

SVEUČILIŠTE U ZAGREBU  
FAKULTET ORGANIZACIJE I INFORMATIKE  
VARAŽDIN

Iva Udovčić

**ANALIZA FILMOVA I SERIJA DOSTUPNIH  
NA HBO PLATFORMI**

**PROJEKT**

**Varaždin, 2024.**

SVEUČILIŠTE U ZAGREBU  
FAKULTET ORGANIZACIJE I INFORMATIKE  
VARAŽDIN

Iva Udovčić

Matični broj: 00161480057

Studij: Organizacija poslovnih sustava

**ANALIZA FILMOVA I SERIJA DOSTUPNIH NA HBO PLATFORMI**

**PROJEKT**

**Mentor/Mentorica:**

mag. educ. inf Maja Cerjan

**Varaždin, svibanj 2024.**

*Iva Udovčić*

**Izjava o izvornosti**

Izjavljujem da je moj završni/diplomski rad izvorni rezultat mojeg rada te da se u izradi istoga nisam koristio drugim izvorima osim onima koji su u njemu navedeni. Za izradu rada su korištene etički prikladne i prihvatljive metode i tehnike rada.

*Autor/Autorica potvrdio/potvrdila prihvaćanjem odredbi u sustavu FOI-radovi*

---

## Sažetak

Ovaj projekt prikazuje ETL proces nad odabranim skupom podataka. Slijedeći korake tog procesa prikazano je kako se kreira skladište, odnosno model zvijezde. Prikazano je i kako različiti izvještaji u alatima za vizualizaciju podataka mogu pomoći u razumijevanju podataka i njihovom lakšem prikazu. Korišteni alati su znatno pomogli u transformaciji podataka te su uz SQL naredbe omogućili nastanak podataka koji su strukturirani i imaju smisla. Učitavanje podataka je treći korak ETL procesa i ujedno korak koji je vrlo važan jer prikazuje koliko su dobro odradjeni prethodna sva koraka. Model zvijezde je prepoznatljiv model koji skladište dijeli na dimenzijske i činjenične tablice. U ovom projektu napravljeno je sedam dimenzijskih i naravno jedan činjenična tablica.

**Ključne riječi:** ETL proces, model zvijezde, naredbe, transformacija, dimenzijske tablice, činjenična tablica

# Sadržaj

1. Uvod .....	1
2. Opis domene .....	2
2.1. Skladišta podataka .....	3
2.2. Korištene tehnologije.....	4
3. Izrada skladišta podataka.....	5
4. ETL postupak .....	6
5. Model zvijezde .....	17
5.1. Dimenzijska tablica <i>age_certification</i> .....	18
5.2. Dimenzijska tablica <i>genres</i> .....	19
5.3. Dimenzijska tablica <i>imdb_data</i> .....	20
5.4. Dimenzijska tablica <i>ostalo</i> .....	21
5.5. Dimenzijska tablica <i>production_countries</i> .....	23
5.6. Dimenzijska tablica <i>tmdb_data</i> .....	24
5.7. Dimenzijska tablica <i>type</i> .....	25
5.8. Činjenična tablica <i>general_table</i> .....	27
5.9. Prikaz strukture skladišta i modela.....	28
6. Analiza u alatu Tabeau .....	29
6.1. Izvještaji .....	30
7. Zaključak .....	35
Popis literature.....	36
Popis slika .....	37

# 1. Uvod

Ovaj projektni rad obuhvaća izradu skladišta podatka te vizualizaciju tih podataka. Za izradu skladišta odabran je skup podataka vezanih za filmove i serije koji su dostupni na HBO platformi. Podaci su preuzeti s interneta te su različitim transformacijama pretvoreni u skup podatka pogodan za analizu i kreiranje izvještaja. Korišteni alati su MySQL Workbench i Tableau.

Rad prikazuje koje transformacije nad podacima su bile potrebne te koje naredbe u MySQL Workbenchu su omogućile stvaranje skladišta i modela zvijezde. Model zvijezde se sastoji od sedam dimenzijskih tablica i od jedne činjenične. Podaci su provedeni kroz ETL postupak (ekstrahiranje, transformiranje i učitavanje podataka). Osim u MySQL Workbenchu podaci su transformirani i u Excelu budući da je u preuzetoj csv datoteci bilo potrebno razdvojiti stupce radi lakšeg daljnog rada s podacima. Na kraju rada se nalaze izvještaji kreirani u alatu Tableau, ukupno je šest izvještaja koji prikazuju različite podatke i njihove ovisnosti i odnose, a sve sa svrhom lakšeg razumijevanja podataka.

U projektu je ukratko objašnjeno što je to ETL, model zvijezde te koja je svrha skladišta podataka. Osim toga, opisane su i korištene tehnologije te opis domene i atributi koji se nalaze u preuzetom skupu podataka. Na samom kraju nalazi se zaključak.

## 2. Opis domene

U ovom projektu analiziran je skup podataka preuzet sa stranice Kaggle. U tom skupu podataka su podaci vezani za serije i filmove na streaming platformi HBO. Podaci su iz svibnja 2022. godine i odnose se na područje SAD-a. Za potrebe kreiranje skladišta korištena je csv datoteka titles.csv koja sadrži sljedeće podatke:

**id:** id naziva filma ili serije na JustWatch

**title:** naziv filma

**type:** označava radi li se o filmu ili seriji

**description:** kratki opis filma ili serije

**release\_year:** godina izdavanja filma

**age\_certification:** sustav oznaka koji određuje je li sadržaj prikladan za određenu starosnu grupu, a koristi se u SAD-u. Postoje filmske i televizijske oznake. Filmske su:

- G – (eng. *General Audiences*) sadržaj prikladan za sve uzraste (filmska oznaka),
- PG – (eng. *Parental Guidance Suggested*) odnosi se na sadržaj koji može sadržavati materijal koji nije prikladan za sve uzraste,
- PG-13 – (eng. *Parents Strongly Cautioned*) – sadržaj nije prikladan za djecu mlađu od 13 godina,
- NC-17 (eng. *No One 17 and Under Admitted*) – vrlo eksplicitan sadržaj i neprikladan za mlađe gledatelje,
- R (eng. *Restricted*) – djeca ispod 17 godina smiju gledati uz prisustvo odrasle osobe.

Televizijske oznake su:

- TV-G (eng. *General Audience*) – prikladno za sve uzraste,
- TV – PG (eng. *Parental Guidance Suggested*) – sadržava materijal koji nije prikladan za sve uzraste, preporučuje se nadzor roditelja,
- TV-Y (eng. *All Children*) – sadržaj namijenjen djeci svih uzrasta,
- TV-Y7 (eng. *Directed to Older Children*) – prikladan za djecu stariju od 7 godina.

**runtime:** duljina trajanja filma ili serije

**genres:** lista žanrova

**production\_countries:** lista zemalja u kojima je koji film ili serija snimljen

**seasons:** broj sezona kao se radi o seriji

**imdb\_id:** id naziva filma ili serije na IMDB

**imdb\_score:** ocjena na IMDB

**tmdb\_popularity:** popularnost filma ili serija na TMDB

**tmdb\_score:** ocjena na TMDB-u.

Svrha ovog skladišta je prikaz dostupnih filmova i serija na platformi, a osim toga analiziranjem podataka može se ustanoviti koji filmovi i serije su najbolje ocijenjeni te se mogu provesti različite usporedbe.

## 2.1. Skladišta podataka

Skladišta podataka omogućuju različite upite i analize nad podacima te podržavaju aktivnosti poslovne inteligencije. Skladišta sadrže velike količine povijesnih podataka te mogu dolaziti iz različitih izvora. Omogućuju organizacijama da dobiju bolji uvid u podatke. Najčešće sadržavaju dolje navedene elemente (Oracle, b.d.):

1. Relacijsku bazu podatka za pohranu i upravljanje podacima
2. ETL za pripremu podatka za analizu
3. Mogućnosti statističke analize, izvještavanja i rudarenja podataka
4. Vizualizaciju i prezentaciju podataka korisnicima.

Prva skladišta podataka javljaju se 80-ih godina prošlog stoljeća, a cilj je bio da omoguće brži protok podatka od operacijskog sustava do sustava za podršku odlučivanju. U to vrijeme postojalo je više DSS-ova koja su koristili različiti korisnici zbog toga što su skladišta tada zahtijevala redundanciju. Različita DSS okruženja najčešće su koristila iste podatke. Nova učinkovitija skladišta su se razvila iz pohrane informacija (Oracle, b.d.).

Umetna inteligencija i veliki podaci omogućili su napredak skladišta te su tako nastala autonomna skladišta koja omogućuju tvrtkama izvlačenje još više vrijednosti iz podataka (Oracle, b.d.).

Za poslovanje su skladišta podataka nužna jer čuvaju bitne podatke, neke od prednosti su dostupnost podataka što povećava učinkovitost. Podaci su dostupni svima te im se pristupa na jednostavan način. Iako, podaci dolaze iz više izvora osigurana je kvaliteta i dosljednost podataka odnosno podatci iz različitih izvora su u jedinstvenom formatu te se eliminira ponavljanje istih podataka. Vrlo važno je spomenuti da se u skladištu nalaze povijesni podaci što omogućava poboljšanja te pokazuje ključne elemente napretka. Osim toga, skladište je od velike važnosti i za donošenje odluka, a samim tim može se utjecati i na prihod poduzeća. Kao što je spomenuto pristup podacima je jednostavan, no radi sigurnosti može se ograničiti

dostupnost podataka određenim skupinama. Sigurnosne mjere se mogu provesti u obliku šifriranja stupaca ili stvaranja prilagođenih korisničkih grupa, a može se korisnicima dopustiti samo čitanje (INCWORX CONSULTING, b.d).

## 2.2. Korištene tehnologije

U ovom projektu korišteno je nekoliko tehnologija. Prvo je korišten MS Excel kako bi se napravile neke osnovne konverzije podataka odnosno kako bi razdvojili stupce. Nakon konverzija u MS Excelu podaci su uvezeni u alat MySQL Workbench u kojem su provedene konverzije podataka te je stvoren model zvijezde. Alat MySQL Workbench priprema podatke za daljnju analizu u alatu Tableau.

Prije stvaranja grafova u alatu Tableau bilo je potrebno povezati bazu podataka iz MySQL Workbencha sa Tableauom. To je napravljeno pomoću MySQL ODBC drivera.

MySQL Workbench je omogućuje modeliranje podataka, razvoj SQL-a, sigurnosno kopiranje i slično. Dostupan je za Windows, Linux i Mac OS X. Ovaj alat omogućuje *Forward i Reverse engineering*. Dakle, preko vizualnih alata koji su dostupni moguće je kreirati i izvršavati SQL upite te postoji vizualna kontrola za jednostavno administriranje MySQL okruženja (MySQL, b.d).

Alat Tableau je jako koristan za vizualizaciju podataka koja se koristi za poslovnu inteligenciju. Pomoću ovog alata mogu se stvoriti različiti dijagrami, grafovi i slično, a sve to sa svrhom boljeg razumijevanja podataka. Jedna od prednosti je što omogućava stvaranje izvještaja spajanjem različitih skupova podataka. Podržava sedam tipova podataka, a to su: ikona, tekst, vrijednosti datuma, vrijednosti datuma i vremena, numeričke podatke, bool vrijednosti te geografske vrijednosti i grupe klastera (Biswal, 2023).

MySQL ODBC driver je API za pristup bazama podataka. MySQL Connector/ODBC je obitelj drajvera koji omogućuju pristup MySQL bazama podataka korištenjem standardnog ODBC API-ja. Omogućava povezivanje s MySQL bazama putem upravljačkog i izvornog sučelja. Ova univerzalna, višeplatformska rješenja podržavaju operativne sustave poput Windowsa, Unixa i macOS-a (Verma 2022).

### 3. Izrada skladišta podataka

Skladište podataka se stvara kako bi se svi podaci nalazili na jednom mjestu te kako bi se s njima moglo manipulirati odnosno upravljati, pohranjivati i analizirati. U nastavku slijedi objašnjenje izrade skladišta i pronašlaska podataka.

Za izradu skladišta prvo je potrebno prikupiti informacije te analizirati kvalitetne izvore podataka odnosno ustanoviti koliko su ti izvori pouzdani. Potom je potrebno odabrati dizajn arhitekture skladišta podataka kao što je jednoslojna, dvoslojna i troslojna arhitektura. Nakon toga slijedi izrada ETL-a, ukratko to je postupak koji uzima podatke iz njihovih izvora i pretvara ih u format prikladan za analitiku. Kad je ETL napravljen slijedi dizajniranje podatkovnog modela i odabir sheme odnosno organizacija podataka unutar skladišta. Dizajn se prikazuje pomoću modela odnosa entiteta ili pomoću dimenzionalnog podatkovnog modela, a shema može biti zvezdasta (model zvijezde) ili snowflake shema (model pahulje). Šesti korak izrade skladišta podataka je izgradnja, testiranje i implementacija gdje se izgrađuje fizičko skladište, testiraju performanse i validiraju podaci. Posljednji korak je održavanje i nadzor a to podrazumijeva ažuriranja, praćenje kvalitete podataka i revizije te izveštavanja (Kropov, 2023).

