WOJSKOWA AKADEMIA TECHNICZNA

IM. JAROSŁAWA DĄBROWSKIEGO W WARSZAWIE

WYDZIAŁ CYBERNETYKI

Przedmiot: Hurtownie Danych

Student: st. szer. pchor. Rafał Guzek

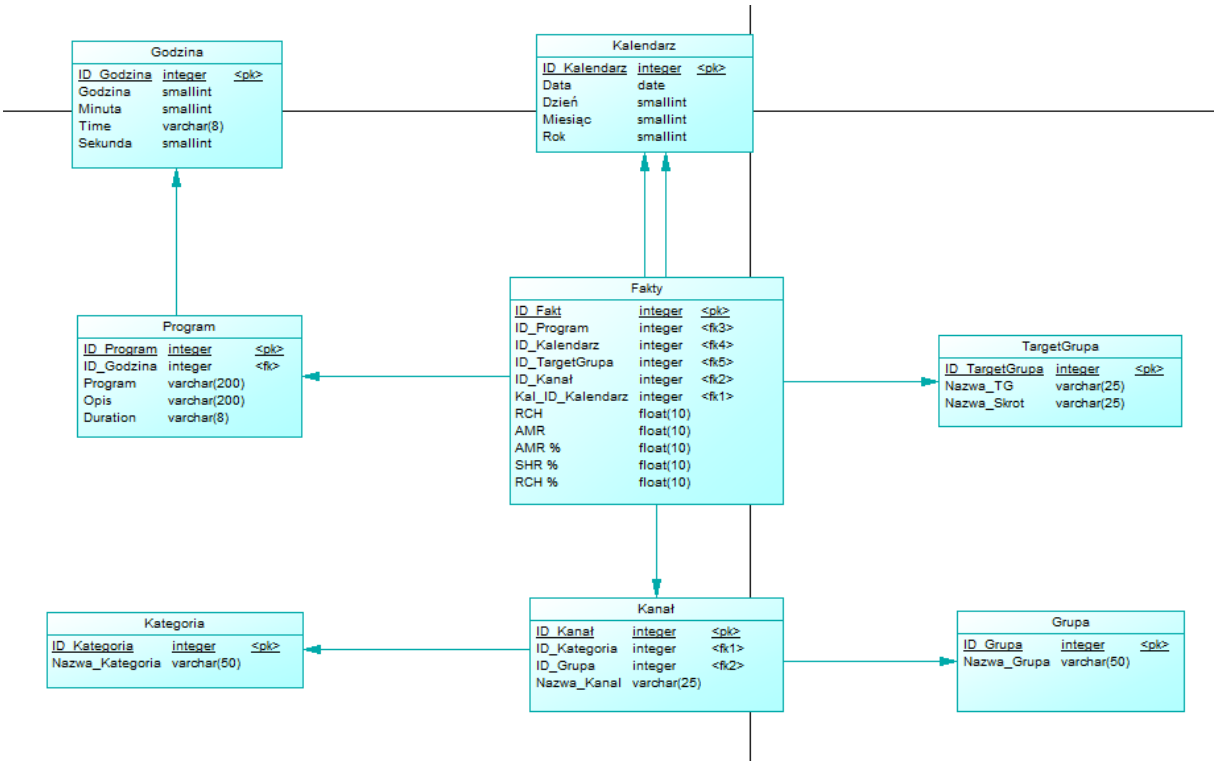
Nr grupy: I5B1S1

Data: 26.06.2018

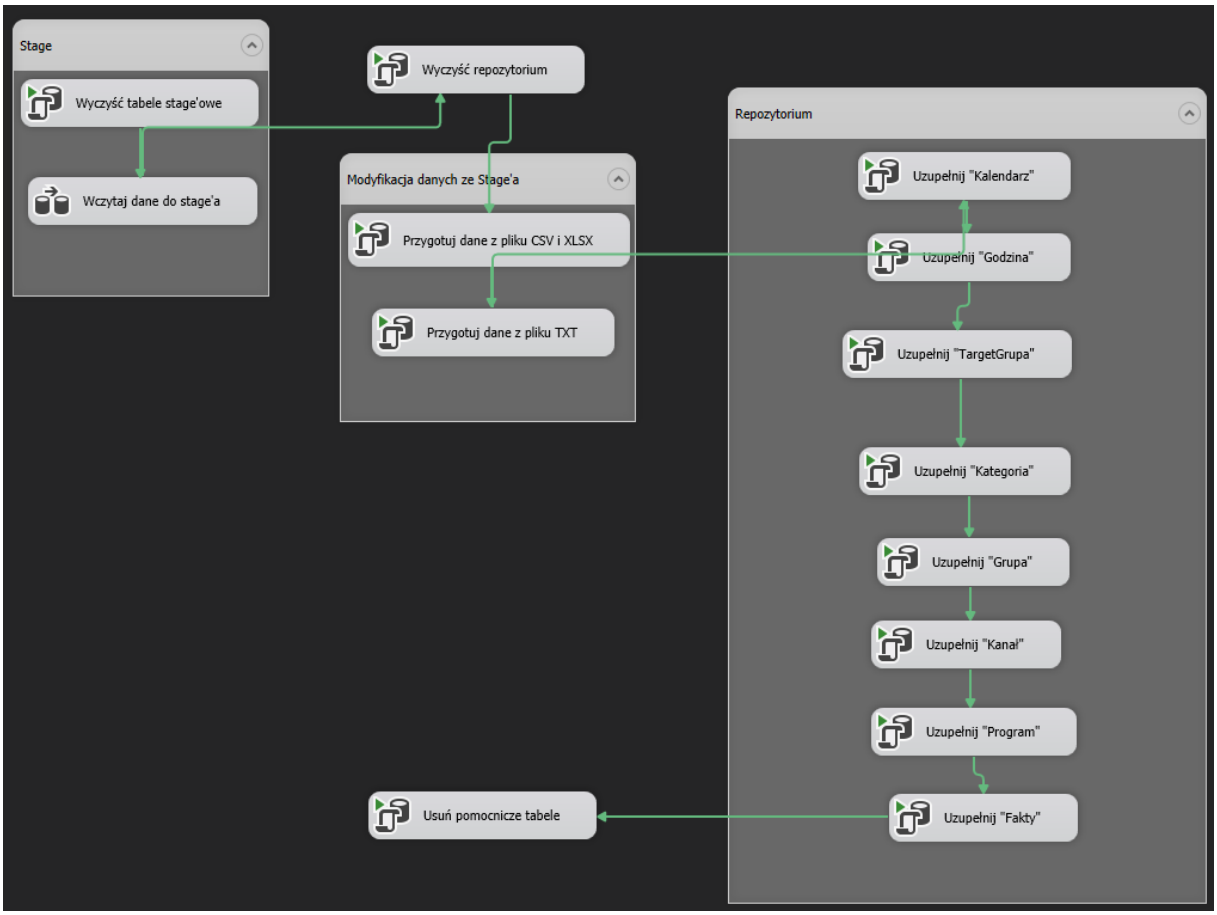
Procedury:

