COMP4124 Group Project Brief

Contents

Introduction

The group project coursework forms the majority of the assessment for COMP4124, with a weight of 75% of your module mark. The aim of this document is to describe:

- The requirements for the group project coursework
- The topics you can choose to work on
- The assessment criteria that will be used to mark your work
- The coursework policies for plagiarism, late submissions, and ECs

Read this document carefully; if you would like clarification on anything, please post your questions on the Moodle Discussion Forum, so the answers are available to everyone. Clarifications/corrections to this document will be posted to Moodle.

Project Overview

The coursework takes the form of a research project with the aim of designing a machine learning solution in Apache Spark and investigating its behaviour when applied to a large dataset. To achieve this, you will need to design and carry out the relevant stages of a data science project, including (but not necessarily limited to):

- Dataset selection and development of research questions
- Data preprocessing
- Method selection, design and implementation
- Experimental design
- Collection and analysis of results

We provide three project topics for you to choose between in Appendix 1. Each topic has some suggested areas to focus on, but you can also come up with your own area of focus **within these topics**.

The coursework has three components:

- Interim milestone, in the form of a short poster discussion (15% of your module mark)
- Final submission, of the report, code, and peer reviews/individual contribution statements (45% of your module mark)
- Final presentation, summarising the project (15% of your module mark)

For more information on these, see the Deliverables section below. Please ensure you check the deadlines for the deliverables on Moodle.

Project Requirements

Successful projects are those that:

- **Have an appropriate focus:** The main subject of the research should be on the behaviour of the techniques when implemented in Spark. "Behaviour" is not limited to, for example, obtained clusters or classification accuracy. You also need to be investigating the scalability of your implementation(s). You should define your research questions appropriately, to ensure your project addresses this requirement.

- **Have some element that is non-trivial to parallelise, that has been designed and implemented by the group:** As shown in the lectures, Spark offers the MLlib library with implementations of a variety of machine learning-related algorithms. You are allowed (and encouraged) to make use of these implementations where appropriate. However, part of your solution should be designed and implemented by the group, rather than using the built-in methods.
  For example: if your project was looking at clustering algorithms, you may choose to use MLlib for preprocessing the data, but then select two clustering algorithms to re-design for a distributed implementation.

- **Demonstrate an ability to critically evaluate the completed work and results:** Your project will not be marked on the actual results you obtain. Instead, your ability to critically discuss your work, experiment set-up, and results will be assessed. As an example, your Discussion and Conclusions could address (amongst other things):
  - Bottlenecks in your designed solution that could contribute to poor scalability
  - Limitations of your experimental study and how this affects the conclusions you can draw
  - What improvements you would make with more resources (e.g. time, amount of data)

Crucially: A project that designs and implements a well-thought out distributed solution for a relatively straightforward method/technique **will receive a higher mark** than a project that implements a more complex technique with heavy reliance on existing libraries and little evidence of consideration of the distributed design. For more information on the criteria that will be used to assess your work, see the assessment criteria for each of the Deliverables in the below sections.

Dataset Selection and Use

It is up to you to make the final selection of the dataset to use. It can be challenging to find large datasets for these projects, and therefore it is not necessary to select too large a dataset. However, you should discuss in your report any limitations for the experiments/ conclusions that are caused by having a smaller dataset. If you would like to discuss the suitability of a dataset, please come and speak to us in the lab.

Throughout the project, we suggest that you only use subsets of the datasets when testing/designing your solutions, to avoid having to wait a long time for your code to run. Only run experiments with the full dataset to provide results for your final experimental study. Note that creating subsets to work with may require some work, as you will want them to be representative of the whole dataset (e.g. preserving the class distribution). Your use of subsets must be included and documented in your final code submission.

Project Guidance and Getting Help

We will be running project guidance sessions in two of our Friday lecture slots – dates will be confirmed on Moodle closer to the time. These sessions aim to expand on the points in this document, and give you some ideas for how to work on the various aspects of your project. The sessions will be recorded on Echo360, but it is strongly recommended that you attend in-person so you can ask questions on the day.

Additionally, we strongly recommend that you make good use of the remaining lab sessions as an opportunity to discuss your project with us and ask any questions. Whilst we cannot give you direct input, tell you what to do, or say whether your ideas are ‘good’ or not, we are more than happy to discuss your project and suggest aspects for you to think about.

