DATA ARCHITECTURE
CA2

DEADLINE / **WEIGHT:** As displayed on Moodle

Dr Peadar Grant

1 Introduction

This CA is designed to showcase and integrate your work in Data Architecture during Semester 2. It is a joint project with Time Series Analysis. You will be using the same dataset for both.

You are required to submit your proposed dataset and a brief description of the project to Moodle before 23:55 on 21st March. There may be a review meeting with each student in the week beginning 22nd March to establish the suitability of the chosen project. *This should be the same as your submission for Time Series Analysis module.*

2 Deliverables

Supply a single ZIP file `ca2.zip` (not .rar) with the necessary files included. File names are case-sensitive. Where multiple formats are accepted, the first matching will be taken and others ignored.

Your submission must have a `README.txt` file with your student number and name in it. If any specific notes regarding setup are required, place them in `README.txt`. It will otherwise be assumed that any `sql` files are for PostgreSQL, `py` files are for Python3.8 as on the shared server and `.sh` files are for bash on Amazon Linux on the shared server.

2.1 Dataset identification (10%)

In the file `dataset.txt` you should provide:

1. The name of the dataset(s)
2. A summary of what is included in the dataset(s)
3. Summary of what types of data they contain and what queries/analyses you plan on performing on that data.

2.2 Database architecture (20%)

In the file `architecture.pdf` (max 1 page) you should show:

1. The database(s) you plan to use for storage.
2. Where the data is loaded from.
3. Any transformations or scripts to load the data to the database (e.g. cleansing in pandas)
4. The database schema (e.g. E-R diagram) itself.
5. Connections to/from the database for Time Series Analysis module

2.3 Database system setup (20%)

Setup commands required to setup your database must be given in `setup.sh/.sql/.py`. This should match the architecture described in subsection 2.2.

2.4 Analysis integration (20%)

You are required to replicate at least two analyses from your Time Series Analysis using Database Programming (e.g. stored procedures). Analysis should be demonstrated from `analysis1.sh/.py/.sql` and `analysis2.sh/.py/.sql`

2.5 Real-time adaptation (20%)

Assume in this section that the dataset you have chosen is a “live” dataset.

In the file `realtime.pdf`, you are required to develop the architecture from subsection 2.2 to incorporate realtime updates.

(You are NOT required to implement this section in code.)

2.6 Reflection (10%)

In the file `reflection.txt` you should cite 3 changes you would make to the technical or organisational aspects of this project if you were completing it again.

3 Demonstration

The lecturer at their sole discretion may require demonstration of the developed system for verification. Where demonstration is required this will take place via Zoom at dates/times as determined by the lecturer. Students who do not attend for demonstration will receive no marks.

4 Feedback

Written feedback will be supplied with your grade. Verbal feedback is available upon request.

5 Queries

Queries on this CA should be addressed to the module Q&A forum on Moodle.