Practical Assignment of Advanced Topics in Databases (2023-24)

The practical assignment involves several of the topics involved in the subject of Advanced Topics in Databases, namely:

- APIs to database management systems;
- Spatial databases, spatial types and spatial functions;
- Visualizations and animations involving Python and Matplotlib.

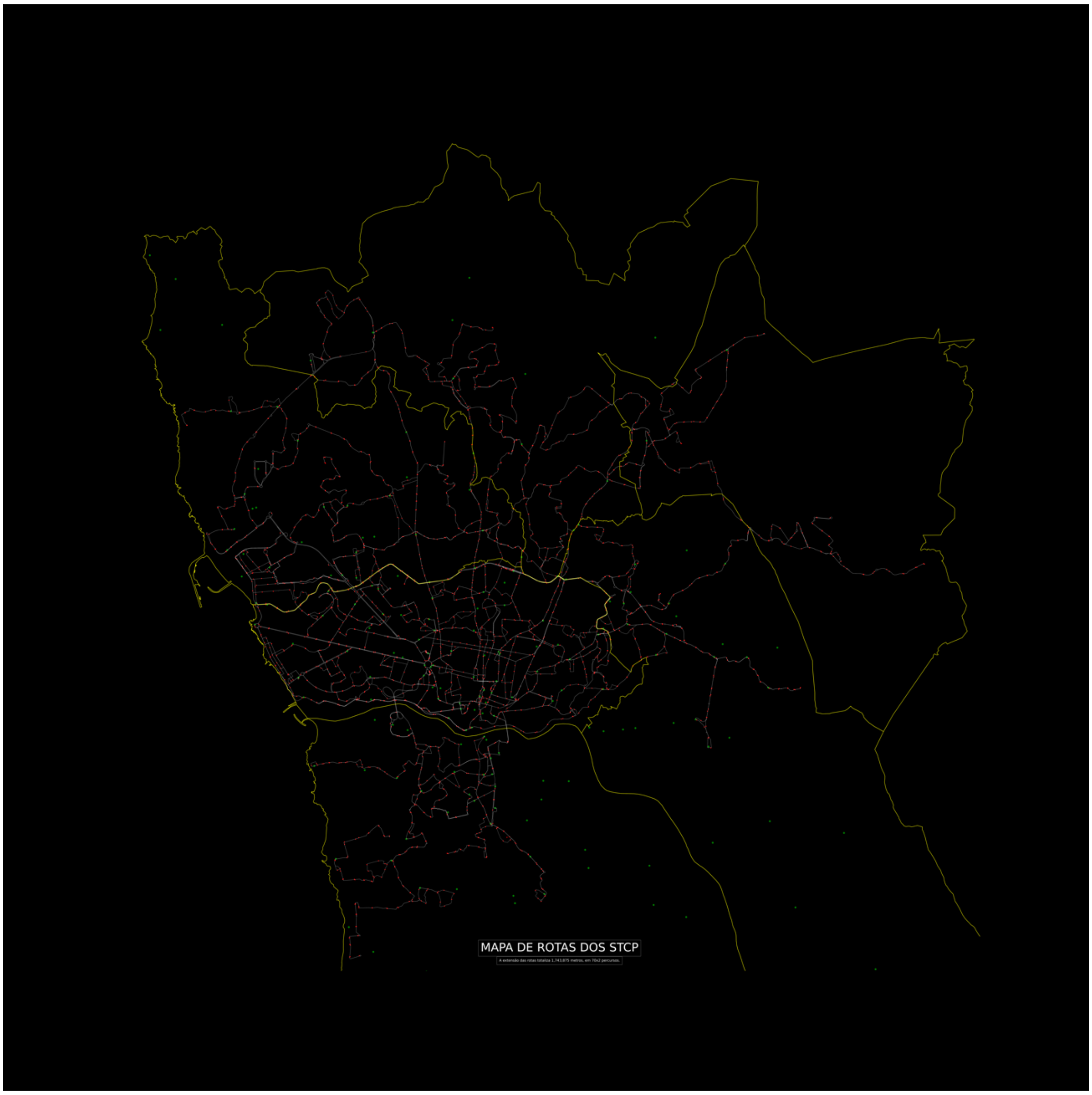
The subject of the practical assignment is the analysis of the operation of STCP. **STCP** (**Sociedade de Transportes Colectivos do Porto, E.I.M., S.A.**, is the public transport company that runs the bus and tram service in Greater Porto, Portugal. Created in 1946, it took over the Porto tram system from its privately owned predecessor and continues to operate it today, but the formerly large tram system now has only three lines, which are heritage tram lines, and the STCP network is now mostly bus service.