Deliverables and Assessment Criteria

Submission of Work

All work will be submitted on Moodle, with the exception of the peer review/individual contribution statement which will be a Microsoft Form. Submission boxes will open closer to the deadlines, and we will let you know via Moodle announcements when this happens.

In the final group allocation, one person from each group has been randomly allocated as the **group leader**. This title does not mean anything, other than that this person will be responsible for submitting the group's work. In order to ensure we only receive one submission per group, only group leaders will be able to see and submit to the assignment submission box on Moodle.

- **Group leaders:** You must communicate clearly with the rest of the group when you are submitting the work. It should be a group decision what is being included in the submission.
- **Other group members:** If you are having problems with contacting your group leader, or have concerns that your work will not be submitted on time, you must contact both the module convenors as soon as possible. In these circumstances, we will make the Moodle submission boxes available to additional group members as well. It is your responsibility to make us aware of these circumstances **in advance of the deadlines**.

Interim Milestone: Poster & Discussion

The interim milestone is intended as an opportunity to present your work and ideas so far to the teaching team, and get feedback to help support you with the rest of the project.

The group will need to:

- Produce a poster, following the guidelines provided below.
- Attend a short (10 minute) discussion with a member of the teaching team. The discussion will follow the format provided below.

Submission: The **group leader** should submit the poster as a PDF file on Moodle **before your allocated timeslot**. Timeslots will be determined closer to the time.

Poster Guidelines

- The poster should be designed so it could be printed legibly on A3 paper. However, there is no need to print the poster; for the demo, it will be viewed on a computer screen.
- You can use any software you like to make your poster.
- The poster should focus on informative figures/diagrams, rather than large amounts of text.
- It is up to you what you include on the poster, but it should clearly communicate your:
    - Research questions and specific area of focus
    - Your chosen dataset
    - Selected methods to investigate
    - Preliminary solution design (illustrated via a diagram)
    - A few key references you have found so far
- You only have limited space, so you need to be selective with what you include.
- Any text or diagrams using ideas from other sources should be cited at the bottom of your poster. If you have used a template for your poster, a reference for this should also be provided.

Discussion Format

The group will attend a 10-minute discussion with a member of the teaching team at a specified time/place. You do not need to prepare a presentation. Instead, we will ask questions about your poster, your work and your plans for finishing the project. Timeslots will be determined closer to the time.

All group members must be present; **any group member not present will receive a mark of 0 for the interim milestone**. If you cannot, for a good reason (e.g. illness), attend the discussion at the specified time, please email both convenors (Rebecca and Joy) and we will work to reschedule the group's poster demo. You do not need to submit an extenuating circumstances claim. Please see the Extenuating Circumstances section towards the end of this document for more information on our policy for the coursework.

Assessment Criteria: Interim Milestone

Note that the assessment criteria are not focused on how much work you have completed so far. Instead, we are looking for how clearly you can communicate your work and ideas, as well as the depth with which you can discuss them. We will also assess the suitability of the focus of your project.

We aim to provide feedback as soon as possible for the poster/discussion, especially regarding the project's suitability. This will give you time to act on the feedback and ensure

your final submission meets the 'appropriate focus' requirement. You will also receive some feedback during the discussion.

Criteria	**Critical questions**
Suitability of the project	Is the project suitably focused on the behaviour of a method/technique when implemented in Spark, rather than just the results produced by the method? Does the project represent sufficient complexity in terms of what the group will implement in Spark (i.e. not just library implementations)?
Evidence of plan for remainder of the project	Have the group thought about the remainder of the project? Have the group thought about problems they may encounter between now and the end of the project?
Overall depth and quality of the discussion	Can the group describe their ideas clearly? Have the group demonstrated understanding of their selected method(s)? Have all group members contributed to the discussion?
Quality of the poster	Does the poster clearly communicate the research questions, dataset, and a preliminary design? Does the poster prioritise visual elements? Are visual elements clear?

Final Submission

The final submission is comprised of the following parts:

(a) A joint research paper, between 6-8 pages in length (including tables, figures, and references) as a PDF file.
(b) A joint code submission in the form of a single Python notebook file.
(c) Peer review and individual contribution statement (maximum 300 words), submitted via a Microsoft Form.

