Hurtownie danych – Projekt

Proces tworzenia hurtowni danych powinien być poprzedzony zrozumieniem „potrzeb biznesu" oraz rzeczywistości (dziedziny problemowej) reprezentowanej przez dostępne zasoby danych. Realizacja poniższego zadania ma uzmysłowić występujące problemy w określonym (wybranym) wycinku rzeczywistości, a następnie umożliwić zidentyfikowanie (określenie) potrzeb, celu i możliwości analiz biznesowych, by wspierać procesy decyzyjne (podejmowanie właściwych decyzji biznesowych).

Projekt końcowy powinien zawierać przynajmniej jedną kostkę Analysis Services, dotyczącą danych wybranych i przetworzonych przez studenta przy użyciu Integration Services. Utworzona kostka powinna:

- zawierać przynajmniej 5 wymiarów, w tym co najmniej dwa o strukturze hierarchicznej (np. czas, miejsce, itp)
- posiadać co najmniej 3 miary, w tym min. jedną nieaddytywną
- odpowiadająca jej tabela faktów powinna posiadać co najmniej 10000 rekordów.

Projekt – etap I

Propozycja tematu

1. Proszę przygotować zakres realizacji projektu zgodnie z poniższą specyfikacją oraz przedyskutować propozycję projektu z osobą prowadzącą zajęcia. Poczynione uzgodnienia zarejestrować w formie wniosków.

Zakres opracowania projektu HD

1.1. Tytuł projektu

Analiza Wypadków Lotniczych 1982-2023 - Narodowa Rada Bezpieczeństwa Transportu.

Aviation Accident Database & Synopses, up to 2023 - National Transportation Safety Board.

1.2. Charakterystyka dziedziny problemowej

Dziedzina problemowa obejmuje wypadki lotnicze z lat 1982-2023 i zawiera informacje takie jak data, kraj, lokalizacja, szczegóły samolotu, linii lotniczej, celu lotu lub fazy, w których wypadki miały miejsce, a także informacje o ofiarach.

1.3. Krótki opis obszaru analizy

Analiza będzie skupiać się na zbadaniu zależności pomiędzy częstotliwością wypadków, a rodzajem linii lotniczej, miejscem wypadku, warunkami pogodowymi, czy producentem lub wyposażeniem samolotów.

1.4. Problemy i potrzeby

Projekt ma na celu zbadanie zależności między częstotliwością wypadków a różnymi czynnikami, takimi jak warunki pogodowe, producent samolotów, faza lotu itp. Urząd lotnictwa cywilnego potrzebuje tych danych do statystyk i kategoryzacji wypadków czy podziału tychże statystyk ze względu na państwa lub same stany w przypadku US.

