BIG DATA ANALYSIS

GROUP PROJECT

DUE 13TH-DECEMBER 23:59H

As you will all be finishing your degree and entering the job market we will simulate a real-world data science project for your assessment. This is a group project and you should form groups of **3-5 people**.

Please read the instructions and grading criteria carefully.

Big Data Analysis in Spark

You and your group are applying to join the data science team at BigDataCompany. You have managed to pass to the final stage of the process, the technical assessment. BigDataCompany is a consulting firm that helps other companies deal with all the problems of big data. To do so, they use the Apache Spark platform on Databricks. They want you to demonstrate your ability to both use Spark technically, but also to apply your skills in a way that makes sense and generates insights from data.

Your task is then as follows:

1) Find up to three databases which you will explore with PySpark on Databricks. Choose a dataset that is both challenging and interesting. Public repositories like UCI Machine Learning Repository, Kaggle, and governmental databases are excellent places to start.

   Please note: The three datasets don't need to relate to each other. The idea is that you will use different functionality in spark, and one dataset might be good for machine learning and the other better for graph analysis.

With this dataset (or datasets) you want to demonstrate your skills as PySpark data scientists.

2) To do so you should generate interesting insights from the data and demonstrate your ability to build useful tools with the data (for forecasting, classifying, etc.) The more ability you can demonstrate across the data scientist tasks, the better (ETL, data cleaning, exploration and summarising, visualisation, machine learning, data engineering, data streaming, graph analytics, etc.). Note: ChatGPT comes up with pretty boring ideas for trying to find interesting insights.
3) **Most importantly,** demonstrate your ability to understand big data and use PySpark. This means demonstrating your ability to work with all of the functionality seen in class. You want to show the ability to use:

a) RDDs and Map-Reduce,

b) DataFrames and SparkSQL,

c) Pipelines, data cleaning, and data engineering,

d) Spark machine learning,

e) Graphframes,

f) Spark streaming (Bonus points).

Important - It also means being very careful, we won't be working with big data but you should be careful to ensure your code would work if your database is huge and spread across multiple partitions. Think about the ordering of data returned from queries and the libraries you use. **You should indicate anywhere explicitly that you have not used big data safe functionality.**

You will need to deliver:

A. **A management report of up to 5 pages** (including references, figures, and tables) with Font-size 11pt, and using Times New Roman as the font. The report should be aimed at management and demonstrate the interesting insights you identified in the data or tools you developed. The report should be less technical and should be used to demonstrate your abilities to deliver value from big data (think about the 4 V's) for a management group without technical knowledge and demonstrate what spark is capable of. The theme of the report should be **"here's how we can make use of big data with spark"** rather than "here's the insights from a machine learning application", remember to focus on demonstrating your skills with spark and the specific problems or value of big data.

B. **All Databricks notebooks** and data should be provided. **Important:** Notebooks should be solved so that the values you have used can be seen without rerunning the code.

C. **We will hold online presentations and interview sessions** where we will ask you to explain **your decisions and code**. You should prepare a 7-minute presentation focusing on explaining your notebooks and the decisions you made in your modelling and how you implemented them in code.
The presentation should focus on the notebooks and not the results. You can focus on explaining the more complex parts of the code and demonstrating that you understand how spark works.

Please keep in mind that you are mostly assessed on your abilities to use Spark correctly, so the more Spark functionality you can correctly demonstrate the better!
Presentations will be scheduled 16th-19th December and will be held online.

Submissions will be done through Moodle.

FAQ's

- **Can I have x number of people in my group?:** Groups who do not have the correct number of people will receive a 0.5 mark penalty.
- **Do I need to use a big database for this big data assessment?** As we only have access to a small cluster on Databricks we cannot deal with a large volume of data, so you are not expected to do so. **However**, all of the analysis you apply should be applicable in the case that you are using a high volume of data. Hint: is the library you're using capable of dealing with big data?
- **Can we extend the deadline?** This question can only be asked in the first 4 weeks of the project. It is also unlikely given constraints with Christmas and exams.

Please ask your practical and theoretical professors if you have any further questions.