Hurtownie danych
Laboratorium Czw 11:15

Projekt

Kajetan Pynka 254495

Spis treści

Etap 1

1. Zakres realizacji projektu

1.1. Tytuł projektu

Analiza systemu rowerów publicznych Bay Area Bike Share w San Francisco.

1.2. Charakterystyka dziedziny problemowej

System rowerów publicznych oferowany przez przedsiębiorstwa prywatne związany jest z następującymi elementami:

- Utrzymywanie rowerów wykorzystywanych przez klientów w stanie nadającym się do użytku
- Zarządzanie i zapewnianie poprawnego działania stacji rowerowych
- Zbieranie anonimowych danych ze stacji / rowerów czy też od klientów
- Zapewnianie klientom możliwości opłaty roweru ze stacji lub wygodnie z aplikacji mobilnej
- Monitorowanie stanu zapełnienia stacji rowerowych i reagowanie w odpowiednim czasie
- Prowadzenie działu obsługi klienta (telefonicznego / internetowego)

1.3. Krótki opis obszaru analizy

W ramach tego projektu skupię się na danych zebranych i udostępnionych przez byłą firmę Bay Area Bike Share. Przedstawiają one użytkowanie poszczególnych rowerów, stacji rowerowych oraz dane pogodowe w okresie między 29 sierpnia 2013r. a 1 września 2015r. Dane dotyczą stacji znajdujących się w regionie Zatoki San Francisco (pochodzą z różnych miast, a same stacje posiadają informacje o długości i szerokości geograficznej).

1.4. Problemy i potrzeby

- Zoptymalizowanie wykorzystania stacji rowerowych
- Zachęcenie użytkowników do zakupienia subskrypcji
- Analiza wpływu pogody na użytkowanie rowerów
- Analiza przychodów pod kątem: regionu (miasta, stacji), czasu (pory dnia, pory roku)
- Wykorzystanie najdłuższych wycieczek rowerowych do wyznaczenia potencjalnych miejsc nowych stacji

1.5. Cel przedsięwzięcia

1.5.1. Oczekiwania

Wykrycie trendów i korelacji pomiędzy danymi, dostarczenie prognoz na kolejne lata funkcjonowania stacji rowerowych, zaproponowanie kroków do podjęcia w celu zwiększenia zysków czy też wydajności.

1.5.2. Zakres analizy – badane aspekty

1. Sumaryczna długość wycieczek ze względu na dzień tygodnia dla każdej stacji.
2. Liczba wycieczek ze względu na zachmurzenie według miast.
3. Procentowy udział klientów niezarejestrowanych oraz subskrybentów ze względu na miasto.
4. Liczba wycieczek podczas mgły ze względu na godzinę i miasto.
5. Liczba wycieczek dla każdej stacji ze względu na opady (=0 – brak, T-nieznaczne, < 0.20 – średnie, > 0.20 – znaczące).
6. Średnia długość wycieczki dla każdej stacji z San Jose i rodzaju klienta.
7. Zestawienie najpopularniejszej stacji docelowej dla każdej stacji ze względu na miesiąc.
8. Liczba wycieczek dla klientów zamieszkujących pod każdym kodem pocztowym ze względu na miesiąc.
9. Sumaryczna długość wycieczek dla każdego roweru ze względu na godzinę.
10. Liczba unikalnych klientów rozpoczynających lub kończących wycieczkę dla każdej stacji ze względu na miesiąc.

1.6. Źródła danych (lokalizacja, format, dostępność)

L.p.	Plik	Typ	Liczba rekordów	Rozmiar [MB]	Opis
1	station.csv	csv	70	0.00565	Łańcuchy znaków w języku angielskim, daty w formacie MM/DD/YYYY, brak znaków specjalnych, liczby całkowite jak i zmiennoprzecinkowe o małej precyzji. Niektóre stacje zmieniły lokalizację i nazwę.
2	status.csv	csv	72.000.000	1990	Małe liczby całkowite mieszczące się w bajcie. Czas w formacie YYYY/MM/DD HH:mm:SS.
3	trip.csv	csv	670.000	80.21	Łańcuchy znaków w języku angielskim, liczby całkowite. Czas w formacie MM/DD/YYYY HH:mm. Rodzaj subskrypcji jako typ wyliczeniowy 2 łańcuchów znakowych: „Subscriber” i „Customer”.
4	weather.csv	csv	3665	0.43806	Data w formacie MM/DD/YYYY. Liczby zmiennoprzecinkowe o małej precyzji, liczby całkowite, łańcuchy znaków w języku angielskim.

