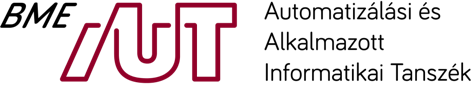
**Házi feladat dokumentáció**

Üzleti Intelligencia

Félév: 2023 ősz

**Online videótéka, a Netflix**

**Pethes Dániel Imre - (EAQXCB)**

**pethesdani.imre@gmail.com** 

# **Projekt elérhetősége**

A projekthez tartozó ETL forrásfájlok, SQL scriptek és Power BI Desktop állomány megtalálható az alábbi Github repository-ban: <https://github.com/pethesdaniel/UzInt-Netflix>

# **Felhasznált technológiák**

A projekthez az alábbi technológiákat használtam:

* Microsoft SQL Server 2022 Developer
* Microsoft SQL Server Management Studio
* Microsoft Visual Studio 2022 Community (with SSIS and SSRS addons)
* R 4.3.2 (CRAN)
* Git (Github)

# Felhasznált adatok

A projekthez az alábbi adatforrásokat használtam fel:

* Netflix Movies and TV Shows
  + <https://www.kaggle.com/datasets/shivamb/netflix-shows>
* Hulu Movies and TV Shows
  + <https://www.kaggle.com/datasets/shivamb/hulu-movies-and-tv-shows>
* Netflix Revenue and Usage Statistics
  + <https://www.kaggle.com/datasets/adnananam/netflix-revenue-and-usage-statistics>
* IMDb Dataset
  + <https://www.kaggle.com/datasets/ashirwadsangwan/imdb-dataset>
* Google Trends: netflix, amazon prime video, hulu
  + <https://trends.google.com/trends/explore?date=all&q=netflix,amazon%20prime%20video,hulu&hl=en-US>

Végül sajnos nem sikerült az Amazon Prime Video ajánlásaival összevetni a Netflix ajánlásait, mivel amit találtam dataset nem volt teljes, nem tartalmazott adatokat a tartalmak platformhoz való hozzáadásáról, más datasetet nem sikerült találnom, így csak a Hulu-val tudtam összevetni a tartalmak átfedését.

# Adatok előkészítése

Az adatok nagy része nem igényelt előkészítést, azonban az IMDb Dataset igen, nagy mérete és alacsony kihasználtsága miatt. A dataset 10 millió bejegyzést tartalmaz, ebből nekünk kb. 10000-re van szükségünk.

Ezen kívül transzformálást igényel, mert a dataset külön tsv fájlban tárolja a nekünk szükséges értékelési adatokat az általános metaadatoktól (cím, kiadás éve stb.).

Az adatok átalakításához létrehoztam egy különálló ETL folyamatot. Ez a folyamat felelős a címek, értékelések és értékelők számának elkülönítésért. Az adott ETL folyat a következőképp néz ki:

