MSc ANALYTICS
BIG DATA PLATFORMS (MSCA 31013)

TEAM PROJECT

Students will work in teams of 3 or 4 members and collaborate on their Big data class project. The goal of the group project is to perform exploratory data analysis and apply machine learning techniques to solve a well-defined business problem on a large dataset. Students will be assessed on their understanding of various concepts taught in the course as well as their ability to leverage big data tools and infrastructure.

Project Requirements

Data
Use **more than one** publicly available or open source (structured and/or semi-structured) dataset. You are highly encouraged to fuse data from multiple sources which could provide insights that may not be obvious from a single source. The total combined raw dataset used in your analysis should be **larger than 5 *GB*.**

Infrastructure
The implementation **must involve use of PySpark** for primary data processing and model building, and possibly other tools in conjunction with PySpark such as BigQuery, Kafka, Airflow, Dremio etc. Scikit-Learn or Pandas may only be used to visualization and reporting on a summarized versio of the data. You may any leverage public cloud provider (GCP, Azure, AWS etc..) as the Big data environment.

Team Size

The team size is **3-4 students per team**. Due to the multitude and complexity of the technologies involved, great teamwork is critical to the success of big data projects.

Grading Rubric

Business Problem (15%)
The final presentation should include a clearly defined business/research problem and the goals and scope of the project. Please include an executive summary, business objectives, data sources and data challenges if any.

Data Analysis and Preparation (35%)
Data engireering which involves data wrangling, data transformation, joining with other data sources, imputation, etc. All data engineering must be performed on the chosen big data infrastructure. Peform the necessary exporatory data analysis (EDA) and visualization on the raw dataset. Bonus points for discovery of interesting insights from the raw data.

Machine Learning Models (25%)
Design and implement **at least two** machine learning problems on the large dataset. The machine learning could be related to a combination of regression, classification, NLP, recommendations, association analysis etc. You may choose to run multiple algorithms for each problem. Bonus points for comparing multiple models and cross-validation for each problem..

Project Execution (25%)

- Combination of multiple big data technologies (including pipelines, scheduling, compression, etc..)
- Demonstrating changes or improvements in design, execution timelines, storage optimization etc.
- Developing customized code/implementation that go beyond basic out of the box functions
- Bonus points if the size of your dataset **significantly exceeds** the required threshold

Project Timelines

- **Week 2** – Form Teams
- **Week 3** – Finalize Business problem and datasets
- **Week 4** – Complete loading of dataset into big data environment (RCC/Cloud)
- **Week 5** – Complete data clean up and processing
- **Week 6** – Complete exploratory data analysis and visualization
- **Week 7** – Complete baseline machine learning models and model metrics
- **Week 8** – Complete challenger/advanced machine learning models
- **Week 9** – Stretch goal: Automation of data pipelines, Graph computing where applicable

Project Submission

- Single submission per team within 1 day of the final presentation
- Upload all source code (ETL, DDL/DML, Notebooks, Pipelines, etc.) as a single zip file
- Upload project presentation as a separate PPT/PDF file