Hurtownia danych filmowych

Kamień milowy 2.

Adrian Kamiński, Konstanty Kraszewski, Piotr Marciniak

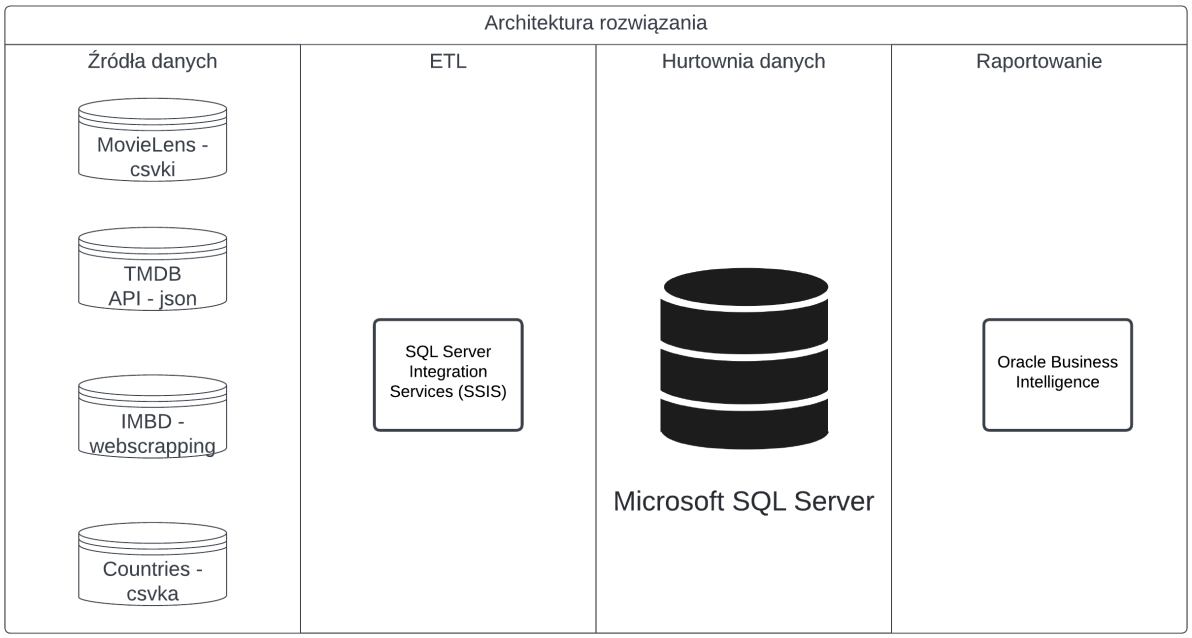
1 Opis celu projektu oraz planowane korzyści z perspektywy odbiorcy rozwiązania

Celem projektu jest zbudowanie hurtowni danych, a następnie utworzenie raportów na temat filmów dla producentów filmowych. Zapewni to dodatkowe informacje na temat:

- trendów filmowych w danych latach,
- aktorów, którzy przynoszą większe zyski,
- krajów, które są dobre do produkcji filmów,
- wpływu, jaki mają reżyserzy na zwrot produkcji.

Te dodatkowe informacje mogą się okazać przydatne przy podejmowaniu strategicznych decyzji na temat nowych produkcji. Potencjalnie mogą one wpływać na duże kwoty, ponieważ zarówno koszt, jak i zwrot z takich produkcji, są liczone w milionach dolarów. Poprawne zinterpretowanie faktów z hurtowni może zapewnić wysoki zwrot z inwestycji.

2 Diagram oraz opis proponowanej architektury rozwiązania



Rysunek 1: Diagram proponowanej architektury rozwiązania.

Źródła danych – wykorzystane zostaną pliki płaskie, pliki JSON zwrócone przez API oraz dane uzyskane z internetu (dokładniejszy opis w opisie zbiorów danych.
ETL – do utworzenia procesu ETL wykorzystane będzie rozwiązanie firmy Microsoft, czyli SSIS (dokładniejszy opis jest przedstawiony w planowanym procesie ETL).
Hurtownia danych – Hurtownię danych stworzymy w narzędziu firmy Microsoft, czyli SQL Server. Model i poszczególne tabele można znaleźć w dziale opisującym to zagadnienie.
Raportowanie – System raportujący stworzony zostanie przy pomocy Oracle Business Intelligence. Dokładniejszy opis raportów można znaleźć w dziale opisującym to zagadnienie.

3 Opis zbiorów danych

Zbiory, które będą wykorzystywane to:

- MovieLens 25M Dataset – jest to zbiór składający z pięciu plików *csv* zawierających informacje na temat ocen, tagów i linków do innych źródeł na temat filmów. Ten zbiór zawiera 25000095 ocen 62423 filmów.
- The Movie DB API – jest to API pozwalające na zapytania do strony *The Movie Database*. API zwraca informację w postaci plików JSON. Pozwala ona z uzyskać dodatkowe informacje na temat filmów, np. kim byli główni aktorzy w filmie, ile wynosił budżet, czy jaki był zysk z filmu.
- IMDB – jest to strona internetowa, z której można pobrać informacje na temat filmów. Jako że bezpośrednio z tej strony będą pozyskiwane informacje, to pierwotne dane będą w formacie HTML. W opisywanym przypadku pobierana będzie głównie informacja o średniej ocen.
- Countries with Regional Codes – jest to zbiór zawierające dane na temat krajów, ładowany z pliku csv. Będzie głównie służył do połączenia kodu kraju z faktyczną informacją o tym kraju. Zbiór ten, będzie tylko raz ładowany do hurtowni.

Dane zostaną podzielone na paczki po 1000 (może więcej) filmów wg zbioru MovieLens wynika, to z tego, że dane zapewniane przez API, czy web scrappingu mają ograniczenia w postaci przepustowości zapytań. Do jednego filmu, często trzeba będzie wykonać około 10 zapytań do API. Dane dotyczące ocen będą agregowane co miesiąc.

4 Planowany proces ETL

Nasz ETL, został podzielony na 3 główne zadania:

- Insert – wstawianie nowych filmów do bazy (wiąże się to z wstawieniem danych do wielu tabel);
- Update Movie – aktualizowanie (insert) tabeli faktowej Movie (periodic snapshot);
- Update People – aktualizowanie (insert) tabeli People (Slowly Changing Dimension Type 2).

Wizualne przedstawienia tych zadań znajdują się poniżej w odpowiednich sekcjach: Insert, Update Movies, Update People.

Insert

Wstawianie działa na podstawie pliku *csv* w odpowiedniej lokalizacji. Dodanie do niego ID filmów, które chcemy dodać do bazy, i uruchomienie zadania Insert wstawi te filmy do naszej hurtowni.

- Insert Countries – uzupełnienie tabeli Countries, transformacje takie jak zmiana daty na odpowiedni typ, lookup, czy dany kraj istnieje.
- Get Existing Movies, People – pobiera ID osób (People), które są w bazie, żeby nie wywoływać zapytania do API o ich dane (oszczędność czasu, istnieje inne zadanie do aktualizacji tabeli People).
- CallToApi – pobieranie danych z API oraz scrappowanie ich ze strony IMDb.
- Cache Language Code – wczytanie tabeli z kodami języków i ich pełnymi nazwami do pamięci podręcznej, aby móc zrobić lookup w dalszym krok.
- Insert Movies Details – zgarniamy plik *csv*, wykonujemy transformacje tj. zmiana daty do postaci YYYYMMDD, zmiana nazewnictwa wartości kolumny (przyjmowała 3 wartości), rzutowanie do odpowiednich typów, lookup do tabeli (w pamięci podręcznej) i obsługa braku wartości w tym lookupie, dodanie danych filmów, których nie było w tabeli.
- Insert People – wczytywanie pliku *csv*, kodowanie kolumny Gender, zmiana dat na odpowiedni format, konwersja typów, obsługa braków danych, tworzenie kolumn technicznych (potrzebnych do SCD 2).
- Insert Movie – wczytanie pliku *csv*, konwersja typów, obsługa braków danych, lookup do tabeli Countries, obsługa Match/No match output, wstawianie do bazy.
- Insert MoviePeople – wczytanie pliku *csv*, lookup do tabeli MovieDetails, zmiana kolumn na odpowiedni typ danych, wstawianie do tabeli.
- Clean up – archiwizuje stworzone w trakcie procesu pliki pomocnicze *csv* (tworzy plik *zip* z datą wywołania).

Dodatkowo wykorzystaliśmy kontenery, żeby zapewnić konkretną kolejność wykonywania działań.

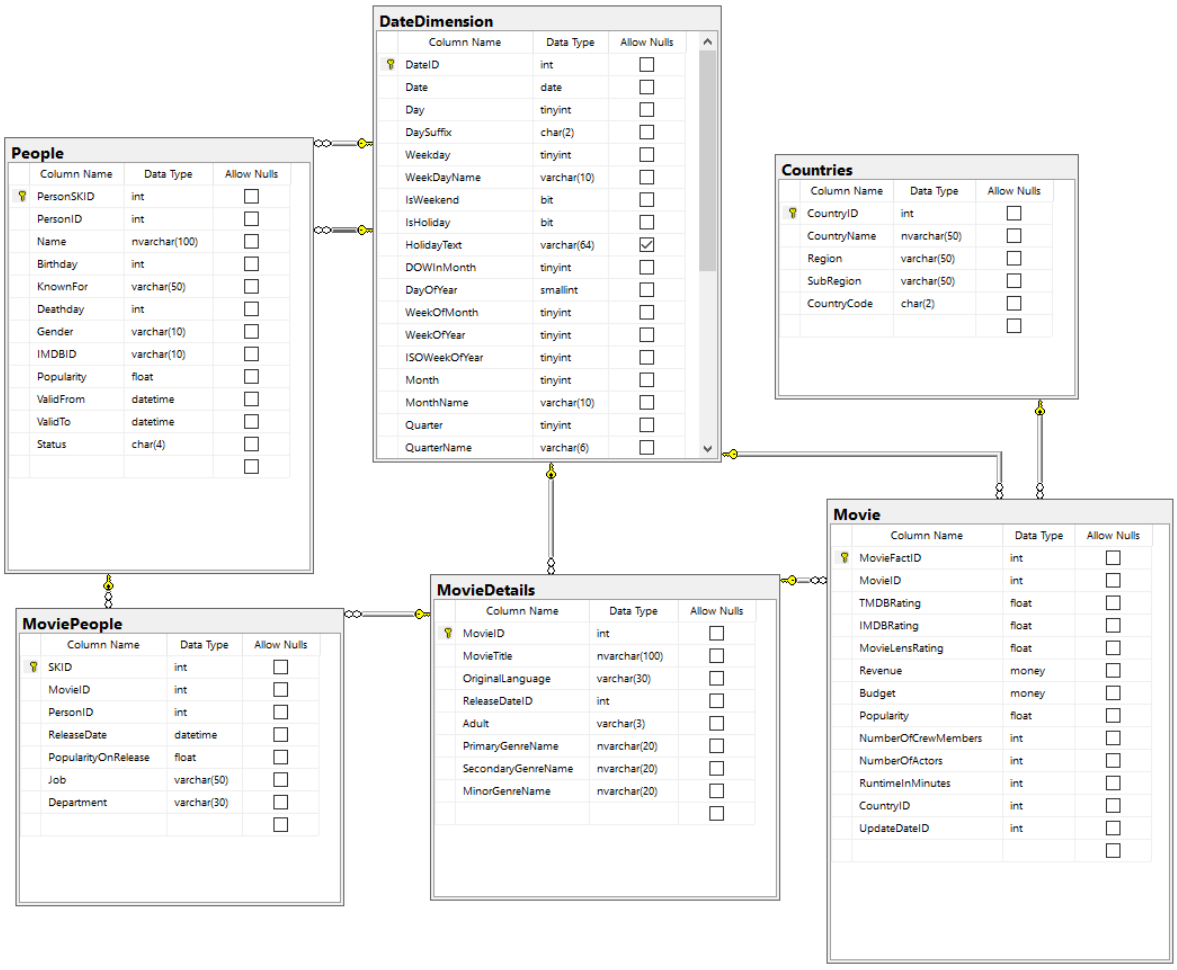
Update Movie

Proces jest bardzo podobny, zaczynamy jednak od pobrania ID filmów, które są w bazie, i następnie pobieramy dla nich nowe dane. Kolejne kroki są analogiczne.