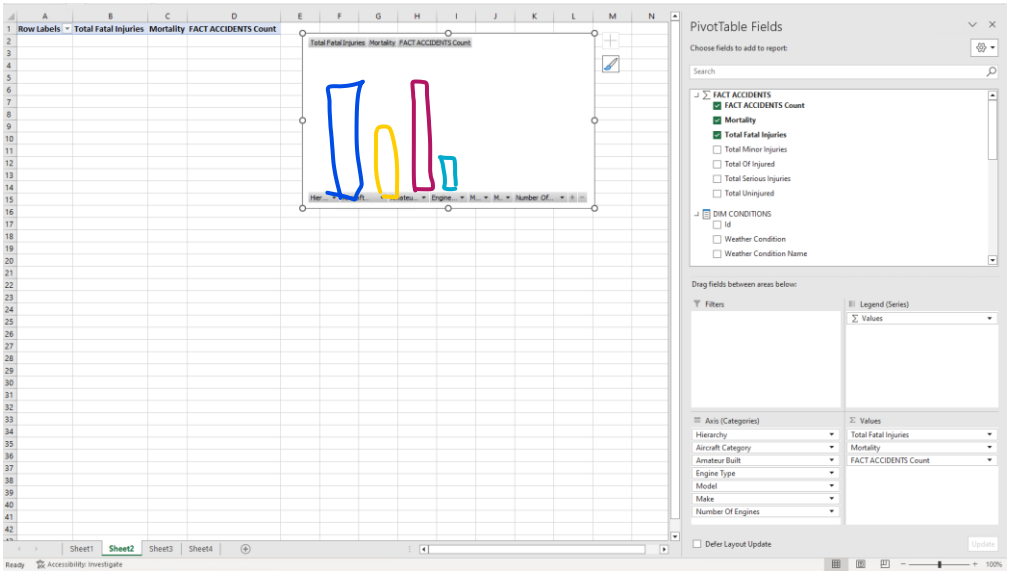
1.5. Cel przedsięwzięcia

1.5.1. Oczekiwania

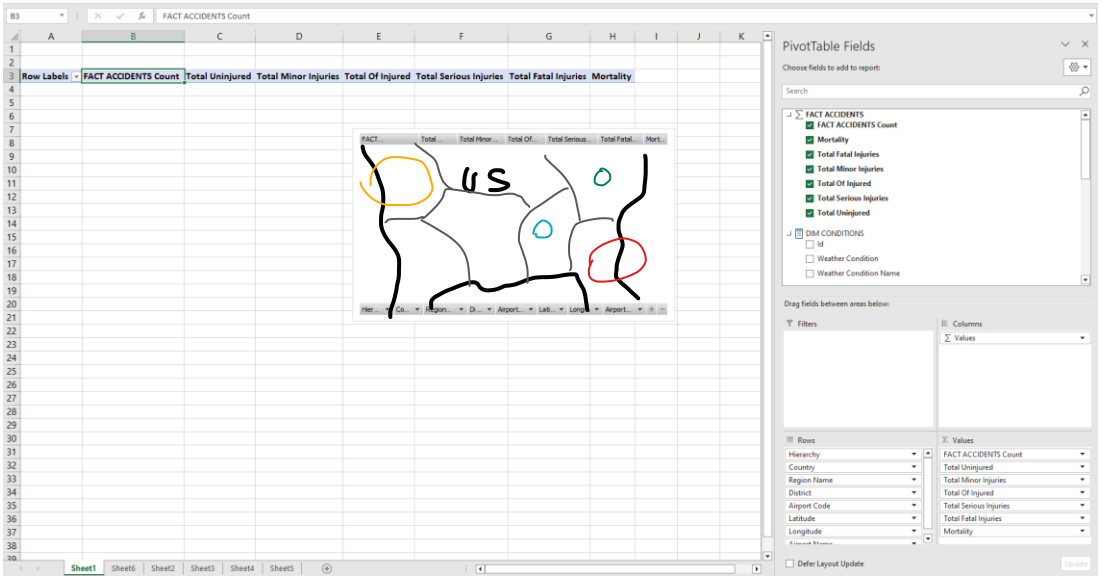
Określenie związku między warunkami lotu a częstotliwością wypadków oraz liczbą i zakresem uszkodzeń samolotu lub uszczerbkiem na zdrowiu ludzi podczas wypadku.

1.5.2. Zakres analizy – badane aspekty (min. 10 wielowymiarowych zestawień, które zostaną utworzone po wdrożeniu kostki) (4 dobre)