Kao što je prije navedeno u ovom projektu se koristi skup podataka pronađen na stranici Kaggle. Skup se sastoji od 3374 zapisa u kojima se nalaze filmovi i serije dostupne na platformi HBO u Sjedinjenim Američkim Državama. Donja slika prikazuje izgled preuzetog skupa:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	<code>id,title,type,description,release_year,age_certification,runtime,genres,production_countries,seasons,imdb_id,imdb_score,imdb_votes,imdb_popularity,tmdb_score</code>																						
2	tm155702,The Wizard of Oz,MOVIE,"Young Dorothy finds herself in a magical world where she makes friends with a lion, a scarecrow and a tin man as they make their way along the yellow brick road to talk with the Wizard and ask for the things they miss most in their lives.",1939,PG-13,[{"genre": "drama"}, {"genre": "fantasy"}, {"genre": "family"}, {"genre": "mystery"}, {"genre": "romance"}, {"genre": "science fiction"}, {"genre": "thriller"}, {"genre": "war"], [{"country": "United States"}], 1, 8.2, 427000, 8.5, 558849.0, 20.0, 67.8, 2																						
3	tm83648,Citizen Kane,MOVIE,"Newspaper magnate, Charles Foster Kane is taken from his mother as a boy and made the ward of a rich industrialist. As a result, every well-meaning, tyrannical or self-destructive move he makes for the rest of his life appears in some way to be aimed at getting back what was taken from him.",1941,PG-13,[{"genre": "drama"}, {"genre": "romance"}, {"genre": "war"], [{"country": "United States"}], 1, 8.5, 558849.0, 20.0, 67.8, 2																						
4	tm77588,Casablanca,MOVIE,"In Casablanca, Morocco in 1941, a cynical American expatriate meets a former lover, with unforeseen complications.",1942,PG-102,[{"genre": "drama"}, {"genre": "romance"}, {"genre": "war"], [{"country": "United States"}], 1, 8.5, 558849.0, 20.0, 67.8, 2																						
5	tm82363,The Big Sleep,MOVIE,"Private Investigator Philip Marlowe is hired by wealthy General Sternwood regarding a matter involving his young daughter Carmen. Before the complex case is over, Marlowe sees murder, blackmail, deception, and what might be love.",1946,PG-13,[{"genre": "drama"}, {"genre": "mystery"}, {"genre": "romance"}, {"genre": "thriller"}, {"genre": "war"], [{"country": "United States"}], 1, 8.0, 156603.0, 12.788, 7.8																						
6	tm84701,The Maltese Falcon,MOVIE,"A private detective takes on a case that involves him with three eccentric criminals, a beautiful lar, and their quest for a priceless statuette.",1941,PG-100,[{"genre": "drama"}, {"genre": "mystery"}, {"genre": "romance"}, {"genre": "thriller"}, {"genre": "war"], [{"country": "United States"}], 1, 8.0, 156603.0, 12.788, 7.8																						
7	tm156463,Gone with the Wind,MOVIE,"The spoiled daughter of a Georgia plantation owner is forced to use every means at her disposal to clav her way out of poverty, following Major William Sherman's destruction of the American Civil War.",1939,PG-13,[{"genre": "drama"}, {"genre": "historical"}, {"genre": "romance"}, {"genre": "war"], [{"country": "United States"}], 1, 8.0, 156603.0, 12.788, 7.8																						
8	ts225761,Tom and Jerry,SHOW,"Tom and Jerry is an American animated franchise, and series of comedy short films created in 1940 by William Hanna and Joseph Barbera. Best known for its 161 theatrical short films by Metro-Goldwyn-Mayer, the series centers on the cat and mouse duo Tom and Jerry.",1940,PG-13,[{"genre": "comedy"}, {"genre": "family"}, {"genre": "science fiction"}, {"genre": "thriller"}, {"genre": "war"], [{"country": "United States"}], 1, 8.0, 156603.0, 12.788, 7.8																						
9	tm50944,The Treasure of the Sierra Madre,MOVIE,"Fred C. Dobbs and Bob Curtin, both down on their luck in Tampico, Mexico in 1935, meet up with a grizzled prospector named Howard and decide to join with him in search of gold in the wilds of central Mexico. Through their search for gold, they find themselves in the middle of a major gold strike.",1940,PG-13,[{"genre": "drama"}, {"genre": "mystery"}, {"genre": "romance"}, {"genre": "war"], [{"country": "United States"}], 1, 8.0, 156603.0, 12.788, 7.8																						
10	tm54459,The Asphalt Jungle,MOVIE,"Recently paroled from prison, legendary burglar \"Doc\" Riedenschneider, with funding from Alonso Emmerich, a crooked lawyer, gathers a small group of veteran criminals together in the Midwest for a big jewel heist.",1950,PG-13,[{"genre": "drama"}, {"genre": "mystery"}, {"genre": "thriller"}, {"genre": "war"], [{"country": "United States"}], 1, 8.0, 156603.0, 12.788, 7.8																						
11	tm81960,The Philadelphia Story,MOVIE,"When a rich woman's ex-husband and a tabloid-type reporter turn up just before her planned remarriage, she begins to learn the truth about herself. Remade as a musical in 1956 as High Society.",1940,PG-102,[{"genre": "drama"}, {"genre": "romance"}, {"genre": "comedy"}, {"genre": "war"], [{"country": "United States"}], 1, 8.0, 156603.0, 12.788, 7.8																						
12	tm58806,Rashomon,MOVIE,"Brimming with action while incisively examining the nature of truth, \"Rashomon\" is perhaps the finest film ever to investigate the philosophy of justice. Through an ingenious use of camera and flashbacks, Kurosawa reveals the complete story of the crime and its aftermath.",1950,PG-102,[{"genre": "drama"}, {"genre": "mystery"}, {"genre": "thriller"}, {"genre": "war"], [{"country": "Japan"}], 1, 8.0, 156603.0, 12.788, 7.8																						
13	tm47077,The Red Shoes,MOVIE,"In this classic drama, Vicki Page is an aspiring ballerina torn between her dedication to dance and her desire to love. While her imperious instructor, Boris Lermontov, urges her to forget anything but ballet, Vicki begins to fall for the man she loves.",1948,PG-102,[{"genre": "drama"}, {"genre": "romance"}, {"genre": "comedy"}, {"genre": "war"], [{"country": "United States"}], 1, 8.0, 156603.0, 12.788, 7.8																						
14	tm160494,Stagecoach,MOVIE,"A group of people travelling on a stagecoach find their journey complicated by the threat of Geronimo, and learn something about each other in the process.",1939,PG-102,[{"genre": "drama"}, {"genre": "western"}, {"genre": "comedy"}, {"genre": "war"], [{"country": "United States"}], 1, 8.0, 156603.0, 12.788, 7.8																						
15	tm47344,M,MOVIE,"In this classic German thriller, Hans Beckert, a serial killer who preys on children, becomes the focus of a massive Berlin police manhunt. Beckert's heinous crimes are so repellent and disruptive to city life that he is even targeted by others in the city.",1928,PG-102,[{"genre": "drama"}, {"genre": "mystery"}, {"genre": "thriller"}, {"genre": "war"], [{"country": "Germany"}], 1, 8.0, 156603.0, 12.788, 7.8																						
16	tm3248,Freaks,MOVIE,"A circus' beautiful trapeze artist agrees to marry the leader of side-show performers, but his deformed friends discover she is only marrying him for his inheritance.",1932,PG-102,[{"genre": "drama"}, {"genre": "horror"}, {"genre": "comedy"}, {"genre": "war"], [{"country": "United States"}], 1, 8.0, 156603.0, 12.788, 7.8																						
17	tm72527,The Great Dictator,MOVIE,"Dictator Adenoid Hynkel tries to expand his empire while a poor Jewish barber tries to avoid persecution from Hynkel's regime.",1940,PG-102,[{"genre": "comedy"}, {"genre": "drama"}, {"genre": "mystery"}, {"genre": "war"], [{"country": "United States"}], 1, 8.0, 156603.0, 12.788, 7.8																						
18	tm2346,The 39 Steps,MOVIE,"Richard Hannay has a rude awakening when a glamorous female spy falls into his bed -- with a knife in her back. Having a bit of trouble explaining it all to Scotland Yard, he heads for the hills of Scotland to try to clear his name by locating the spy.",1935,PG-102,[{"genre": "drama"}, {"genre": "mystery"}, {"genre": "thriller"}, {"genre": "war"], [{"country": "United Kingdom"}], 1, 8.0, 156603.0, 12.788, 7.8																						
19	tm62322,The Lady Vanishes,MOVIE,"On a train headed for England a group of travelers is delayed by an avalanche. Huddled up in a hotel in a fictional European country, young Iris befriends elderly Miss Froy. When the train resumes, Iris suffers a bout of unconsciousness.",1938,PG-102,[{"genre": "drama"}, {"genre": "mystery"}, {"genre": "comedy"}, {"genre": "war"], [{"country": "United Kingdom"}], 1, 8.0, 156603.0, 12.788, 7.8																						
20	tm5668,Bicycle Thieves,MOVIE,"A working man's livelihood is threatened when someone steals his bicycle.",1948,PG-102,[{"genre": "drama"}, {"genre": "European"}, {"genre": "war"}, {"genre": "comedy"}, {"genre": "thriller"}, {"genre": "war"], [{"country": "Italy"}], 1, 8.0, 156603.0, 12.788, 7.8																						
21	tm32933,The Passion of Joan of Arc,MOVIE,"A classic of the silent age, this film tells the story of the doomed but ultimately canonized 15th-century teenage warrior. On trial for claiming she'd spoken to God, Jeanne d'Arc is subjected to inhumane treatment and scarecrows.",1928,PG-102,[{"genre": "drama"}, {"genre": "mystery"}, {"genre": "thriller"}, {"genre": "war"], [{"country": "France"}], 1, 8.0, 156603.0, 12.788, 7.8																						
22	tm52402,City Lights,MOVIE,"In this sound-era silent film, a tramp falls in love with a beautiful blind flower seller.",1931,PG-102,[{"genre": "drama"}, {"genre": "comedy"}, {"genre": "romance"}, {"genre": "war"], [{"country": "United States"}], 1, 8.0, 156603.0, 12.788, 7.8																						
23	tm87737,Häxan,MOVIE,"Grave robbing, torture, possessed nuns, and a satanic Sabbath: Benjamin Christensen's legendary film uses a series of dramatic vignettes to explore the scientific hypothesis that the witches of the Middle Ages suffered the same hysteria as women in modern times.",1922,PG-102,[{"genre": "drama"}, {"genre": "mystery"}, {"genre": "thriller"}, {"genre": "war"], [{"country": "Denmark"}], 1, 8.0, 156603.0, 12.788, 7.8																						
24	tm137796,King Kong,MOVIE,"Adventurous filmmaker, Carl Denham, sets out to produce a motion picture unlike anything the world has seen before. Alongside his leading lady Ann Darrow and his first mate Jack Driscoll, they arrive on an island and discover a legend of a giant ape.",1933,PG-102,[{"genre": "drama"}, {"genre": "mystery"}, {"genre": "thriller"}, {"genre": "war"], [{"country": "United States"}], 1, 8.0, 156603.0, 12.788, 7.8																						
25	tm88256,Black Narcissus,MOVIE,"A group of Anglican nuns, led by Sister Clodagh, are sent to a mountain in the Himalayas. The climate in the region is hostile and the nuns are housed in an old odd palace. They work to establish a school and a hospital, but slowly the nuns begin to go mad.",1947,PG-102,[{"genre": "drama"}, {"genre": "mystery"}, {"genre": "thriller"}, {"genre": "war"], [{"country": "United Kingdom"}], 1, 8.0, 156603.0, 12.788, 7.8																						
26	tm75346,To Be or Not to Be,MOVIE,"During the Nazi occupation of Poland, an acting troupe becomes embroiled in a Polish soldier's efforts to track down a German spy.",1942,PG-102,[{"genre": "drama"}, {"genre": "mystery"}, {"genre": "thriller"}, {"genre": "war"], [{"country": "Poland"}], 1, 8.0, 156603.0, 12.788, 7.8																						
27	tm88077,Great Expectations,MOVIE,"In this Dickens adaptation, orphan Pip discovers through lawyer Mr. Jaggers that a mysterious benefactor wishes to ensure that he becomes a gentleman. Reunited with his childhood patron, Miss Havisham, and his first love, Estella, Pip begins to realize his own potential.",1946,PG-102,[{"genre": "drama"}, {"genre": "mystery"}, {"genre": "thriller"}, {"genre": "war"], [{"country": "United Kingdom"}], 1, 8.0, 156603.0, 12.788, 7.8																						

Slika 1. Početni skup podataka (samostalna izrada)

## 4. ETL postupak

ETL postupak (eng. *Extract, transform, load*) podrazumijeva stvaranje skladišta podataka iz više izvora podataka, skladište se stvara poštujući poslovna pravila za čišćenje i organiziranje sirovih podataka te nakon toga pohranu i analizu. Ovaj postupak premješta podatke u intervalima iz izvora u odredište, kao što se može zaključiti i iz imena radi u tri koraka, a to je: ekstrahiranje podataka iz izvorene baze podataka, transformiranje podataka tako da ih je jednostavnije analizirati te učitavanje u odredišnu bazu (aws, b.d).

Ekstrakcija je kopiranje i izdvajanje podataka iz više izvora u pripremno područje. Pripremno područje (eng. *staging area*) je privremeno mjesto podataka i sadržaj tog područja se briše nakon ekstrakcije, no postoji mogućnost arhiviranja. Podaci su neobrađeni (aws, b.d).

Transformacija podataka podrazumijeva uklanjanje pogreški, praznih redova ili nadopunjavanje nedostajućih vrijednosti. Ukratko transformacija povećava kvalitetu podataka te ih transformira u željeni format (aws, b.d).

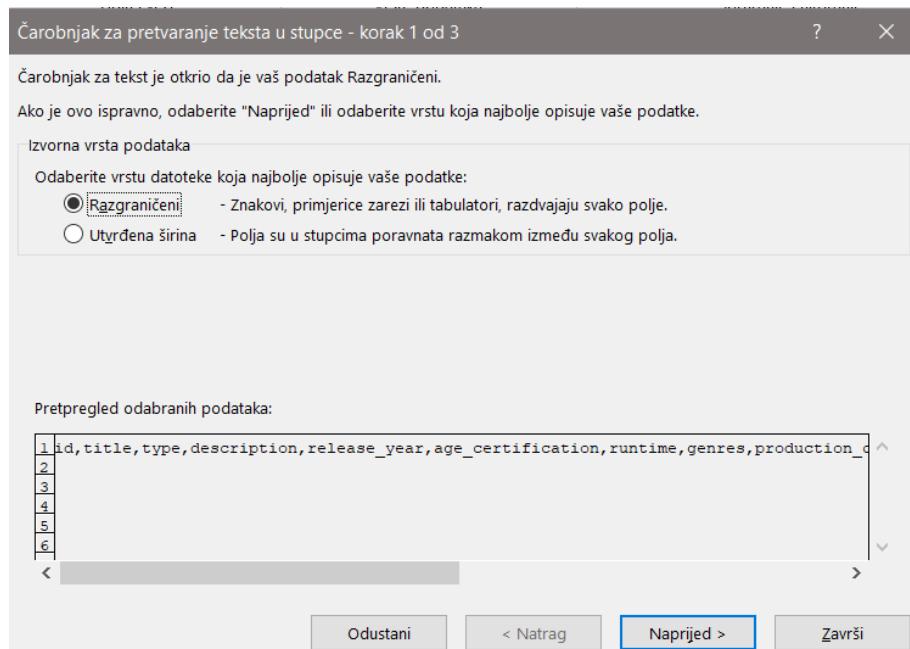
Učitavanje znači premještanje podataka u odredišnu bazu. Postoje dva načina učitavanja, a to su puno opterećenje i inkrementalno opterećenje. Kod punog opterećenja cijelokupni podaci se transformiraju i učitavaju u ciljnu bazu podatka, ova vrsta učitavanja koristi se kada se podaci učitavaju prvi put. Inkrementalno opterećenje znači da se podaci učitavaju u redovitim intervalima odnosno učitava se razlika između izvornog i ciljnog sustava. Prilikom svakog učitavanja pohranjuje se datum posljednjeg izdvajanja tako da se učitaju samo oni zapisi dodani nakon tog datuma (aws, b.d).



Slika 2. ETL postupak (Integrate.Io, 2023)

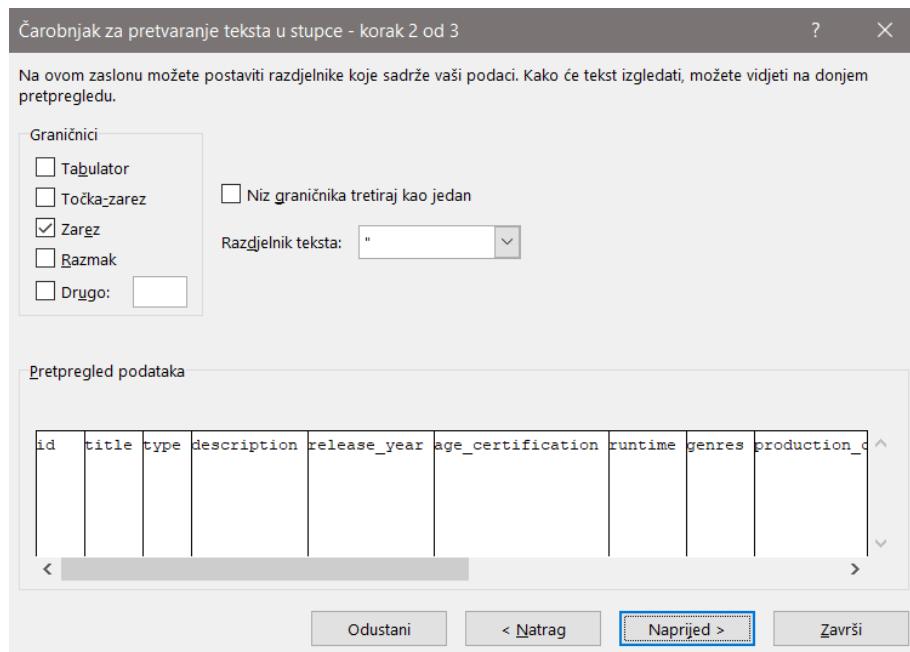
Kao što je već spomenuto proces ekstrakcije podataka je napravljen na način da su podaci preuzeti sa izvorišnog mesta odnosno s interneta te su uvezeni u pripremno područje to jest Excel. U nastavku su prikazane sve transformacije podataka.

Iz slike 1 vidimo kako su podaci loše strukturirani te je potrebno provesti neke transformacije. Prva transformacija je u Excelu. Potrebno je razdvojiti stupce, tako da svaki sadržava po jedan atribut. To se napravi tako da se u Excelu odabere kartica „Podaci“ te „Tekst u stupce“ potom se otvara dijaloški okvir kako je prikazano na donjoj slici.