- Stage – Procedura odpowiada za tworzenie tabel potrzebnych do wprowadzenia danych z plików przy wykorzystaniu Visual Studio
- Repozytorium – Skrypt tworzący repozytorium główne.
- Przygotuj_CSV_XLSX – Procedura odpowiada za przetworzenie surowych danych znajdujących się w stage'u i doprowadzenie ich do takiej postaci, aby można je było umieścić w repozytorium. Przy wykorzystaniu kursora ustawionego na tabeli w stage, pobieram każdy rekord do zmiennych i na zmiennych wykonuję operacje z wykorzystaniem funkcji „LEFT", „CHARINDEX", „REPLACE". Dzięki pozbywam się zbędnych informacji i otrzymuję przejrzystą zawartość danych ze stage'a z plików CSV i XLSX. Zawierają one dane dotyczące kategorii, kanałów i grup kanałów.
- Przygotuj_TXT – Przy użyciu tej procedury tworzę dwie tabele pomocnicze na fakty. Do pierwszej kopiuję zawartość stage'a, następnie przy wykorzystaniu funkcji „UPDATE" dokonuję oczyszczenia rekordów z niepotrzebnych znaków takich jak cudzysłów, kropki, spacje, znak procentów itp. Następnie w tak wyczyszczonej tabeli, uzupełniam puste miejsca w których powinny znajdować się dane. Do tego celu używam kursora, który umieszczony na wspomnianej tabeli kopiuje zawartość wypełnionych wierszy i wprowadza w miejsca gdzie również dana wartość powinna się znajdować. Rekordy te zostają umieszczone w drugiej pomocniczej tabeli. Po zakończeniu, wykorzystuję funkcję PIVOT i UNPIVOT, aby odwrócić kolumny z wierszami do postaci jaką zalecił Pan na zajęciach.
- R_Kalendarz – Procedura wypełnia tabelę kalendarza wartościami data, dzień, miesiąc i rok, dla roku 2015, czyli wprowadza 365 rekordów do tabeli.
- R_Godzina – Procedura uzupełnia tabelę godzin wszystkimi możliwymi godzinami w granulacji sekundowej. Dodatkowo dla każdej godziny wprowadzona zostaje jej wartość godzinna, minutowa i sekundowa. Pozwala to na agregację po czasie rozpoczęcia programu dla poszczególnych godzin. Do pobierania poszczególnych składowych godziny wykorzystałem funkcję „DATEPART" a w celu inkrementacji czasu użyłem funkcji „DATEADD"
- R_Grupa - Procedura wprowadza wyczyszczone dane do tabeli Grupa z wykorzystaniem kursora na pomocniczej tabeli z poprawionymi danymi.
- R_Kanał – Procedura uzupełia tabelę Kanał nazwami kanałów oraz przyporządkowuje klucze do tabel Grupa i Kategoria zgodnie z przynależnością danego kanału. W przypadku gdy dla danego kanału nie istnieje kategoria, zostaje on przyporządkowany do kategorii „Unknown".
- R_Programy – Procedura przy wykorzystaniu tabeli z faktami po użyciu funkcji PIVOT, wprowadza wszystkie istniejące kombinacje programów i opisów z godziną i czasem trwania
- R_Fakty – Procedura uzupełnia tabelę faktów na podstawie danych w tabeli Pivot_Fakty. Przy użyciu JOIN'ów z innymi tabelami otrzymuję w postaci wyniku zapytania postać gotową do wprowadzenia do repozytorium. Dodatkowo przed tym uzupełniam tabelę TargetGrupa o brakujące wartości, które nie były zamieszczone w pliku. W tym przypadku są to „TotalIndividuals" oraz „Podgrupa".