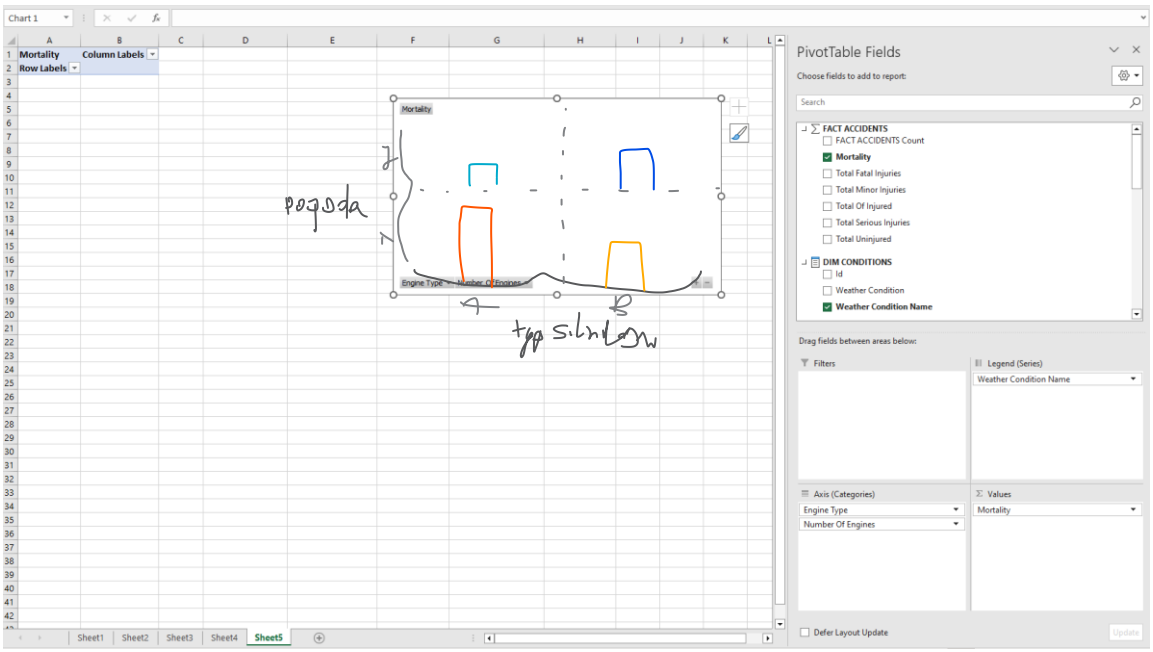
1. Analiza badająca liczbę wypadków w zależności od typu „samolotu”, producenta, modelu, liczby silników i typu silnika, a także o fazę lotu. Taka analiza pozwoliłaby na identyfikację szczególnie niebezpiecznych modeli lub producentów „samolotów”, które są bardziej podatne na wypadki.



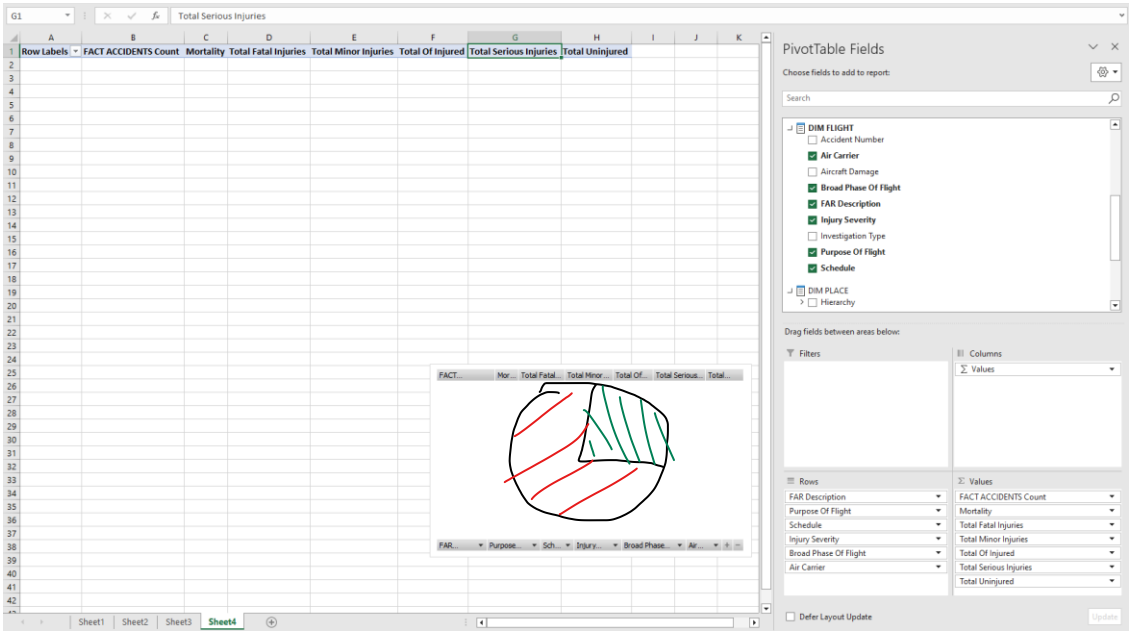
2. Analiza wypadków lotniczych w zależności od miejsca zdarzenia (kod lotniska, kraj, szerokość i długość geograficzna). Taka analiza pozwoliłaby na identyfikację obszarów, w których dochodzi najczęściej do wypadków lotniczych i pozwoliłaby na podjęcie odpowiednich działań w celu poprawy bezpieczeństwa lotów w tych obszarach. Można zaprezentować na mapie. Dla Stanów Zjednocznonych jest znaczna ilość danych, która pozwoli dokładną anazlię w tym obszarze. Wskaźniki: liczba wypadków, śmiertelność, liczba poszkodowanych: Wymiary – Według Hierarchi wymiariu DIM_PLACE.