A képen képernyőkép, szöveg, multimédia, szoftver látható

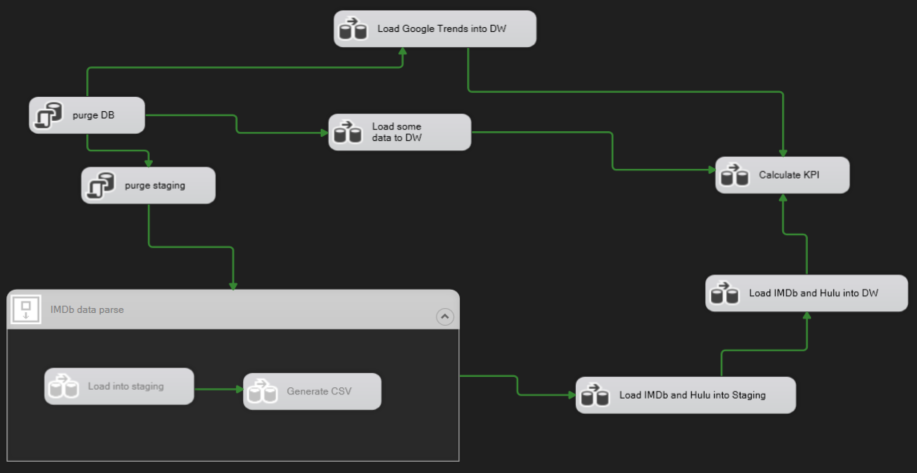
Automatikusan generált leírás

A folyamat során betöltöm az adatokat a staging adatbázisba, majd az adott táblákon SQL lekérdezés segítségével összejoinolom az adatokat, melyeket kimentek egy CSV fájlba.

Ez a teljes ETL folyamaton belül egy szekvenciában található, mellyel könnyen kikapcsolhatjuk a folyamat ezen részét, ezzel időt és erőforrásokat megtakarítva.

# ETL folyamat

A teljes ETL folyamat az alábbi ábrán látható:



A folyamat első 2 lépésben törlöm a DW és staging adatbázis tábláinak tartalmát. Egyes adatokon a transzformációk könnyen elvégezhetőek, ezért azokat közvetlenül a DW-ba töltöm be. Ezek pl. a sorozatok adatai és a feliratkozó/bevétel adatok, illetve a Google Trends adatok és a KPI számítás. Ezeken a SSIS által szolgáltatott adat átalakító műveletekkel, valamint C# scriptekkel végzek transzformációkat.

Más adatokon komplexebb transzformációkat végzek, vagy nagyobb adatmennyiséggel dolgozom, ahol pl. az SSIS által nyújtott MergeJoin nem elég optimalizált, ez esetben az adatokat a staging adatbázis tábláiba töltöm be és SQL scriptek segítségével manipulálom őket. Ilyen adatok az IMDb dataset adatainak feldolgozása és összehasonlítása a Netflix adatokkal, illetve a Hulu tartalmának feldolgozása és összehasonlítása.