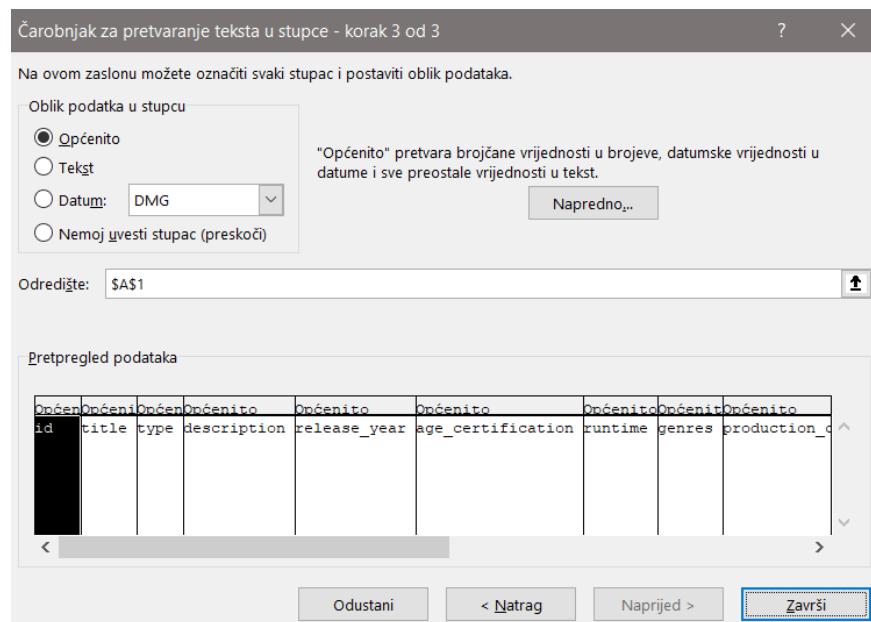
Slika 3. Pretvaranje teksta u stupce 1. korak (samostalna izrada)

Excel sam prepoznaće što u podacima dijeli podatke. U ovim podacima je prepoznao da se radi o razgraničenoj datoteci. Nakon odabira vrste datoteke slijedi odabir razdjelnika, kako donja slika prikazuje u ovom slučaju su to zarezi.



Slika 4. Pretvaranje teksta u stupce 2. korak (samostalna izrada)

Slika 5 prikazuje još neke dodatne mogućnosti, kao što je odabir oblika podataka. Podaci su ostavljeni u općenitom obliku.



Slika 5. Pretvaranje teksta u stupce 3. korak (samostalna izrada)

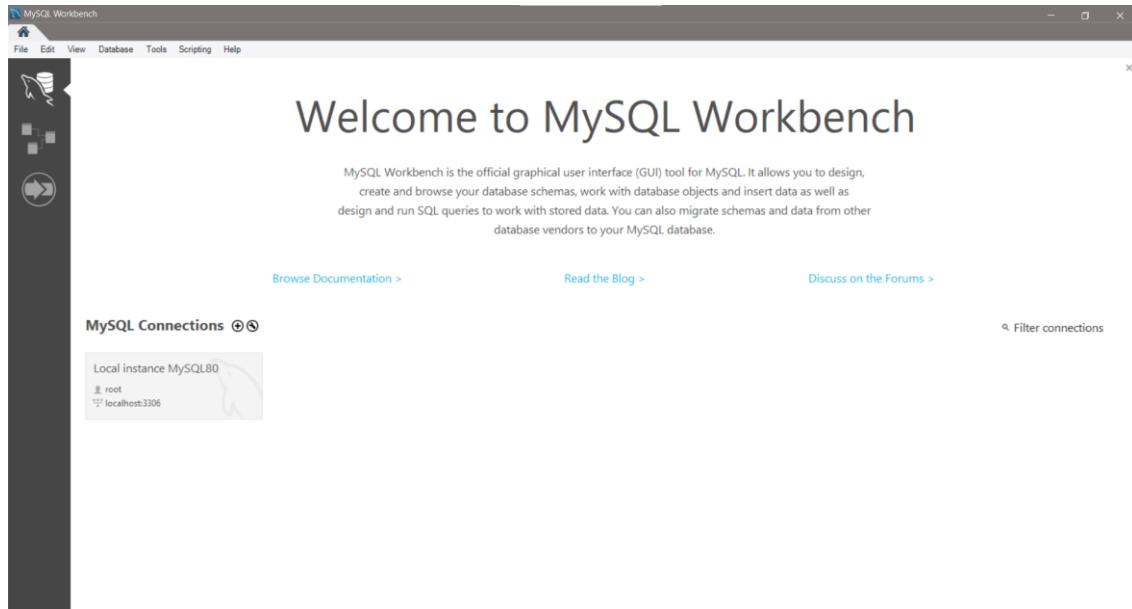
Nakon provedene transformacije podataka u Excelu csv datoteka izgleda kao na slici

6.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	id	title	type	descriptor	release_ye	age_certifi	runtime	genres	production	seasons	imdb_id	imdb_scor	imdb_vote:tmdb.popi	tmdb_score	
2	tm155702	The Wizard	MOVIE	Young Dor	1939	G	102	['fantasy', 'US']			tt0032138	8.sij	389774.0	41.442	7.lip
3	tm83640	Citizen Kar	MOVIE	Newspape	1941	PG	119	['drama']	['US']		tt0033467	8.ožu	433804.0	14.383	8.0
4	tm77584	Casablanca	MOVIE	In Casabla	1942	PG	102	['drama', 'rom']	['US']		tt0034583	8.svi	558849.0	20.087	8.vi
5	tm82363	The Big Sie	MOVIE	Private Inv	1946		116	['thriller', 'c']	['US']		tt0038355	7.ruj	84494.0	12.911	7.srp
6	tm84701	The Maltes	MOVIE	A private d	1941		100	['thriller', 'r']	['US']		tt0033670	8.0	156603.0	12.788	7.kol
7	tm156463	Gone with	MOVIE	The spoiler	1939	G	233	['war', 'rom']	['US']		tt0031381	8.vlj	309856.0	24.092	8.0
8	ts225761	Tom and Je	SHOW	Tom and Je	1940		8	['animation']	['US']	16.0	tt1215899	7.srp	853.0	14.202	10.0
9	tm5094	The Treast	MOVIE	Fred C. Bo	1948		126	['western']	['US']		tt0040897	8.vlj	122971.0	14.006	8.0
10	tm54459	The Asphai	MOVIE	Recently pi	1950		112	['drama', 't']	['US']		tt0042206	7.kol	26557.0	9.809	7.svi
11	tm811960	The Philad	MOVIE	When a ric	1940		113	['romance']	['US']		tt0032904	7.ruj	68337.0	11.587	7.srp
12	tm58806	Rashomon	MOVIE	Brimming v	1950		88	['drama', 'c']	['JP']		tt0042676	8.vlj	165278.0	12.359	8.vlj
13	tm4707	The Red St	MOVIE	In this clas	1948		133	['romance']	['GB']		tt0040725	8.sij	34367.0	9.378	8.sij
14	tm160494	Stagecoach	MOVIE	A group of	1939		96	['western']	['US']		tt0031971	7.kol	48149.0	11.786	7.srp
15	tm47834	M	MOVIE	In this clas	1931	PG-13	117	['thriller', 'e']	['DE']		tt0022100	8.ožu	155068.0	9.165	8.sij
16	tm3248	Freaks	MOVIE	A circus b	1932		62	['drama', 'r']	['US']		tt0022913	7.ruj	45726.0	8.294	7.kol
17	tm72527	The Great I	MOVIE	Dictator Ac	1940		125	['comedy']	['US']		tt0032553	8.tra	219871.0	15.335	8.tra
18	tm2346	The 39 Ste	MOVIE	Richard He	1935		86	['thriller', 'c']	['GB']		tt0026029	7.lip	56259.0	10.794	7.tra
19	tm8232	The Lady V	MOVIE	On a train l	1938		96	['thriller']	['GB']		tt0030341	7.kol	51847.0	23.621	7.vi
20	tm5608	Bicycle Thi	MOVIE	A working r	1948		88	['drama', 'e']	['IT']		tt0040522	8.ožu	160558.0	12.024	8.ožu
21	tm3293	The Passio	MOVIE	A classic o	1928		110	['drama', 'e']	['FR']		tt0019254	8.vlj	53460.0	11.213	8.sij
22	tm52402	City Lights	MOVIE	In this sour	1931	G	87	['romance']	['US']		tt0021749	8.svi	181367.0	13.647	8.tra
23	tm87737	H&Raxx	MOVIE	Grave robb	1922		105	['horror', 'e']	['SE']		tt0013257	7.lip	41108.0	6.312	7.lip
24	tm137798	King Kong	MOVIE	Adventuroi	1933		104	['fantasy']	['US']		tt0024216	7.ruj	84414.0	17.542	7.lip
25	tm86256	Black Narc	MOVIE	A group of.	1947		101	['drama']	['GB']		tt0039192	7.kol	24474.0	6.685	7.srp
26	tm75346	To Be or N	MOVIE	During the	1942	G	99	['comedy']	['US']		tt0035446	8.vlj	35911.0	10.077	7.ruj
27	tm88077	Great Expe	MOVIE	In this Dick	1946		118	['romance']	['GB']		tt0038574	7.kol	24201.0	10.109	7.vlj

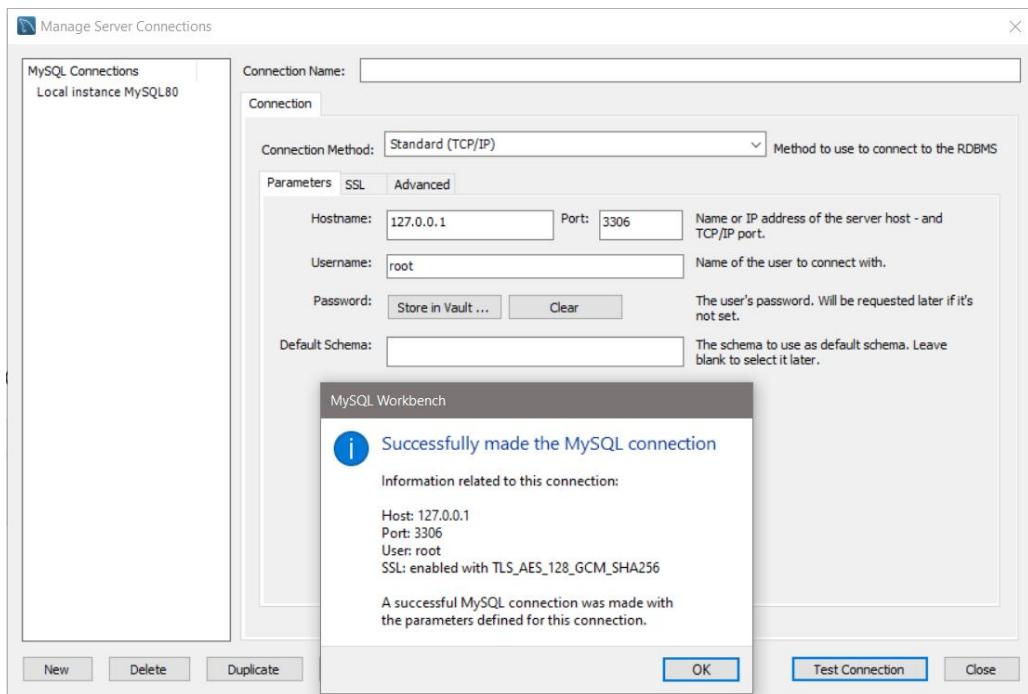
Slika 6. Finalni izgled podataka u Excelu (samostalna izrada)

Nakon transformacija u Excelu slijedi transformacija u alatu MySQL Workbench. Kako bi se kreiralo skladište podataka potrebno je spojiti MySQL Workbench na lokalni poslužitelj te uvesti podatke i napraviti potrebne transformacije. Slika 7 pokazuje početnu stranicu alata te dostupnu konekciju na bazu. Vidimo da je ime korisnika *root*, *hostname* je *localhost*, a 3306 je *port*.



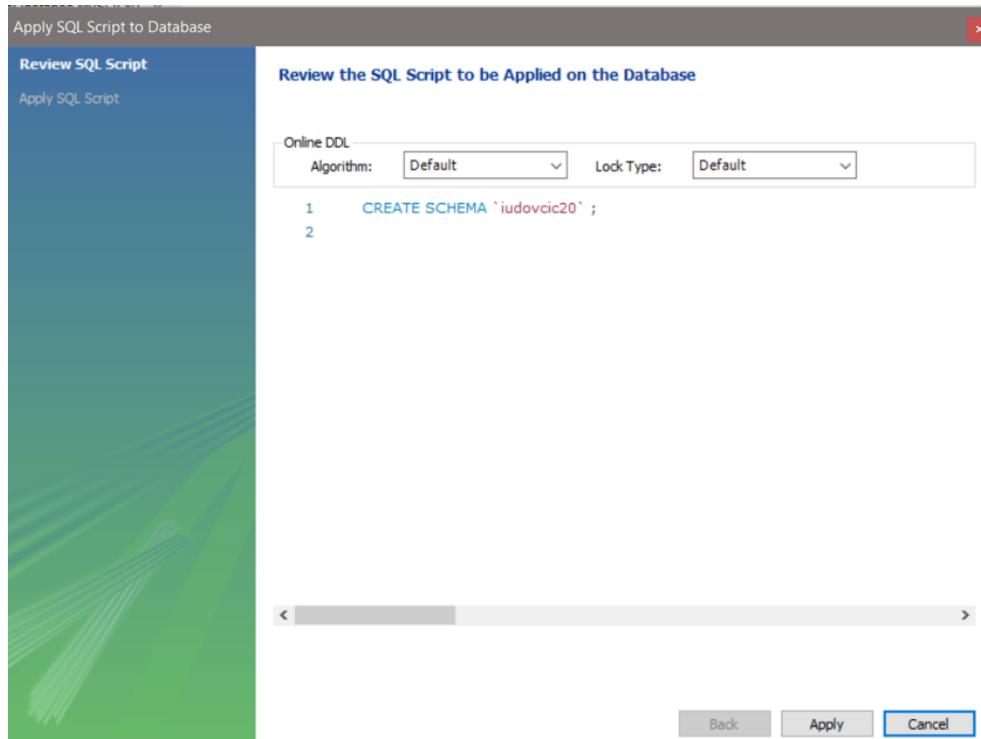
Slika 7. Početni zaslon MySQL Workbench (samostalna izrada)

Ako kliknemo na *Database* na alatnoj traci te potom na *Manage Connections* dobit ćemo dodatne informacije o vezi te ukoliko kliknemo na *Test Connection* dobijemo informacije prikazane kao u manjem dijaloškom okviru na slici 8.



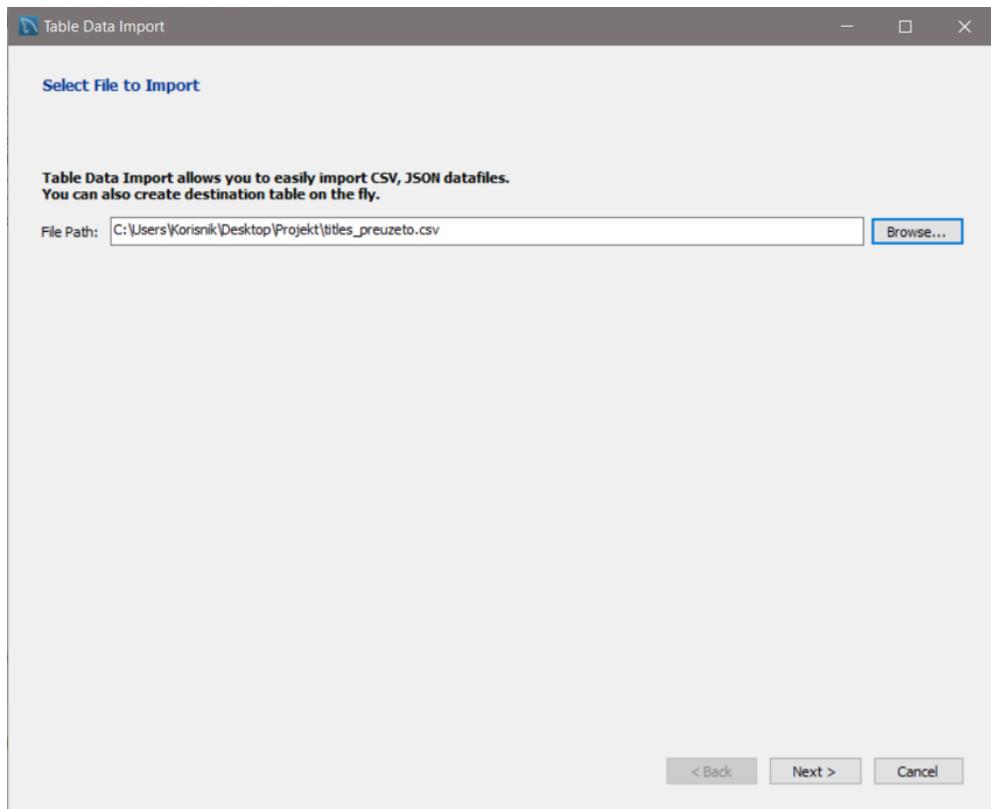
Slika 8. Povezivanje na bazu (samostalna izrada)

Nakon uspješnog povezivanja potrebno je kreirati *schema* odnosno skladište. U ovom slučaju naziv *schema* je iudovcic20.