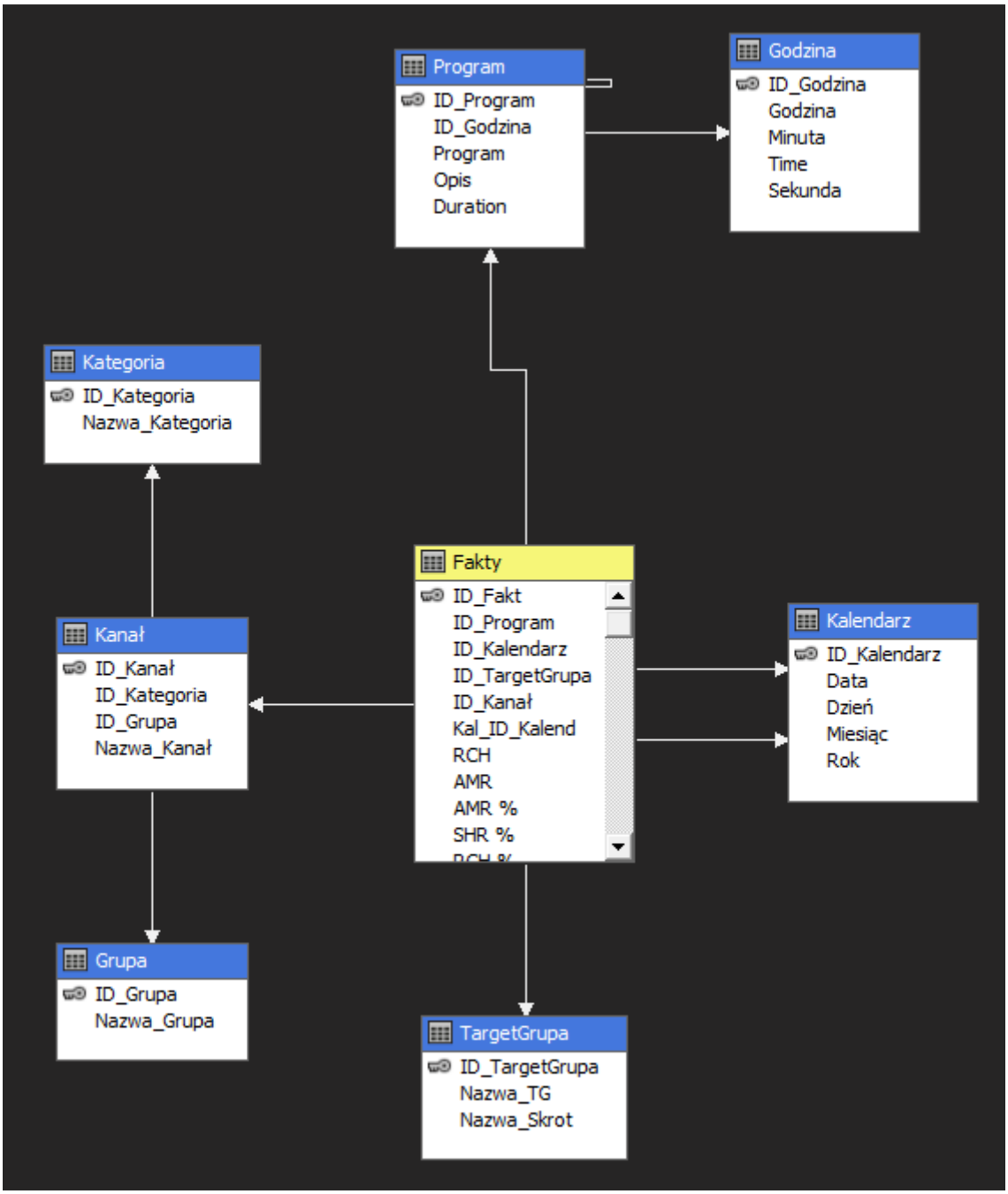
Model repozytorium:



SSIS:

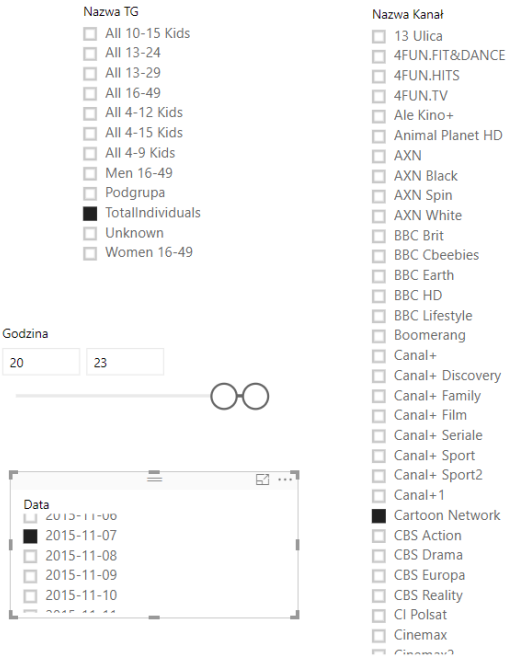


Cube:

