|  |
| --- |
| ***Project: Building a Media Streaming Platform with***  ***IBM Cloud Video Streaming*** |

**INTRODUCTION:**

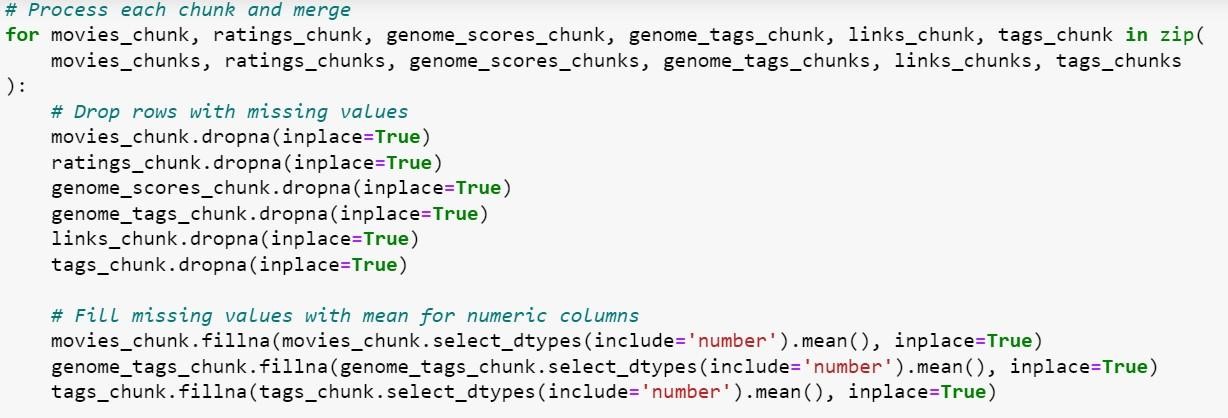
* The rapid growth of digital content consumption has led to a corresponding increase in the demand for media streaming platforms. However, many existing streaming platforms offer a similar user experience, with limited differentiation.
* This project aims to build a media streaming platform using IBM Cloud Video Streaming.
* The first step in this process is to load and preprocess the dataset. This document will provide a step-by-step explanation about loading and preprocessing the dataset in python.

# Step 1: Load the datasets

* The first step is to load the necessary datasets using the Pandas library in Python. The datasets are read in chunks to efficiently manage large data files. The datasets may include:
  + movies.csv
  + ratings.csv
  + genome-scores.csv
  + genome-tags.csv
  + links.csv
  + tags.csv
* To load the datasets, we can use the following code:

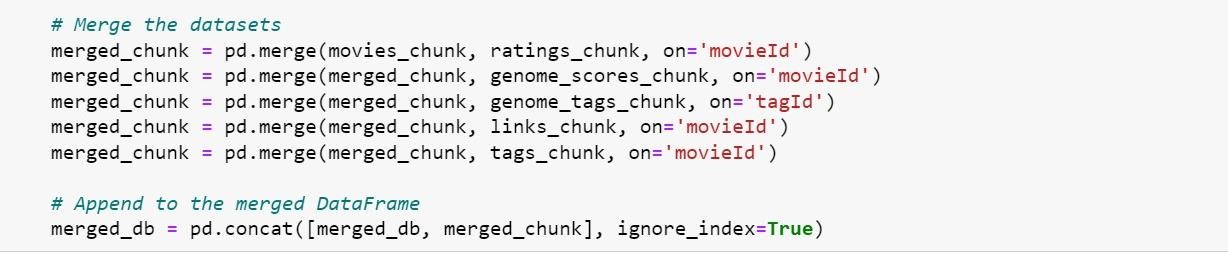
# Step 2: Preprocess the data

* Once the datasets have been loaded, we can begin preprocessing them. This involves cleaning the data, removing duplicate rows, and filling in missing values.
* To clean the data, we can use the following code:



# Step 3: Merge the datasets

* Once the datasets have been cleaned, we can merge them together. This will create a unified dataset that contains all of the information about the movies, ratings, genome scores, genome tags, links, and tags.
* To merge the datasets, we can use the following code:



# Step 4: Remove duplicate rows

* Once the datasets have been merged, we need to remove any duplicate rows.

This will ensure that the dataset is clean and accurate.

* To remove duplicate rows, we can use the following code:



# Step 5: Save the preprocessed dataset

* Once the preprocessed dataset has been created, we can save it to a CSV file. This will make it easy to load the dataset into IBM Cloud Video Streaming for use in our media streaming platform.
* To save the preprocessed dataset, we can use the following code:



**CONCLUSION:**

* + In this document, we have covered the essential steps for cleaning and preprocessing data from dataset using Python.
  + The provided Python script efficiently handles data cleaning and merging of multiple data sources to create a unified, high-quality dataset.