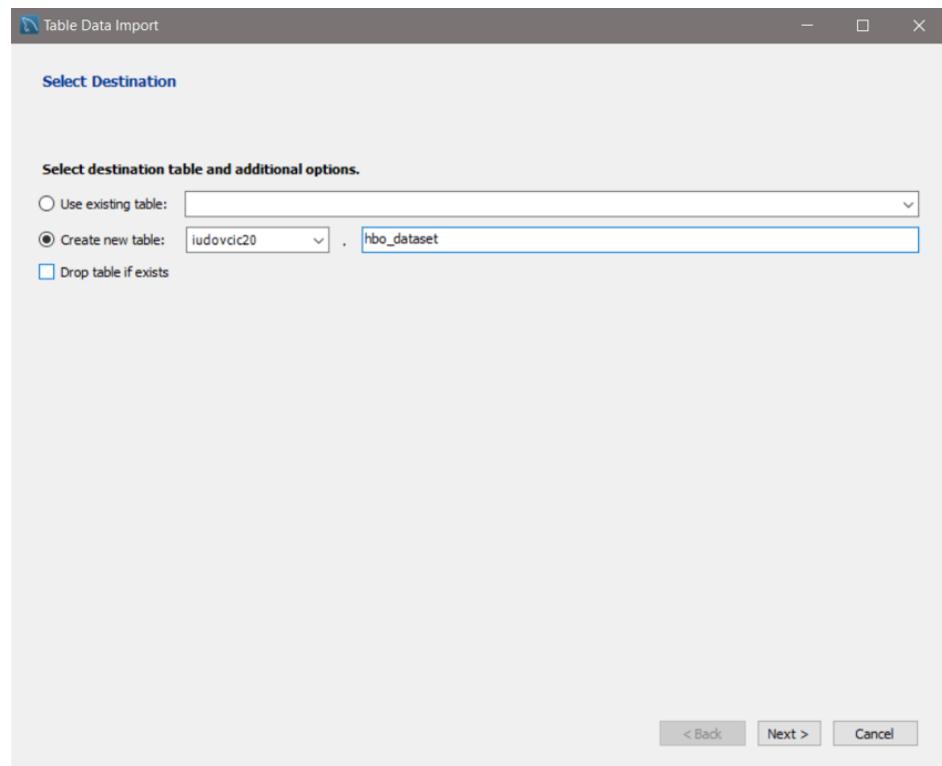
Slika 9. Kreiranje scheme (samostalna izrada)

U kreirano skladište uvozimo podatke tako da kliknemo desni klik na *schema* i odaberemo *Table Data Import*. Potom se otvara prozor kao na slici 10. Na slici 10 može se vidjeti da su podaci u datoteci koja se zove *titles\_preuzeto.csv*.



Slika 10. Uvoz podataka 1. korak (samostalna izrada)

Sljedeći korak uvoza podataka je odabir imena tablice i skladišta gdje će biti smještena ta tablica. Kako je vidljivo na slici 11, tablica se zove *hbo\_dataset*, a skladište *iudovcic20*.



Slika 11. Odabir skladišta i imena tablice (samostalna izrada)

Kada smo odabrali željeni skup podataka potrebno je odabrati stupce koji će tvoriti skladište. Na slici 12 prikazan je odabir stupaca.

The screenshot shows the 'Table Data Import' application window titled 'Configure Import Settings'. It indicates the file format is 'csv'. Under 'Columns', several columns are mapped to field types: id (text), title (text), type (text), description (text), release\_year (int), and age\_certification (text). Below this, a preview table shows movie data from the 'hbo\_dataset' table, including columns like id, title, type, description, release\_year, age\_certification, runtime, genres, production..., and seasons. The preview table has scroll bars and navigation arrows at the bottom.

Slika 12. Odabir stupaca (samostalna izrada)

Nakon uspješnog uvoza podataka možemo vidjeti tipove podatka koji se nalaze u stupcima kreiranog skladišta, ali vidimo i druge stavke.

The screenshot shows the MySQL Workbench interface with the 'hbo\_dataset' table selected. The table has 14 columns: id, title, type, description, release\_year, age\_certification, runtime, genres, age\_certification (repeated), runtime (repeated), genres (repeated), and type (repeated). The 'age\_certification', 'runtime', and 'genres' columns are repeated three times in the table structure. The 'age\_certification' column is defined as 'text' with a default value of 'YES', character set 'utf8mb4', and collation 'utf8mb4\_0900'. The 'runtime' column is defined as 'int' with a default value of 'YES', character set 'utf8mb4', and collation 'utf8mb4\_0900'. The 'genres' column is defined as 'text' with a default value of 'YES', character set 'utf8mb4', and collation 'utf8mb4\_0900'. The 'type' column is defined as 'text' with a default value of 'YES', character set 'utf8mb4', and collation 'utf8mb4\_0900'. The 'Extra' and 'Comments' columns are both empty.

Slika 13. Tablica hbo\_dataset (samostalna izrada)

Na slici 14 vidimo podatke koji su uvezeni.

The screenshot shows the MySQL Workbench interface with the 'hbo\_dataset' table populated with data. The table contains 20 rows of movie information, each with an 'id' and a corresponding title, type, description, release year, age certification, runtime, and genres. The data includes classics like 'The Wizard of Oz', 'Citizen Kane', and 'Casablanca', as well as more recent entries like 'The Big Sleep' and 'The Maltese Falcon'. The 'genres' column lists various categories such as 'fantasy', 'family', 'drama', 'romance', 'war', 'thriller', 'crime', 'western', 'comedy', 'horror', 'romantic', 'drama', 'music', and 'documentary'.

Slika 14. Izgled uvezenih podataka (samostalna izrada)

Kada su podaci učitani potrebno ih je transformirati, u nastavku su navedene sve provedene transformacije u alatu MySQL Workbench.

Prva provedena promjena popunjava prazne retke atributa *age\_certification* i postavlja ih na „*none*“. U nastavku su prikazani naredba za promjenu te izgled tablice nakon promjene.

```
1 UPDATE iudovcic20.hbo_dataset SET age_certification="none" WHERE hbo_dataset.age_certification="" ;
```

Slika 15. Update nad age\_certification (samostalna izrada)

	type	description	release_year	age_certification	runtime	genres	production_countries	seasons	imdb_id
▶	MOVIE	Young Dorothy finds herself in a magical world ...	1939	G	102	['fantasy', 'family']	['US']		tt0032138
	MOVIE	Newspaper magnate, Charles Foster Kane is ta...	1941	PG	119	['drama']	['US']		tt0033467
	MOVIE	In Casablanca, Morocco in December 1941, a c...	1942	PG	102	['drama', 'romance', 'war']	['US']		tt003458
	MOVIE	Private Investigator Philip Marlowe is hired by w...	1946	none	116	['thriller', 'crime']	['US']		tt0038351
	MOVIE	A private detective takes on a case that involve...	1941	none	100	['thriller', 'romance', 'crime']	['US']		tt0033870
	MOVIE	The spoiled daughter of a well-to-do plantation ...	1939	G	233	['war', 'romance', 'drama', 'history']	['US']		tt003138
	SHOW	Tom and Jerry is an American animated franchis...	1940	none	8	['animation', 'comedy', 'family', 'action']	['US']	16.0	tt1215895
▶	MOVIE	Fred C. Dobbs and Bob Curtin, both down on th...	1948	none	126	['western', 'drama']	['US']		tt004089
	MOVIE	Recently paroled from prison, legendary burglar...	1950	none	112	['drama', 'thriller', 'crime']	['US']		tt0042208
	MOVIE	When a rich woman's ex-husband and a tabloid...	1940	none	113	['romance', 'comedy']	['US']		tt0032904
	MOVIE	Brimming with action while incisively examining t...	1950	none	88	['drama', 'crime']	['JP']		tt0042876
	MOVIE	In this classic drama, Vicki Page is an aspiring b...	1948	none	133	['romance', 'drama', 'music']	['GB']		tt0040721
	MOVIE	A group of people traveling on a stagecoach fin...	1939	none	96	['western', 'drama']	['US']		tt003197
	MOVIE	In this classic German thriller, Hans Beckert, a s...	1931	PG-13	117	['thriller', 'european', 'crime']	['DE']		tt0022100
	MOVIE	A circus' beautiful trapeze artist agrees to marr...	1932	none	62	['drama', 'horror']	['US']		tt0022911
	MOVIE	Dictator Adenoid Hynkel tries to expand his emp...	1940	none	125	['comedy', 'war', 'drama']	['US']		tt003255
	MOVIE	Richard Hannay has a rude awakening when a ...	1935	none	86	['thriller', 'crime']	['GB']		tt0026025
	MOVIE	On a train headed for England a group of travel...	1938	none	96	['thriller']	['GB']		tt003034
	MOVIE	A working man's livelihood is threatened when s...	1948	none	88	['drama', 'european']	['IT']		tt0040522
	MOVIE	A classic of the silent era, this film tells the stor...	1928	none	110	['drama', 'european', 'historical']	['FR']		tt0019754

Slika 16. Izgled tablice nakon update upita nad age\_certification (samostalna izrada)

Druga provedena promjena popunjava prazne retke atributa seasons tako da ih postavlja na 0. U nastavku je prikazana naredba i izgled tablice nakon provedene naredbe.

```
1 UPDATE iudovcic20.hbo_dataset SET seasons="0" WHERE hbo_dataset.seasons="" ;
```

Slika 17. Update nad seasons (samostalna izrada)

	release_year	age_certification	runtime	genres	production_countries	seasons	imdb_id	imdb_score	imdb_votes	tmdb_popularity	tmdb_score
▶	1939	G	102	['fantasy', 'family']	['US']	0	tt0032138	8.sij	389774	41.442	7.lip
	341	PG	119	['drama']	['US']	0	tt0033467	8.ožu	433804	14.383	8.0
	342	PG	102	['drama', 'romance', 'war']	['US']	0	tt0034583	8.svi	558849	20.087	8.vlj
	346	none	116	['thriller', 'crime']	['US']	0	tt0038355	7.ruj	84494	12.911	7.srp
	341	none	100	['thriller', 'romance', 'crime']	['US']	0	tt0033870	8.0	156603	12.788	7.kol
	339	G	233	['war', 'romance', 'drama', 'history']	['US']	0	tt0031381	8.vlj	309856	24.092	8.0
	340	none	8	['animation', 'comedy', 'family', 'action']	['US']	16.0	tt12158944	7.srp	853	14.202	10.0
	348	none	126	['western', 'drama']	['US']	0	tt0040897	8.vlj	122971	14.006	8.0
	350	none	112	['drama', 'thriller', 'crime']	['US']	0	tt0042208	7.kol	26557	9.809	7.svi
	340	none	113	['romance', 'comedy']	['US']	0	tt0032904	7.ruj	68337	11.587	7.srp
	350	none	88	['drama', 'crime']	['JP']	0	tt0042876	8.vlj	165278	12.359	8.vlj
	348	none	133	['romance', 'drama', 'music']	['GB']	0	tt0040725	8.sij	34367	9.378	8.sij
	339	none	96	['western', 'drama']	['US']	0	tt0031971	7.kol	48149	11.786	7.srp
	331	PG-13	117	['thriller', 'european', 'crime']	['DE']	0	tt0022100	8.ožu	155068	9.165	8.sij
	332	none	62	['drama', 'horror']	['US']	0	tt0022913	7.ruj	45726	8.294	7.kol
	340	none	125	['comedy', 'war', 'drama']	['US']	0	tt0032553	8.tra	219871	15.335	8.tra
	335	none	86	['thriller', 'crime']	['GB']	0	tt0026029	7.lip	56259	10.794	7.tra
	338	none	96	['thriller']	['GB']	0	tt0030341	7.kol	51847	23.621	7.svi
	348	none	88	['drama', 'european']	['IT']	0	tt0040522	8.ožu	160558	12.024	8.ožu
	328	none	110	['drama', 'european', 'historical']	['FR']	0	tt0019754	8.vlj	53460	11.213	8.oži

Slika 18. Izgled tablice nakon update nad seasons (samostalna izrada)

Potrebno je još pretvoriti stupce imdb\_score i tmdb\_score u decimalni broj, ali prije toga je potrebno mjesecu zamijeniti brojem budući da se prilikom transformacije podataka u Excelu dogodila promjena koja broj nakon točke pretvara u mjesec. Kako se može vidjeti na slici 13 ta dva stupca su tipa text.

Zbog duljine naredbe u nastavku je prikazana naredba samo za tmdb\_score atribut, a ista takva naredba je korištena i za transformaciju stupca imdb\_score samo je na

odgovarajućim mjestima stavljeni *imdb\_score* umjesto *tmdb\_score*. Također, u nastavku je i izgled tablice te rezultati *Table Inspector*a nakon provedenih svih transformacija.

```

1 UPDATE iudovcic20.hbo_dataset
2 SET tmdb_score = CAST(
3     CASE
4         WHEN tmdb_score LIKE '%.sij%' THEN REPLACE(tmdb_score, 'sij', '1')
5         WHEN tmdb_score LIKE '%.vlij%' THEN REPLACE(tmdb_score, 'vlij', '2')
6         WHEN tmdb_score LIKE '%.ožu%' THEN REPLACE(tmdb_score, 'ožu', '3')
7         WHEN tmdb_score LIKE '%.tra%' THEN REPLACE(tmdb_score, 'tra', '4')
8         WHEN tmdb_score LIKE '%.svi%' THEN REPLACE(tmdb_score, 'svi', '5')
9         WHEN tmdb_score LIKE '%.lip%' THEN REPLACE(tmdb_score, 'lip', '6')
10        WHEN tmdb_score LIKE '%.srp%' THEN REPLACE(tmdb_score, 'srp', '7')
11        WHEN tmdb_score LIKE '%.kol%' THEN REPLACE(tmdb_score, 'kol', '8')
12        WHEN tmdb_score LIKE '%.ruj%' THEN REPLACE(tmdb_score, 'ruj', '9')
13        WHEN tmdb_score LIKE '%.lis%' THEN REPLACE(tmdb_score, 'lis', '10')
14        WHEN tmdb_score LIKE '%.stu%' THEN REPLACE(tmdb_score, 'stu', '11')
15        WHEN tmdb_score LIKE '%.pro%' THEN REPLACE(tmdb_score, 'pro', '12')
16        ELSE tmdb_score
17    END AS DECIMAL (5,1)
18 )
19 ;

```

Slika 19. Update tmdb\_score i imdb\_score (samostalna izrada)

release_year	age_certification	runtime	genres	production_countries	seasons	imdb_id	imdb_score	imdb_votes	tmdb_popularity	tmdb_score
939	G	102	[fantasy, 'family']	[US]	0	tt0032138	8.1	389774	41.442	7.6
941	PG	119	[drama]	[US]	0	tt003467	8.3	433804	14.383	8.0
942	PG	102	[drama, 'romance', 'war']	[US]	0	tt0034583	8.5	558849	20.087	8.2
946	none	116	[thriller, 'crime']	[US]	0	tt0038355	7.9	84494	12.911	7.7
941	none	100	[thriller, 'romance', 'crime']	[US]	0	tt0033870	8.0	156603	12.788	7.8
939	G	233	[war, 'romance', 'drama', 'history']	[US]	0	tt0031381	8.2	309856	24.092	8.0
940	none	8	[animation, 'comedy', 'family', 'action']	[US]	16.0	tt12158994	7.7	853	14.202	10.0
948	none	126	[western, 'drama']	[US]	0	tt0040897	8.2	122971	14.006	8.0
950	none	112	[drama, 'thriller', 'crime']	[US]	0	tt0042208	7.8	26557	9.809	7.5
940	none	113	[romance, 'comedy']	[US]	0	tt0032904	7.9	68337	11.587	7.7
950	none	88	[drama, 'crime']	[JP]	0	tt0042876	8.2	165278	12.359	8.2
948	none	133	[romance, 'drama', 'music']	[GB]	0	tt0040725	8.1	34367	9.378	8.1
939	none	96	[western, 'drama']	[US]	0	tt0031971	7.8	48149	11.786	7.7
931	PG-13	117	[thriller, 'european', 'crime']	[DE]	0	tt0022100	8.3	155068	9.165	8.1
932	none	62	[drama, 'horror']	[US]	0	tt0022913	7.9	45726	8.294	7.8
940	none	125	[comedy, 'war', 'drama']	[US]	0	tt0032553	8.4	219871	15.335	8.4
935	none	86	[thriller, 'crime']	[GB]	0	tt0026029	7.6	56259	10.794	7.4
938	none	96	[thriller]	[GB]	0	tt0030341	7.8	51847	23.621	7.5
948	none	88	[drama, 'european']	[IT]	0	tt0040522	8.3	160558	12.024	8.3
928	none	110	[drama, 'mystery', 'history']	[FR]	0	Hn19254	8.2	52460	11.713	8.1