3. Analiza zależności wypadków lotniczych od warunków pogodowych oraz od rodzaju i liczby silników. Taka analiza pozwoliłaby na zidentyfikowanie sytuacji, w których warunki pogodowe wpływają na zwiększenie ryzyka wypadków lotniczych dla konkretnych samolotów posiadających dany typ silników i pozwoliłaby na podejmowanie odpowiednich działań zapobiegawczych. Wskaźniki: śmiertelność Wymiary: Stan pogody, rodzaje silników, liczba silników.



4. Analiza badającą liczbę osób poszkodowanych w zależności od przewoźnika lotniczego, harmonogramu lotów i celu lotu. Taka analiza pozwoliłaby na zidentyfikowanie przewoźników lub typów lotów, które są bardziej narażone na wypadki lotnicze i pozwoliłaby na podejmowanie odpowiednich działań, takich jak zwiększenie nadzoru nad tymi przewoźnikami.



1.6. Źródła danych (lokalizacja, format, dostępność)

Wstępna analiza źródeł danych

Dane są dostępne na stronie kaggle:

https://www.kaggle.com/datasets/khsamaha/aviation-accident-database-synopses

Do pobrania są dostępne dwa pliki AviationData.csv oraz USState_Codes.csv.

Dane pochodzą z NTSB – National Transportation Safety Board:

https://www.ntsb.gov/

Lp.	Plik	Typ	Liczba rekordów	Rozmiar [MB]	Opis
1.	AviationData.csv	csv	88889	21.4 MB	Tabela zawierająca informacje o wypadkach, samolotach, czasie, miejscu i osobach rannych/zabitych
2.	USState_Codes.csv	csv	62	0,0009 MB	Tabela zawierająca poszczególne stany Stanów Zjednoczonych/regiony i ich kody

2. Profilowanie danych

2.1. Analiza danych

Plik: Aviation_Data.csv				
Lp.	Atrybut	Typ danych	Zakres wartości	Uwagi – ocena jakości danych
1.	Event.Id	nvarchar(50)		Indeks zdarzenia. NULL = 0.0%
2.	Investigation.Type	nvarchar(20)	{Accident, Incident}	Typ zdarzenia. NULL = 0.0 %
3.	Accident.Number	nvarchar(20)		Numer zdarzenia. NULL = 0.0 %
4.	Event.Date	DATE	24/10/1948 do 29/12/2022	Data zdarzenia. NULL = 0.0 %
5.	Location	nvarchar(100)		Lokalizacja, w której wypadek miał miejsce. NULL=0.06%
6.	Country	nvarchar(50)		Kraj, w którym wypadek miał miejsce. NULL=0.25%
7.	Latitude	nvarchar(20)		Wartość podawana w postaci dwóch notacji DMS, DD (wymaga preprecessingu). NULL = 61.33%
8.	Longitude	nvarchar(20)		Wartość podawana w postaci dwóch notacji DMS, DD (wymaga preprecessingu). NULL = 61.33%
9.	Airport.Code	nvarchar(10)		Kod lotniska, w obrębie którego miał miejsce wypadek. NULL=43.17%