Information about STCP operation in GTFS format is available in the following URL:
https://opendata.porto.digital/dataset/horarios-paragens-e-rotas-em-formato-gtfs-stcp

The General Transit Feed Specification (GTFS) is an Open Standard used to distribute relevant information about transit systems to riders. It allows public transit agencies to publish their transit data in a format that can be consumed by a wide variety of software applications. Today, the GTFS data format is used by thousands of public transport providers.

A GTFS feed is composed of a series of text files collected in a ZIP file. Each file models a particular aspect of transit information: stops, routes, trips, and other schedule data. The details of each file are defined in the GTFS reference.

The figure below displays information about the STCP operation, showing the routes and stops. This figure was built using Python's Matplotlib module, colleting data diretly from a PostGIS database that was loaded based in the information available from the above URL.


MAPA DE ROTAS DOS STCP
A extensão das rotas totaliza 1.743.875 metros, em 70x2 percursos.

1. Part 1 of the practical assignment

The first part of the project involves creating a spatial database in PostGIS that stores the information from the GTFS files, including routes, stops and schedules.

You should create visualizations using Python's Matplotlib of the data you store in the database.

An optional component of this Part 1 is to animate the schedule of a particular route (or more than one route) based of the schedule information, using the Matplotlib animation module (as will be seen in the classes).

This part accounts for 40% of the total grade of the practical assignment. The correct creation of the tables and the correct spatial representations of the spatial objects is the main criteria used in the evaluation. The way information is visualized (or animated) is also a relevant criterion for the evaluation.

2. Part 2 of the practical assignment

The second part of the practical assignment involves modelling a simple data warehouse from the operation of STCP, joined with the taxi_services table that was also studied during classes. The basic idea is to have simple dimensions such as Time (hour, day of the week, etc), Route and/or Initial Stop and Final Stop, from the STCP operation. The population of the measures in a Fact Table should be obtained from the taxi_services table, where a particular taxi service is matched to the nearest STCP bus stops in terms of both the initial point and the final point. This matching should take into account the BUS routes as operated by STCP (there should be a way to connect, using one or more routes, the initial bus stop with the final bus stop).

Optionally, you can use the pg_routing module of PostGIS to determine the connectivity between bus stops, but it is not necessary, as this connectivity analysis is simple for the date involved. Documentation on the pg_routing module is available here: https://pgrouting.org/, and some additional points will be awarded if pg_routing is used.

Note that the main measure on the data warehouse will be the number of taxi trips, as obtained from the taxi_services table. But you can include additional measures such as the total amount of time between initial point and end point (also from the taxi_services table).

You should add some SQL queries and their result over your data warehouse. This should be analytical queries, using group by cube and group by rollup constructors (as will be seen in the classes), to build the summarization reports.

This part accounts for 60% of the practical assignment and the main criteria used for evaluation is the correct modelling of the data warehouse, the correct joining with the taxi_services table (using spatial functions), and the correct design of SQL queries to summarize the information with multi-dimensional grouping constructors.

3. Groups and dates

This practical assignment is to be done in groups of 2 or 3 students. Groups shall be created in Moodle. There will be a presentation of the practical assignment that also counts for the final grade of each student, as was previously defined.

The following dates are defined:

a. Submit an archive through Moodle with all the files, including a report (up to 6 pages) until 23:59:59 May 30th 2023;
b. Presentations will be done of Friday, May 31st in time slots to be defined to each group.

Good luck!