Big Data Platforms – Final Project

# Project Description

In this project you can choose to work on one of the following categories:

1. Implementing a distributed algorithm – implement a distributed algorithm in PySpark. Apply it on a dataset (at least one), evaluate it against a similar non-distributed algorithm in terms of performance, time, etc.
2. Analyze massive databases - Analyze relevant databases in PySpark (at least three with over 10K records each), research them (apply at least two different algorithms), find interesting patterns and present them.
3. Create/generate databases – Create a large-scale framework that create datasets in a constant manner.
4. Write an academic survey (only with approval) – Summarize 15-30 academic paper in a relevant big data domain.

# Project requirements

* Groups of three members.
* A statement document detailing the contribution of each group member must be attached.
* Every class and function declared in your project should be described using [docstring](https://peps.python.org/pep-0257/).
* The code in the project should follow [Style code pep 8](https://peps.python.org/pep-0008/).
* The code will be published through GitHub, which presents the work findings under an open code license.
* The final paper (at least 4 pages) will be submitted in a pdf with ACM format (attached a report example), you should also provide the LATEX project zip.
* The subject of the project must be submitted for approval by 15 Dec 23:59.
* Do not copy/improve other projects, such as projects that are open source e.g., Kaggle, and GitHub.

# Project grading

* The degree of innovation of the research results.
* The level of an academic report.
* The level of the data analysis (priority will be given to analyzing data from many different sources).
* Making the research accessible to other people/researchers.
* The level of the published code (the code of the work will be open source) and the accessibility of the code.
* Comparison with other projects in the course.

# Project Milestones

1. Submit the literature review by 29 Dec 23:59, the review must be comprehensive in your domain at least one page of two columns format.
2. Submit the project presentation by 12 Jan 23:59, you will need to present your project in the last two weeks of the Semester.
3. Submit the final project followed all requirements by 2 Feb 23:59.

# Project Reference

(1) Implementing a distributed algorithm:

* Kiveris, R., Lattanzi, S., Mirrokni, V., Rastogi, V., & Vassilvitskii, S. (2014, November). Connected components in mapreduce and beyond. In *Proceedings of the ACM Symposium on Cloud Computing* (pp. 1-13).
* Ran, X., Zhou, X., Lei, M., Tepsan, W., & Deng, W. (2021). A novel k-means clustering algorithm with a noise algorithm for capturing urban hotspots. *Applied Sciences*, *11*(23), 11202.

(2) Analyze massive databases:

* Kagan, D., Chesney, T., & Fire, M. (2020). Using data science to understand the film industry’s gender gap. *Palgrave Communications*, *6*(1), 1-16.

(3) Create/generate database:

* Rajasinghe, N., Samarabandu, J., & Wang, X. (2018, May). INSecS-DCS: a highly customizable network intrusion dataset creation framework. In *2018 IEEE Canadian Conference on Electrical & Computer Engineering (CCECE)* (pp. 1-4). IEEE.
* Borah, P., Bhattacharyya, D. K., & Kalita, J. K. (2020, December). Malware Dataset Generation and Evaluation. In *2020 IEEE 4th Conference on Information & Communication Technology (CICT)* (pp. 1-6). IEEE.