More details of these are provided in the sections below.

Research Paper and Code Submission

Research paper: You must write your work up as a research paper, 6-8 pages in length. This includes tables, figures, and references. You should use the IEEE format (template available here: https://www.ieee.org/conferences/publishing/templates.html).

You have limited space in the research paper, so you will need to be selective with what you include. The structure of the paper may vary depending on the project, but generally should include the following:

1. Title: a representative name that describes what you have done.

2. Abstract: briefly outlines the problem statement of the project, and includes information on the method, results, and conclusions.
3. Introduction: providing context for the project and clearly stating the research questions.
4. Literature review: discussing a few key references related to (a) the problem and/or dataset you are using; and (b) the design of your scalable solution. **Do not** include a section describing MapReduce, Apache Spark, or any other big data framework – we assume this as existing knowledge of the audience (i.e. the people marking your work!)
5. Methodology: brief description of your overall pipeline (preprocessing steps, chosen techniques), plus a more in-depth explanation of the implementation of your designed solution. This should be supported by diagrams/pseudocode where appropriate.
6. Experimental study: describing your experimental set-up and reporting your results.
7. Discussion: Critically discuss your obtained results, comparing different methods/ strategies. You should also compare with previous work where appropriate.
8. Conclusions and suggestions for future work.
9. References: accurate references, using a consistent referencing style.

Code: Your code should be submitted as a single Python notebook file. The notebook must be able to be run from start to finish. You should use Spark for all aspects of the project, from data preprocessing right through to the experimental study.

You should make good use of markdown cells to document the code and explain the purpose of different parts of the notebook. The notebook should contain clear comments that attributing parts of the code to group members. Each part of the code should be attributed to one or more group members.

Submission: The **group leader** should submit two files (the research paper as a PDF file, and the code as a single .ipynb file) by the final submission deadline. There is no need to submit your dataset, but the source you got it from should be clearly linked/cited in both the paper and code notebook.

Assessment Criteria: Paper and Code

The group will receive a mark for the final submission based on the research paper and code submission, according to the below assessment criteria. Marks for individual group members will then be determined by considering the group mark, code attribution comments, peer reviews, and individual contribution statements.

Criteria	**Critical questions**
Project requirements	Is the project suitably focused on the behaviour of a method/technique when implemented in Spark, rather than just the results produced by the method? Does the project represent sufficient complexity in terms of what the group has implemented in Spark (i.e. not just library implementations)? Has the group demonstrated the ability to critically discuss their results and work?
Introduction and research questions	Has the project been well-contextualised? Have specific research questions been defined? Do the research questions give sufficient opportunity to investigate the behaviour of the method(s) when applied in a distributed environment?
Literature review	Have relevant papers been identified and concisely summarised? Have the authors related the papers to their own project?
Methodology – overall pipeline	Are all stages clearly described? Are the choices of methods/techniques clearly justified?
Methodology – design of a distributed solution in Spark	Has some part of the project been implemented by the group members themselves using Spark? Is the distributed solution clearly explained and supported where appropriate with diagrams showing data movement? Does the explanation highlight the consideration the group has put into reducing data movement? Have approximations from the original algorithms been clearly described and justified?
Experimental study and results	Have the experiments been well-designed to address the research questions? Are the experiments designed to investigate scalability? Are the results organised in a sensible way? Are the results appropriately presented in tables and graphs?
Discussion and conclusions	Does the discussion appropriately interpret the results? Does the discussion include commentary on limitations and shortcomings of the completed work? Does the conclusion provide a good summary of the work? Has appropriate future work been identified and justified?
References	Have good quality references been selected? Is the referencing style consistent throughout? Is the reference list accurate?
Overall quality of writing	Is the writing consistent throughout? Is the writing easy to follow? Is the paper well-structured? Have visual elements (tables, figures) been effectively used throughout, with informative captions?
Code – data processing	Does the code consistently use PySpark for data processing? Has the group made efforts to minimise data movement, either in shuffles or in the amount of data brought back to the driver? Is the code in the notebook consistent with what has been described in the paper?
Code – general quality	Is the notebook well-structured and easy to follow? Has the code been well-documented in the markdown cells/comments? Have informative variable/function names been used? Would the results be reproducible based on the descriptions and code in the notebook?