Slika 20. Izgled tablice nakon updatea tmdb\_score i imdb\_score (samostalna izrada)

iudovcic20 - Schema										
hbo_dataset		iudovcic20.hbo_dataset		hbo_dataset		iudovcic20.hbo_dataset		iudovcic20.hbo_dataset		
Info	Columns	Indexes	Triggers	Foreign keys	Partitions	Grants	DDL			
	Column	Type	Default Value	Nullable	Character Set	Collation	Privileges	Extra	Comments	
	age_certification	text		YES	utf8mb4	utf8mb4_0900_...	select,insert,update,references			
	description	text		YES	utf8mb4	utf8mb4_0900_...	select,insert,update,references			
	genres	text		YES	utf8mb4	utf8mb4_0900_...	select,insert,update,references			
	id	text		YES	utf8mb4	utf8mb4_0900_...	select,insert,update,references			
	imdb_id	text		YES	utf8mb4	utf8mb4_0900_...	select,insert,update,references			
	imdb_score	decimal(5,1)		YES			select,insert,update,references			
	imdb_votes	double		YES			select,insert,update,references			
	production_countries	text		YES	utf8mb4	utf8mb4_0900_...	select,insert,update,references			
	release_year	int		YES			select,insert,update,references			
	runtime	int		YES			select,insert,update,references			
	seasons	text		YES	utf8mb4	utf8mb4_0900_...	select,insert,update,references			
	title	text		YES	utf8mb4	utf8mb4_0900_...	select,insert,update,references			
	tmdb_popularity	double		YES			select,insert,update,references			
	tmdb_score	decimal(5,1)		YES			select,insert,update,references			
	type	text		YES	utf8mb4	utf8mb4_0900_...	select,insert,update,references			

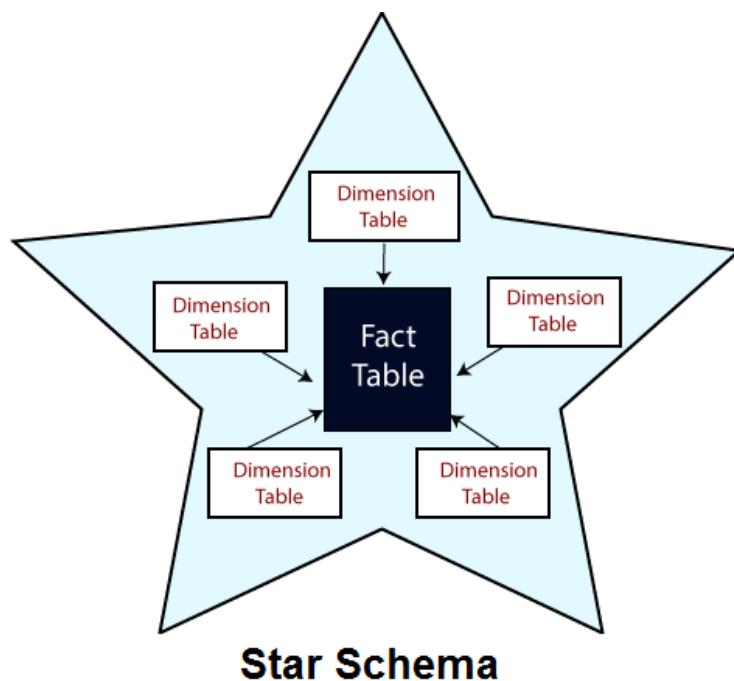
Slika 21. Table Inspector nakon svih updata (samostalna izrada)

Nakon svih provedenih transformacija, može se prijeći na slijedeći korak, a to je učitavanje. U trećem koraku ETL procesa stvara se skladište podataka prema modelu zvijezde.

## 5. Model zvijezde

Nakon izvršenog ETL procesa, slijedi kreiranje modela zvijezde, odnosno kreiranje dimenzijskih i činjeničnih tablica. To je model koji predstavlja podatke na intuitivan i strukturiran način. U činjeničnoj tablici se nalaze mjere, a u dimenzijskim tablicama su opisi atributa. Dimenzijske i činjenične tablice spojene su vanjskim ključevima koji se nalaze u činjeničnoj tablici. Takva struktura omogućava upite koji se kreiraju pomoću atributa iz dimenzijskih tablica. Ovo je jednostavna i lako razumljiva tehnika koja omogućava brzo kreiranje upita. Osim toga, lako se proširi jer se uvijek može dodati još dimenzijskih tablica. Ukoliko model zvijezde ima više dimenzijskih tablica naziva se model stonoge. Model zvijezde je denormalizirana struktura što znači da je redundancija dopuštena (GeeksforGeeks, 2023).

Model zvijezde je relacijska shema čiji dizajn predstavlja multidimenzionalni model. U dimenzijskim tablicama nalaze se opisi atributa koji daju kontekst podacima u činjeničnoj tablici. Svaka dimenzija je jedno gledište za analizu podataka. Dimenzijske tablice su manje te su prikladnije za postavljanje upita od činjeničnih tablica (Gheorghiu, 2023).



Slika 22. Model zvijezde (GeeksforGeeks, 2023)

U nastavku slijedi prikaz modela zvijezde za izabrani skup podataka. Važno je napomenuti da su svi primarni ključevi postavljeni kao *integer*.

## 5.1. Dimenzijska tablica *age\_certification*

Sadrži dva atributa, a to su *idage\_certification* i *age\_certification*. Atribut *idage\_certification* je primarni ključ postavljen na *auto increment* i *not null*, a *age\_certification* je tipa *VARCHAR(45)* i *not null*. U nastavku je prikazano kreiranje tablice *age\_certification*, upit koji je popunio tu tablicu te izgled popunjene tablice.

The screenshot shows the 'Create Table' dialog in MySQL Workbench. The 'Table Name' is set to 'age\_certification', 'Schema' to 'iudovcic20', and 'Engine' to 'InnoDB'. The 'Charset/Collation' dropdowns are set to 'Default Charset' and 'Default Collation'. The 'Comments' field is empty. The 'Columns' section contains two entries: 'idage\_certification' (INT, PK, NOT NULL) and 'age\_certification' (VARCHAR(45), NOT NULL). The 'Default/Expression' column for 'age\_certification' is highlighted in blue.

Slika 23. Kreiranje tablice *age\_certification* 1. dio (samostalna izrada)

The screenshot shows the 'Apply SQL Script to Database' dialog. The left panel is titled 'Review SQL Script' and shows the 'Apply SQL Script' button. The right panel is titled 'Review the SQL Script to be Applied on the Database' and displays the following SQL code:

```
CREATE TABLE `iudovcic20`.`age_certification` (
  `idage_certification` INT NOT NULL AUTO_INCREMENT,
  `age_certification` VARCHAR(45) NOT NULL,
  PRIMARY KEY (`idage_certification`));
```

Below the code are buttons for 'Back', 'Apply', and 'Cancel'.

Slika 24. Kreiranje tablice *age\_certification* 2. dio (samostalna izrada)

```
1  INSERT INTO iudovcic20.age_certification (age_certification)
2  SELECT type FROM iudovcic20.hbo_dataset;
```

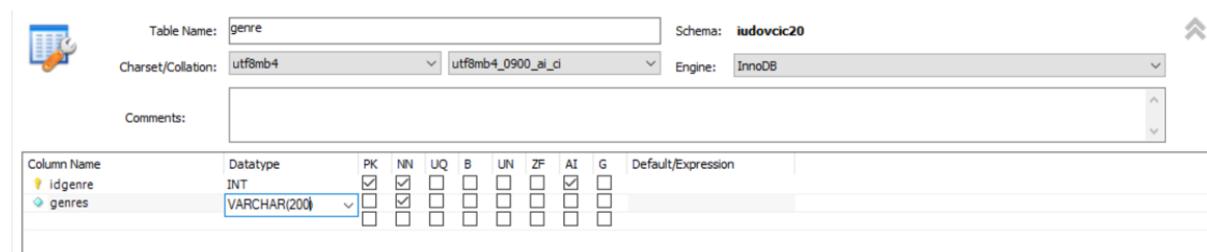
Slika 25. Naredba za popunjavanje *age\_certification* tablice (samostalna izrada)

	<i>idage_certification</i>	<i>age_certification</i>
1		G
2		PG
3		PG
4		none
5		none
6		G

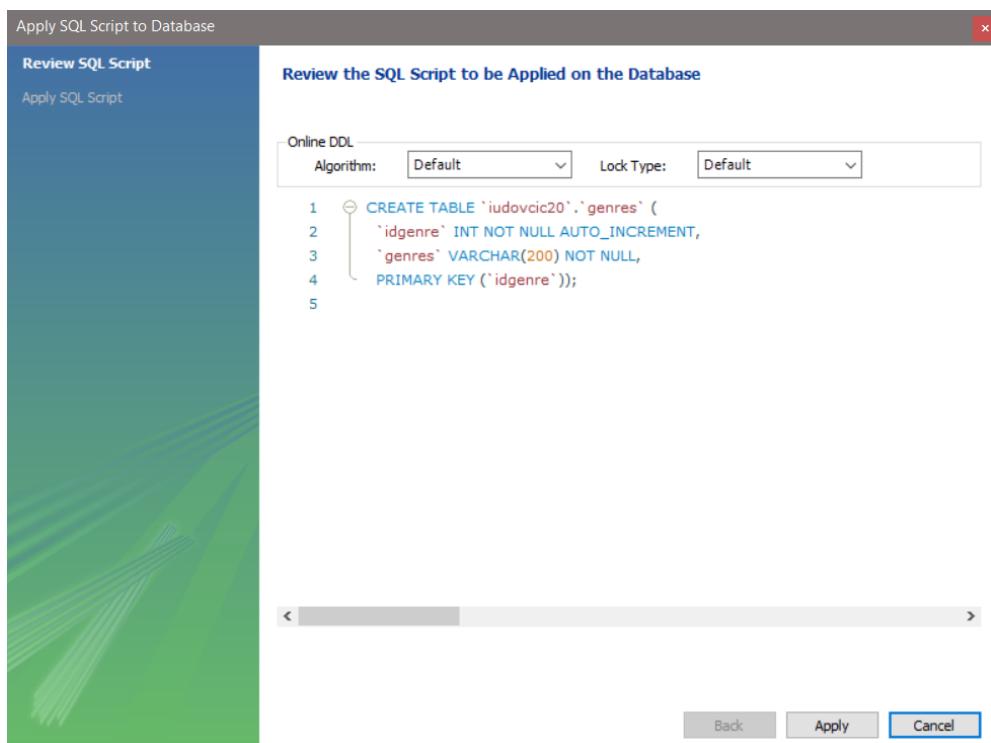
Slika 26. Izgled tablice *age\_certification* (samostalna izrada)

## 5.2. Dimenzijska tablica genres

U tablici se nalaze dva atributa, a to su *idgenre* koji je postavljen kao primarni ključ, *auto increment* i *not null*, a drugi atribut je *genre* koji je postavljen na *not null*, a tip podatka je *VARCHAR(200)*. Slika 27 prikazuje kreiranje tablice, slika 28 prikazuje naredbu koju generira MySQL Workbench za kreiranje tablice, slika 29 prikazuje naredbu za popunjavanje novokreirane tablice podacima, a slika 30 prikazuje izgled popunjene tablice.



Slika 27. Kreiranje tablice genres 1. dio (samostalna izrada)



Slika 28. Kreiranje tablice genres (samostalna izrada)

```
1 INSERT INTO iudovcic20.genre (genres)
2 SELECT genres FROM iudovcic20.hbo_dataset;
```

Slika 29. Naredba za popunjavanje genres (samostalna izrada)

	idgenres	genre
▶	1	['fantasy', 'family']
	2	['drama']
	3	['drama', 'romance', 'war']
	4	['thriller', 'crime']
	5	['thriller', 'romance', 'crime']
	6	['war', 'romance', 'drama', 'history']

Slika 30. Izgled tablice genres (samostalna izrada)

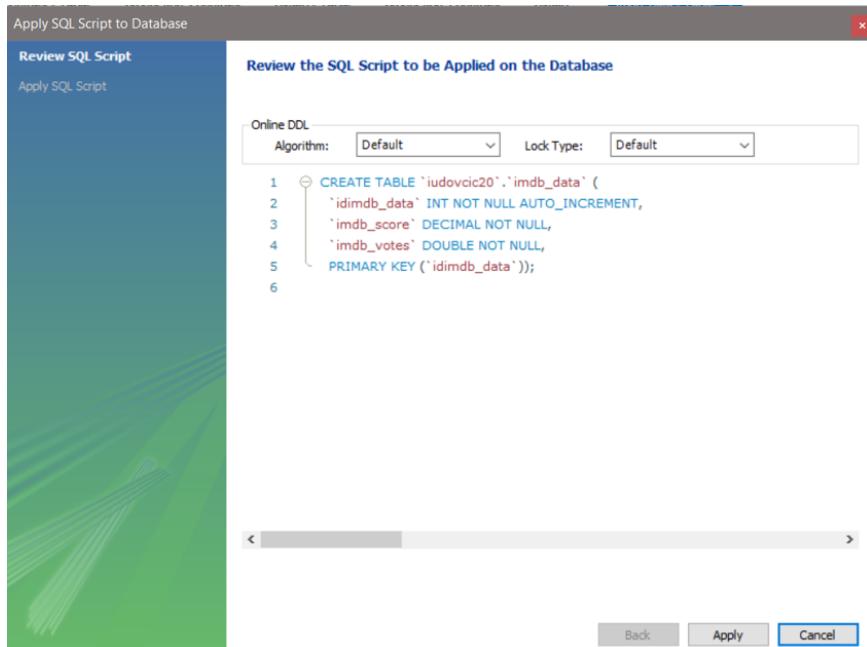
### 5.3. Dimenzijska tablica *imdb\_data*

U tablici se nalaze tri atributa, a to su *idimdb\_data*, *imdb\_score* i *imdb\_vote*. Primarni ključ je *idimdb\_data* koji je *auto increment* i *not null*, a *imdb\_score* je *decimal(5, 1)* što znači da ima jedno decimalno mjesto, a ukupno najviše može biti pet brojeva. Posljednji atribut *imdb\_vote* je *double*. Svi atributi su postavljeni na *not null*. Slika 31. prikazuje kreiranje dimenzijske tablice.

Column Name	Datatype	PK	NN	UQ	B	UN	ZF	AI	G	Default/Expression
idimdb_data	INT	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>					
imdb_score	DECIMAL	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>					
imdb_votes	DOUBLE	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>					

Slika 31. Kreiranje tablice *imdb\_data* 1. dio (samostalna izrada)

Na slici 32 prikazan je drugi korak kreiranja odnosno prikazana je naredba za kreiranje tablice.



Slika 32. Kreiranje tablice *imdb\_data* 2. dio (samostalna izrada)

Slika 33 prikazuje naredbu koja popunjava novokreiranu tablicu *imdb\_data* s podacima iz tablice *hbo\_dataset* u kojoj se nalaze svi uvezeni podaci.

```

1 •  INSERT INTO iudovcic20.imdb_data (imdb_score, imdb_votes)
2   SELECT imdb_score, imdb_votes FROM iudovcic20.hbo_dataset;

```

Slika 33. Naredba za popunjavanje *imdb\_data* (samostalna izrada)

Slika 34 prikazuje izgled tablice nakon izvršene naredbe sa slike 33.

<i>idimdb_data</i>	<i>imdb_score</i>	<i>imdb_votes</i>
1	8.1	389774
2	8.3	433804
3	8.5	558849
4	7.9	84494
5	8.0	156603
6	8.2	309856

Slika 34. Izgled tablice *imdb\_data* (samostalna izrada)

## 5.4. Dimenzijska tablica *ostalo*

U tablici *ostalo* nalaze se atributi *title*, *seasons*, *runtime* i *release\_year*. Osim njih tu je i primarni ključ *idostalo* koji je *auto increment* i *not null*. Što se tiče ostalih atributa i njihovih tipova, *title* je tipa *VARCHAR (200)* i *not null*, *seasons* je *VARCHAR (45)* i *not null*, *runtime* je tipa *INT* i *not null*, a *release\_year* je *INT* i *not null*.

Slika 35 prikazuje kreiranje tablice, a slika 36 prikazuje naredbu koju kreira alat. Na slici 37 prikazana je naredba za popunjavanje tablice podacima iz tablice hbo\_dataset, a slika 38 prikazuje izgled tablice nakon popunjavanja.