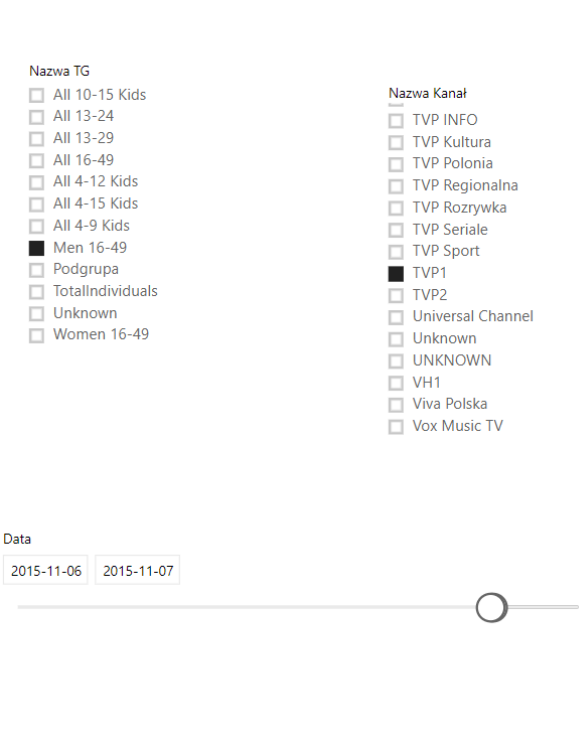

Program
ID_Program
ID_Godzina
Program
Opis
Duration
Godzina
ID_Godzina
Godzina
Minuta
Time
Sekunda
Kategoria
ID_Kategoria
Nazwa_Kategoria
Fakty
ID_Fakt
ID_Program
ID_Kalendarz
ID_TargetGrupa
ID_Kanał
Kal_ID_Kalend
RCH
AMR
AMR %
SHR %
Kanał
ID_Kanał
ID_Kategoria
ID_Grupa
Nazwa_Kanał
Kalendarz
ID_Kalendarz
Data
Dzień
Miesiąc
Rok
Grupa
ID_Grupa
Nazwa_Grupa
TargetGrupa
ID_TargetGrupa
Nazwa_TG
Nazwa_Skrot

Przykładowe raporty na dst:

Nazwa Kanał	Program	Time	Duration	AMR
Cartoon Network	PROMO	21:09:02	00:00:08.0000000	75 221,00
Cartoon Network	PROMO	20:01:01	00:00:09.0000000	66 542,00
Cartoon Network	PROMO	20:56:29	00:00:16.0000000	57 952,00
Cartoon Network	PROMO	21:38:15	00:00:16.0000000	66 381,00
Cartoon Network	PROMO	20:39:58	00:00:38.0000000	57 122,00
Cartoon Network	PROMO	20:18:21	00:00:41.0000000	47 482,00
Cartoon Network	PROMO	20:43:29	00:02:25.0000000	57 918,00
Cartoon Network	PROMO	21:13:48	00:02:44.0000000	72 058,00
Cartoon Network	ZWYCZAJNY SERIAL	20:45:54	00:21:14.0000000	61 970,00
Cartoon Network	NINJAGO MISTRZOWIE SPINJITZU	21:16:32	00:45:17.0000000	61 521,00
Suma				**624 167,00**