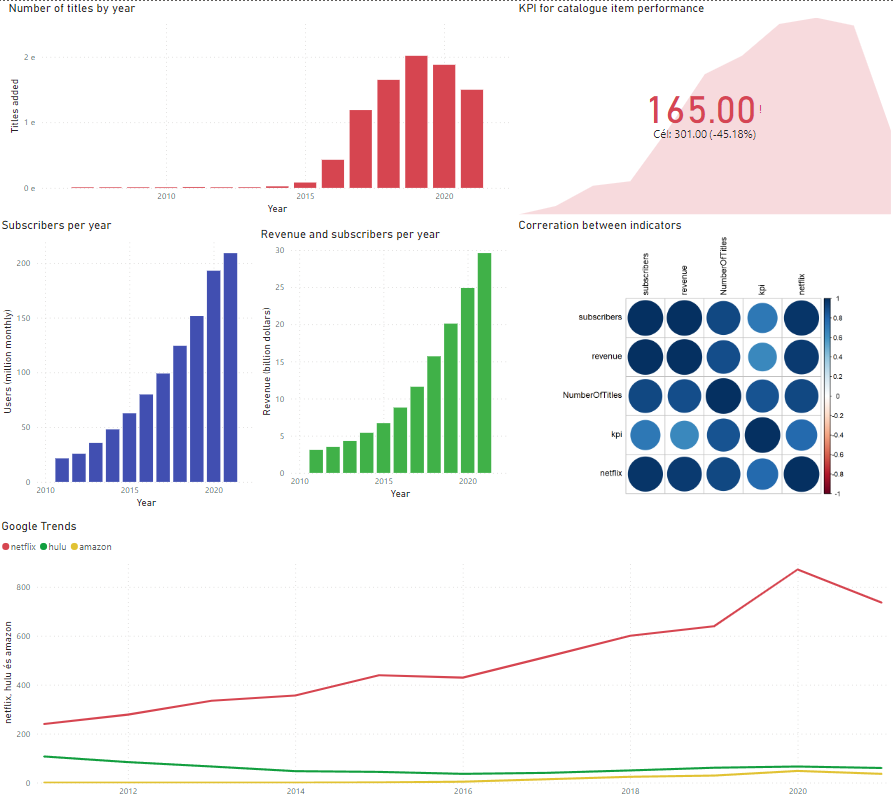
# Power BI Desktop

A vizualizációkat Power BI Desktop segítségével készítettem. Az adatok az SQL adatbázisból Importálás lekérdezési módot használva kerülnek betöltésre. Az elkészült projekt adatstruktúrája az alábbi ábrán látható.

A képen szöveg, képernyőkép, Téglalap, diagram látható

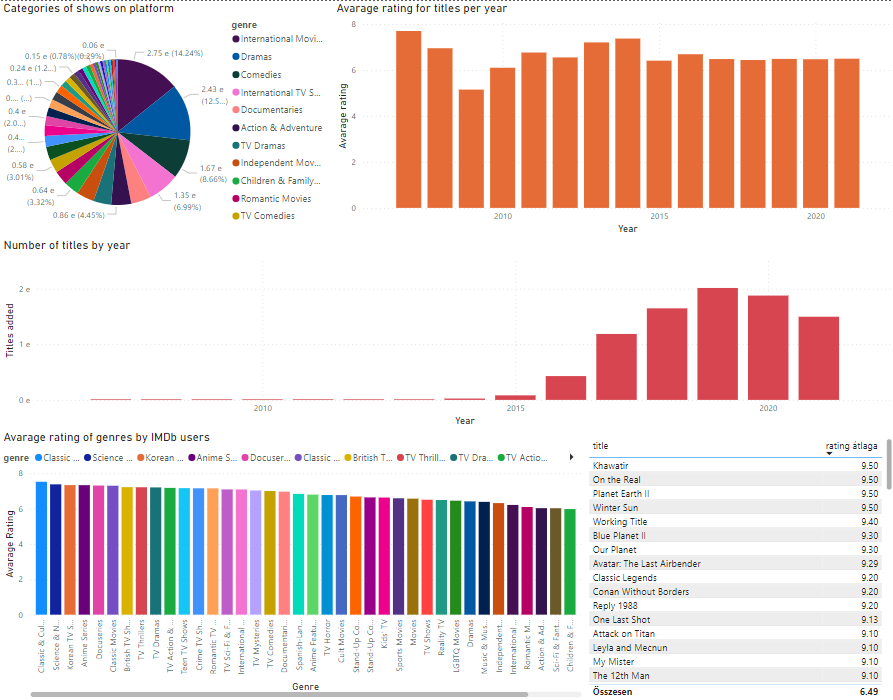
Automatikusan generált leírás

Az elkészült fájl 3 nézetből áll, az oldalak mérete változó a diagrammok számától függően.



Az első oldalon az üzleti adatok láthatóak, ezek a platformon található tartalmak számossága és az ebből származó KPI, a feliratkozók száma, a bevétel, a Google Trends adatok, mind évekre lebontva.

Ezen kívül megtalálható még egy korrelációs diagram, mely összeveti ezeket az adatokat. Ezt egy R szkript generálja a corrplot könyvtár segítségével.



A második oldalon a platformon található tartalmakat és azok értékeléseit tudjuk megtekinteni évek és kategóriák alapján. Itt szűrhetünk különböző kategóriákra, azok értékeléseire, egyes évekre, vagy akár konkrét sorozatokra.

A képen szöveg, képernyőkép, diagram, Diagram látható

Automatikusan generált leírás

A harmadik oldalon egy összevetést láthatunk a Hulu streaming szolgáltató ajánlásaival. A csúszka segítségével beállíthatjuk a vizsgált időtartam végét, illetve a waterfall diagrammon láthatjuk az összesítést.