Update People

Pobieramy ID osób, które są w bazie, następnie tworzymy pliki *csv* z nowymi informacjami na temat tych osób. Oprócz transformacji, które były użyte podczas kroku Insert, dodatkowo aktualizujemy ostatnią aktualną informację na temat danej osoby (status, validTo), nowa obserwacja dostaje validFrom o 10ms późniejszy niż ustawiony validTo i taką obserwację wstawiamy do bazy.

5 Model fizyczny hurtowni



Rysunek 2: Model hurtowni danych.

W naszym rozwiązaniu są dwie tabele faktowe, czyli tabele Movie oraz MoviePeople. Tabela Movie jest tabelą typu Periodic Snapshot, co miesiąc pobierane będą nowe fakty na temat filmów w hurtowni danych. Natomiast MoviePeople przechowuje zwykłe fakty odnośnie osób biorących udział w produkcji filmu.

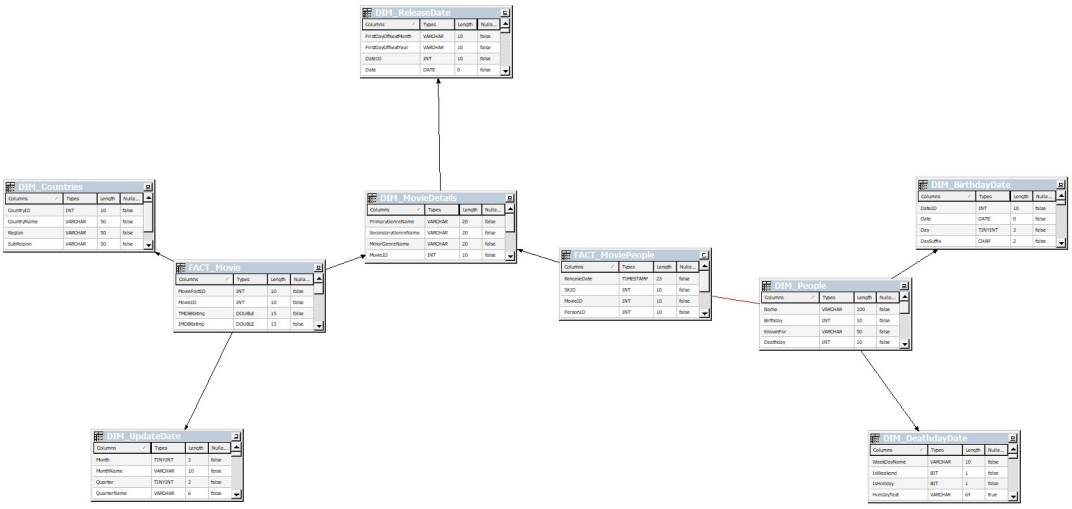
6 Opis kluczowych miar i atrybutów w modelu

Kluczowymi miarami w modelu są dane liczbowe dotyczące wyników finansowych filmów, czyli przede wszystkim budżet oraz zysk, które pozwalają stwierdzić, czy dany tytuł odniósł sukces z biznesowego punktu widzenia.

W drugiej kolejności są to miary określające popularność filmu oraz jego oceny pokazujące, jak został odebrany przez publiczność. Poza tym istotne są również czas trwania i liczba zaangażowanych w pracę nad filmem ludzi mające wpływ na koszt produkcji.
Pozostałymi atrybutami są kraj produkcji mogący mieć wpływ na zasięg filmu oraz najważniejsi ludzie biorący udział w procesie tworzenia: reżyser i główni aktorzy. Popularność i inne dane dotyczące poszczególnych osób mogą także mieć niemały wpływ na końcowy wynik danego tytułu.

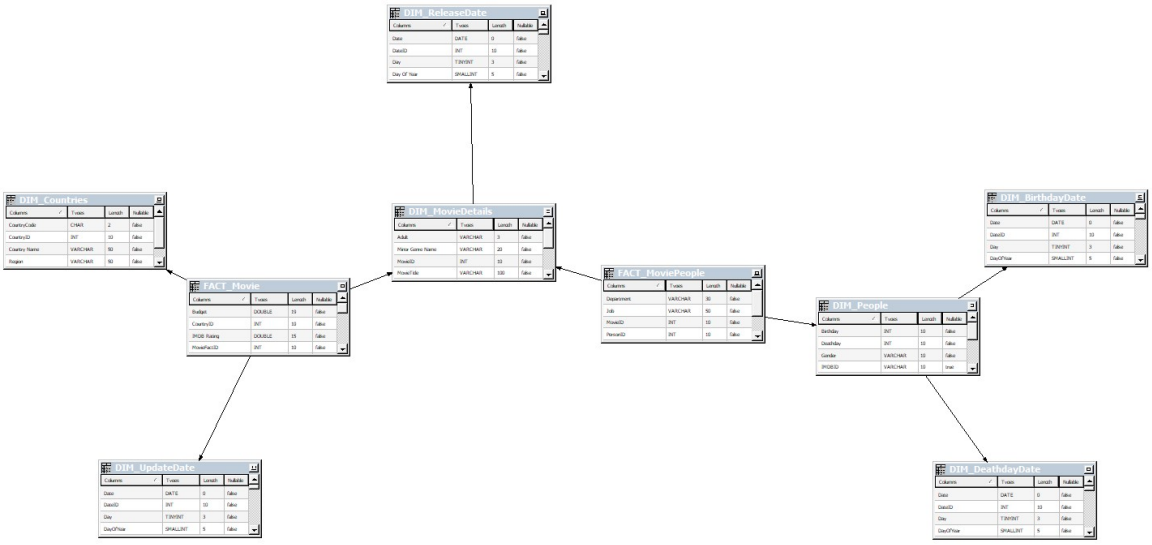
7 Warstwa raportowa

Warstwa raportowa dostaje się do hurtowni danych za pomocą portu 1433 korzystając z protokołu TCP. Hurtownia została założona lokalnie przy skorzystaniu z bazy MS SQL. Konfiguracja OBIEE przebiegała standardowo zgodnie z instrukcją z zajęć.
Poniżej mamy przedstawiony diagram fizyczny modelu. Jest on dosyć standardowy. Większość łączeń odbywa się w parach klucz-klucz obcy oprócz łączenia pomiędzy tabelą faktową MoviePeople a tabelą wymiarową People. Łączenie to jest złożonym joinem, zakładamy w nim (co wynika z ETL'a), że łączymy te dwie tabele po równości PersonId oraz po tym, że data wypuszczenia filmu musi być w przedziale pomiędzy ValidFrom a ValidTo.



Rysunek 3: Diagram fizyczny modelu.

Model biznesowy odzwierciedla jego fizyczną implementację i jest pokazany poniżej. Wszystkie połączenia są wewnętrznymi joinami o kardynalności N – 0,1.



Rysunek 4: Diagram biznesowy modelu.

W ramach tego modelu zdefiniowaliśmy nowe miarki:

- Ratio Revenue to Budget, czyli stosunek Revenue do Budget;
- Universal Rating – miarkę, która sprowadzała wszystkie sensowne miarki (w przypadku IMDB rating równy -1 oznaczał brak) do przedziału $[0, 10]$, po czym je uśredniała.

Stworzyliśmy także hierarchie:

- Genre – przedstawiała (3-stopniową) hierarchię, która przechodząc od rodzica zawierała Total Genre, Primary Genre Name oraz na ostatnim poziomie MovieID, MovieName. Początkowo miała odzwierciedlać kolejne poziomy gatunków filmu. Jednak z powodu tego, że wizualizacje nie działały zgodnie z oczekiwaniami, ponieważ kolejne poziomy gatunków mają te same nazwy, ostatecznie zawarte zostały tylko jej najważniejsze poziomy;
- Geography – przedstawiała hierarchię składającą się począwszy od najwyższego poziomu z Total Geography, Region, Subregion oraz na ostatnim poziomie CountryID, CountryName.

Oprócz tego uporządkowaliśmy nazwy atrybutów i miarek do nazw, które bardziej odpowiadały biznesowi. Wszystkie miarki ustawiliśmy domyślnie na średnią jako rodzaj agregacji. Jeśli zaistniała potrzeba zastosowania innego sposobu agregacji, zmienialiśmy ją już na poziomie wizualizacji danych (rzadki przypadek). Oprócz tego stworzyliśmy kilka wyliczonych miarek:

- top 3 języków z najbardziej popularnymi aktorami;
- top 3 języków z najmniej popularnymi aktorami;
- zliczanie filmów - zmiana sposobu agregacji ID;

oraz wyliczonych obiektów:

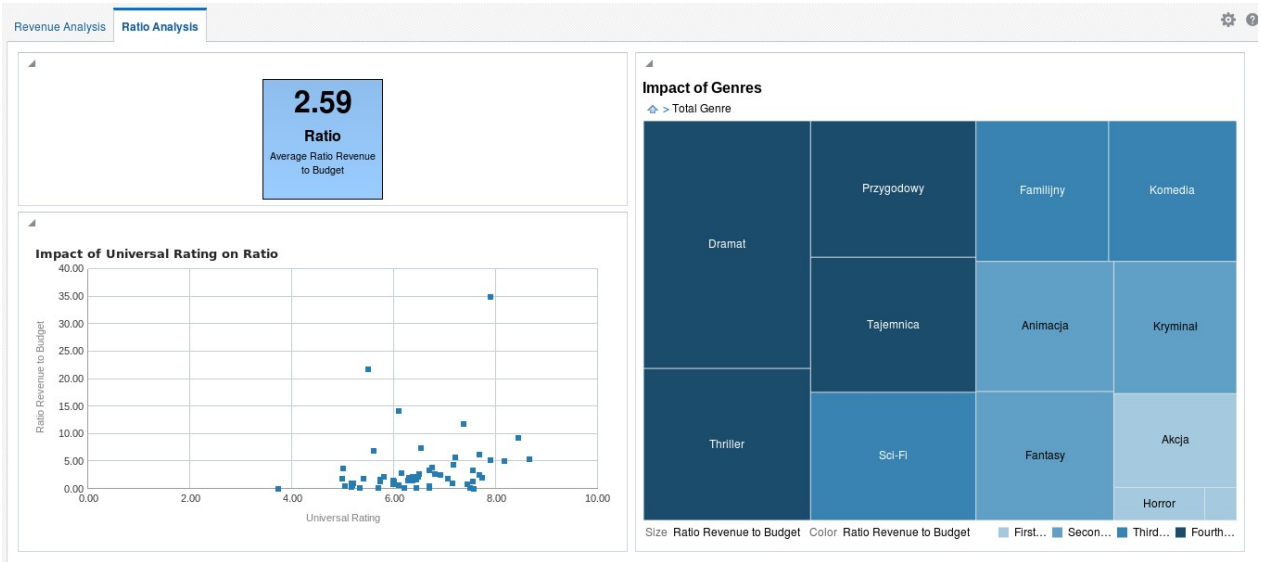
- link – zwraca kolumna z linkiem do IMDB;
- stworzenie koszyków z długościami filmów.