2. Profilowanie danych

2.1. Analiza danych

Plik: station.csv				
L.p.	Atrybut	Typ danych	Zakres wartości	Uwagi – ocena jakości danych
1	id	int	2-84	W pełni poprawne dane
2	name	varchar(45)	9-45 znaków	W pełni poprawne dane, nastąpiła zmiana nazw niektórych stacji
3	lat	float(6, 4)	37.3297-37.8048	W pełni poprawne dane, nastąpiła zmiana położenia niektórych stacji
4	long	float(7, 4)	-122.419 do -121.8773	W pełni poprawne dane, nastąpiła zmiana położenia niektórych stacji
5	dock_count	int	11-27	W pełni poprawne dane
6	city	varchar(13)	8-13 znaków	W pełni poprawne dane
7	installation_date	datetime	8/5/2013 – 4/9/2014	W pełni poprawne dane, format MM/DD/YYYY

Plik: status.csv				
L.p.	Atrybut	Typ danych	Zakres wartości	Uwagi – ocena jakości danych
1	station_id	int	2-84	W pełni poprawne dane
2	bikes_available	int	0-27	W pełni poprawne dane
3	docks_available	int	0-27	W pełni poprawne dane
4	time	datetime	2013/08/29 12:06:01 – 2015/08/31 12:06:01	W pełni poprawne dane, czas w formacie YYYY/MM/DD HH:mm:SS.

Plik: trip.csv				
L.p.	Atrybut	Typ danych	Zakres wartości	Uwagi – ocena jakości danych
1	id	int	4079-913459	W pełni poprawne dane
2	duration	int	60-17270400	W pełni poprawne dane, czas mierzony w sekundach
3	start_date	datetime	8/29/2013 9:38 – 8/31/2015 23:26	W pełni poprawne dane, data w formacie MM/DD/YYYY HH:mm.
4	start_station_name	varchar(45)	9-45	W pełni poprawne dane, nazwa niektórych stacji uległa zmianie w czasie
5	start_station_id	int	2-84	W pełni poprawne dane
6	end_date	datetime	8/29/2013 9:41 – 8/31/2015 23:39	W pełni poprawne dane, data w formacie MM/DD/YYYY HH:mm.
7	end_station_name	varchar(45)	9-45	W pełni poprawne dane, nazwa niektórych stacji uległa zmianie w czasie
8	end_station_id	int	2-84	W pełni poprawne dane
9	bike_id	int	9-878	W pełni poprawne dane
10	subscription_type	varchar(10)	8-10 znaków	W pełni poprawne dane, przyjmuje zasadniczo dwie wartości: „Subscriber” oraz „Customer”
11	zip_code	varchar(11)	1-11 znaków	Około 1% rekordów posiada wartości puste, około 2% rekordów posiada nieprawidłowe wartości: zdecydownie za mało albo za dużo cyfr, występują też przypadki liter.

Plik: weather.csv				
L.p.	Atrybut	Typ danych	Zakres wartości	Uwagi – ocena jakości danych
1	date	datetime	8/29/2013 – 8/31/2015	W pełni poprawne dane, data w formacie MM/DD/YYYY
2	Max_temperature_f	int	44-102	Występują 4 rekordy puste (mniej niż 1% wszystkich)
3	Mean_temperature_f	int	38-84	Występują 4 rekordy puste (mniej niż 1% wszystkich)
4	Min_temperature_f	int	25-75	Występują 4 rekordy puste (mniej niż 1% wszystkich)
5	Max_dew_point_f	int	20-68	Występują 54 rekordy puste (około 1% wszystkich)
6	Mean_dew_point_f	int	13-65	Występują 54 rekordy puste (około 1% wszystkich)
7	Min_dew_point_f	int	2-63	Występują 54 rekordy puste (około 1% wszystkich)
8	Max_humidity	int	24-100	Występują 54 rekordy puste (około 1% wszystkich)
9	Mean_humidity	int	24-96	Występują 54 rekordy puste (około 1% wszystkich)
10	Min_humidity	int	4-93	Występują 54 rekordy puste (około 1% wszystkich)
11	Max_sea_level_pressure_inches	float(4,2)	29.5-30.65	Występuje 1 rekord pusty (mniej niż 1% wszystkich)
12	Mean_sea_level_pressure_inches	float(4,2)	29.43-30.41	Występuje 1 rekord pusty (mniej niż 1% wszystkich)
13	Min_sea_level_pressure_inches	float(4,2)	28.98-30.37	Występuje 1 rekord pusty (mniej niż 1% wszystkich)
14	Max_visibility_miles	int	5-20	Występuje 13 rekordów pustych (mniej niż 1% wszystkich)
15	Mean_visibility_miles	int	4-20	Występuje 13 rekordów pustych (mniej niż 1% wszystkich)
16	Min_visibility_miles	int	0-20	Występuje 13 rekordów pustych (mniej niż 1% wszystkich)
17	Max_wind_speed_mph	int	0-128	Występuje 1 rekord pusty (mniej niż 1% wszystkich)
18	Mean_wind_speed_mph	int	0-23	Występuje 1 rekord pusty (mniej niż 1% wszystkich)
19	Max_gust_speed_mph	int	6-114	25% rekordów jest pustych
20	Precipitation_inches	varchar(4)	1-4 znaków	Występuje 1 rekord pusty. W 96% przypadków jest to float(4,2) natomiast dla 4% rekordów znak 'T' oznaczający nieznaczne opady.
21	Cloud_cover	int	0-8	Występuje 1 rekord pusty (mniej niż 1% wszystkich)
22	Events	varchar(17)	3-17 znaków	86% rekordów jest pustych, pozostałe posiadają jedną z pięciu wartości: „Rain", „Fog", „Fog-Rain", „Rain-Thunderstorm", „rain".