The screenshot shows the 'Create Table' dialog in MySQL Workbench. The table name is 'ostalo', schema is 'iudovcic20', charset is 'Default Charset', collation is 'Default Collation', and engine is 'InnoDB'. The table structure includes columns: idostalo (INT, PK, NOT NULL, AUTO\_INCREMENT), title (VARCHAR(200) NOT NULL), seasons (VARCHAR(45) NOT NULL), runtime (INT NOT NULL), and release\_year (INT NOT NULL). Primary key is defined as 'idostalo'.

Slika 35. Kreiranje tablice ostalo 1. dio (samostalna izrada)

The screenshot shows the 'Review SQL Script' dialog in MySQL Workbench. It displays the SQL code for creating the 'ostalo' table:

```

CREATE TABLE `iudovcic20`.`ostalo` (
  `idostalo` INT NOT NULL AUTO_INCREMENT,
  `title` VARCHAR(200) NOT NULL,
  `seasons` VARCHAR(45) NOT NULL,
  `runtime` INT NOT NULL,
  `release_year` INT NOT NULL,
  PRIMARY KEY (`idostalo`));

```

Slika 36. Kreiranje tablice ostalo 2.dio (samostalna izrada)

---

```

1 •  INSERT INTO iudovcic20.hbo_dataset (title, seasons, release_year, runtime)
2   SELECT title, seasons, release_year, runtime FROM iudovcic20.hbo_dataset;

```

Slika 37. Naredba za popunjavanje ostalo (samostalna izrada)

idostalo	title	seasons	runtime	release_year
1	The Wizard of Oz	0	102	1939
2	Citizen Kane	0	119	1941
3	Casablanca	0	102	1942
4	The Big Sleep	0	116	1946
5	The Maltese Falcon	0	100	1941
6	Gone with the Wind	0	233	1939

Slika 38. Izgled tablice ostalo (samostalna izrada)

## 5.5. Dimenzijska tablica *production\_countries*

U ovoj dimenzijskog tablici su dva atributa, a to su *idproduction\_countries* i *production\_countries*. Prvi atribut je primarni ključ, postavljen je na *auto increment* i *not null*, drugi atribut je tipa *VARCHAR(100)* te je *not null*. Dolje prikazane slike prikazuju kreiranja tablice, naredbu koju generira MySQL Workbench za kreiranje, naredbu za popunjavanje tablice te izgled tablice nakon te naredbe.

Slika 39. Kreiranje tablice production\_countries 1. dio (samostalna izrada)

Slika 40. Kreiranje tablice production\_countries 2. dio (samostalna izrada)

```

1 •  INSERT INTO iudovcic20.production_countries (production_countries )
2   SELECT production_countries   FROM iudovcic20.hbo_dataset;

```

Slika 41. Naredba za popunjavanje production\_countries (samostalna izrada)

	idproduction_countries	production_countries
1		['US']
2		['US']
3		['US']
4		['US']
5		['US']
6		['US']

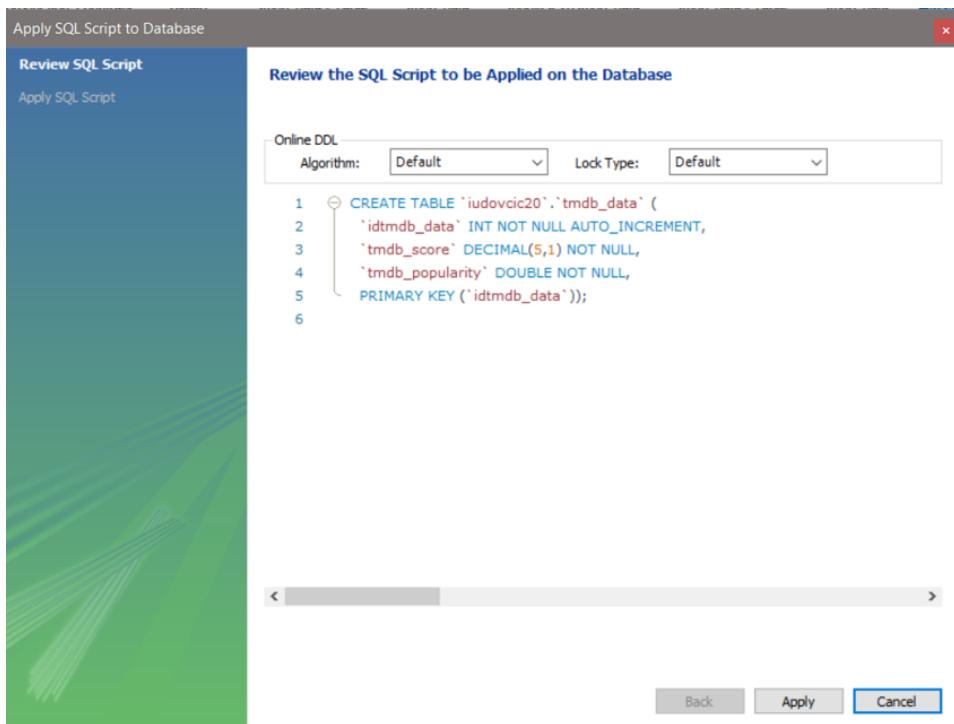
Slika 42. Izgled tablice production\_countries (samostalna izrada)

## 5.6. Dimenzijska tablica tmdb\_data

Ova tablica sadržava tri atributa, a to su *idtmdb\_data* koji je primarni ključ, postavljen na *auto increment* te na *not null*, drugi atribut je *tmdb\_score* koji je *decimal(5,1)* što znači da može imati ukupno pet brojeva, a iza decimalne točke je samo jedan broj. Treći atribut je *tmdb\_popularity* koji je tipa *double*. Donje slike prikazuju kako je tablica kreirana, kako je popunjena te kako izgleda nakon naredbe insert.

Column Name	Datatype	PK	NN	UQ	B	UN	ZF	AI	G	Default/Expression
idtmdb_data	INT	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>					
tmdb_score	DECIMAL(5,1)	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>					
tmdb_popularity	DOUBLE	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>					

Slika 43. Kreiranje tablice tmdb\_data 1. dio (samostalna izrada)



Slika 44. Kreiranje tablice tmdb\_data 2. dio (samostalna izrada)

```

1 • INSERT INTO iudovcic20.tmdb_data (tmdb_score, tmdb_popularity)
2   SELECT tmdb_score, tmdb_popularity FROM iudovcic20.hbo_dataset;

```

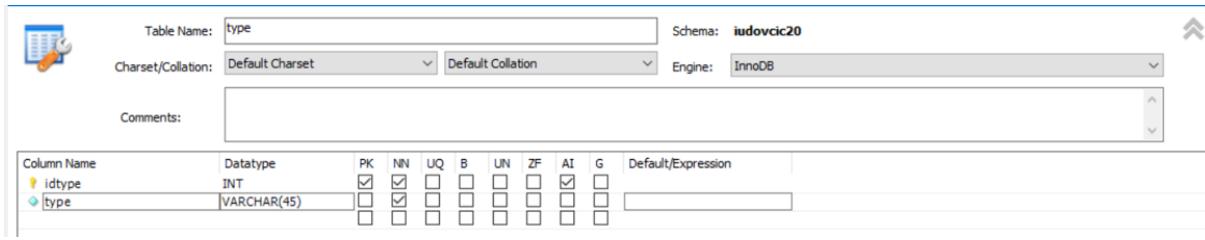
Slika 45. Naredba za popunjavanje tmdb\_data (samostalna izrada)

	idtmdb_data	tmdb_score	tmdb_popularity
1	1	7.6	41.442
2	2	8.0	14.383
3	3	8.2	20.087
4	4	7.7	12.911
5	5	7.8	12.788
6	6	8.0	24.092

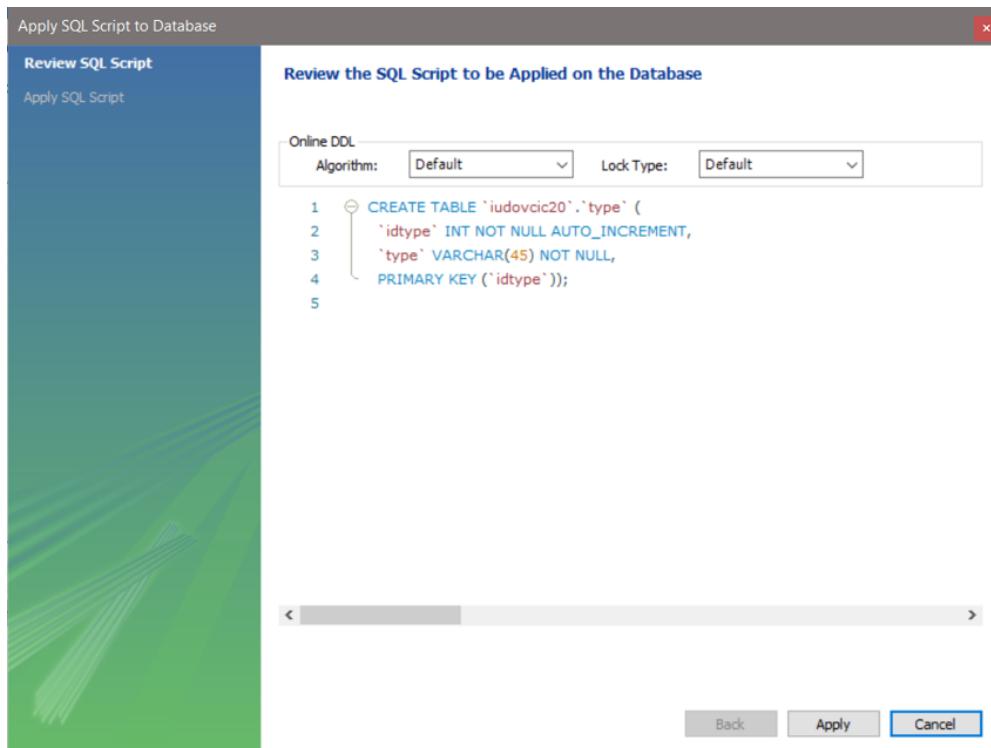
Slika 46. Izgled tablice tmdb\_data (samostalna izrada)

## 5.7. Dimenzijska tablica type

U ovoj tablici su dva atributa: *idtype* koji je primarni ključ i postavljen na *auto increment* i *not null*, a drugi atribut je *type* koji je *VARCHAR (45)* i postavljen je na *not null*. Donje slike prikazuju kako je tablica napravljena, kako su u nju podaci umetnuti te njezin finalni izgled.



Slika 47. Kreiranje tablice type 1. dio (samostalna izrada)



Slika 48. Kreiranje tablice type 2. dio (samostalna izrada)

```

1 INSERT INTO iudovcic20.type (type)
2 SELECT type FROM iudovcic20.hbo_dataset;
  
```

Slika 49. Naredba za popunjavanje type (samostalna izrada)

idtype	type
1	MOVIE
2	MOVIE
3	MOVIE
4	MOVIE
5	MOVIE
6	MOVIE
7	SHOW

Slika 50. Izgled tablice type (samostalna izrada)

## 5.8. Činjenična tablica `general_table`

Činjenična tablica se sastoji od sedam vanjskih ključeva (po jedan ključ za svaku dimenziju): `idage_certification`, `idgenre`, `idimdb_data`, `idostalo`, `idproduction_countries`, `idtmdb_data`, `idtype`. Primarni ključ je `idgeneral_table` koji je postavljen na *not null* i *auto increment*, a ostali atributi su *not null*, a tip je *INT*. Sve veze su 1:N.

Slika 51 prikazuje kako je stvorena činjenična tablica.

Column Name	Datatype	PK	NN	UQ	B	UN	ZF	AI	G	Default/Expression
idgeneral_table	INT	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>					
idage_certification	INT	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>					
idgenre	INT	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>					
idimdb_data	INT	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>					
idostalo	INT	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>					
idproduction_countries	INT	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>					
idtmdb_data	INT	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>					
idtype	INT	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>					

Slika 51. Kreiranje tablice general\_table 1. dio (samostalna izrada)

Slika 52 prikazuje naredbu koju generira alat, a stvara činjeničnu tablicu.

```
CREATE TABLE `iudovicic20`.`general_table` (
  `idgeneral_table` INT NOT NULL AUTO_INCREMENT,
  `idage_certification` INT NOT NULL,
  `idgenre` INT NOT NULL,
  `idimdb_data` INT NOT NULL,
  `idostalo` INT NOT NULL,
  `idproduction_countries` INT NOT NULL,
  `idtmdb_data` INT NOT NULL,
  `idtype` INT NOT NULL,
  PRIMARY KEY (`idgeneral_table`),
  INDEX `idage_certification_idx` (`idage_certification` ASC) VISIBLE,
  INDEX `idgenre_idx` (`idgenre` ASC) VISIBLE,
  INDEX `idimdb_data_idx` (`idimdb_data` ASC) VISIBLE,
  INDEX `idostalo_idx` (`idostalo` ASC) VISIBLE,
  INDEX `idproduction_countries_idx` (`idproduction_countries` ASC) VISIBLE,
  INDEX `idtmdb_data_idx` (`idtmdb_data` ASC) VISIBLE,
  INDEX `idtype_idx` (`idtype` ASC) VISIBLE,
  CONSTRAINT `idage_certification`
    FOREIGN KEY (`idage_certification`)
    REFERENCES `iudovicic20`.`age_certification`(`idage_certification`),
  CONSTRAINT `idgenre`
    FOREIGN KEY (`idgenre`)
    REFERENCES `iudovicic20`.`genre`(`idgenre`),
  CONSTRAINT `idimdb_data`
    FOREIGN KEY (`idimdb_data`)
    REFERENCES `iudovicic20`.`imdb_data`(`idimdb_data`),
  CONSTRAINT `idostalo`
    FOREIGN KEY (`idostalo`)
    REFERENCES `iudovicic20`.`ostalo`(`idostalo`),
  CONSTRAINT `idproduction_countries`
    FOREIGN KEY (`idproduction_countries`)
    REFERENCES `iudovicic20`.`production_countries`(`idproduction_countries`),
  CONSTRAINT `idtmdb_data`
    FOREIGN KEY (`idtmdb_data`)
    REFERENCES `iudovicic20`.`tmdb_data`(`idtmdb_data`),
  CONSTRAINT `idtype`
    FOREIGN KEY (`idtype`)
    REFERENCES `iudovicic20`.`type`(`idtype`)
) ENGINE=InnoDB;
```

Slika 52. Kreiranje tablice general\_table 2. dio (samostalna izrada)

Slika 53 prikazuje naredbu za popunjavanje činjenične tablice, a slika 54 izgled popunjene tablice.

```
INSERT INTO iudovicic20.general_table (idage_certification, idgenre, idimdb_data, idtmdb_data, idostalo, idproduction_countries, idtype)
SELECT idage_certification, idgenre, idimdb_data, idtmdb_data, idostalo,
       idproduction_countries, idtype
FROM iudovicic20.age_certification, iudovicic20.genre, iudovicic20.imdb_data,
     iudovicic20.tmdb_data, iudovicic20.ostalo, iudovicic20.production_countries, iudovicic20.type
WHERE
  idage_certification=idgenre AND idgenre=idimdb_data AND idimdb_data=idtmdb_data AND idtmdb_data=idostalo AND
  idostalo=idproduction_countries AND idproduction_countries=idtype;
```