8 Opis raportów dla użytkowników

Generowane raporty będą dotyczyły następujących zagadnień:

- wpływ wymienionych czynników na dochód z filmu i jego stosunek dochodu do poniesionych kosztów:
  - Gatunek filmu;
  - Universal Rating;

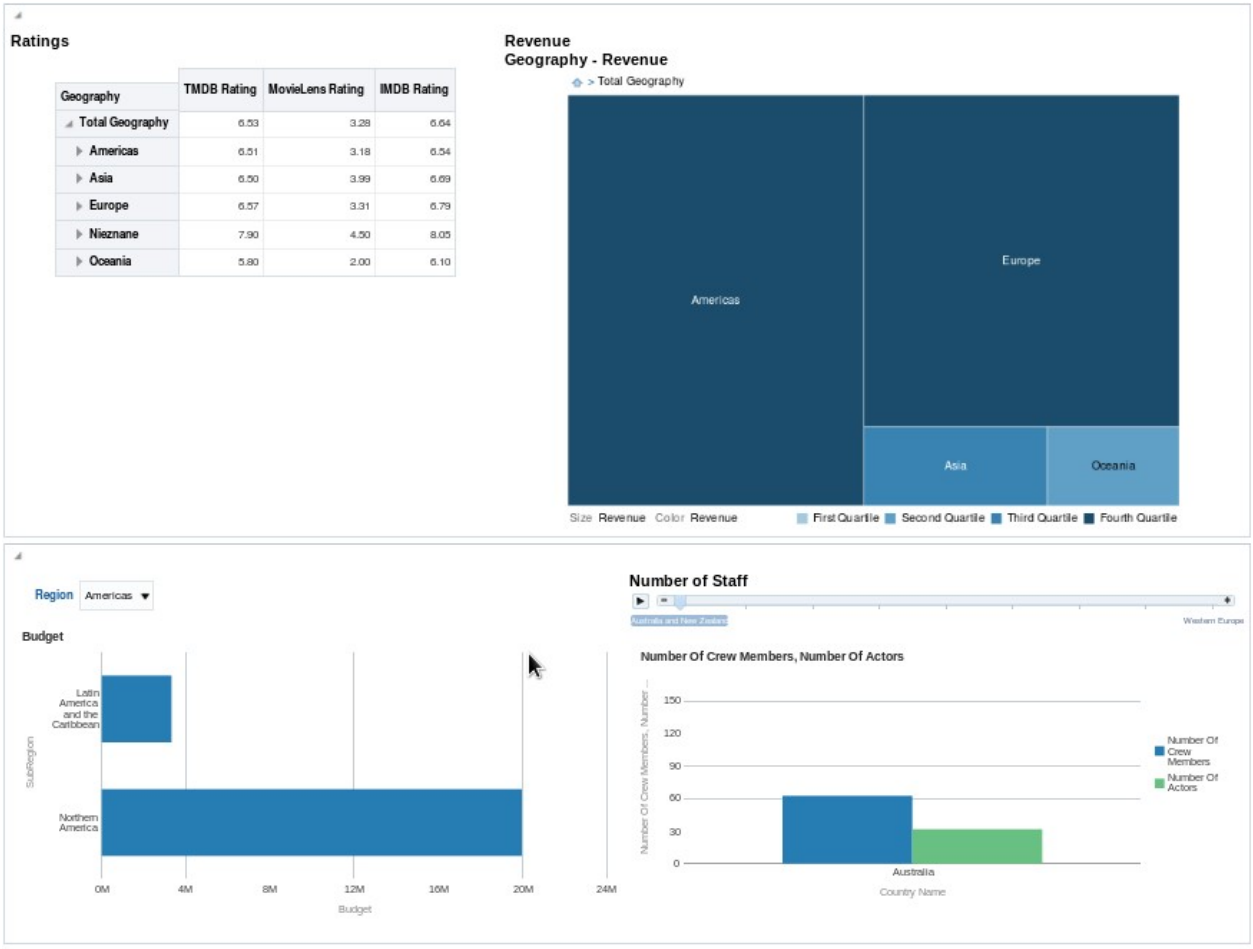
Rysunek 5: Pierwsza strona raportu przedstawiająca analizę dochodu z filmu.



Rysunek 6: Druga strona raportu przedstawiająca analizę współczynnika dochodu do budżetu.

- wpływ lokalizacji na:
  - Ocenę;
  - Dochód;

- Budżet;
- Liczbę ludzi zaangażowanych w produkcję;

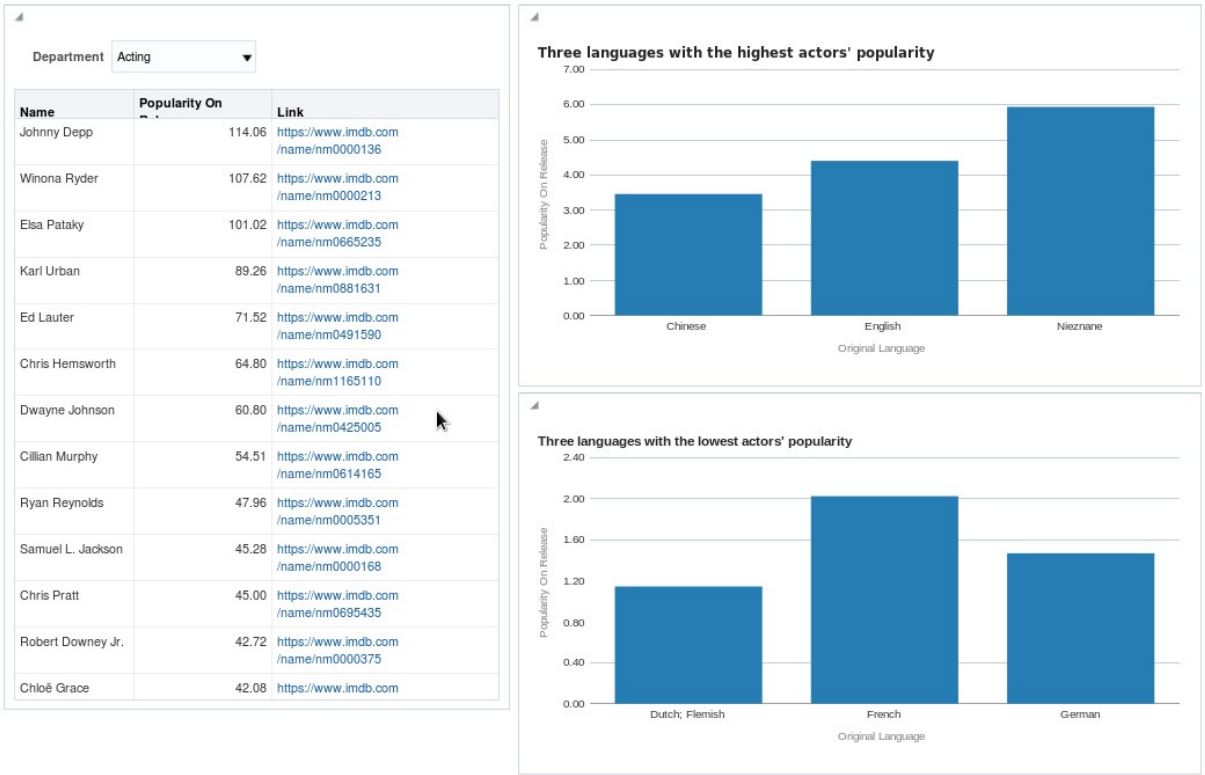


Rysunek 7: Raport przedstawiający analizę wpływu lokalizacji na inne parametry filmów.

- predykcja wielkości zapotrzebowania na ludzi w zależności od:
  - Budżetu;
  - Gatunku;
  - Daty wydania;

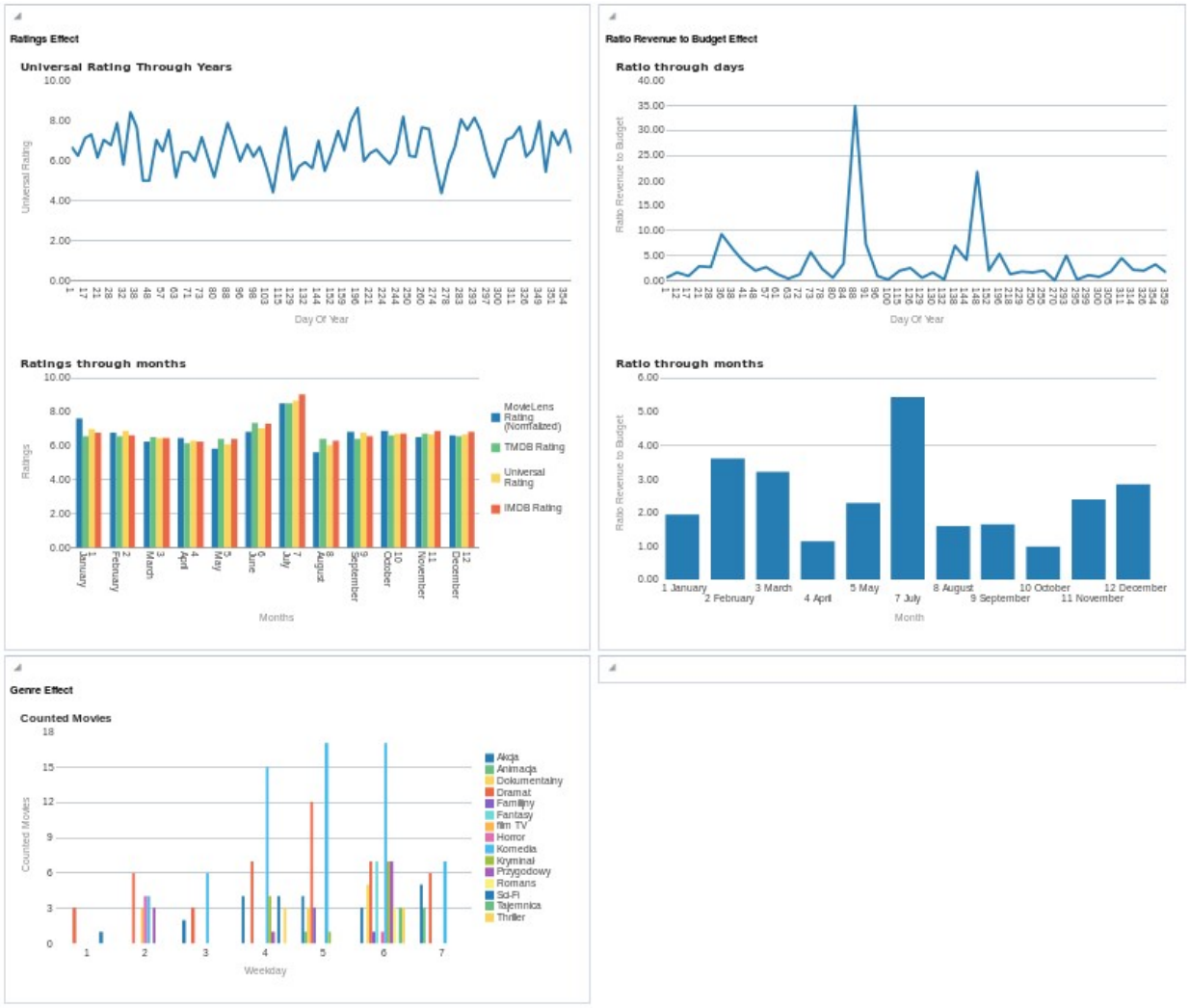
Rysunek 8: Raport przedstawiający analizę zapotrzebowania na ekipę i aktorów.