23	Wind_dir_degrees	int	0-2772	Występuje 1 rekord pusty (mniej niż 1% wszystkich)
24	Zip_code	int	94041-95113	Dane w pełni poprawne, występuje pięć kodów pocztowych: 94107, 94063, 94301, 94091, 95113

2.2. Ocena przydatności danych

L.p.	Plik	Ocena jakości danych
1	station.csv	Brak pustych pól, wszystkie są poprawne. Dla niektórych stacji zmieniła się nazwa oraz położenie (nie jest to problemem ponieważ dalej obowiązuje ten sam identyfikator).
2	status.csv	W pełni poprawne dane, 3 niewielkie liczby całkowite wraz z czasem co do sekundy (w praktyce każdy zapis był dokonywany co minutę). Format YYYY/MM/DD HH:mm jest akceptowalny. Ogólnie jednak dane nie są przydatne jeśli o chodzi o założenia projektowe, więc można zignorować ten plik.
3	trip.csv	Wszystkie kolumny, poza jedną, są wypełnione poprawnymi danymi i są w pełni użyteczne. Należy pamiętać o tym, że również występuję tu kwestia zmiany nazw niektórych stacji. Kod pocztowy po oczyszczeniu będzie się nadawał do dalszej analizy.
4	weather.csv	W większości kolumn występują marginalne brakujące dane, nieistotne dane atmosferyczne możemy odrzucić na potrzeby dalszych analiz. Liczba opadów pozostanie jako łańcuch znakowy, należy pamiętać o znaku ‘T’ jako jednej z możliwości tej kolumny. Należy oczyścić kolumnę zdarzeń atmosferycznych i połączyć „Rain” oraz „rain” w jedno zdarzenie. Kod pocztowy do przekształcenia na miasto (wtedy odpowiada miastu ze stacji).

2.3. Definicja typów encji/klas oraz związków

Encje:

Encja: **DIM_TIME**			
Nazwa atrybutu	Opis atrybutu	Typ	OBL(+) OPC(-)
PK_TIME	Klucz główny, identyfikator w formie łańcucha znaków	varchar(12)	+
Year	Rok zapisany jako liczba całkowita	integer	+
Month	Miesiąc zapisany jako liczba całkowita	integer	+
Day	Dzień zapisany jako liczba całkowita	integer	+
Hour	Godzina zapisana jako liczba całkowita	integer	-
Minute	Minuta zapisana jako liczba całkowita	integer	-

Encja: **DIM_STATION**			
Nazwa atrybutu	Opis atrybutu	Typ	OBL(+) OPC(-)
STATION_ID	Klucz główny, identyfikator pojedynczej stacji	integer	+
Name	Nazwa stacji rowerowej zapisana jako łańcuch znaków	varchar(45)	+
Lat	Szerokość geograficzna położenia stacji	float(6,4)	+
Long	Długość geograficzna położenia stacji	float(7,4)	+
Dock_Count	Liczba możliwych rowerów do zaparkowania w stacji	integer	+
City	Ciąg znaków oznaczający nazwę miasta, w której znajduje się stacja	varchar(13)	+
Installation_Date	Klucz obcy, referencja do czasu przechowywanego w DIM_TIME	varchar(12)	+

Encja: **DIM_BIKE**			
Nazwa atrybutu	Opis atrybutu	Typ	OBL(+) OPC(-)
BIKE_ID	Klucz główny, identyfikator roweru	integer	+