Time	Duration	Program	RCH_P
00:09:05	00:00:21.0000000	PROMO	0,73
00:09:26	00:00:38.0000000	OGLOSZENIE PLATNE	0,73
00:16:12	01:43:48.0000000	KATYN	1,34
00:58:10	00:00:26.0000000	PROMO	0,00
00:58:36	00:20:40.0000000	NASZAARMIA.PL	0,40
01:01:14	00:00:41.0000000	PROMO	0,00
01:32:06	00:27:54.0000000	SPRAWA DLA REPORTERA	0,00
02:00:00	00:13:07.0000000	KATYN	0,00
02:00:00	00:19:13.0000000	SPRAWA DLA REPORTERA	0,00
02:19:15	00:00:30.0000000	PROMO	0,00
02:25:16	00:00:26.0000000	PROMO	0,00
02:25:43	01:43:56.0000000	EGOISCI	0,64
02:31:43	00:50:13.0000000	KLINIKA SPALONYCH TWARZY	0,00
03:26:04	00:11:10.0000000	NOTACJE	0,00
03:41:07	00:10:50.0000000	NOTACJE	0,00
03:56:05	00:41:24.0000000	SWIAT SIE KRECI	0,19
04:15:47	00:00:30.0000000	PROMO	0,00
04:22:15	00:40:25.0000000	SWIAT SIE KRECI	0,19
04:44:37	00:29:58.0000000	ZAGADKOWA JEDYNKA	0,19
05:09:48	00:01:00.0000000	ZEGAR	0,00
05:14:35	00:01:00.0000000	ZEGAR	0,00
05:17:43	00:01:00.0000000	ZEGAR	0,00
05:22:36	00:44:26.0000000	EGZAMIN Z ZYCIA	0,00
05:23:01	00:01:01.0000000	ZEGAR	0,00
05:31:05	00:21:36.0000000	KLAN	0,28
05:52:41	00:05:04.0000000	TELESPRZEDAZ	0,28
05:57:45	00:10:00.0000000	TELESPRZEDAZ	0,00
06:07:45	00:10:00.0000000	TELESPRZEDAZ	0,00
06:10:10	00:25:14.0000000	SLOWNIK POLSKO@POLSKI	0,00
06:17:45	00:05:04.0000000	TELESPRZEDAZ	0,00
06:28:57	00:52:51.0000000	NOSOROZCE KLATWA MAGICZNEGO ROGU	0,00
06:38:32	00:15:03.0000000	PELNOSPRAWNI	0,00
06:53:35	00:00:30.0000000	PROMO	0,00
Suma			**310,05**



Wnioski

Projekt składa się z wielu etapów i każdy z nich jest bardzo czasochłonny. Początkowo trudnością było poruszanie się po nowym oprogramowaniu. Następnie po przejściu do tworzenia hurtowni, pojawiły się problemy od strony SQL'a, którego musiałem odświeżyć i przypomnieć sobie tworzenie w nim zapytań itp. Od strony wczytywania danych, problemem było wyczyszczenie rekordów z niepotrzebnych znaków, znalezienie wszystkich błędów aby dane były poprawne, uporządkować i przekopiować tak aby tworzyły spójną całość. Następnie trzeba było dobrze dopasować do siebie wartości z różnych plików aby wyciągnąć z nich np. kategorie i grupy kanałów a następnie dopasować je do kanałów. Dużo problemów przysporzyła funkcja PIVOT, która niekiedy działała a niekiedy nie, lecz ostatecznie udało się ja doprowadzić do prawidłowego działania. Pracując z plikiem z przykładowymi danymi optymalizacja kodu wydawała się nie mieć dużego znaczenia, jednak próbując wczytać plik główny, czas wydłużał się do kilku godzin i musiałem poprawiać kod tak aby wykonywał się szybciej, co również zajęło dużo czasu. Prawidłowe zaprojektowanie repozytorium było dużym wyzwaniem i często musiałem dokonywać modyfikacji, aby poprawie móc wprowadzić dane a następnie wygenerować raporty.