- analiza popularności głównej obsady filmu:



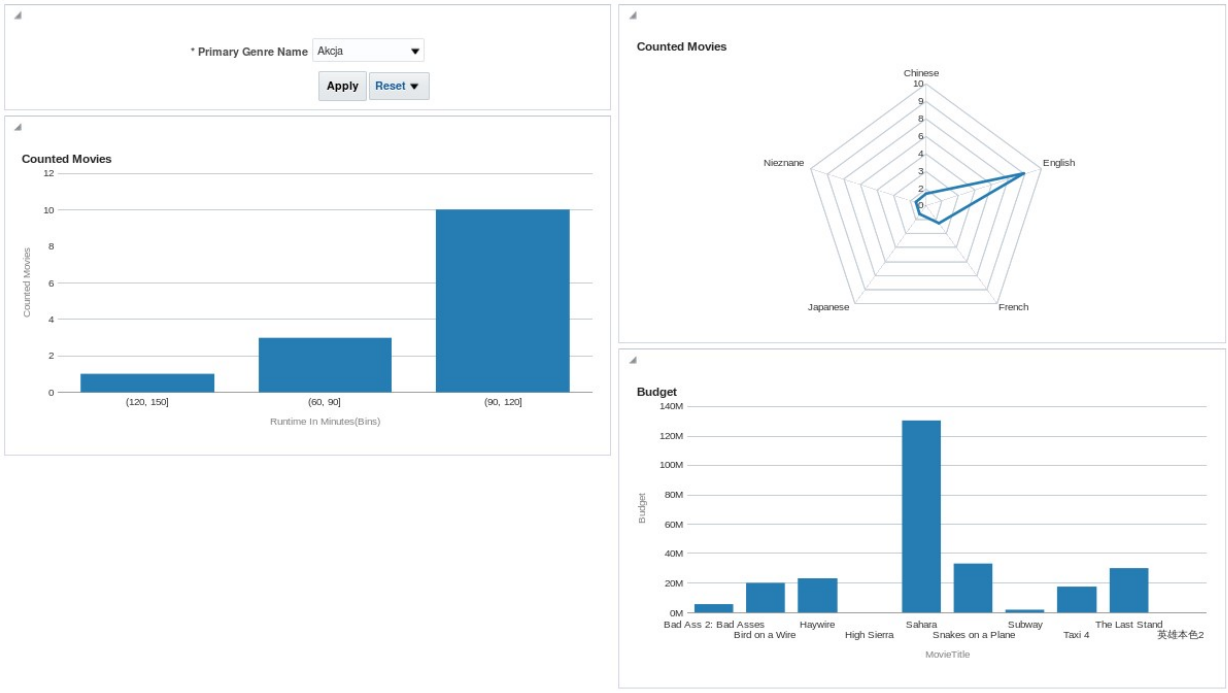
Rysunek 9: Raport przedstawiający analizę popularności aktorów grających w filmie.

- wpływ daty premiery na:
  - Ocenę;
  - Stosunek dochodu do kosztów;
  - Gatunek;



Rysunek 10: Raport przedstawiający analizę dni wydania filmów.

- wpływ gatunku filmu na:
  - Liczbę nakręconych filmów;
  - Język;
  - Budżet.



Rysunek 11: Raport przedstawiający analizę gatunku filmowego.

9 Podsumowanie rezultatów projektu

Po zakończeniu prac nad projektem zostajemy z w pełni działającym systemem. Zgodnie z początkowymi założeniami pozwala on na generowanie sześciu różnych rodzajów raportów dotyczących różnych aspektów filmów z hurtowni danych. Sama hurtownia jest przygotowana do regularnego dodawania i aktualizacji zawartych w niej danych, co gwarantuje aktualność informacji przedstawianych w tworzonych raportach. Powstały system jest przystępny i intuicyjny, dzięki czemu może być używany przez różne osoby związane z przemysłem filmowym i pomagać im zarówno w codziennej pracy, jak i w podejmowaniu istotnych decyzji dotyczących wysokich kwot pieniędzy.

10 Opis podziału pracy w zespole

Adrian Kamiński

- Przygotowanie kodu wołającego API.
- ETL – implementacja całości rozwiązania.
- Pomoc przy raportach (obserwator).
- Test/dokumentacja dotycząca ETL.

Konstanty Kraszewski

- Tworzenie modelu hurtowni danych.
- Pomoc przy raportach (obserwator).
- Praca nad dokumentacją.

Piotr Marciniak

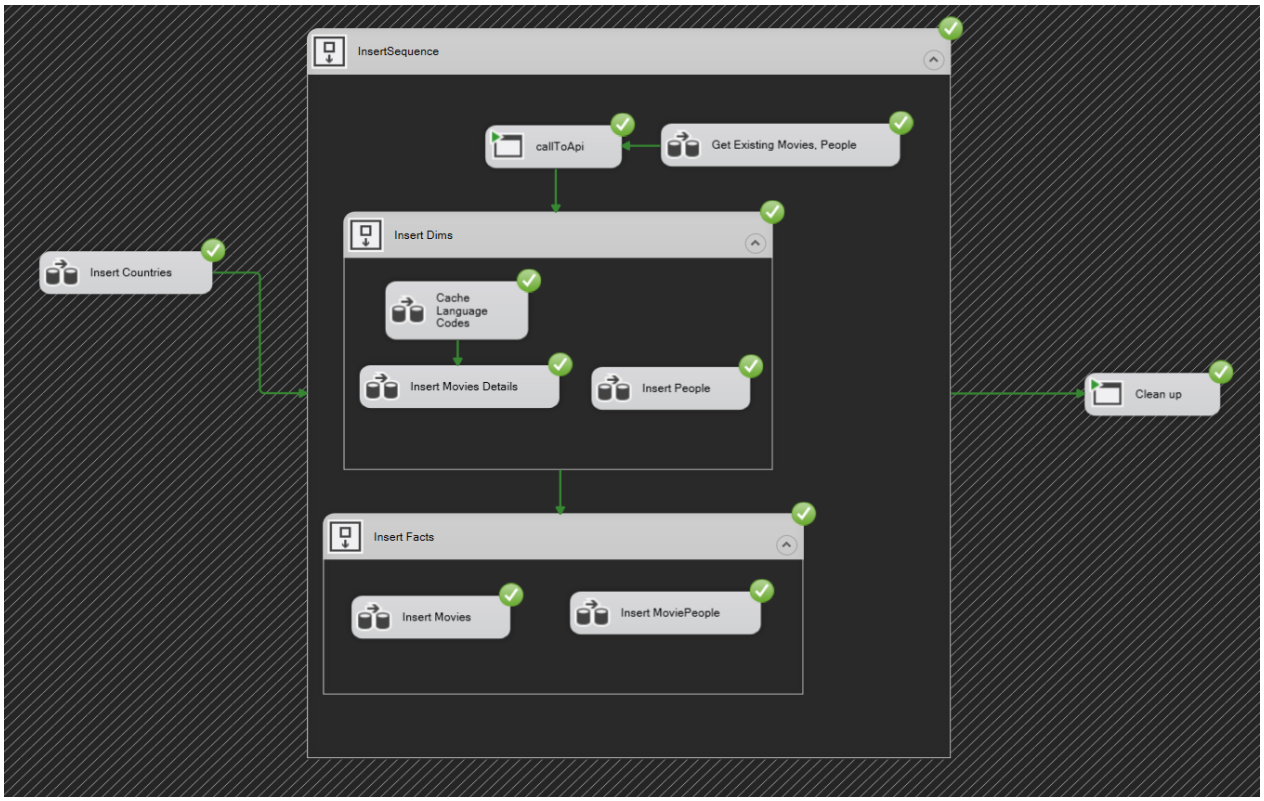
- Przygotowanie raportów.
- Tworzenie dokumentacji.
- Scrapowanie imdb.
- Testowanie warstwy raportowej.

11 Testy funkcjonalne

Inicjalizacja hurtowni

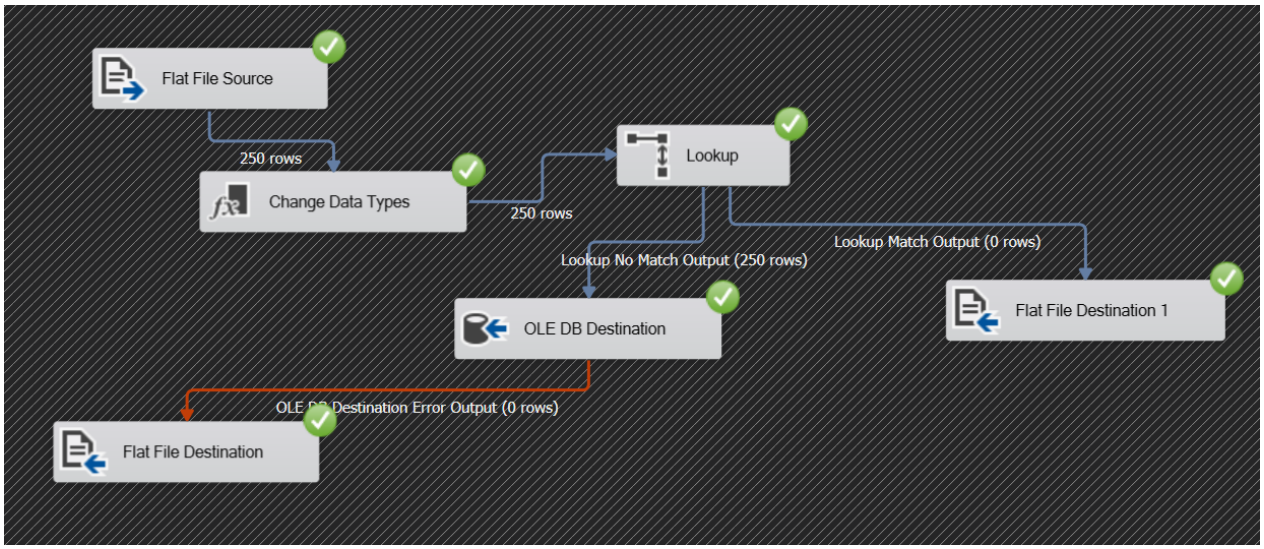
	Count_Countries	Count_DateDimension	Count_Movie	Count_MovieDetails	Count_MoviePeople	Count_People
1	0	105923	0	0	0	0

Rysunek 12: Stan hurtowni bezpośrednio po jej utworzeniu.

