MSc/PGDip Data Analytics

Database and Analytics Programming (H9DAP)

Project 2021

WEIGHT: 70% of overall marks
COURSE: MSc / PGDip in Data Analytics

TASKS:

You are required to identify and carry out a series of analyses on a collection of datasets (that are somehow related or complement each other) utilising appropriate programming languages, programming environments and database systems.

This will be a group project, with teams comprising of 3-4 participants.

Although this is a team project substantial elements of the overall project should be undertaken on an individual basis and then integrated into the overall deliverables for the project.

Your project must incorporate the following elements/tasks:

1. Three or four semi-structured datasets must be used, depending on whether there are 3 or 4 members in each group.
2. Datasets must be programmatically stored in appropriate database(s) prior to processing.
3. Programmatic pre-processing, transformation, analysis and visualisation of the data.
4. Programmatically storing the processed output data in appropriate databases.
5. Programmatically create a dataset that joins together the initial datasets (or data resulting from processing the initial datasets) for a further analysis of the resultant dataset.
6. Report writing.

Choice of datasets and subsequent processing of chosen datasets

It is important that 3-4 distinct datasets are chosen. Each of these datasets should be structurally distinct (i.e., they should not share the same structure in terms of field types etc.). The datasets should be chosen, however, such that there are potential linkages or joins across the datasets (e.g., datasets contain data across corresponding time periods/ datasets have a shared reference to a common entity). Team coordination will be required to identify the datasets that each team member will work on in order to establish that there are potential linkages across the datasets.

Each group member will then be responsible for one of the chosen datasets. Each group member must individually complete elements/tasks 2, 3, and 4 (as listed above for their chosen dataset).

For example, you may use Python to programmatically retrieve a semi-structured dataset (XML or JSON or web-scraped or streaming data) and store this data in MongoDB. You may then read the data from MongoDB, to-process and transform it, in the process creating some structured datasets that you store in PostgreSQL for later usage. Following that you may use Python or R to conduct further analysis of the data to find interesting patterns my applying knowledge from other modules (e.g., statistical analysis), and generate visualisation plots for better presentation of the results.

All project artefacts should be managed through a private github repository. One of the team members should elect to create the repository and access to the repository should be provided to all other team members and to your lecturer.

Project elements/tasks 5 and 6 as listed above are team-based tasks. Each individual team member is responsible for the portion of the report that documents the details relating to their chosen dataset.

The report should also document the related work associated with task 5.

Teams should also consider creating a cloud based server for hosting any required databases that will be used to store data.

Each dataset should contain at least 5,000 records. Some appropriate datasets may be found at:

https://catalog.data.gov/dataset?res_format=XML

http://aiweb.cs.washington.edu/research/projects/xmltk/xmldata/

https://data.gov.ie/dataset?res_format=JSON

https://catalog.data.gov/dataset?res_format=JSON

https://data.worldbank.org/

Other sources of data may be used as well.

PROJECT REPORT:

The results of your analysis should be included in a project report. The project report should discuss the programming and data processing challenges that you encountered and the means and mechanisms you implemented to overcome these challenges.
The report should be around 4000 words in length (excluding references), should follow the IEEE format, as well as appropriate referencing and academic style. MS Word and LaTeX templates can be downloaded from:
https://www.ieee.org/conferences_events/conferences/publishing/templates.html

The report should contain the following sections:

Abstract:

- A 150-200 words executive summary of the project objectives and achievements
- **Note:** Look at abstracts in your literature review to get an idea of what makes a good/bad abstract

Introduction:

- Statement of the project objectives
- Motivation of the problem
- Relevance of chosen topic
- Elicitation of appropriately formed research question(s)

Related Work:

- Summarise relevant (academic) works that addressed similar problems or guided your decisions
- Critical evaluation (i.e., go beyond just a summary of the works and present their limitations and implications)

Methodology:

- Description of the underlying dataset(s) and justification of choosing them
- Descriptions and justifications of the data gathering and handling activities carried out (e.g., use of APIs, databases, etc.)
- Descriptions and justifications of the implemented data processing algorithms.
- Justifications for the choice of technologies used (i.e., programming languages, databases, etc.)
- Diagrams highlighting the data gathering, processing and analysis flow will be useful here.

Results:

- Presentation of results by making appropriate use of figures, tables, etc.
- Evidence of how the project objectives were met
- Discussion of the research findings, their interpretation(s) and implications

Conclusions and Future Work:

- What (in general) others can/could learn from your work
- Explicit discussion of research question(s) in the context of your findings
- Limitations of your solutions (critical self-evaluation)
- Future work directions (i.e., if you had more time what would you do differently or in addition to extend your work?)

References:

- A complete list of academic works and/or online materials used in the project. References should be included as in-text citations according to the IEEE citation style.

PRESENTATION:

There will be a presentation, which will act as a discussion point for your work, it should be used to provide references to **what** you did, **how** you did it, and **why**. If appropriate, you may (but are not obliged to) demonstrate your approach or key parts of it at the presentation. Presentations should be recorded and not exceed a total of 10 minutes in duration.