Peer Review and Individual Contribution

There are no marks awarded for the peer review and individual contribution itself; however, **failure to submit these will result in an individual mark of 0 for the final submission.**

The Microsoft Form for the peer review and individual contribution will be made available towards the end of the project. You must fill out the form according to its instructions, scoring your peers on their contributions to the project in various areas.

The form will also include a box to enter an individual contribution statement (maximum 300 words). This should explain your contribution to the project.

Submission: All students must submit the peer review and individual contribution form by the final submission deadline.

Final Presentation

Groups will be required to give a short presentation summarising their completed project. You only have limited time, so do not try to mention everything – focus on the key aspects! A general rule of thumb is no more than 1 slide per minute, so you should have no more than 12 slides.

Format: 12 minutes to present, 5 minutes for Q&A. Timings will be strictly adhered to.

Submission: The **group leader** should submit the materials used for the presentation **in advance of your presentation timeslot**. Timeslots will be determined closer to the time.

All group members must be present and contribute to the presentation. A**ny group member not present will receive a mark of 0 for presentation**. If you cannot attend the presentation due to Extenuating Circumstances, please submit an EC claim. The group will need to present at the original time, and we will schedule an individual presentation for the student with an approved extension at a later date. Please see the Extenuating Circumstances section towards the end of this document for more information on our policy for the coursework.

Assessment Criteria: Presentation

Criteria	Critical questions
Overall quality of the presentation	Was the presentation engaging? Was the presentation well-structured? Was the presentation supported by appropriate and well-designed slides/other materials? Did the group time the presentation well?
Group participation	Did the group members equally participate in the presentation?
Response to questions	Did the group engage with the questions and provide suitably detailed answers?
Summary of the project	Were the key aspects of the project presented? Were clear descriptions and explanations of the work done given? Did the group demonstrate a good understanding of their completed work in the presentation and response to questions?

Coursework Policies

Academic Misconduct

The work submitted for this coursework must be the work of the group members only. While it is expected (and encouraged) that you look at other code repositories and papers as part of your research and learning for this project, any ideas or code that you take should be properly cited. We will be using software and other methods to detect plagiarism in the submissions. Where we suspect plagiarism to have occurred, we will submit cases to the academic misconduct committee.

Generative AI

For the purposes of this coursework:

- You are **not allowed** to use generative AI to write your report or code for you, or to create your poster or presentation.
- You are allowed to make use of generative AI for assisted coding, e.g. via inline code completion in Colab.

Late Submission

The default late submission policy applies to all components of the group project. That is, there is a 5 percentage point penalty for each late working day. This is applied separately for each component.

Extenuating Circumstances (ECs)

Students should only submit an EC claim if there is an impact on their ability to meet the **final report deadline or participate in the final presentation**. In these cases, the group will need to submit their report at the original deadline. The student with an approved extension will be able to submit a revised version with a list of amendments clearly describing their work since the original deadline. The student with ECs will then be required to do a short individual presentation after their extended deadline.

Please refer to the University EC Policy for circumstances that constitute Extenuating Circumstances.

If you are unable to attend the interim poster discussion for a good reason (e.g. illness): You do not need to submit an EC claim. Instead, contact both convenors (Rebecca and Joy), and we will try to reschedule the group's discussion. In the situation where a longer extension may be required, the group's discussion will go ahead at the original time (to ensure timely feedback to the group). We will then reschedule an individual discussion with the student with the extended deadline at a later date.

Reassessment

Reassessment (first sits and resits) will take place over the summer. This will take the form of a similar research project to be completed **individually**. The scope of the reassessment project will therefore be smaller. The reassessment will not have an interim poster discussion. The only deliverables will be the final report, code, and a short individual presentation.

Appendix 1: Project Topics

This section describes the three project topics. You can choose to work on **any one** of these topics. Each topic gives you a general area to work within. However, the choice of the final area of focus for the project is up to you, and you are expected to do the necessary research in order to define this. This includes selecting an appropriate dataset – and remember, these are 'real' data, so will usually not be in a sensible format, and you will need to appropriately clean and preprocess them.