10.	Airport.Name	nvarchar(100)		Nazwa lotniska, w obrębie którego miał miejsce wypadek. NULL=40.49%
11.	Injury.Severity	nvarchar(20)		Informacja na temat liczby zmarłych pasażerów. NULL=1.12%
12.	Aircraft.Damage	nvarchar(15)		Informacja na temat powagi obrażeń statku lotniczego. NULL= 3.59%
13.	Aircraft.Category	nvarchar(30)		Rodzaj statku. NULL=63.68%
14.	Registration.Number	nvarchar(15)		Numer rejestracyjny statku. NULL=1.48%
15.	Make	nvarchar(50)		Producent samolotu. NULL=0.7%
16.	Model	nvarchar(50)		Model samolotu, NULL=0.1%
17.	Amateur.Built	nvarchar(3)	{Yes, No, NULL}	Informacja na temat budowy samolotu. NULL=0.11%
18.	Number.of.Engines	int	{NULL, 0, 1, 2, 3, 4, 6, 8}	Liczba silników. NULL= 6.84%
19.	Engine.Type	nvarchar(30)		Rodzaj silników. NULL = 7.96%
20.	FAR.Description	nvarchar(200)		Opis genezy lotu. NULL = 63.97%
21.	Schedule	nvarchar(10)	{NULL, UNK, SCHD,NSCH}	Kod harmonogramu lotu. NULL = 85.85%
22.	Purpose.of.flight	nvarchar(30)		Cel lotu. NULL = 6.97%
23.	Air.carrier	nvarchar(100)		Linia lotnicza. NULL = 81.27%
24.	Total.Fatal.Injuries	int	0 - 349	Liczba zmarłych. NULL = 12.83%
25.	Total.Serious.Injuries	int	0 - 161	Liczba poważnie rannych. NULL = 14.07%
26.	Total.Minor.Injuries	int	0 - 380	Liczba lekko rannych. NULL = 13.42%
27.	Total.Uninjured	int	0 - 699	Suma niezranionych, NULL = 6.65%
28.	Weather.Condition	nvarchar(5)	{NULL, VMC, UNK, IMC}	Kod warunków pogodowych, NULL = 5.05%
29.	Broad.phase.of.flight	nvarchar(20)		Faza lotu, w której miał miejsce wypadek, NULL = 30.56%
30.	Report.Status	nvarchar(MAX)		Bardzo różne opisy powodów wypadków. NULL = 7.18%
31.	Publication.Date	nvarchar(10)	1982-01-01 do 2022-10-31	Data publikacji wypadku. Format daty zapisany w postaci ciągu znaków „YYYY-MM-DD” – wymaga pre processingu. NULL = 15.49%

Plik: USState_Codes.csv				
Lp.	Atrybut	Typ danych	Zakres wartości	Uwagi – ocena jakości danych
1.	US_State	nvarchar(20)		Kod stanu dotyczący Stanów Zjednoczonych.
2.	Abbreviation	nvarchar(2)		Rozwinięcie stanu.

2.2. Ocena przydatności danych w pliku do tworzenia hurtowni danych

Lp.	Plik	Ocena jakości danych
1.	Aviation_Data.csv	Niezbędne do utworzenia hurtowni, obejmuje lata 1948-2022. Dość duża ilość pól jest niekompletna. Głównie obejmuje region Stanów Zjednoczonych. Wymagane ujednolicenie danych w, niektórych kolumnach.
2.	USState_Codes.csv	Opcjonalne – zawiera jedynie pełne nazwy regionów lub stanów

2.3. Definicja typów encji/klas (wraz z własnościami) oraz związków pomiędzy nimi

Encje:

1. Time
    - Event.Date
2. Place
    - Country
    - Location
    - Latitude
    - Longitude
    - Airport.Code
    - Airport.Name
3. Accident
    - Accident.Number
    - Investigation.Type
    - Injury.Severity
    - Aircraft.damage
    - FAR.Description
    - Schedule
    - Purpose.of.flight
    - Air.carrier
    - Broad.phase.of.flight
    - Total.Fatal.Injuries
    - Total.Serious.Injuries
    - Total.Minor.Injuries
    - Total.Uninjured
4. Plane
    - Make
    - Model
    - Amateur.Built
    - Number.of.Engines
    - Engine.Type
    - Aircraft.Category
5. Conditions
    - Weather.Condition

Związki:

Wyznacza(ACCIDENT(0,N) : TIME(1,1))

Encja Accident musi dotyczyć tylko jednej encji Time.

Encja Time może dotyczyć wielu encji Accident.

Określa(ACCIDENT(0,N) : PLACE(1,1))

Encja Accident musi dotyczyć tylko jednej encji Place.

Encja Place może dotyczyć wielu encji Accident.

Wskazuje(ACCIDENT(1,N) : PLANE(1,1))

Encja Accident musi dotyczyć tylko jednej encji Plane.

Encja Plane może dotyczyć wielu encji Accident.

Opisuje(ACCIDENT(0,N) : CONDITIONS(1,1))

Encja Accident musi dotyczyć tylko jednej encji Conditions.

Encja Conditions może dotyczyć wielu encji Accident.

