDB (<https://duckdb.org/>) lässt sich ohne großen Aufwand einbinden. Daten (DB, CSV, Parquet etc.) können eingelesen und weiterverarbeitet werden. Die verarbeiteten Daten können in einer lokalen Datenbank abgelegt werden.

- Laden Sie das Notebook DuckDB-Aufgabe.ipynb aus dem Moodle herunter. Lösen Sie die noch offenen Aufgaben im Notebook.

### 3 Window Functions in SQL

- Analysieren Sie die folgende Tabelle mittels Windowfunctions. Sie können die Abfragen im DB Fiddle ausführen (<https://www.db-fiddle.com/>).

```
WITH Mitarbeiter (persnr, name, stufe, raum, gehalt, steuerklasse) AS (
VALUES (2125, 'Dumais', 'EG10', 226, 85000, 1),
(2126, 'Franklin', 'EG10', 232, 80000, 3),
(2127, 'Tsukamoto', 'EG11', 310, 65000, 5),
(2128, 'Meitner', 'EG10', 250, 85000, 1),
(2133, 'Hamilton', 'EG11', 52, 68000, 1),
(2134, 'Lovelace', 'EG11', 309, 55000, 5),
(2136, 'Curie', 'EG10', 36, 95000, 3),
(2137, 'Noether', 'EG10', 7, 98000, 1)
)
```

1. Geben Sie alle Werte aus und fügen Sie zusätzlich das Durchschnittsgehalt zu jedem Mitarbeiter hinzu.
2. Ermitteln Sie für jeden Mitarbeiter das Durchschnittsgehalt aller Mitarbeiter partitioniert nach Stufe.

3. Geben Sie die wachsende Summe (das Quantil) des Gehaltes aller Mitarbeiter partitioniert nach Stufe und absteigend sortiert nach ihrem Gehalt. Gleich verdienende Mitarbeiter sind im selben Quantil.
4. Errechnen Sie das gleitende Durchschnittsgehalt aus den 2 weniger und 2 mehr verdienenden Mitarbeiter in Bezug auf den aktuellen Mitarbeiter.
5. Ermitteln Sie die drei bestverdienenden Mitarbeiter.

## 4 Rekursion in SQL (optional)

Die folgende Tabelle enthält die Fahrdauern und Teilstrecken einer Ingolstädter Buslinie. Teilstrecken sind dabei die Abschnitte zwischen zwei benachbarten Haltestellen.

von	nach	zeit
ZOB	Technische Hochschule	2
Technische Hochschule	Stadttheater	4
Stadttheater	Rathausplatz	4
Rathausplatz	Brückenkopf	5

- Geben Sie ein SQL-Statement an, dass alle möglichen Fahrstrecken und die dazugehörigen Fahrzeiten angibt. Die Haltezeit pro Haltestelle beträgt 2 Minuten.
- Welche Auswirkung hätte es, wenn man eine Teilstrecke Technische Hochschule → ZOB einfügt?
- Falls das bisherige Statement zur Fahrstreckengenerierung ein Problem hat, wie lässt sich dieses beheben?