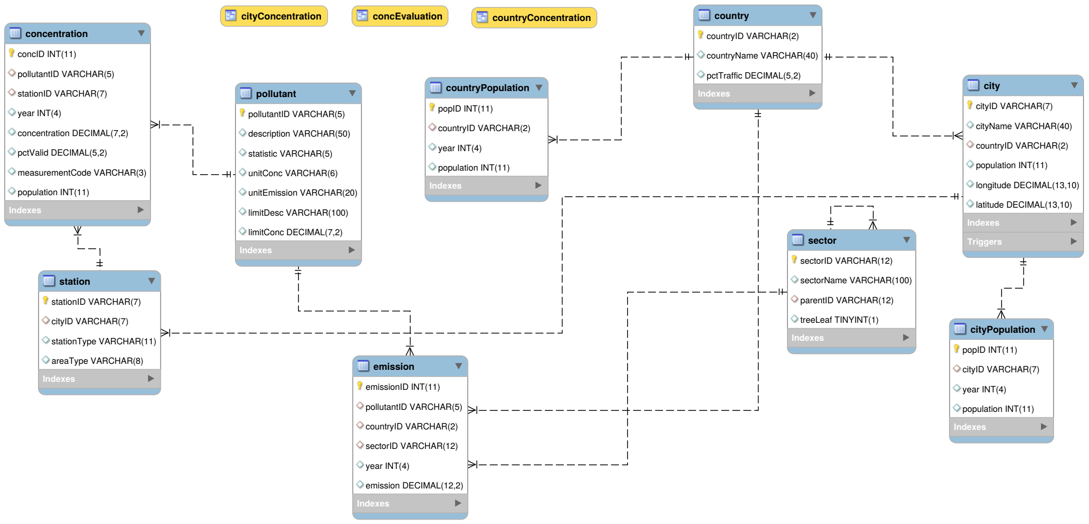
Milestone III - Database

Computing Project 2016/17
14D003/14D004

Carlos Isaac Rodriguez Prado
Hans-Peter Höllwirth
Veronika Kyuchukova

The installation script (setup.sh install) sets up the database in several steps: [1] Setting up the tables and other database elements, [2] migrating the data from the source files, and [3] optimizing the performance of the database.

[1] The entity relationship diagram of the normalized database *airpollution*, created with script *ddl_performance.sql*, looks as follows:



[2] The database combines datasets from 5 different sources. The air pollution data comes in two Excel files (separate file for 2013 data) and is loaded into the database with R scripts *migrateConc2012ToDB.R* and *migrateConc2013ToDB.R*, populating tables ***country***, ***city***, ***station***, and ***concentration***. Data inconsistencies (e.g. different name formats) are fixed upon insertion, using database triggers. City geo data for map visualizations is appended with R script *migrateGeoDataToDB.R*. A separate data file provides both annual country and city population counts. R script *migratePopulationsToDB.R* loads this dataset into tables ***countryPopulation*** and ***cityPopulation***. Finally, the national annual emission data by sectors is loaded into tables ***sector*** and ***emission*** with R script *migrateEmissionsToDB.R*. Data migration takes roughly 1 minute on a t2.medium instance.

[3] In a final step, the installation script executes *ddl_performance.sql* which contains table indexes (in order to avoid regular full table scans) and general purpose views for regression analysis (such as aggregating pollution data to city or country level).