2.4. Propozycja wymiarów, hierarchii, miar (w tym nieaddytywnych)

DIM_TIME:

Id	int	PK, NOT NULL
Year	int	NOT NULL
Quarter	int	NOT NULL
Month	int	NOT NULL
Month In Words	nvarchar(10)	NOT NULL
Day	int	NOT NULL
Day In Words	nvarchar(10)	NOT NULL

DIM_PLACE:

Id	int	PK, NOT NULL
Country	nvarchar(50)	NULL
District	nvarchar(35)	NULL
Region_Name	nvarchar(15)	NULL
Latitude	decimal(18,6)	NULL
Longitude	decimal(18,6)	NULL
Airport_Code	nvarchar(10)	NULL
Airport_Name	nvarchar(100)	NULL

DIM_FLIGHT:

Accident_Number	nvarchar(20)	PK, NOT NULL
Investigation_Type	nvarchar(20)	NOT NULL
Injury_Severity	nvarchar(20)	NULL
Aircraft_damage	nvarchar(15)	NULL
FAR_Description	nvarchar(200)	NULL
Schedule	nvarchar(10)	NULL
Purpose_of_flight	nvarchar(30)	NULL
Air_carrier	nvarchar(100)	NULL
Broad_phase_of_flight	nvarchar(20)	NULL

DIM_PLANE:

Id	int	PK, NOT NULL
Make	nvarchar(50)	NULL
Model	nvarchar(50)	NULL
Amateur_Built	nvarchar(3)	NULL
Number_of_Engines	int	NULL
Engine_Type	nvarchar(30)	NULL
Aircraft_Category	nvarchar(30)	NULL

DIM_CONDITIONS:

Id	int	PK, NOT NULL
Weather_Condition	nvarchar(5)	NOT NULL
Weather_Condition_Name	nvarchar(30)	NOT NULL

FACT_ACCIDENTS:

Flight_Id	nvarchar(20)	PK, NOT NULL
Time_Id	int	NOT NULL
Place_Id	int	NOT NULL
Plane_Id	int	NOT NULL
Weather_Conditions_Id	int	NOT NULL
Total_Fatal_Injuries	int	NULL
Total_Serious.Injuries	int	NULL
Total_Minor.Injuries	int	NULL
Total_Uninjured	int	NULL
TotalOfInjured	int	NULL
Mortality	decimal(18,6)	NULL

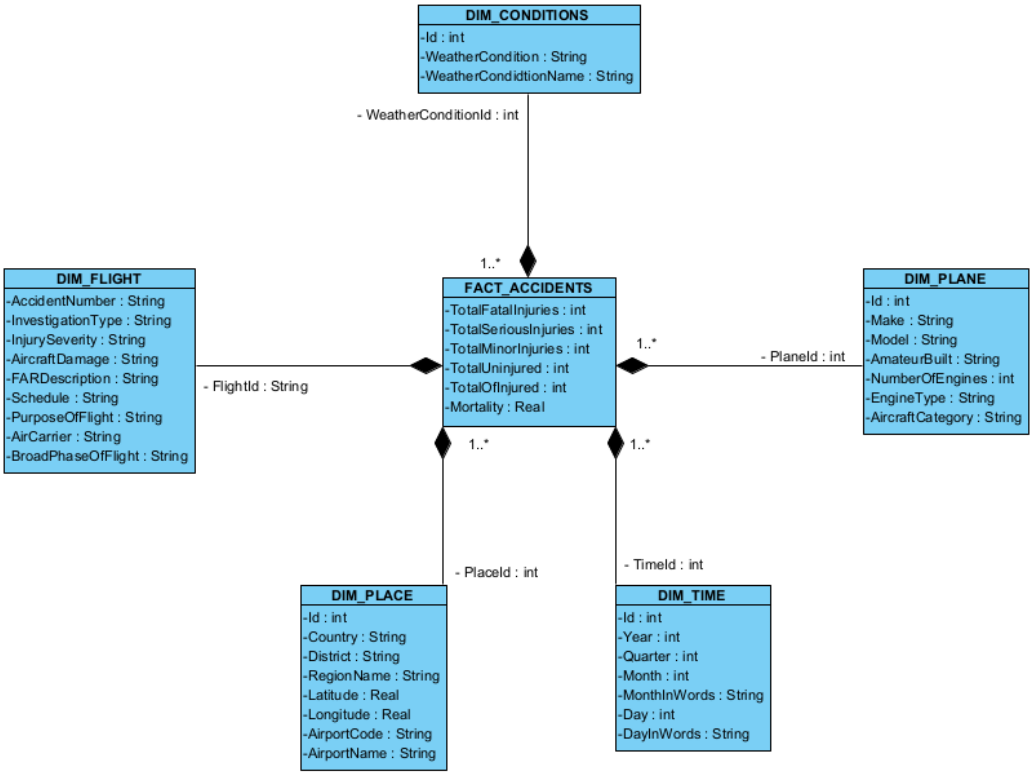
Hierarchie:

