**Basic Information:**

Project Title: MovieLens Data Exploration and Analysis

Student: Giodampalis Evangelos

Email: [giodampalis.e@live.unic.ac.cy](mailto:giodampalis.e@live.unic.ac.cy)

**Overview**

For my final project in the "Big Data Management and Processing" course, I conducted an analysis on the "MovieLens" dataset. I obtained the data from the MovieLens website, downloaded it, and then inserted it into a MongoDB (NoSQL database). Subsequently, I utilized Apache Spark to extract and analyze the data, aiming to derive valuable insights.

To execute this project, I set up MongoDB on my computer. Additionally, I utilized MongoDB Compass, the user interface for MongoDB. I also installed Apache Spark-3.4.2, configured to run with Apache Hadoop3. To establish a connection between Spark and MongoDB, I integrated and installed the Mongo-Spark connector. This integration allowed me to execute PySpark code from Jupyter notebook, providing access to and analysis of the data stored in the database.

In the process of data analysis, I loaded information from the database to identify the best movies based on user ratings. Furthermore, I explored whether the highest-rated movies were also the most frequently rated, attempting to discern if users consistently rate movies they genuinely enjoy. Lastly, I delved into movie genres, identifying the most prevalent genres across all movies and determining user preferences based on their ratings of movies.

**Background and Motivation:**

The initiation of the MovieLens dataset project is rooted in my profound interest in the convergence of data science and the entertainment industry. Films serve as entertainment and reflect societal preferences and cultural trends. The focal problem of this project revolves around reviling the best movies based on users rating and understand how users' rate. Also, comprehending user preferences genres of movies and extracting insightful findings that hold the potential to influence recommendation systems and the broader film industry.

Understanding user preferences can wield influence not only over online streaming platforms but also across the expansive landscape of the film industry. Insights into user ratings and preferences can serve as a guiding force for filmmakers and studios in crafting content that resonates with their target audience.

Analyzing the MovieLens dataset presents a formidable technological challenge owing to its voluminous nature. Effectively managing and processing this data necessitates the application of big data management and processing solutions, rendering it an ideal undertaking to explore the integration of these technologies.

**Data Sources:**

I acquired the MovieLens movie ratings dataset, which contains 100836 ratings and 3683 tag applications across 9742 movies. The dataset is accessible from the MovieLens website, where it can be downloaded. The substantial volume of data renders it conducive to a Big Data project.

**Big Data Dimensions:**

The MovieLens dataset introduces big data challenges primarily in terms of its considerable volume. With thousands of ratings and diverse movie genres, the dataset necessitates effective big data management and processing solutions to derive meaningful insights.

**Tools/Libraries:**

1. NoSQL Database:

MongoDB for efficient data storage and retrieval.

2. Big Data Processing:

Apache Sparke for scalable and parallel processing.

3. Data Analysis:

Jupyter notebook using Python code with pyspark ,Pandas, NumPy for exploratory data analysis. Visualization facilitated by Matplotlib.