Encja: **DIM_CUSTOMER**			
Nazwa atrybutu	Opis atrybutu	Typ	OBL(+) OPC(-)
CUSTOMER_ID	Klucz główny, identyfikator klienta w formie liczby całkowitej	integer	+
Cust_Zip_Code	Łańcuch znaków reprezentujący kod pocztowy klienta, zły kod pocztowy zastąpiony przez NULL	varchar(11)	-
Subscription_Type	Łańcuch znaków reprezentujący status klienta: „Customer” albo „Subscriber”	varchar(10)	+

Encja: **DIM_WEATHER**			
Nazwa atrybutu	Opis atrybutu	Typ	OBL(+) OPC(-)
Measure_Date	Klucz główny, klucz obcy z referencją do czasu przechowywanego w DIM_TIME	varchar(12)	+
Measure_City	Ciąg znaków oznaczający miasto, w którym dokonanu pomiaru	varchar(13)	+
Max_Temperature	Liczba całkowita oznaczająca maksymalną osiągniętą temperaturę	integer	-
Mean_Temperature	Liczba całkowita oznaczającą średnią osiągniętą temperaturę	integer	-
Min_Temperature	Liczba całkowita oznaczającą minimalną osiągniętą temperaturę	integer	-
Max_Humidity	Liczba całkowita oznaczająca maksymalną osiągniętą wilgotność	integer	-

Mean_Humidity	Liczba całkowita oznaczająca średnią osiągniętą wilgotność	integer	-
Min_Humidity	Liczba całkowita oznaczająca minimalną osiągniętą wilgotność	integer	-
Max_Pressure	Liczba zmiennoprzecinkowa oznaczająca maksymalne osiągnięte ciśnienie	float(4,2)	-
Mean_Pressure	Liczba zmiennoprzecinkowa oznaczająca średnie osiągnięte ciśnienie	float(4,2)	-
Min_Pressure	Liczba zmiennoprzecinkowa oznaczająca minimalne osiągnięte ciśnienie	float(4,2)	-
Max_Visibility	Liczba całkowita oznaczająca maksymalną widoczność w milach	integer	-
Mean_Visibility	Liczba całkowita oznaczająca średnią widoczność w milach	integer	-
Min_Visibility	Liczba całkowita oznaczająca minimalną widoczność w milach	integer	-
Precipitation_Inches	Łańcuch znaków określający liczbę opadów w calach lub znak ‘T’ gdy opady były niewielkie.	varchar(4)	-
Cloud_Cover	Liczba całkowita oznaczająca zachmurzenie w skali 0-8	integer	-
Events	Zdarzenie atmosferyczne zapisane jako ciąg znaków, jedno z 4 wydarzeń: „Rain”, „Fog”, „Fog-Rain”, „Rain-Thunderstorm”	varchar(17)	-

Encja: **FACT_TRIP**			
Nazwa atrybutu	Opis atrybutu	Typ	OBL(+) OPC(-)
TRIP_ID	Klucz główny, liczba całkowita oznaczająca identyfikator wycieczki	integer	+
Start_Time	Klucz obcy, referencja czasu przechowywanego w DIM_TIME	varchar(12)	+
End_Time	Klucz obcy, referencja czasu przechowywanego w DIM_TIME	varchar(12)	+
Start_Station	Klucz obcy, referencja do identyfikatora stacji z DIM_STATION	integer	+
End_Station	Klucz obcy, referencja do identyfikatora stacji z DIM_STATION	integer	+
Start_Station_Name	Ciąg znaków reprezentujący nazwę stacji startowej	varchar(45)	+
End_Station_Name	Ciąg znaków reprezentujący nazwę stacji końcowej	varchar(45)	+
Trip_Customer	Klucz obcy, referencja do klienta z DIM_CUSTOMER	integer	+
Trip_Bike	Klucz obcy, referencja do roweru z DIM_BIKE	integer	+
Trip_Weather	Klucz obcy, referencja do pogody z DIM_WEATHER na podstawie Start_Time	varchar(12)	+
Duration	Liczba całkowita oznaczająca długość wycieczki w sekundach	integer	+

Związki:

- DIM_TIME(1) – (0..*)DIM_WEATHER
- DIM_TIME(1) – (0..*)DIM_STATION
- DIM_TIME(2) – (0..*)FACT_TRIP
- DIM_STATION(2) – (0..*)FACT_TRIP
- DIM_WEATHER(1) – (0..*)FACT_TRIP
- DIM_CUSTOMER(1) – (0..*)FACT_TRIP
- DIM_BIKE(1) – (0..*)FACT_TRIP

