## **Final Project - Twitter Insight Engine using Kubernetes**

## Guidelines:

### Submissions:

* The assignment must be submitted by 22.2.24 at 23:59.
* The solution will be submitted through the model in zip file containing a pdf file and a link to a GitHub project. DO NOT SHARE YOUR CODE FILES.
* The summary file should be written in a word processor (e.g., Word or Pages) and should be exported as a pdf file.
* The submission is in pairs only, the ID number and names of each one must be stated at the start of the summary file.
* The zip file name will be of the form FP\_ID1\_ID2.zip.
* Only one group member is required to upload the solution.
* Any deviation from these principals, without a written approval of the course staff, will resolve in reducing points from the assignment.

### Questions and solutions:

* Read the questions carefully and answer exactly what you were asked to do.
* The style of the solutions should be in line with what is learned on the course.
* Do not copy solutions! Copying solutions will resolve of a zero grade for the task.
* Administrative emails on work-related questions should be sent to or.benson@post.runi.ac.il (no technical questions).
* Submit your technical questions on Piazza, please check if there is already the same question you want to ask.
* For clarifications or any other assistance, reception hours are also available by appointment.

### Checking:

* The code should be well-organized, readable, and include comments where necessary. Proper naming conventions for variables and functions should be followed.
* The code should be optimized for performance, avoiding unnecessary computations.
* The code should include appropriate error handling and data validation to manage unexpected inputs or situations. It’s ok if not all edge cases are covered.
* A clear explanation or documentation of the approach used, especially for complex problems. Include any assumptions made during problem-solving.
* The solution will be checked for correctness, completeness, and adherence to the given instructions and format.
* The solution should be original work by the students.
* The solution will be checked with a specific scoring key, that will be posted after grading.

In this project, you will leverage Kubernetes to build a Twitter Insight Engine. You'll use containerized applications to process and analyze Twitter data, aiming to derive valuable insights and present them in a clear executive summary.

**Tasks:**

1. **Twitter Data Acquisition: (10 points)**

Source real-time Twitter data using the Twitter API. Ensure the dataset includes diverse data types (text, timestamps, user metrics) for comprehensive analysis. More on Twitter API v2 - [Twitter API Documentation | Docs | Twitter Developer Platform](https://developer.twitter.com/en/docs/twitter-api). For testing or if you didn’t able to use the API, use the following dataset - [Tweets Dataset (kaggle.com)](https://www.kaggle.com/datasets/mmmarchetti/tweets-dataset)

1. **Kubernetes Cluster Configuration: (10 points)**

Design a Kubernetes cluster setup to efficiently manage and process the Twitter data. Focus on pod configurations, resource allocation, and network policies to optimize data throughput and processing.

1. **Data Ingestion and Processing: (20 points)**

Develop scripts (in Python) to ingest Twitter data into the system. Address data cleaning, transformation, and ensure data integrity within the Kubernetes environment.

1. **Data Analysis and Insight Extraction: (30 points)**

Create microservices in Kubernetes to analyze Twitter data. Emphasize Kubernetes capabilities like scalability and efficient resource management to handle large-scale data.

1. **Insight visualizations: (10 points)**

Integrate your data processing and analysis logic within a Python environment running on Kubernetes. Utilize Python libraries (e.g., NumPy, pandas) for additional data analysis or visualization tasks.

1. **Executive Summary: (20 points)**

Compose an executive summary, limited to two pages, including: A brief description of the Twitter data and the rationale behind its selection. An overview of the Kubernetes cluster configuration and design rationale. Key insights derived from the data with examples of how Kubernetes enhanced the analysis. Challenges encountered during the project and the solutions implemented. Visualizations (graphs, charts) that underscore significant findings.

**Deliverables**:

1. **Code Repository:**

A GitHub repository containing: All scripts for data ingestion, processing, and analysis in Kubernetes. Configuration files and deployment scripts for the Kubernetes environment. Any additional scripts or files utilized in the project.

1. **Executive Summary:**

A concise two-page PDF document summarizing the findings, methodology, and insights gained from the Twitter data using Kubernetes.

**Enjoy coding! 😊**