Each member of the team is expected to be able to present all aspects of the work individually and without assistance from other group members.

SUBMISSION:

Your submission must include your **project report document** along with any **programming code**, **data** and **system configuration elements**. Only **one** submission is required by each group.

An additional document detailing the distribution of work performed by the team (submitted by **every** member of the team). You should also note any work (with evidence) undertaken that did not make it into the final paper.

A link to your team's presentation video should be submitted via Moodle.

The final report must be submitted to Moodle (TurnItIn) by the published deadline.

Late submissions will only be accepted if a student was approved an extension due to personal circumstances by the School Office.

All submissions will be electronically screened for evidence of academic misconduct (plagiarism and collusion)!

Grade **Criterion**	**Solid H1** **> 80%**	**H1** **> 70%**	**H2.1** **> 60%**	**H2.2** **> 50%**	**PASS** **> 40%**	**FAIL** **< 40%**
Project Objectives (10%)	Challenging project objectives are well presented, met, and thoroughly discussed.	Challenging project objectives are well presented, met, and thoroughly discussed.	Reasonable project objectives are well presented, met, and discussed.	Reasonable project objectives are clear, and at mostly met.	There are clear objectives, which are at least partially met.	Cannot discern project objectives, and/or if project objectives were met.
Literature Review (10%)	Excellent critical analysis of substantive and relevant literature.	Very good critical analysis of substantive and relevant literature.	Good analysis of relevant literature.	Adequate analysis of mostly relevant literature.	Some review of some relevant literature but limited evidence of understanding.	Little relevant literature re-viewed, very limited evidence of understanding.
Data Complexity and Handling (20%)	The datasets have been well prepared and meaningfully explored. All datasets were stored in appropriate data-bases before and after processing. At least two datasets have a high degree of complexity. At least one dataset was programmatically retrieved (e.g., through API or web scraping).	The datasets have been well pre-pared and meaningfully explored. All datasets were stored in appropriate databases before and after processing. At least two datasets have a high degree of complexity.	The datasets have been well prepared and explored. At least one dataset was stored in appropriate databases. At least one dataset has a high degree of complexity.	The datasets have been appropriately prepared for analysis. At least one dataset was stored in databases. At least one of the datasets is non-trivial.	The datasets were appropriately handled fit-ting for the objectives. The use of databases is very basic. The datasets are probably somewhat trivial.	Only one somewhat trivial dataset was used. No database was used to store the datasets. No obvious development was conducted.
Data Processing Implementation (20%)	The data processing algorithms play a well-conceived and essential role in meeting the project objectives. The implementation significantly exceeds the stated minimum requirements.	The data processing algorithms play a well-conceived and essential role in meeting the project objectives. Multiple data processing technologies / languages were used.	The use of data processing algorithms is well-thought and appropriate for the project objectives. Comprehensive use of at least one data programming language.	The use of data processing algorithms is meaningful and appropriate for the project objectives. Appropriate use of at least one data programming language.	Appropriate but basic use of data processing algorithms. Basic use of data programming languages.	No implementation or inappropriate use of data processing algorithms.
Level of Automation (10%)	Everything is automated within one process control flow. Every run probably results in different results as new data is extracted and subsequently included (i.e., through an API).	Everything is automated within one process control flow.	Most core components are automated within one process control flow.	Some components are connected within a larger process. Yet some aspects are run as separate pro-cesses.	Individually all project components are auto-mated, but not necessarily connected together.	None or little automation
Results and Conclusions (20%)	3 or more insightful findings are presented and thoroughly discussed with appropriate references to existing work.	3 or more interesting non-arbitrary findings are presented and thoroughly discussed with appropriate references to existing work.	3 or more interesting non-arbitrary findings are presented and thoroughly discussed.	2 or more interesting non-arbitrary findings are presented and appropriately discussed	2 or more interesting non-arbitrary findings are presented	Little to no non-arbitrary results and/or findings

Quality of Writing (10%)	Very well written, with no language errors. All figures are well conceived and read-able. The IEEE template is strictly adhered to. Report does not exceed the length limits. References are appropriately and correctly used.	Well written, with no (large) language errors. All figures are well conceived and read-able. The IEEE template is adhered to. Report does not exceed the length limits. References are appropriately and correctly used.	Main document has a few language and/or style errors. Figures are well presented. IEEE template and length limit are adhered to. References are complete, and correctly used.	Main document has a few language and/or style errors. Some figures are may be hard to read. IEEE template and length limit are largely adhered to. References are complete, and correctly used.	Main document is readable with some language and/or style errors. Figures may be hard to read or presented in a suboptimal manner IEEE template may have been broken. References are mostly complete and correctly used.	Littered with typos, and/or poor use of English. IEEE template not used. Figures may be hard to read. References (if any) are probably incomplete.