PROJECT PROPOSAL

Evros Vasileiou
vasileiou.e8@live.unic.ac.cy

Project Title:
Netflix Big Data Analytics: Catalog Trends and NoSQL Batch Processing

Background & Motivation:

In this project, I will examine the Netflix catalog using Big Data concepts. Streaming platforms such as Netflix rely on large, semi-structured datasets to make business and recommendation decisions. The public Netflix dataset available on Kaggle reflects many real-world data challenges, including multi-valued fields, inconsistent formats, missing values, and duplicated rows. These characteristics make it an ideal dataset for demonstrating the techniques taught in the course.

I chose this topic because Netflix is widely recognizable, and analyzing its catalog can reveal interesting insights about global content, genres, and trends. At the same time, the dataset is manageable while still allowing me to build a complete data-processing workflow with cleaning, modeling, storage, and analytics. Beyond the academic interest, I also enjoy watching movies and TV shows in my free time. This made the Netflix dataset especially appealing, because it allows me to explore a platform I use personally while applying Big Data principles in a meaningful way.

Project Objectives:

This project will aim to answer several questions about the Netflix catalog by building a complete analytics workflow. I will investigate how the size and composition of the catalog changes over time, what the distribution is between Movies and TV Shows, which genres appear most frequently, and which countries contribute the most content. I will also analyze patterns in ratings and durations to better understand differences between various categories of titles.

To support these goals, I will develop a NoSQL document model suitable for storing semi-structured metadata. I will then perform queries and batch analytics on this model using MongoDB. The results will provide insights into catalog trends, global diversity, and historical growth patterns. The overall objective is not only to analyze the dataset but also to demonstrate how Big Data principles can be applied to real metadata.

Data Sources:

The main dataset I will use is the “Netflix Movies and TV Shows” dataset from Kaggle, stored locally as a CSV file. It contains around eight thousand titles, each with structured fields such as title, release year, rating, and type, as well as semi-structured fields like multiple genres and multiple countries listed per title. These features will allow me to demonstrate data cleaning, normalization, and schema design.

The Kaggle dataset itself already reflects several key Big Data dimensions. In terms of volume, the dataset contains thousands of entries, and the same schema could scale to millions. The variety dimension is visible in multi-valued attributes, free-form text fields, inconsistent duration strings, and mixed data types. Veracity challenges include missing values, ambiguous country labels, and duplicated show identifiers. Conceptual velocity is represented through the date_added field, which indicates when titles were added to the platform. These characteristics make the dataset an appropriate choice for modeling in a NoSQL environment.

Big Data Dimensions:

This project will address multiple Big Data dimensions presented in Lecture 1. Volume will be demonstrated through the dataset size and the scalability of the chosen data model. Variety will appear in the mix of structured and semi-structured fields and in the inconsistent formatting of attributes such as durations and country lists. Veracity will be evident through the need to clean missing values, standardize fields, and remove duplicates. Conceptual velocity will be illustrated by the timeline of catalog additions and the real-world behavior of streaming platforms where data updates occur continuously. Finally, the value dimension will be reflected in the meaningful insights the analysis will generate for understanding catalog trends.

By explicitly mapping these characteristics to Big Data dimensions, I will justify why the problem requires approaches beyond simple in-memory processing and why NoSQL systems are suitable for such semi-structured datasets.

Solution Overview:

To address these objectives, I will build a complete data-processing pipeline. I will begin by inspecting the dataset and identifying quality issues such as missing values and formatting inconsistencies. Then, I will perform cleaning steps, including parsing dates, splitting multi-valued fields, normalizing duration strings, and removing duplicate entries. This preparation will make the data ready for both analysis and storage.

Next, I will conduct exploratory data analysis to visualize catalog patterns such as content distribution over time, the prevalence of Movies versus TV Shows, the most common genres, and the leading countries. These visualizations will help contextualize the results of the later NoSQL analysis.

After the data has been cleaned and transformed, I will design a document-oriented schema and store the dataset in MongoDB, where fields such as genres and countries will be represented as arrays inside each document. This approach will follow denormalization principles discussed in the lectures and will eliminate the need for relational joins that would otherwise occur if genres and countries were stored in separate tables.

Once the data is stored in MongoDB, I will implement queries to retrieve titles by country, genre, release period, and duration characteristics. I will also create aggregation pipelines to compute grouped statistics such as yearly title counts, top genres, most common countries, rating distributions, and average durations. These pipelines will demonstrate batch analytics capabilities in a Big Data context.

Tools and Technologies:

For this project, I will use Python for data processing and cleaning, relying on pandas and NumPy for transformation tasks. I will use matplotlib to create visualizations during exploratory analysis. The NoSQL backend will be MongoDB, accessed through the PyMongo library. All work will be carried out in a Jupyter Notebook environment to ensure reproducibility and clarity.

Expected Outcome:

By the end of the project, I will deliver a fully functional data-processing workflow that demonstrates Big Data concepts in practice. The final result will include cleaned and structured Netflix metadata, a NoSQL schema stored persistently in MongoDB, practical queries for data exploration, and aggregation pipelines for batch analytics. Visualizations will help illustrate the key findings. I will also prepare a short video presentation summarizing the process, results, and Big Data principles involved.