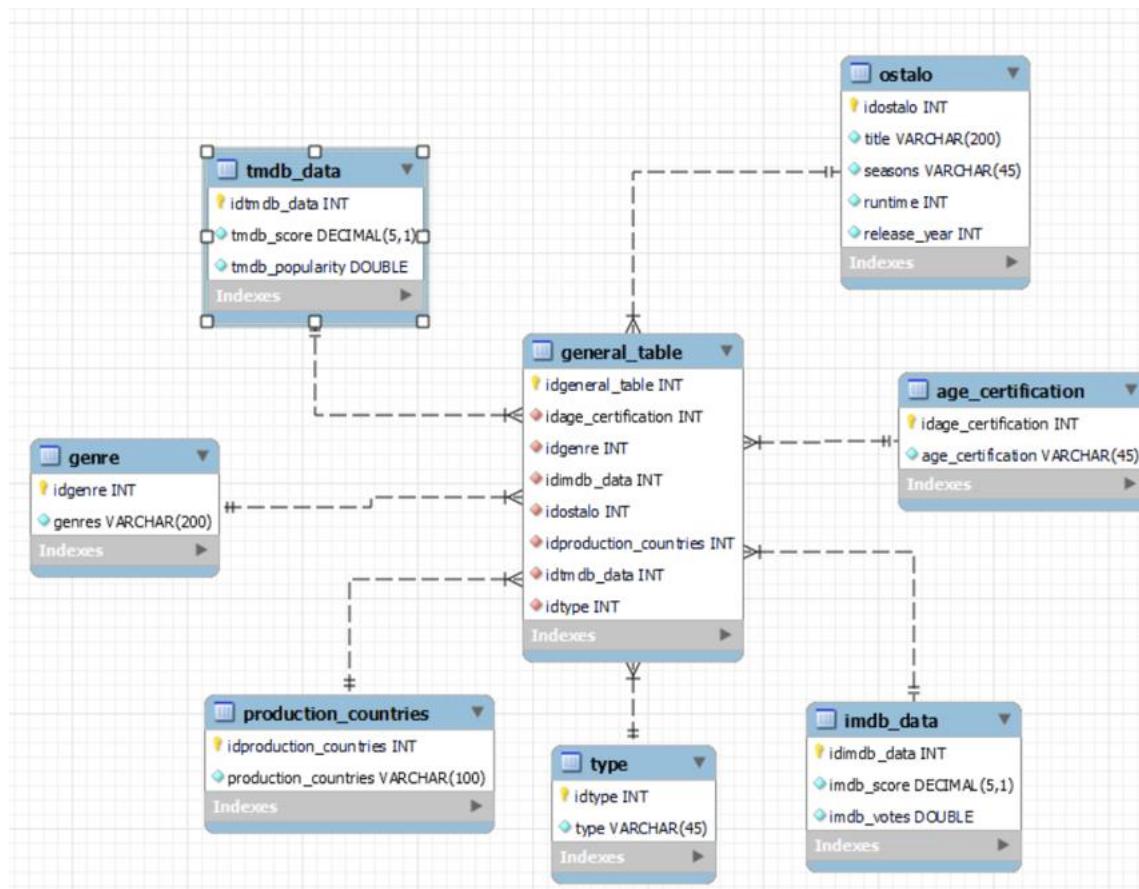
Slika 53. Naredba za popunjavanje general\_table (samostalna izrada)

	idgeneral_table	idage_certification	idgenres	idimdb_data	idostalo	idproduction_countries	idtmdb_data	idtype
▶	1	1	1	1	1	1	1	1
	2	2	2	2	2	2	2	2
	3	3	3	3	3	3	3	3
	4	4	4	4	4	4	4	4
	5	5	5	5	5	5	5	5
	6	6	6	6	6	6	6	6

Slika 54. Izgled tablice general\_table (samostalna izrada)

## 5.9. Prikaz strukture skladišta i modela

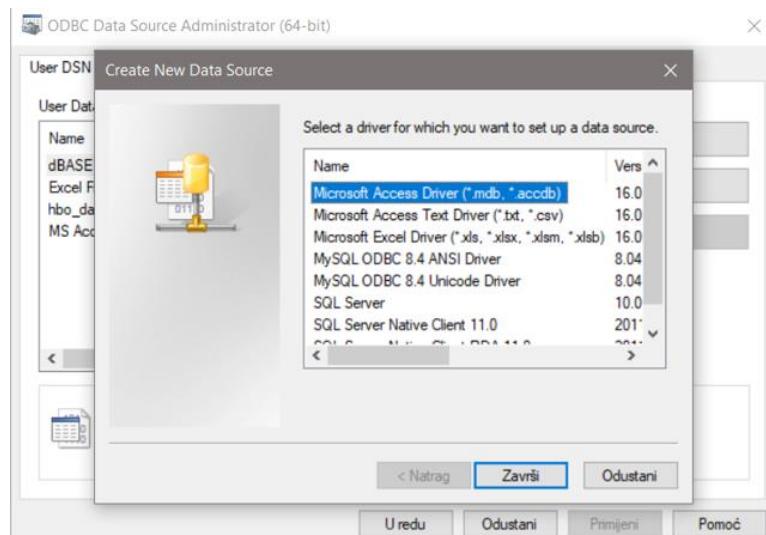
Donja slika prikazuje model zvijezde sastavljen od sedam dimenzijskih tablica i jedne činjenične tablice. Sve veze su 1:N.



Slika 55. Struktura skladišta (samostalna izrada)

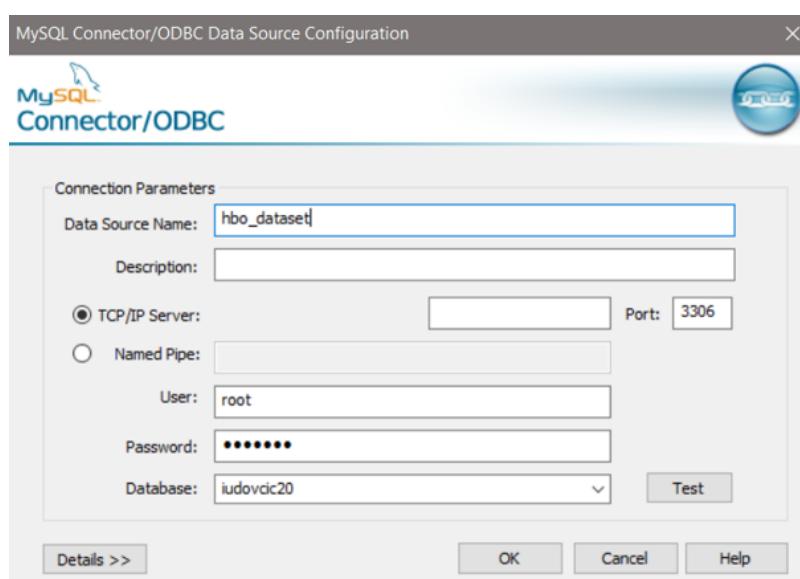
## 6. Analiza u alatu Tableau

Da bi se provela analiza potrebno je povezati bazu s alatom Tableau, to je napravljeno koristeći ODBC Data Connector tako što je izabran novi izvor podataka. Slika 57 prikazuje izvore podataka. Za povezivanje sa Tableau-om potrebno je izbrati ili *MySQL ODBC 8.4 ANSI Driver* ili *MySQL ODBC 8.4 Unicode Driver*. Izabran je *MySQL ODBC 8.4 ANSI Driver*.



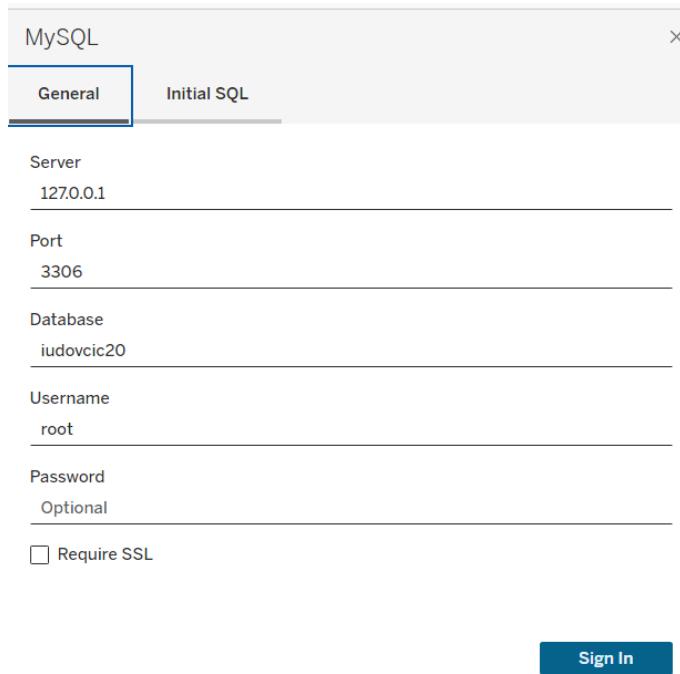
Slika 56. Odabir novog izvora podataka (samostalna izrada)

Nakon što se odabere izvor podataka potrebno je upisati parametre za povezivanje.



Slika 57. ODBC Connector (samostalna izrada)

Nakon toga u alatu Tableau odabrana je veza na server MySQL i upisani su podaci za povezivanje kako je prikazano na donjoj slici.

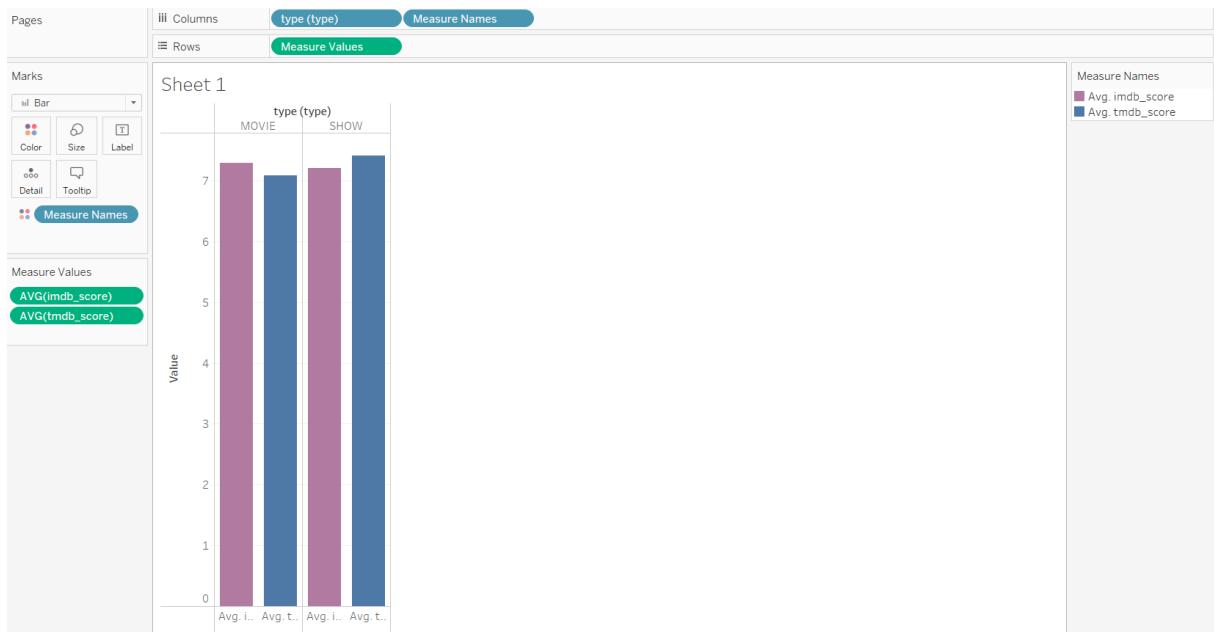


Slika 58. Povezivanje s Tableau-om (samostalna izrada)

## 6.1. Izvještaji

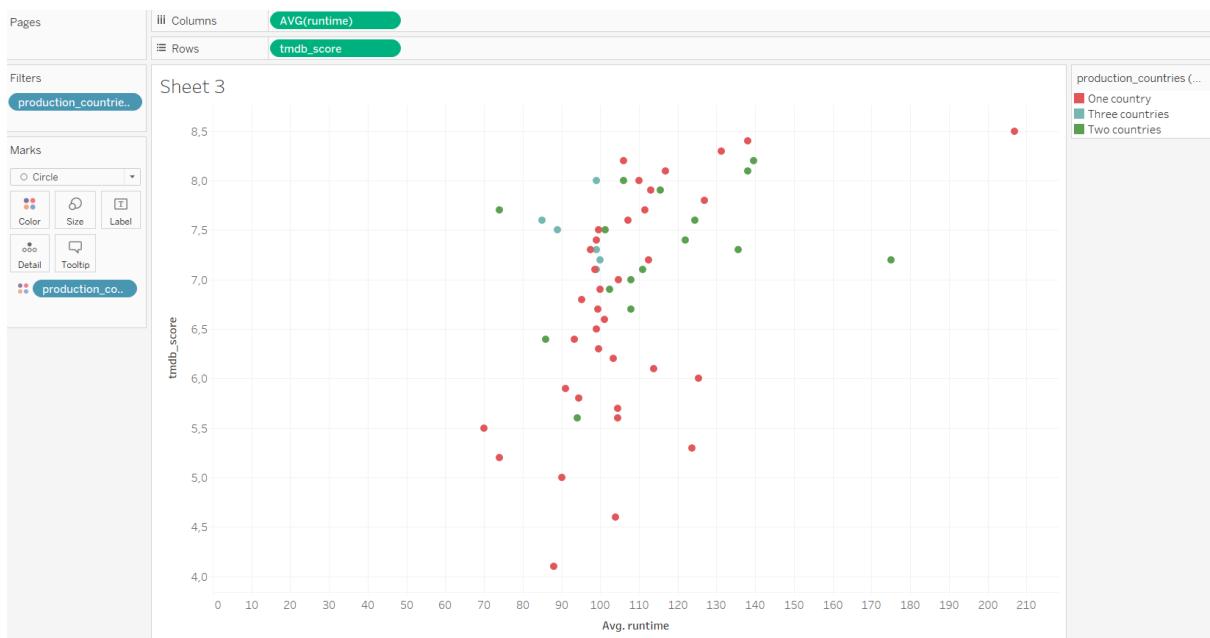
Radi boljeg razumijevanja podataka Tableau omogućuje različite usporedbe prikazane u obliku izvještaja koji su lako razumljivi i jednostavnvi za napraviti.

Prvi izvještaj prikazuje prosječnu ocjenu prema *imdb*-u i prema *tmdb*-u za filmove i serije. Možemo vidjeti kako serije imaju veću prosječnu ocjenu prema *tmdb*-u, a filmovi prema *imdb*-u. Što se tiče prosječne *tmdb* ocjene kod filmova nešto je niža nego *imdb* ocjena, a kod filmova je obrnuto.



Slika 59. Izvještaj 1 (samostalna izrada)

Drugi izvještaj prikazuje ovisnost trajanja filma (prosjek) i zemlje u kojoj je film snimljen prema *tmdb* ocjeni. Može se zaključiti da filmovi i serije u prosjeku traju između 70 i 140 minuta uz dva stršila. Iako, u podacima postoje serije koje znatno manje traju od prosjeka tu je prikazan prosjek radi jednostavnije usporedbe. Zemlje produkcije grupirane su u grupe od jedne zemlje, dvije i tri, a grupa od četiri zemlje je izostavljena iz ovog grafa. Dakle, grupa *One country* sadrži samo one filmove i serije koji su snimljeni u jednoj zemlji, a grupa *Two country* filmove i serije koji su snimljeni u dvije zemlje. Vidimo s grafa da je najviše filmova i serija snimljeno u jednoj zemlji, a najmanje ih je snimljeno u tri zemlje. Filmovi i serije snimljeni u jednoj zemlji imaju šarolike ocjene, dok oni snimljeni u dvije zemlje su između 5.5 i 8.5. Filmovi i serije koji su snimljeni u tri zemlje imaju ocjenu između 7 i 8.



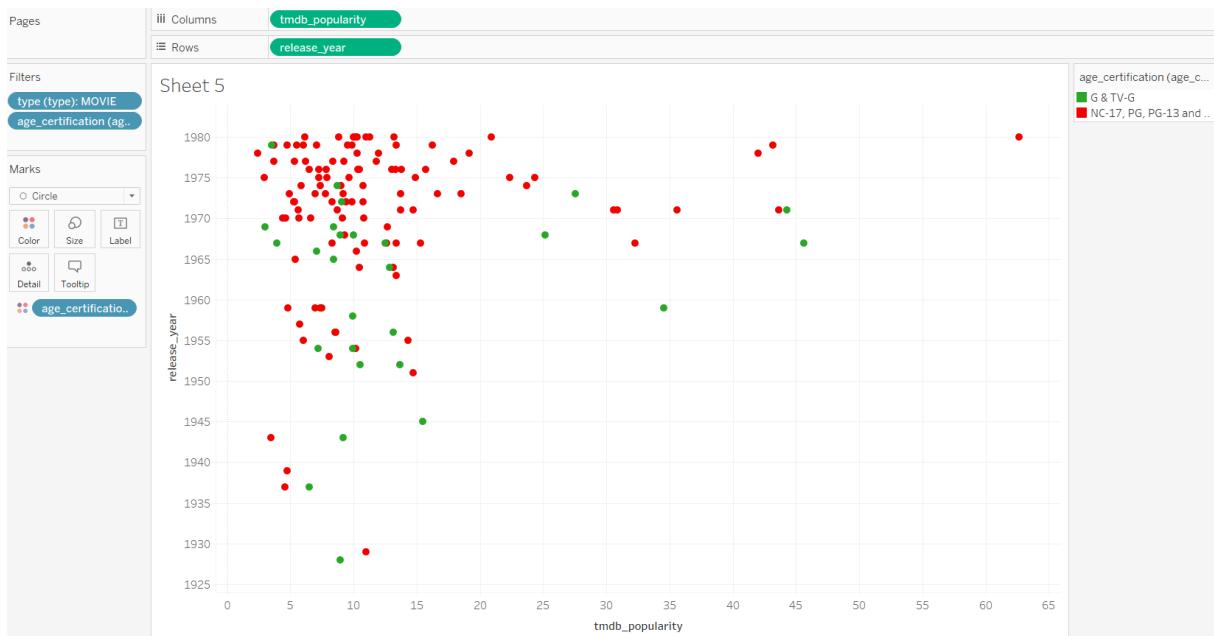
Slika 60. Izvještaj 2 (samostalna izrada)

Graf 3 prikazuje emisije, broj sezona, prosječnu *imdb* ocjenu. Svaki krug prikazuje broj sezona i prosječnu ocjenu za seriju koja ima taj broj sezona.