DIM_TIME: Year, Quarter, Month In Words, Day, Day In Words.

DIM_PLACE: Country, Region_Name, District

DIM_PLANE: Aircraft_Category, Engine_Type, Number_Of_Engines, Make, Model

2.5. Diagram klas – model danych utworzony na podstawie danych zgromadzonych w plikach



3. Utworzyć bazę danych zgodnie z zaproponowanym konceptualnym modelem danych (p. 2.3. i 2.4.)

```
CREATE TABLE DIM_TIME
(
	Id INT PRIMARY KEY,
	"Year" INT NOT NULL,
	"Quarter" INT NOT NULL,
	"Month" INT NOT NULL,
	"Month In Words" NVARCHAR(10) NOT NULL,
	"Day" INT NOT NULL,
	"Day In Words" NVARCHAR(10) NOT NULL
);

CREATE TABLE DIM_PLACE
(
	Id INT PRIMARY KEY,
	Country NVARCHAR(50) NOT NULL,
	District NVARCHAR(35) NULL,
	Region_Name NVARCHAR(15) NULL,
	Latitude DECIMAL(18,6) NULL,
	Longitude DECIMAL(18,6) NULL,
	Airport_Code NVARCHAR(10) NULL,
	Airport_Name NVARCHAR(100) NULL
);

CREATE TABLE DIM_CONDITIONS
(
	Id INT PRIMARY KEY,
	Weather_Condition NVARCHAR(5) NOT NULL,
	Weather_Condition_Name NVARCHAR(30) NOT NULL
);
```

```
CREATE TABLE DIM_PLANE
(
	Id INT PRIMARY KEY,
	Make NVARCHAR(50) NULL,
	Model NVARCHAR(50) NULL,
	Amateur_Built NVARCHAR(3) NULL,
	Number_Of_Engines INT NULL,
	Engine_Type NVARCHAR(30),
	Aircraft_Category NVARCHAR(30)
);

CREATE TABLE DIM_FLIGHT
(
	Accident_Number NVARCHAR(20) PRIMARY KEY,
	Investigation_Type NVARCHAR(20) NULL,
	Injury_Severity NVARCHAR(20) NULL,
	Aircraft_damage NVARCHAR(15) NULL,
	FAR_Description NVARCHAR(200) NULL,
	Schedule NVARCHAR(10) NULL,
	Purpose_of_flight NVARCHAR(30) NULL,
	Air_Carrier NVARCHAR(100) NULL,
	Broad_phase_of_flight NVARCHAR(20) NULL
);


CREATE TABLE FACT_ACCIDENTS
(
	Flight_Id NVARCHAR(20) PRIMARY KEY,
	Time_Id INT NOT NULL,
	Place_Id INT NOT NULL,
	Plane_Id INT NOT NULL,
	Weather_Conditions_Id INT NOT NULL,
	Total_Fatal_Injuries INT NULL,
	Total_Serious_Injuries INT NULL,
	Total_Minor_Injuries INT NULL,
	Total_Uninjured INT NULL,
	TotalOfInjured INT NULL,
	Mortality DECIMAL(18,6) NULL
);

ALTER TABLE FACT_ACCIDENTS
	ADD CONSTRAINT CONDITIONS_FOREIGN_KEY FOREIGN KEY(Weather_Conditions_Id)
REFERENCES DIM_CONDITIONS(Id),
		CONSTRAINT ACCIDENT_FOREIGN_KEY FOREIGN KEY(Flight_Id) REFERENCES
DIM_FLIGHT(Accident_Number),
		CONSTRAINT PLACE_FOREIGN_KEY FOREIGN KEY(Place_Id) REFERENCES
DIM_PLACE(Id),
		CONSTRAINT PLANE_FOREIGN_KEY FOREIGN KEY(Plane_Id) REFERENCES
DIM_PLANE(Id),
		CONSTRAINT EVENT_DATE_FOREIGN_KEY FOREIGN KEY(Time_Id) REFERENCES
DIM_TIME(Id);
```