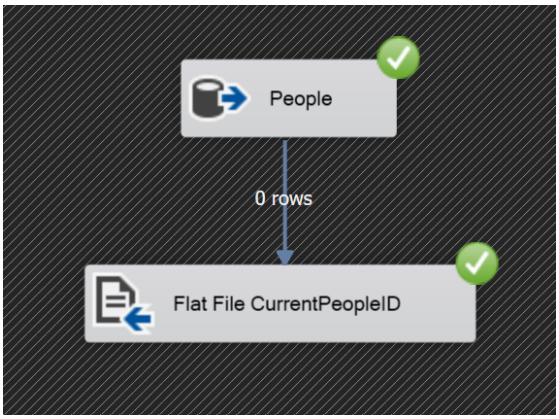


Rysunek 13: Pierwsze dodanie danych do hurtowni.

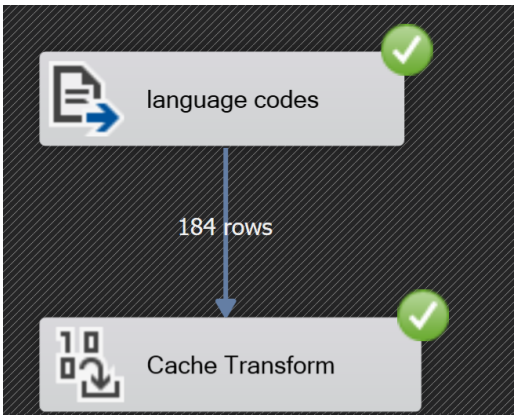
Potwierdzenie poprawności wykonania każdego z etapów (zgadzają się liczby wierszy przechodzące dalej).



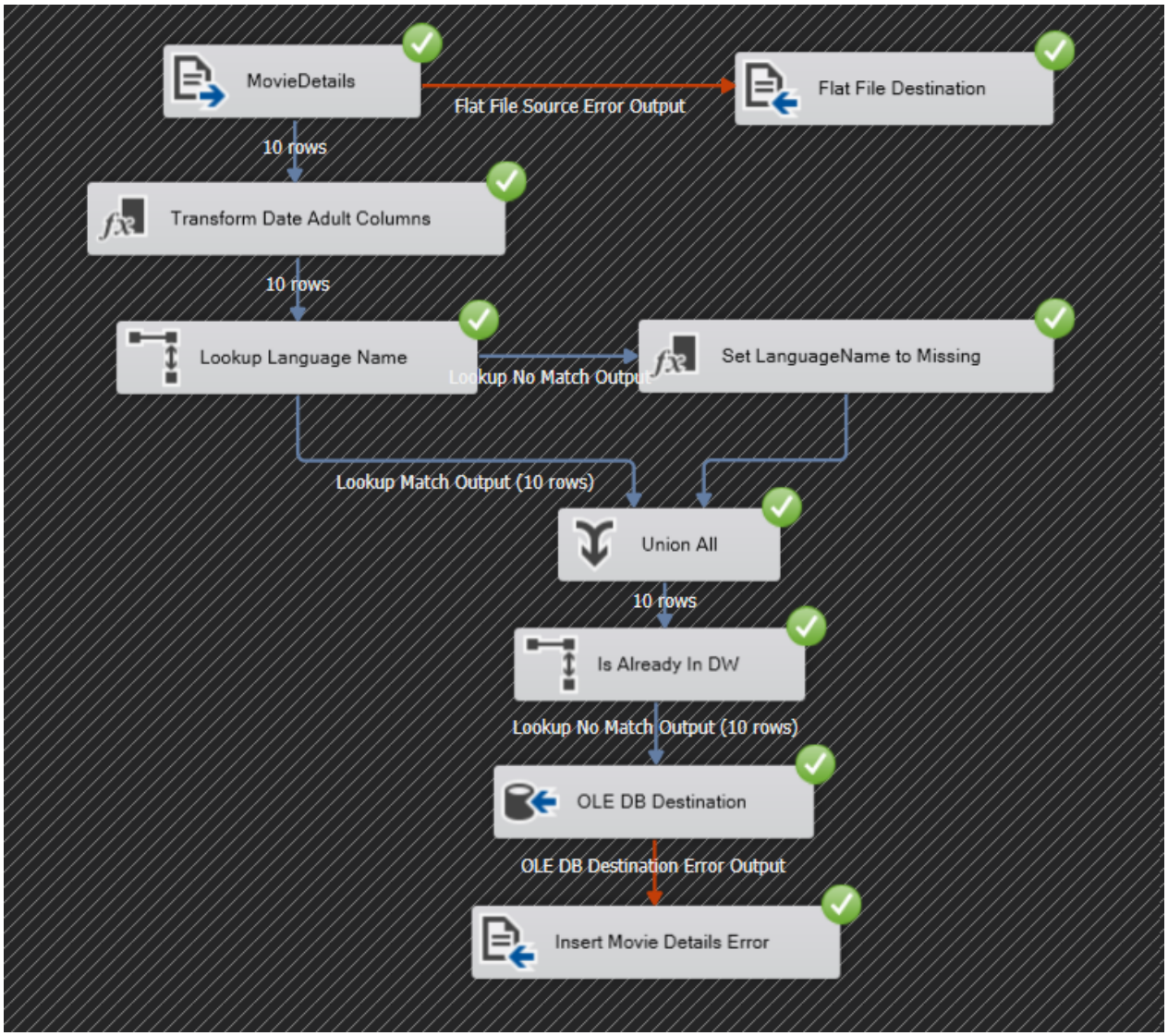
Rysunek 14: Dodawanie krajów (z *csv*).



(a) Zapisanie istniejących ID osób do pliku *csv*.

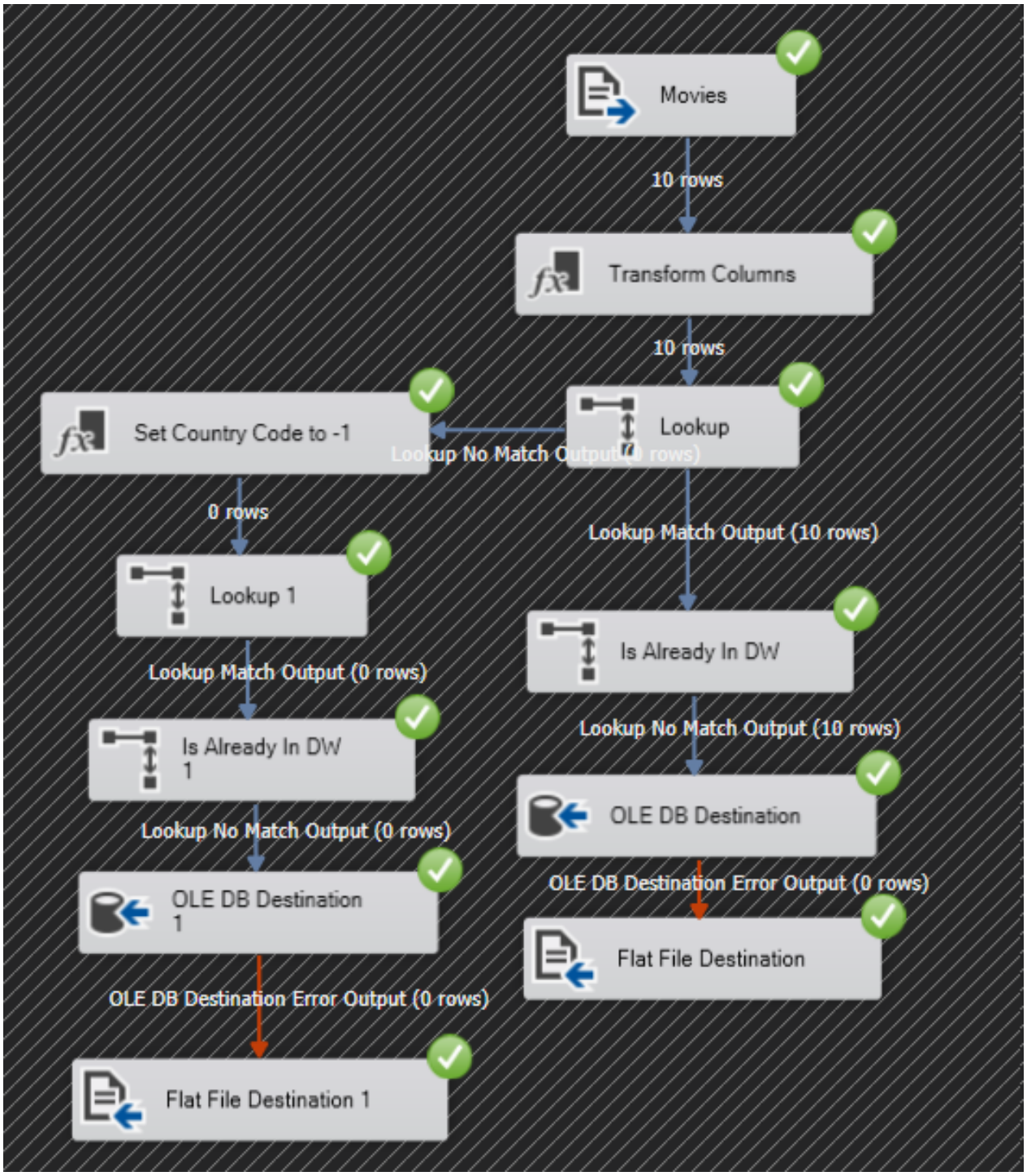


(b) Wczytanie pliku *csv* do cache'a w celu lookupu.