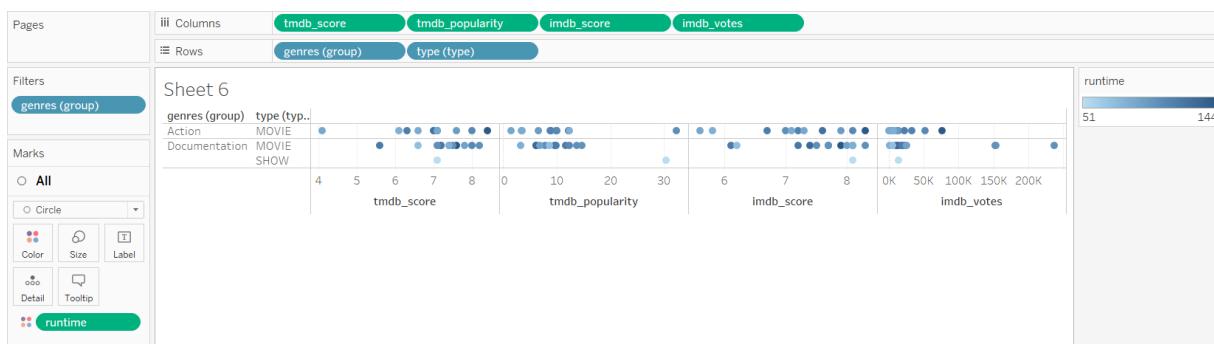
Slika 61. Izvještaj 3 (samostalna izrada)

Četvrti izvještaj prikazuje imaju li veću popularnost na *tmdb*-u filmovi koji imaju oznaku G i TV-G odnosno pogodni su za sve uzraste ili veću ocjenu imaju filmovi koji su za djecu i oni koji imaju drugačija ograničenja (nisu za maloljetne ili za djecu ispod 13 godina). U obzir su uzeti filmovi i serije koji su izšli između 1925. i 1980. Možemo zaključiti kako je više filmova koji imaju neko ograničenje te kako većina filmova ima popularnost između 5 i 15.



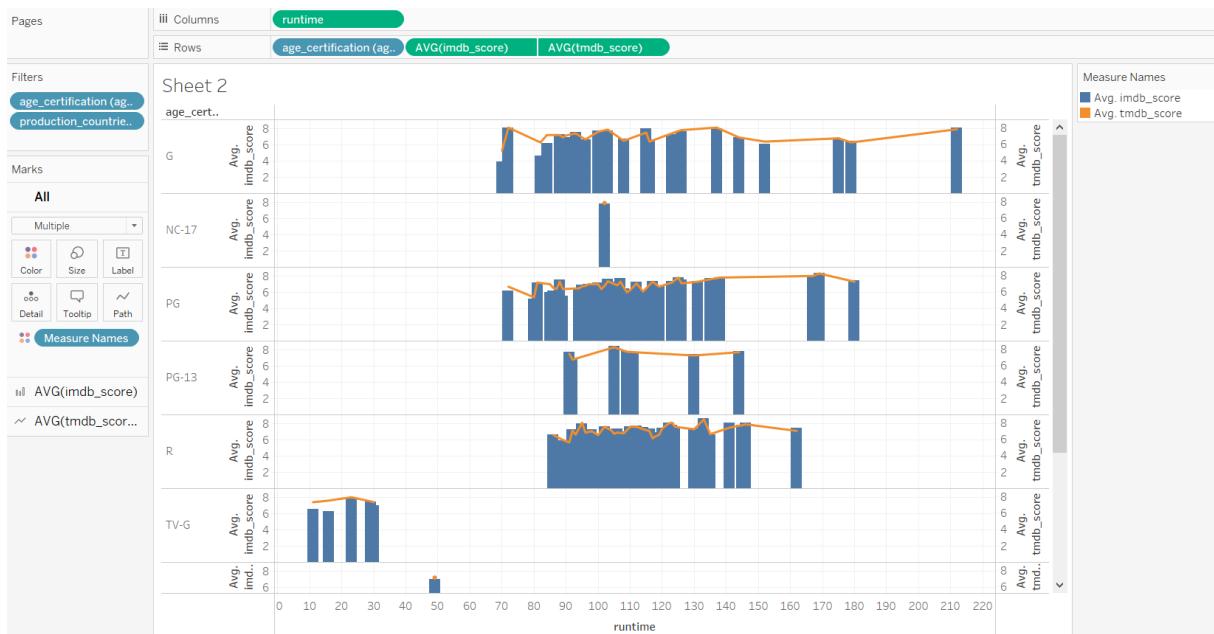
Slika 62. Izvještaj 4 (samostalna izrada)

Izvještaj pet prikazuje dokumentarne i akcijske filmovi i emisije po ocjeni na *tmdb*-u, *imdb*-u te koliko imaju *imdb* glasova i kolika im je popularnost na *tmdb*-u. Također, graf pokazuje i trajanje tih filmova i serija. Vidimo da akcijskih serija nema, a dokumentarnih je vrlo malo. Ocjene su približno jednake, a popularnost je nešto niža kod akcijskih filmova nego kod dokumentarnih. Glasovi na *imdb*-u za dokumentarne filmove (ako se izuzmu dva stršila) su nešto manji nego za akcijske. Ako kod *imdb* ocjena izmazemo stršila i kod akcijskih i kod dokumentarnih filmova vidimo da su ocjene vrlo blizu. Ako kod *tmdb* ocjena izuzmemo stršila vidimo kako malo veću ocjenu imaju akcijski filmovi.

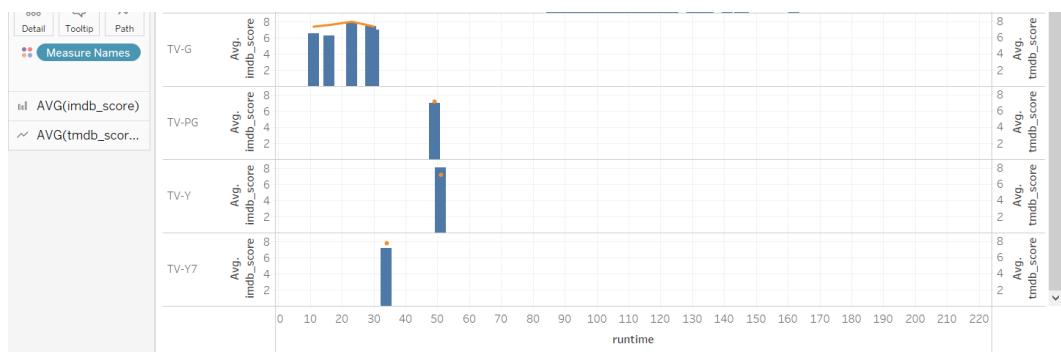


Slika 63. Izvještaj 5 (samostalna izrada)

U izvještaju šest uspoređuju se prosječne ocjene na *imdb* i *tmdb* po svakoj oznaci za dobnu skupinu za koju je sadržaj namijenjen. Osim toga, prikazano je i vrijeme trajanja filma ili serije. Vidimo da između ocjena na *imdb*-u i *tmdb*-u ima manjih odstupanja te ukoliko osoba želi gledati neki film ili seriju može pogledati samo jednu ocjenu kako bi dobila realnu predodžbu o ocjeni filma ili serije. Osim toga, uočljivo je da filmovi imaju duže vrijeme trajanja.



Slika 64. Izvještaj 6. dio (samostalna izrada)



Slika 65. Izvještaj 6.2. dio (samostalna izrada)

## 7. Zaključak

Cilj ovog projekta je bio analizirati podatke o filmovima i serijama koji su dostupni na HBO platformi. Skup podataka koji je korišten besplatno je preuzet te su nad njim provedene različite transformacije, poput onih u Excelu i MySQL Workbenchu. Osim ta dva alata korišten je i alat za poslovnu inteligenciju. Alat Tableau omogućio je prikaz podataka i stvaranje različitih grafova. Prije učitavanja podataka u alat Tableau bilo je potrebo izvršiti neke transformacije kao što je popunjavanje praznih redova i mijenjanje tipova podataka.

Na kraju rada prikazano je šest izvještaja koji pobliže prikazuju podatke te daju bolji uvid o filmovima i serijama kao što je ocjena filma ili serije, oznaka dobne skupine za koju je namijenjen, trajanje, broj sezona, godina kad je film ili serija izašao te u kojoj zemlji je snimljen.

Rad u alatima koji su korišteni u ovom radu je prilično jednostavan, a posebno rad u alatu Tableau.

# Popis literature

Oracle, (bez dat). *What Is a Data Warehouse?*, Preuzeto 21.5.2024. s <https://www.oracle.com/database/what-is-a-data-warehouse/>

INCWORX CONSULTING (bez dat). *Zašto je skladištenje podataka važno?*, Preuzeto 22.5.2024. s <https://www.incworx.com/blog/why-is-data-warehousing-important>

MySQL, (bez dat) *MySQL Workbench Enhances Dana Migration*, Preuzeto 22.5.2024. s <https://www.mysql.com/products/workbench/>

Biswal, A. (21.7.2023). *What is Tableau: Power-packed Tutorial For Beginners*, Preuzeto 22.5.2024. s <https://www.simplilearn.com/tutorials/tableau-tutorial/what-is-tableau>

Verma R. (1.6.2022). *What is MySQL ODBC Connector? – A Comprehensive Guide*, Preuzeto 22.5.2024. s <https://hevodata.com/learn/mysql-odbc-connector/>

Kropov V. (1.11.2023). *Building a dana warehouse: A step-by-step guide*, Preuzeto 22.5.2024. s <https://www.n-ix.com/building-a-data-warehouse/>

Aws (bez dat). *What is ETL (Extract Transform Load)?*, Preuzeto 22.5.2024. s <https://aws.amazon.com/what-is/etl/>

Integrate.io (20.10.2023), *ETL and Dana Warehousing Explained: ETL Tool Basics*, Preuzeto 22.5.2023 s <https://www.integrate.io/blog/etl-data-warehousing-explained-etl-tool-basics/>

GeeksforGeeks (8.5.2023). *Star Schema in Dana Warehouse modeling*, Preuzeto 23.5.2024. s <https://www.geeksforgeeks.org/star-schema-in-data-warehouse-modeling/>

Gheorghiu (23.5.2023). *What Is the Star Schema Data Model? An Explanation with 3 Examples*, Preuzeto 23.5.2024. s <https://vertabelo.com/blog/star-chema-data-model/>

Kaggle – dataset (2022). *HBO Max TV Shows and Movies*, Preuzeto 20.5.2024. s <https://www.kaggle.com/datasets/victorsoeiro/hbo-max-tv-shows-and-movies?select=titles.csv>

# Popis slika

Slika 1. Početni skup podataka (samostalna izrada) .....	5
Slika 2. ETL postupak (Integrate.Io, 2023).....	6
Slika 3. Pretvaranje teksta u stupce 1. korak (samostalna izrada).....	7
Slika 4. Pretvaranje teksta u stupce 2. korak (samostalna izrada).....	8
Slika 5. Pretvaranje teksta u stupce 3. korak (samostalna izrada).....	8
Slika 6. Finalni izgled podataka u Excelu (samostalna izrada).....	9
Slika 7. Početni zaslon MySQL Workbench (samostalna izrada) .....	9
Slika 8. Povezivanje na bazu (samostalna izrada).....	10
Slika 9. Kreiranje scheme (samostalna izrada) .....	10
Slika 10. Uvoz podataka 1. korak (samostalna izrada) .....	11
Slika 11. Odabir skladišta i imena tablice (samostalna izrada).....	12
Slika 12. Odabir stupaca (samostalna izrada) .....	12
Slika 13. Tablica hbo_dataset (samostalna izrada) .....	13
Slika 14. Izgled uvezenih podataka (samostalna izrada).....	13
Slika 15. Update nad age_certification (samostalna izrada) .....	14
Slika 16. Izgled tablice nakon update upita nad age_certification (samostalna izrada).....	14
Slika 17. Update nad seasons (samostalna izrada) .....	14
Slika 18. Izgled tablice nakon update nad seasons (samostalna izrada) .....	14
Slika 19. Update tmdb_score i imdb_score (samostalna izrada) .....	15
Slika 20. Izgled tablice nakon updata imdb_score i tmdb_score (samostalna izrada) .....	15
Slika 21. Table Inspector nakon svih updata (samostalna izrada) .....	15
Slika 22. Model zvijezde (GeeksforGeeks, 2023).....	17
Slika 23. Kreiranje tablice age_certification 1.dio (samostalna izrada).....	18
Slika 24. Kreiranje tablice age_certification 2. dio (samostalna izrada).....	18
Slika 25. Naredba za popunjavanje age_certification tablice (samostalna izrada) .....	18
Slika 26. Izgled tablice age_certification (samostalna izrada) .....	18
Slika 27. Kreiranje tablice genres 1. dio (samostalna izrada) .....	19
Slika 28. Kreiranje tablice genres (samostalna izrada) .....	19
Slika 29. Naredba za popunjavanje genres (samostalna izrada) .....	19
Slika 30. Izgled tablice genres (samostalna izrada) .....	20
Slika 31. Kreiranje tablice imdb_data 1. dio (samostalna izrada).....	20
Slika 32. Kreiranje tablice imdb_data 2. dio (samostalna izrada).....	21
Slika 33. Naredba za popunjavanje imdb_data (samostalna izrada) .....	21
Slika 34. Izgled tablice imdb_data (samostalna izrada) .....	21
Slika 35. Kreiranje tablice ostalo 1. dio (samostalna izrada) .....	22
Slika 36. Kreiranje tablice ostalo 2.dio (samostalna izrada) .....	22
Slika 37. Naredba za popunjavanje ostalo (samostalna izrada) .....	22
Slika 38. Izgled tablice ostalo (samostalna izrada) .....	23
Slika 39. Kreiranje tablice production_countries 1. dio (samostalna izrada).....	23
Slika 40. Kreiranje tablice production_countries 2. dio (samostalna izrada).....	23
Slika 41. Naredba za popunjavanje production_countries (samostalna izrada) .....	24
Slika 42. Izgled tablice production_countries (samostalna izrada) .....	24
Slika 43. Kreiranje tablice tmdb_data 1. dio (samostalna izrada).....	24
Slika 44. Kreiranje tablice tmdb_data 2. dio (samostalna izrada).....	25
Slika 45. Naredba za popunjavanje tmdb_data (samostalna izrada) .....	25
Slika 46. Izgled tablice tmdb_data (samostalna izrada) .....	25
Slika 47. Kreiranje tablice type 1. dio (samostalna izrada).....	26

Slika 48. Kreiranje tablice type 2. dio (samostalna izrada).....	26
Slika 49. Naredba za popunjavanje type (samostalna izrada) .....	26
Slika 50. Izgled tablice type (samostalna izrada) .....	26
Slika 51. Kreiranje tablice general_table 1. dio (samostalna izrada) .....	27
Slika 52. Kreiranje tablice general_table 2. dio (samostalna izrada) .....	27
Slika 53. Naredba za popunjavanje general_table (samostalna izrada) .....	27
Slika 54. Izgled tablice general_table (samostalna izrada) .....	28
Slika 55. Struktura skladišta (samostalna izrada).....	28
Slika 56. Odabir novog izvora podataka (samostalna izrada) .....	29
Slika 57. ODBC Connector (samostalna izrada).....	29
Slika 58. Povezivanje s Tableau-om (samostalna izrada) .....	30
Slika 59. Izvještaj 1 (samostalna izrada) .....	31
Slika 60. Izvještaj 2 (samostalna izrada) .....	32
Slika 61. Izvještaj 3 (samostalna izrada) .....	32
Slika 62. Izvještaj 4 (samostalna izrada) .....	33
Slika 63. Izvještaj 5 (samostalna izrada) .....	33
Slika 64. Izvještaj 6 1. dio (samostalna izrada) .....	34
Slika 65. Izvještaj 6 2. dio (samostalna izrada) .....	34