SMU Data science Bootcamp

ETL project - Movies

Doris Batalia, Justin Schalenky and Madhuram S

19-Feb-2019 to 23-Feb-2019

# Overview

## Project Background and Description

In this project, we are planning to extract, transform and load data into a SQL/NoSQL Database with goal to build analytics and visualization for Movies data obtained from TMDB (The Movies DataBase) and OMDB (Open Movies DataBase).

We will use only movies from 1-Jan-2009 to 31-Dec-2017

Below are key questions/areas we will be using analyze the data,

* Movie Genre and its profitability
* Movie popularity against its production cost
* How does Critics rating compare with audience ratings?

## Project Scope

Scope of activities include: -

* Define objectives (Why, what and how).
* Define Sources of Data and Destination where it will be stored
* Define Data **E**xtraction, **T**ransformation and **L**oad rules
* Perform Data **E**xtraction, **T**ransformation and **L**oad
* Data Analysis & Visualization
* Presentation through a website (Nice to have: Flask based)

## High-Level Process Steps / Activities

The high process and steps are provided below,

Data obtained from 2 data sources

* TMDB-5000.csv from Kaggle
* JSON from OMDB API

Python script created for data extraction. Libraries used will be,

* Pandas – extraction
* Requests – API





Explore and Identify,

* Datasets info – size, columns and datatypes
* Categorical / non-categorical columns
* Primary column(s) for merging
* Data Issues like, missing values
* Required columns for analytics



Using Python,

* Handle missing data
* Apply target formatting
* Rename Columns
* Filter and Group
* Derive new fields
* Merge the 2 datasets.
* Prepare for loading



**MongoDB** will be used. Using Python and PyMongo,

* Create Database
* Load data to Collections
* Verify Load success using Query
* Verify Data Integrity



**Visualize**: -

Using Matplotlib

* Generate plots that answers our queries

**Publish: -**

* With HTML & CSS, publish findings on Internet
* Flask (TBD)



## Data Definitions

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source Data: TMDB  File Name: tmdb\_5000\_movies.csv  Format: CSV | | Destination Database: MoviesDB  Destination Collection name: TMDB\_source\_data  Location: @localhost:27017 | | | |
| **Field Name** | **DataType** | **Field Name** | **DataType** | **Req?** | **Format (Target)** |
| budget | Numeric | budget | Numeric | Y | Numeric (0.00) |
| genres | Text | genres | Text | Y | Text |
| homepage | Text | homepage | Text | N | Text |
| id | int | id | int | N |  |
| keywords | Text | keywords | Text | Y | Text |
| original\_language | Text | original\_language | Text | N |  |
| original\_title | Text | original\_title | Text | N |  |
| overview | Text | overview | Text | Y | Text |
| popularity | Int | popularity | Int | Y | Int |
| production\_companies | Text | production\_companies | Text | Y | Text |
| production\_countries | Text | production\_countries | Text | Y | Text |
| release\_date | Text | release\_date | Text | Y | YYYY-MM-DD |
| revenue | Numeric | revenue | Numeric | Y | Numeric (0.00) |
| runtime | Numeric | runtime | Numeric | N | Numeric (0.00) |
| spoken\_languages | Text | spoken\_languages | Text | N |  |
| status | Text | status | Text | Y | Text |
| tagline | Text | tagline | Text | Y | Text |
| title | Text | title | Text | Y | Text |
| vote\_average | Numeric | vote\_average | Numeric | Y | Numeric (0.00) |
| Vote\_count | Numeric | Vote\_count | Numeric | Y | Int |

## Extraction Process

|  |  |
| --- | --- |
|  | Details of Extraction Process |

## Data Exploration

|  |  |
| --- | --- |
|  | Details on the data set obtained |

TMDB :

## Data Transformation

|  |  |
| --- | --- |
|  | Details on transformation process |

## Visualization

|  |  |
| --- | --- |
|  | Details on plots we will build |

## Publishing

|  |  |
| --- | --- |
|  | Screenshots of pages |

## Conclusion

|  |  |
| --- | --- |
|  | Explain Results / Findings |

# Approval and Authority to Proceed

We approve the project as described above, and authorize the team to proceed.

|  |  |  |
| --- | --- | --- |
| Name | Title | Date |
|  |  |  |
|  |  |  |
|  |  |  |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |
| Approved By |  |  | Date |  | Approved By |  |  | Date |