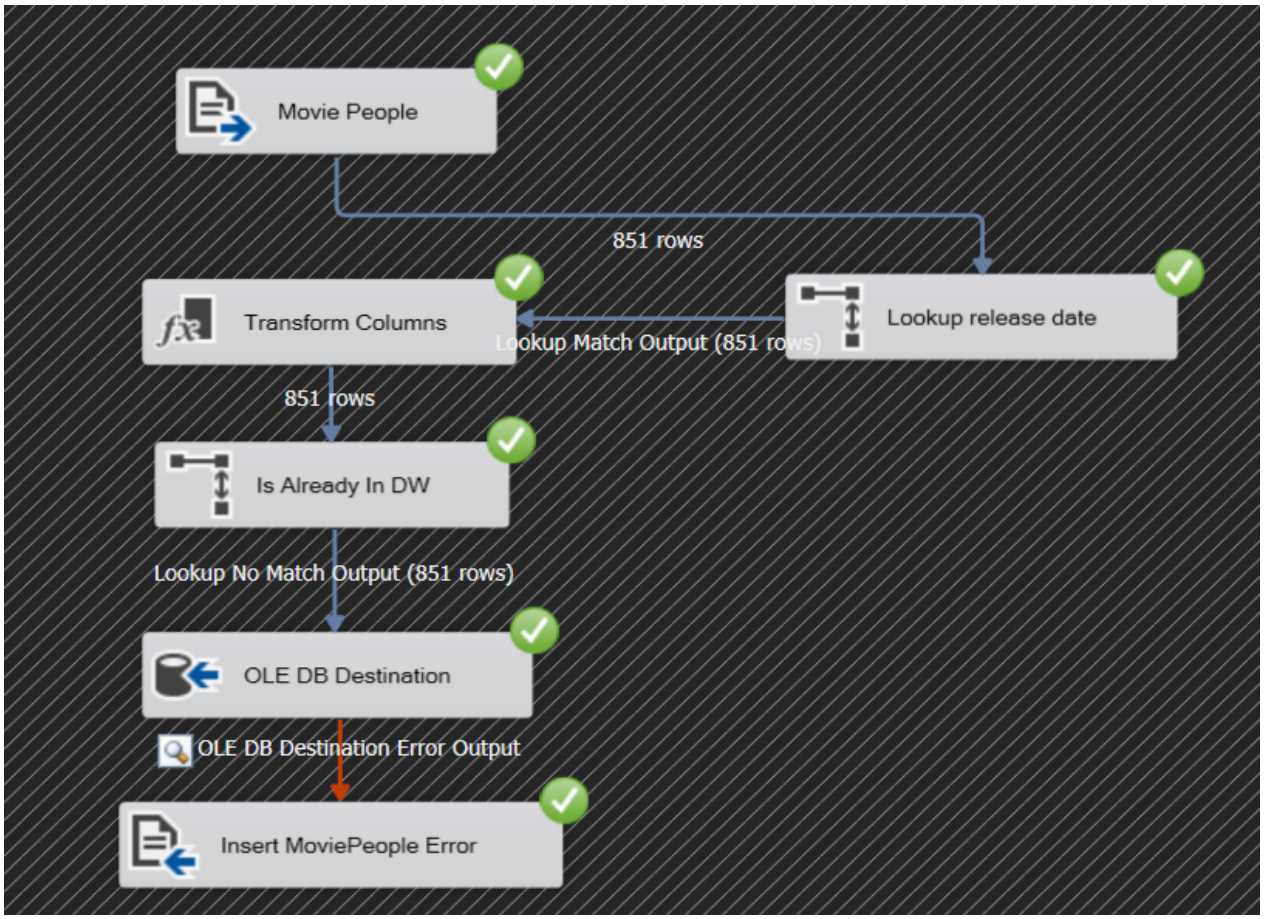


Rysunek 15: Wstawianie danych do tabeli MovieDetails.

Rysunek 16: Wstawianie danych do tabeli People.



Rysunek 17: Wstawianie danych do tabeli Movie.



Rysunek 18: Wstawianie danych do tabeli MoviePeople.

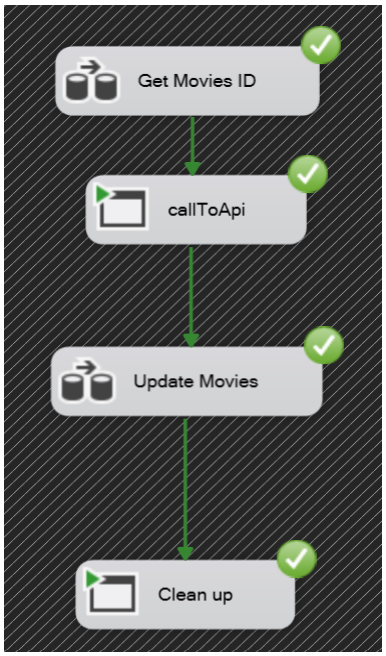
	Count_Countries	Count_DateDimension	Count_Movie	Count_MovieDetails	Count_MoviePeople	Count_People
1	250	105923	10	10	851	851

Rysunek 19: Stan hurtowni bezpośrednio po wstawieniu pierwszych filmów.

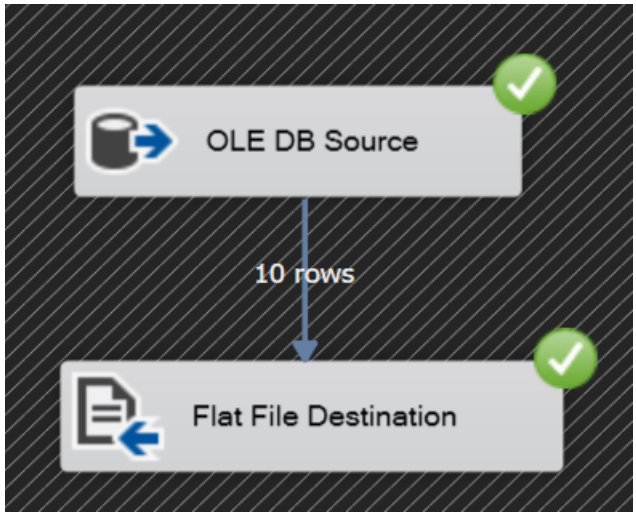
Widzimy, że liczba wierszy w hurtowni zgadza się z liczbą wierszy, które ETL wczytywał z plików *csv*, i które deklarował, że wstawia do naszej bazy.

Aktualizacja hurtowni

Aktualizując tabelę Movie spodziewamy się, że ilość wierszy w tej tabeli podwoi się przy pozostałych tabelach pozostałych bez zmian.



(a) Globalne spojrzenie na update Movie.



(b) Zgarnianie ID filmów które są w bazie.

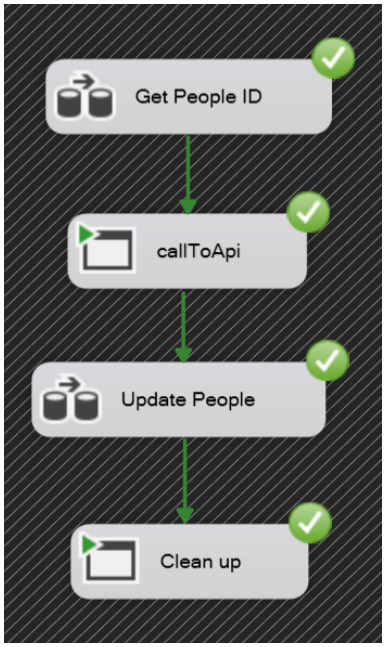
Rysunek 20: Zadanie aktualizujące tabelę Movie.

	Count_Countries	Count_DateDimension	Count_Movie	Count_MovieDetails	Count_MoviePeople	Count_People
1	250	105923	20	10	851	851

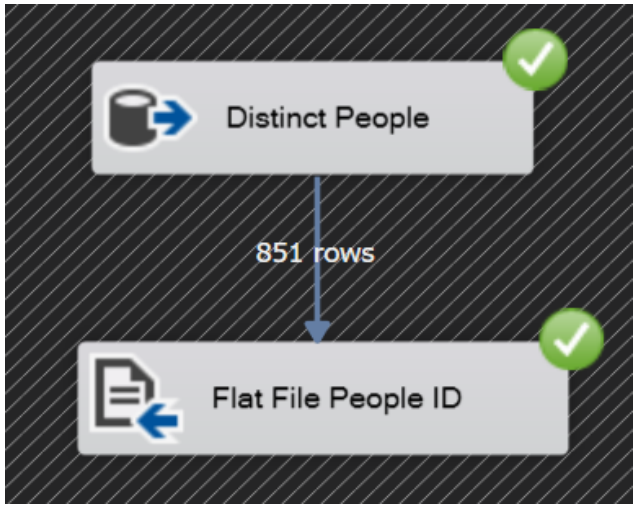
Rysunek 21: Stan hurtowni danych po aktualizacji.

Zgodnie z tym, czego się spodziewaliśmy, liczba obserwacji w tabeli Movie podwoiła się.

Przy aktualizacji tabeli People oczekujemy podwojenia się w niej liczby wierszy.



(a) Globalne spojrzenie na aktualizację tabeli People.



(b) Pobieranie ID osób które są w bazie.

Rysunek 22: Zadanie aktualizujące tabelę People.

	Count_Countries	Count_DateDimension	Count_Movie	Count_MovieDetails	Count_MoviePeople	Count_People
1	250	105923	20	10	851	1702

Rysunek 23: Stan hurtowni danych po aktualizacji.

Wszystko się zgadza, liczba obserwacji w tabeli People podwoiła się.

Poprawność wykonanych transformacji

Zajrzyjmy do tabel i sprawdźmy, czy wszystkie tabele zawierają to, czego byśmy się spodziewali.

	MovieID	MovieTitle	OriginalLanguage	ReleaseDateID	Adult	PrimaryGenreName	SecondaryGenreName	MinorGenreName
1	1125	Dreamgirls	English	20061225	Nie	Dramat	Muzyczny	Nieznane
2	1381	The Fountain	English	20061122	Nie	Dramat	Przygodowy	Sci-Fi
3	2268	The Golden Compass	English	20071204	Nie	Przygodowy	Fantasy	Nieznane
4	10003	The Saint	English	19970403	Nie	Thriller	Akcja	Romans
5	11484	Rollerball	English	19750625	Nie	Akcja	Sci-Fi	Nieznane
6	19426	Le notti di Cabiria	Italian	19571003	Nie	Dramat	Nieznane	Nieznane
7	27932	Airport 1975	English	19741018	Nie	Akcja	Przygodowy	Dramat
8	37292	Broken Arrow	English	19500801	Nie	Western	Nieznane	Nieznane
9	40454	Follow Me, Boys!	English	19661201	Nie	Dramat	Familijny	Nieznane
10	72387	Safe	English	20120416	Nie	Akcja	Kryminał	Thriller

Rysunek 24: Zawartość tabeli MovieDetails.

Widzimy, że lookup do pomocniczej tabeli z językami (kod, nazwa, plik *csv*, który był w pamięci podręcznej) zadział, data również posiada odpowiedni format, a kolumna Adult przyjmuje informacyjne wartości (zamiast 0, 1).