2.4. Propozycja wymiarów, hierarchii, miar

Wymiary:

- DIM_TIME
- DIM_STATION
- DIM_BIKE
- DIM_CUSTOMER
- DIM_WEATHER

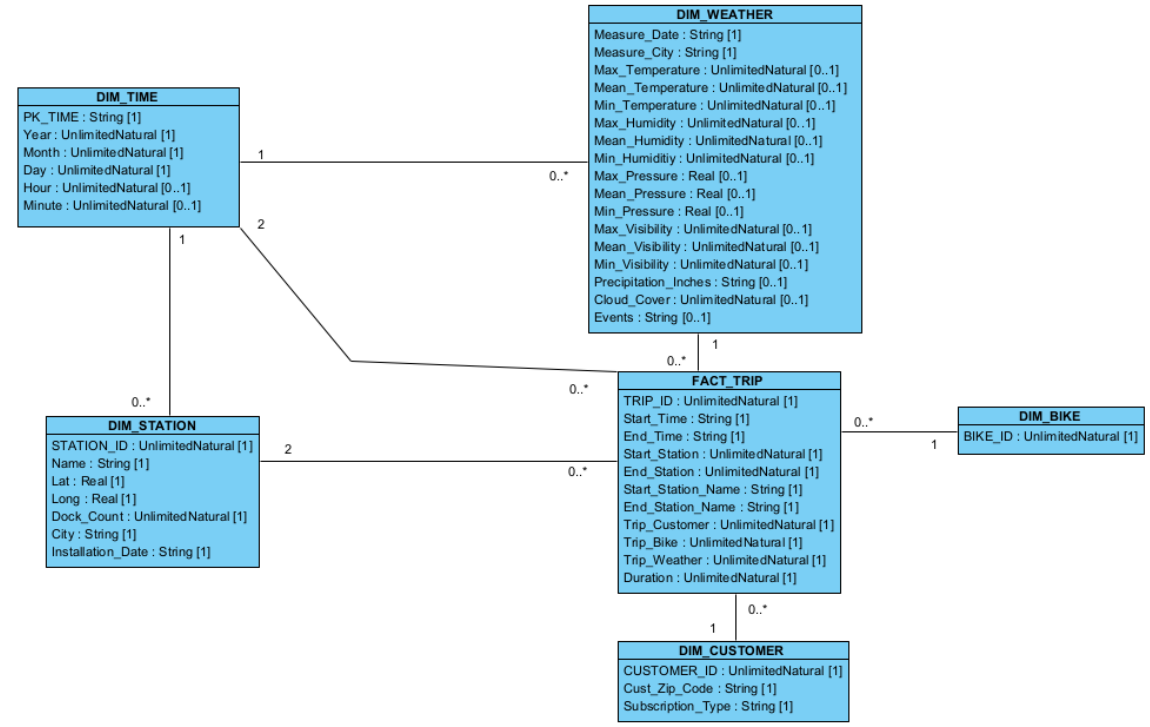
Hierarchie:

- DIM_TIME: Year -> Month -> Day -> Hour -> Minute
- DIM_STATION: City -> Name

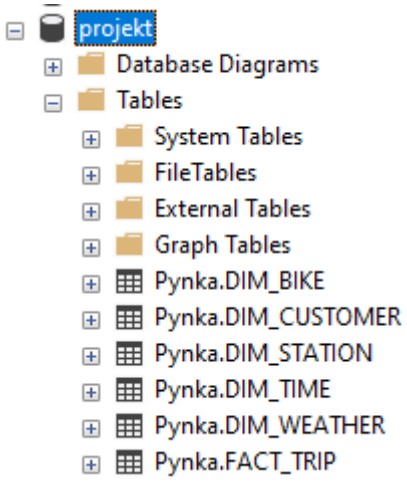
Miary:

- Długość wycieczki
- Liczba unikalnych klientów
- Liczba unikalnych rowerów
- Liczba wycieczek

2.5. Diagram klas



3. Utworzenie bazy danych



Rysunek przedstawia utworzoną strukturę bazy danych.

Wnioski:

W celu przygotowania odpowiedniej hurtowni danych musimy być świadomi wymagań biznesu by wiedzieć po co nam ma służyć dana hurtownia. Następnie, po dogłębnej analizie danych możemy wstępnie zaplanować naszą hurtownię, tak aby móc w przyszłości wytworzyć kostkę, a co za tym idzie by móc skutecznie dokonywać analizy danych znajdujących się w hurtowni.