Project 1: Clustering of large-scale GPS trajectory data

Trajectory clustering aims to group together similar trajectories, according to some pre-defined similarity criteria. The similarity criteria chosen depends on the problem being tackled, but may include consideration of location of the trajectory, shape of the trajectory, temporal attributes, as well as additional features associated with the trajectories such as speed and elevation. Trajectory clustering methods can be challenging to apply to large trajectory datasets due to the computational complexity that arises from computing similarities between many trajectories. The general aim of this project is to investigate trajectory clustering methods in the context of big data.

Possible aspects to focus on:

- Investigation of different partitioning strategies based on geographic area
- Investigation of different distributed implementations of the distance calculations between trajectories
- Investigation of distributed implementations of multiple trajectory clustering methods

Preliminary resources:

- **TRACLUS**, well-known trajectory clustering method: Lee, J.G., Han, J. and Whang, K.Y., 2007, June. **Trajectory clustering: a partition-and-group framework.** In Proceedings of the 2007 ACM SIGMOD international conference on Management of data (pp. 593-604).
- Not clustering focused, but this paper is an example that partitions Spark RDDs based on spatial and temporal attributes: Yu, J., Fu, Z. and Sarwat, M., 2020. **Dissecting geosparksim: a scalable microscopic road network traffic simulator in apache spark.** Distributed and Parallel Databases, 38(4), pp.963-994.

- Microsoft trajectory data mining overview - https://www.microsoft.com/en-us/research/project/trajectory-data-mining/
    - The paper linked in the 'Overview' section may be a useful starting point for trajectory data mining in general

Potential dataset sources:

- Microsoft GeoLife dataset - https://www.microsoft.com/en-us/download/details.aspx?id=52367
- Beijing taxi trajectories - https://www.kaggle.com/datasets/arashnic/tdriver/data
- The Movebank repository contains tracking data from studies of animals, you can search it for open datasets - https://www.movebank.org/cms/movebank-main

Project 2: Distance-based classification of big time series data

Time series classification is the task of learning a model capable of classifying time series. Aside from deep learning, distance-based classifiers have shown success for this, using time series-specific distance/similarity measures. However, there are challenges in applying distance-based methods to big time series data due to their computational complexity. The general aim of this project is to investigate distance-based time series classification methods in the context of big data.

Possible aspects to focus on:

- Investigation of distributed implementations of data reduction techniques for time series classification
- Investigation of different distributed implementations of the distance/similarity calculations between time series
- Investigation of distributed implementations of multiple time series classification algorithms

Preliminary resources:

- Aeon (scikit-learn compatible library) gives an overview of some common algorithms which you could choose from - https://www.aeon-toolkit.org/en/latest/examples/classification/distance_based.html
- Lucas, B., Shifaz, A., Pelletier, C., O'Neill, L., Zaidi, N., Goethals, B., Petitjean, F. and Webb, G.I., 2019. **Proximity forest: an effective and scalable distance-based classifier for time series.** Data Mining and Knowledge Discovery, 33(3), pp.607-635.

  - Section 2.1 gives a short overview of distance-based time series classification methods and related issues

Potential dataset sources:

- The UCR time series classification archive is a good place to start - https://www.timeseriesclassification.com/index.php

Project 3: Techniques for class imbalance in big data

Even within very large datasets, it is still possible to suffer from data scarcity. One form this can take is *class imbalance*, where a dataset has many more instances of one class (or some classes) than of others. This can lead to classifiers favouring the majority class (i.e. the most frequently occurring class). The general aim of this project is to investigate the use of techniques for handling class imbalance in the context of big data.

Possible aspects to focus on:

- Investigation of partitioning strategies for distributed implementations of undersampling and/or oversampling techniques
- Investigation of different distributed implementations of undersampling and/or oversampling techniques
- Investigation of different distributed implementations of ensemble classifiers combined with sampling

Preliminary resources:

- Imbalanced-learn user guide, provides some nice introductions to potential algorithms - https://imbalanced-learn.org/stable/user_guide.html
- Galar, M., Fernandez, A., Barrenechea, E., Bustince, H. and Herrera, F., 2011. **A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches**. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(4), pp.463-484.

Potential dataset sources:

- Credit card fraud detection - https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud
- Imbalanced-learn datasets - https://imbalanced-learn.org/stable/datasets/index.html
  - You can load datasets from imbalanced-learn, but most are quite small

- Imbalanced-learn also includes a function ‘make_imbalance()’ that you could use to artificially create imbalanced datasets from balanced datasets you have found