Wnioski:

Pierwszy etap projekut okazał się być zaskakująco wymagający, być może dlatego, iż jest to najważniejszy etap projektowania hurtownii danych, który już na samym początku określa cel, problem, potrzebę dogłębnej analizy danych.

Wybrany przez nas zbiór danych, zaczerpnięty z rządowej strony, przedstawiał się jako godny zaufania i będący na najwyższym poziomie zestaw danych. Jednakże po przyjrzeniu się wartościom atrybutów, okazało się, że mamy doczynienia z dużą niekompletnością danych.

Po zdefiniowaniu typów encji oraz związków, doszliśmy do wniosku, iż rozsądnym krokiem będzie rozdzielenie encji Accident na wymiar DIM_FLIGHT oraz na tabelę faktów FACT_ACCIDENTS, ponieważ pozwoli to nam na dodatkowe filtrowanie po uzyskanych wynikach.

Analizując poszczególne atrybuty natknęliśmy się na potrzebę przetworzenia niektórych wartości atrybutów. Zauważyliśmy niespójności co do formatu zapisanych wartości dla szerokości i długości geograficznych, niektóre zapisane w formacie DMS – Degrees Minutes Seconds, inne w DD – Decimal Degrees. Dodatakowo zauważliśmy, inny format zapisu daty w atrybucie dotyczącym publikacji wypadku. W celu późniejszego wytworzenia hierarchii ze względu na lokalizację wypadków, podzieliliśmy atrybut Location na nazwę okręgu oraz kod jego stanu, ponieważ zauważyliśmy, że kod danego stanu określa większy obszar niż nazwa okręgu.

Projekt – etap II

Proces ETL

1. Utworzone tabele w poprzednim punkcie wypełnić danymi zgodnie z ustalonymi założeniami projektowymi wykorzystując zapytania SQL lub inne narzędzia dostępne w Integration Services.

Przy ocenie będą brane następujące elementy pakietu(ów):

- właściwa struktura procesu ETL (odpowiednie rozbicie procesu ETL na zadania/pakiety, dobrze dobrane nazwy poszczególnych zadań, wprowadzona automatyzacja, obsługa błędów, itp.)
- stabilność i prawidłowe, bezbłędne wykonanie
- złożoność przeprowadzonych operacji. Przykładowo, jeżeli dane źródłowe już są w pełni zdenormalizowane proszę nie spodziewać się maksymalnej liczby punktów za ten element
- dokumentacja powinna zawierać krótki opis dotyczący każdego zadania, które pozwoli zorientować się, jaki jest jego cel (np. zadanie Z kopiuje dane z tabeli X i Y do tabeli T dokonując denormalizacji) oraz mapę logiczną procesu ETL.

Wnioski:

Projekt – etap III

Kostka:

1. Przygotować projekt kostki, edytować wymiary, dodać miary kalkulowane. Przygotować zestawienia z p. 1.5.2. oraz pokazać inne ciekawe zależności w analizowanych danych (analiza w głąb, a nie tylko tabele przestawne).

Przy ocenie będą brane następujące elementy kostki:

- prawidłowa struktura kostki – model kostki powinien analitykowi na intuicyjne i łatwe korzystanie z danych
- miary kalkulowane
- dokumentacja, która powinna zawierać krótki opis wszystkich wymiarów, wszystkich ich atrybutów oraz wszystkich miar

Wnioski:

Projekt – etap IV

Prezentacja

Prezentacja powinna zawierać 4-8 slajdów (trwać ok. 8 minut) i wyjaśniać jakie dane są przedmiotem analizy. Prezentacja powinna być zakończona, krótką demonstracją, która pokaże najciekawsze związki między danymi znajdującymi się w kostce.

Uwaga. Projekt będzie ostatecznie zaliczony po złożeniu pisemnego sprawozdania zawierającego opisy poszczególnych etapów pracy.