	PersonSKID	PersonID	Name	Birthday	KnownFor	Deathday	Gender	IMDBID	Popularity	ValidFrom	ValidTo	Status
1	1	4091	Fred MacMurray	19080830	Acting	19911105	Nieznana	nm0534045	3,185	1777-01-01 00:00:00.000	2022-06-12 10:29:53.367	Hist
2	217	2151819	Caridad Angus	89991231	Crew	89991231	Mezczyzna	Nieznane_	0,6	1777-01-01 00:00:00.000	2022-06-12 10:29:53.370	Hist
3	897	4091	Fred MacMurray	19080830	Acting	19911105	Nieznana	nm0534045	3,756	2022-06-12 10:29:53.377	8999-12-31 00:00:00.000	Curr
4	1632	2151819	Caridad Angus	89991231	Crew	89991231	Mezczyzna	Nieznane_	0,6	2022-06-12 10:29:53.380	8999-12-31 00:00:00.000	Curr

Rysunek 25: Zawartość tabeli People.

Widać, że daty przyjmują odpowiedni format, Gender jest odpowiednio przekodowana, tabele techniczne zawierają odpowiednie wartości oraz, że prawidłowo zaimplementowaliśmy SCD 2.

	MovieFactID	MovieID	TMDBRating	IMDBRating	MovieLensRating	Revenue	Budget	Popularity	NumberOfCrewMembers	NumberOfActors	RuntimeInMinutes	CountryID	UpdateDateID
1	1	40454	6,4	8,3	0,5	394436586,00	30000000,00	2,142	8	19	131	236	20220611
2	2	27932	5,6	8,3	3,5	47000000,00	3000000,00	8,365	13	72	107	236	20220611
3	3	10003	6,1	8,3	3,21666666666667	118063304,00	68000000,00	11,559	20	29	116	236	20220611
4	4	2268	6	8,3	3,11111111111111	372234864,00	180000000,00	23,411	161	57	113	235	20220611
5	5	1125	6,8	8,3	3,1	154937680,00	70000000,00	11,974	89	33	134	236	20220611
6	6	11484	6,2	8,3	3,5	30000000,00	30000000,00	8,162	19	19	125	236	20220611
7	7	37292	6,7	8,3	3	394436586,00	30000000,00	12,189	25	32	93	236	20220611
8	8	1381	6,9	8,3	3,5	15304890,00	35000000,00	9,516	84	18	96	41	20220611
9	9	72387	6,5	8,3	3,5	40346186,00	30000000,00	29,512	89	49	94	236	20220611
10	10	19426	8,1	8,3	4,28571428571429	752045,00	30000000,00	10,646	27	24	110	111	20220611
11	11	1125	6,8	8,3	3,1	154937680,00	70000000,00	11,974	89	33	134	236	20220612
12	12	1381	6,9	8,3	3,5	15304890,00	35000000,00	9,516	84	18	96	41	20220612
13	13	2268	6	8,3	3,11111111111111	372234864,00	180000000,00	23,411	161	57	113	235	20220612
14	14	10003	6,1	8,3	3,21666666666667	118063304,00	68000000,00	11,559	20	29	116	236	20220612
15	15	11484	6,2	8,3	3,5	30000000,00	30000000,00	8,162	19	19	125	236	20220612
16	16	19426	8,1	8,3	4,28571428571429	752045,00	30000000,00	10,646	27	24	110	111	20220612
17	17	27932	5,6	8,3	3,5	47000000,00	3000000,00	7,755	13	72	107	236	20220612
18	18	37292	6,7	8,3	3	394436586,00	30000000,00	12,189	25	32	93	236	20220612
19	19	40454	6,4	8,3	0,5	394436586,00	30000000,00	2,142	8	19	131	236	20220612
20	20	72387	6,5	8,3	3,5	40346186,00	30000000,00	29,512	89	49	94	236	20220612

Rysunek 26: Zawartość tabeli Movie.

Daty są w odpowiednim formacie, CountryID zostało dopasowane na podstawie lookupu do tabeli Countries.

	SKID	MovieID	PersonID	ReleaseDate	PopularityOnRelease	Job	Department
1	1	40454	4091	1966-12-01 00:00:00.000	3,185	Actor	Acting
2	217	2268	2151819	2007-12-04 00:00:00.000	0,6	Actor	Acting

Rysunek 27: Zawartość tabeli MoviePeople.

Widzimy, że wartości, których się spodziewaliśmy, znajdują się w tabeli.

Poprawność ETL po kilku iteracjach

	Count_Countries	Count_DateDimension	Count_Movie	Count_MovieDetails	Count_MoviePeople	Count_People
1	250	105923	199	99	6470	12412

Rysunek 28: Stan hurtowni danych.

	MovieID	MovieTitle	OriginalLanguage	ReleaseDateID	Adult	PrimaryGenreName	SecondaryGenreName	MinorGenreName
1	87	Indiana Jones and the Temple of Doom	English	19840523	Nie	Przygodowy	Akcja	Nieznane
2	155	The Dark Knight	English	20080714	Nie	Dramat	Akcja	Kryminał
3	292	Dave Chappelle's Block Party	English	20050912	Nie	Komedia	Dokumentalny	Muzyczny
4	320	Insomnia	English	20020524	Nie	Kryminał	Tajemnica	Thriller
5	326	Snakes on a Plane	English	20060817	Nie	Akcja	Kryminał	Thriller
6	583	Life of Brian	English	19790817	Nie	Komedia	Nieznane	Nieznane
7	598	Cidade de Deus	Portuguese	20020205	Nie	Dramat	Kryminał	Nieznane
8	978	Seven Years in Tibet	English	19970912	Nie	Przygodowy	Dramat	Historyczny

	PersonSKID	PersonID	Name	Birthday	KnownFor	Deathday	Gender	IMDBID	Popularity	ValidFrom	ValidTo	Status
1	2000	4091	Fred MacMurray	19080830	Acting	19911105	Nieznana	nm0534045	2,287	1777-01-01 00:00:00.000	2022-06-05 15:06:01.917	Hist
2	3640	4091	Fred MacMurray	19080830	Acting	19911105	Nieznana	nm0534045	2,287	2022-06-05 15:06:01.927	2022-06-09 19:50:10.807	Hist
3	6611	4091	Fred MacMurray	19080830	Acting	19911105	Nieznana	nm0534045	4,276	2022-06-09 19:50:10.817	8999-12-31 00:00:00.000	Curr
4	9028	2160	Carlo Varini	19460816	Camera	20140518	Nieznana	nm0002299	0,84	1777-01-01 00:00:00.000	8999-12-31 00:00:00.000	Curr

	MovieFactID	MovieID	TMDBRating	IMDBRating	MovieLensRating	Revenue	Budget	Popularity	NumberOfCrewMembers	NumberOfActors	RuntimeInMinutes	CountryID	UpdateDateID
1	1	62395	7	7,1	3,5	-1,00	-1,00	26,24	7	10	91	111	20220605
2	2	13649	6,5	5,1	2	7000000,00	7000000,00	25,358	26	75	107	236	20220605
3	3	73884	6,7	7,3	5	1827125,00	-1,00	2,082	4	8	75	101	20220605
4	4	19955	6,4	6,7	4	30178449...	-1,00	8,457	7	16	90	77	20220605
5	5	77801	7,5	7,5	3,5	-1,00	750000,00	6,008	14	20	85	236	20220605
6	6	34560	6,5	6,1	3,75	-1,00	-1,00	9,835	8	26	82	236	20220605

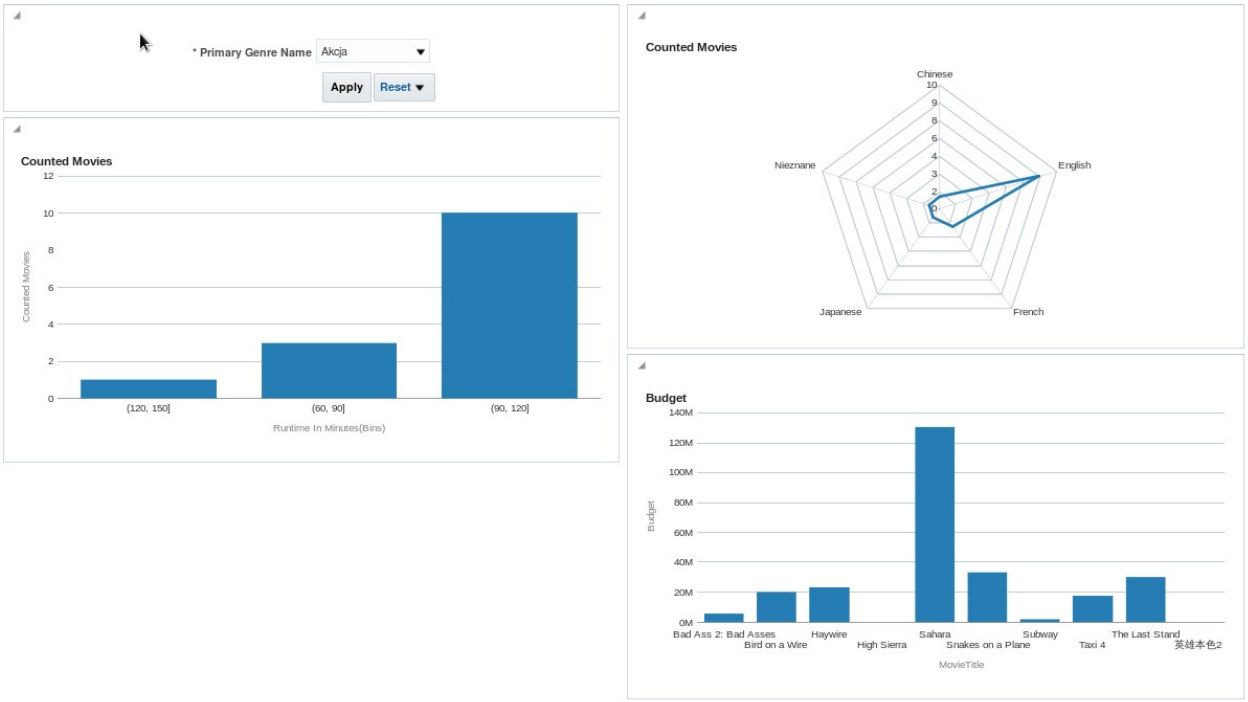
	SKID	MovieID	PersonID	PopularityOnRelease	Job	Department	ReleaseDate
1	2000	15944	4091	2,287	Actor	Acting	1959-03-19 00:00:00.000
2	13583	10656	2160	0,84	Director of Photography	Camera	1985-04-10 00:00:00.000

Rysunek 29: Zawartość bazy danych.

