Clinical Trials Analytics via PySpark

Homework 1 - *Big Data Engineering, 2025 @ UNINA*

Introduction to Clinical Trials

Clinical trials are scientific studies conducted on human subjects to evaluate the safety and effectiveness of medical treatments, drugs, devices, or procedures. Each trial follows a defined protocol and may span several phases (Phase I, II, III, IV), sometimes involving thousands of participants across multiple countries.

The Dataset

The dataset used in this exercise comes from **Dimensions.ai**, a platform that aggregates data on global scientific research. Each row in the dataset represents a clinical trial; informations about the columns can be found in the provided `legend.csv` file.

Some columns contain *nested or structured data*, such as lists of conditions, organizations, or locations.

Task

Perform at least five analytics using PySpark on the provided clinical trials dataset. The results must be compiled and presented in a structured PDF report. For each analysis, the report should include the following components:

- **Objective:** The goal of the analysis.
- **Description:** A brief description of the methodology used.
- **Code:** The PySpark code used to perform the analysis.

Include analyses of varying complexity, from basic aggregations to more complex operations.

Examples of Analytics

1. Number of studies started per year

2. Average number of participants per study title

3. Top 10 most frequent medical conditions

4. Countries with the highest average number of participants per study