**Big Data Tools and Technologies**

This project involved the use of python, SQL, databricks and Apache PySpark to analyze large datasets and to train machine learning models.

**Data Analysis and Visualization of clinical trials datasets**

The purpose of this task was to prepare, analyze and gain insights from a clinical trial dataset. The following questions were answered using Resilient DIstributed Datatsets (RDD), Dataframe (DF) and SQL:

1. The number of distinct studies in the dataset.
2. Types of studies in the dataset along with the frequencies of each type. These should be ordered from most frequent to least frequent.
3. The top 5 conditions with their frequencies.
4. Find the 10 most common sponsors that are not pharmaceutical companies, along with the number of clinical trials they have sponsored.
5. Plot number of completed studies for each month in 2023. The results can be found in the noteboooks: Ifiok\_Inuk\_sql.sql, Ifiok\_Inuk\_df.ipynb and Ifiok\_Inuk\_rdd.ipynb

**Machine Learning - Game Recommender System**

The purpose of this task was to train a collaborative filtering recommender system (using implicit user feedback) on the dataset provided by Steam, an online video game distribution service. The provided dataset was “steam-200k.csv”. The ML model was trained using Alternating Least Squares (ALS) matrix factorization. This can be found in the Ifiok\_Inuk\_ml.ipynb notebook.