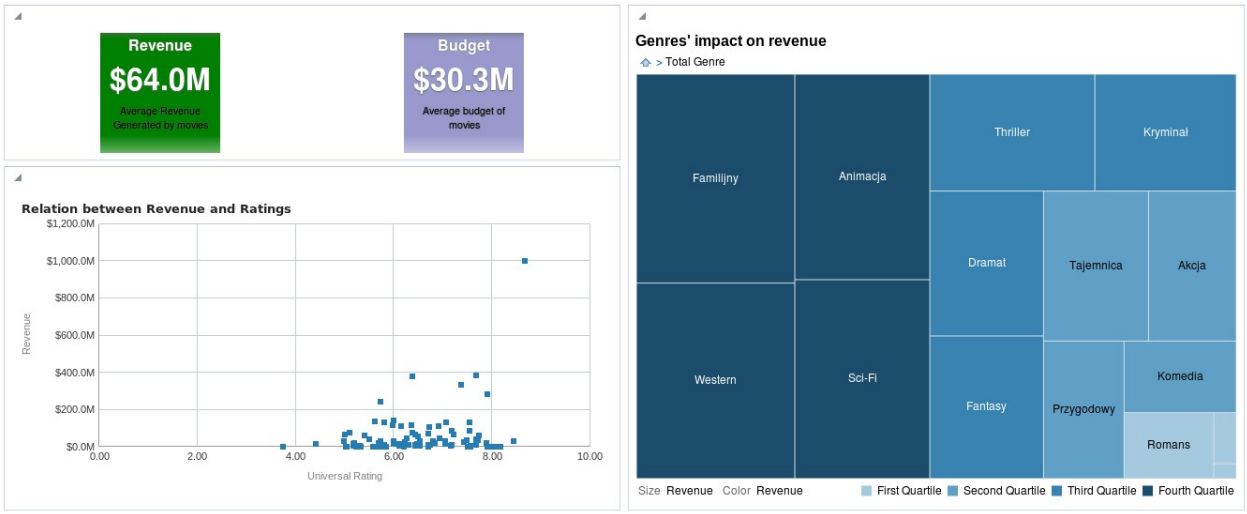
Jak poprzednio, widzimy, że nasza hurtownia jest wciąż spójna po kilku iteracjach dodawania filmów i aktualizacji naszej bazy danych.

Raporty po aktualizacji hurtowni

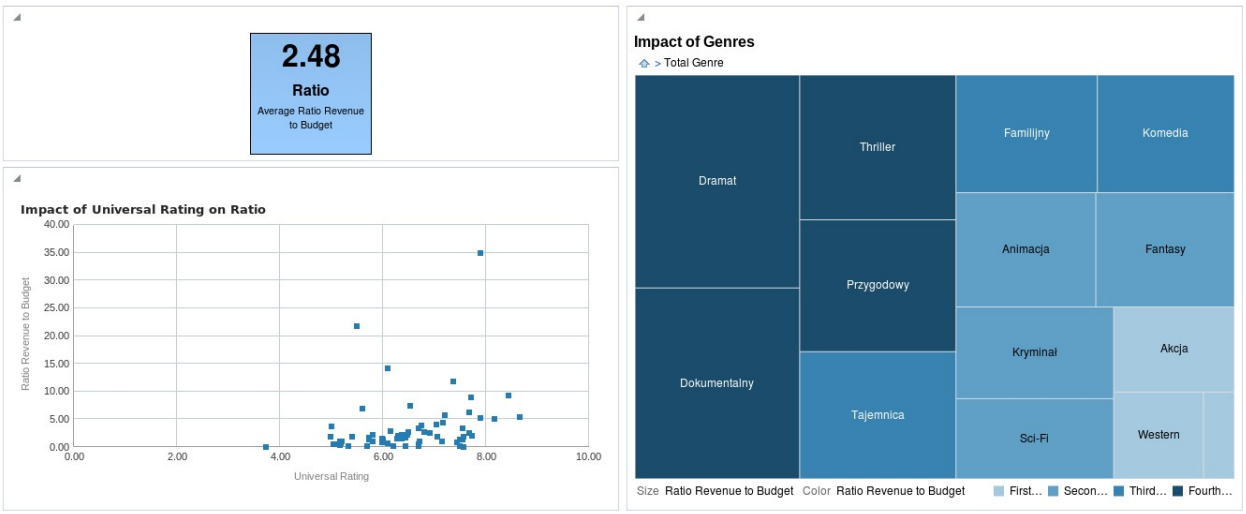
Poniżej zaprezentowane są raporty po aktualizacji hurtowni danych. Różnią się one od tych przedstawionych wcześniej w sekcji 8, gdyż do hurtowni zostały dodane informacje o nowych filmach (oprócz raportu z gatunkiem, gdyż nie został dodany film akcji).



Rysunek 30: Zaktualizowany raport przedstawiający analizę gatunku filmowego.



Rysunek 31: Zaktualizowana pierwsza strona raportu przedstawiająca analizę dochodu z filmu.



Rysunek 32: Zaktualizowana druga strona raportu przedstawiająca analizę współczynnika dochodu do budżetu.

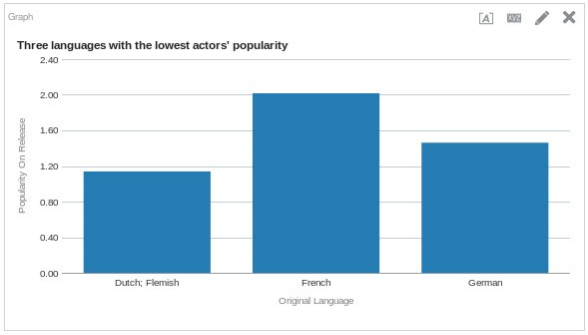
Rysunek 33: Zaktualizowany raport przedstawiający analizę zapotrzebowania na ekipę i aktorów.

Poprawność wyników na raportach

Poniżej mamy testy sprawdzające poprawność danych na raportach. Zostały one wykonane w ten sposób, że napisaliśmy zapytanie SQL o to, co chcielibyśmy przedstawić. Następnie sprawdziliśmy na wizualizacjach, czy dane się zgadzają. Dane uzyskane z hurtowni są naszym oczekiwanym wynikiem, natomiast zrzuty ekranu potwierdzają ich poprawność.

	OriginalLanguage	Popularity
1	Dutch; Flemish	1,14178125
2	German	1,46359259259259
3	French	2,01681382978723

Dane uzyskane bezpośrednio z hurtowni.



Dane prezentowane na raporcie.

	Region	avg_TMDB	avg_MovieLensRating	avg_IMDB
1	Americas	6,51048387096774	3,2113085195586	6,53467741935484
2	Asia	6,53529411764706	3,98403361344538	6,74117647058824
3	Europe	6,57619047619048	3,3365937098195	6,78571428571429
4	Nieznane	7,9	4,5	8,05
5	Oceania	5,8	2	6,1

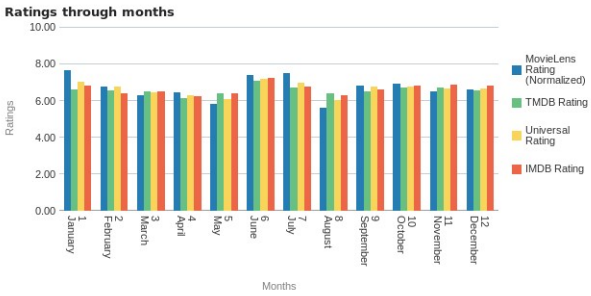
Dane uzyskane bezpośrednio z hurtowni.

Geography	TMDB Rating	MovieLens Rating	IMDB Rating
Total Geography	6.54	3.31	6.64
Americas	6.51	3.21	6.53
Asia	6.54	3.98	6.74
Europe	6.58	3.34	6.79
Nieznane	7.90	4.50	8.05
Oceania	5.80	2.00	6.10

Dane prezentowane na raporcie.

	Month	avg_TMDB	avg_IMDB	avg_MovieLensRating
1	1	6,65555555555556	6,78888888888889	7,64252645502645
2	2	6,80666666666667	6,78666666666667	6,77032167832168
3	3	6,544	6,556	6,29755961938745
4	4	6,10526315789474	6,29473684210526	6,46992481203008
5	5	6,20952380952381	6,26666666666667	5,80536715324851
6	6	7,1	6,98571428571429	7,40595238095238
7	7	6,7	6,76666666666667	7,49217002237136
8	8	6,37142857142857	6,10714285714286	5,64309165954431
9	9	6,61071428571428	6,85	6,84100185528757
10	10	6,60833333333333	6,75	6,90423850574713
11	11	6,65555555555556	6,81666666666667	6,52954144620811
12	12	6,55882352941176	6,85294117647059	6,63346515076619

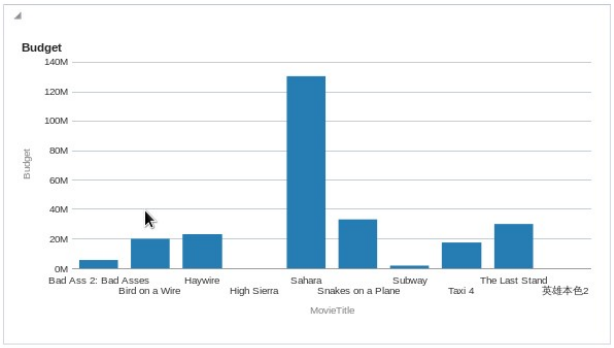
Dane uzyskane bezpośrednio z hurtowni.



Dane prezentowane na raporcie.

	MovieTitle	s
1	Bad Ass 2: Bad Asses	5500000,00
2	Bird on a Wire	20000000,00
3	Haywire	23000000,00
4	High Sierra	455000,00
5	Sahara	130000000,00
6	Snakes on a Plane	33000000,00
7	Subway	2000000,00
8	Taxi 4	17500000,00
9	The Last Stand	30000000,00
10	英雄本色2	150000,00

Dane uzyskane bezpośrednio z hurtowni.



Dane prezentowane na raporcie.

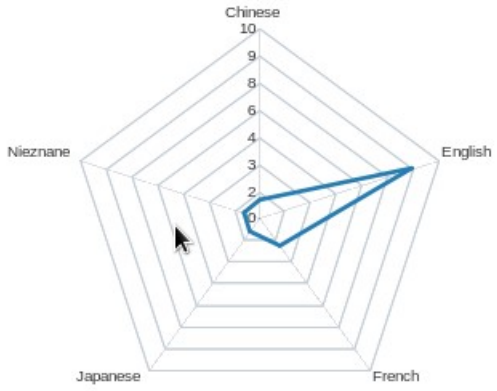
	avg_Budget
1	30251956,9459

Dane uzyskane bezpośrednio z hurtowni.

Dane prezentowane na raporcie.

	OriginalLanguage	counted
1	Chinese	1
2	English	9
3	French	2
4	Japanese	1
5	Nieznane	1

Dane uzyskane bezpośrednio z hurtowni.



Dane prezentowane na raporcie.

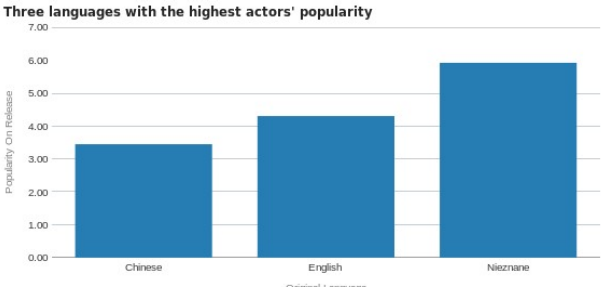
	avg_Revenue
1	64033390,0581

Dane uzyskane bezpośrednio z hurtowni.

Dane prezentowane na raporcie.

	OriginalLanguage	Popularity
1	Nieznane	5,90714285714286
2	English	4,29967928877182
3	Chinese	3,453725

Dane uzyskane bezpośrednio z hurtowni.



Dane prezentowane na raporcie.