# HW 1 – Apache Cassandra

## Guidelines:

### Submissions:

* The assignment must be submitted by 25.01.24 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 HW1\_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 in 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 solution will be checked with a specific scoring key, that will be posted after grading.

In this homework, you will leverage Apache Cassandra and Python to analyze chosen datasets. In addition, you will derive meaningful insights and present findings in a concise executive summary.

**Tasks:**

1. **Dataset selection: (10 points)**
   * Choose dataset from Kaggle or any other preferred source.
   * The dataset should have at least three different tables/CSV files to make it more challenging. Make sure to provide the link to the dataset in your documentation.
   * The dataset should be complex enough to enable insightful analysis (preferably containing multiple types of data fields like text, numbers, dates, etc.).
2. **Cassandra Database Design: (20 points)**
   * Design a schema that efficiently stores your selected dataset in Apache Cassandra.
   * Consider partition keys, clustering columns, and appropriate data types to optimize queries.
3. **Data Ingestion: (10 points)**
   * Write Python scripts to ingest the dataset into your Cassandra database.
   * Ensure data integrity and handle any data cleaning or transformation that might be necessary.
4. **Data Analysis: (30 points)**
   * Develop CQL (Cassandra Query Language) queries to extract meaningful insights from the data.
   * Queries should demonstrate understanding of Cassandra’s strengths, such as fast reads, scalable writes, and efficient handling of large datasets.
5. **Python Integration: (10 points)**
   * Integrate your Cassandra queries within a Python environment.
   * Use Python libraries (e.g., pandas, matplotlib) for additional data processing or visualization if needed.
6. **Executive Summary: (20 points)**
   * Prepare an executive summary, ensuring it does not exceed two pages, and should include the following:
     + A brief description of the selected dataset and why it was chosen.
     + An overview of the database schema and rationale for the design.
     + Key insights derived from the data with supporting queries and results.
     + Any challenges faced during the assignment and how they were overcome.
     + Visualizations (charts, graphs) that highlight the most interesting findings.

**Deliverables**:

1. **Code Repository:** 
   * A GitHub repository containing:
   * All Python scripts for data ingestion and analysis.
   * CQL scripts for database schema and queries.
   * Any additional scripts or files used in the project.

1. **Executive Summary:** 
   * At most two pages pdf file summarizing the findings and approach.