HURTOWNIE DANYCH

Projekt

Maciej Kopiński 254578

Projekt – etap I (28.04./10.05.2022 r.)

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 – **Baza danych i streszczeń wypadków lotniczych (Aviation Accident Database & Synopses)**

1.2. Charakterystyka dziedziny problemowej

Dziedzina problemowa opisuje wypadki lotnicze od roku 1948 do 2007 oraz udostępnia informacje o nich, takie jak m.in.: data zdarzenia, kraj zdarzenia, szerokość oraz długość geograficzną, informacje o samolocie, linii lotniczej, celu lotu bądź etapu lotu, w którym nastąpił wypadek, czy najważniejsze – informacje o rannych czy zabitych osobach.

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. Analizie zostanie poddana także zależność liczby osób rannych/zabitych w zależności od producenta samolotów, czy fazy lotu. Dodatkowo będzie można sprawdzić zależności pomiędzy informacjami o wypadku, a datą opublikowania informacji o nim.

1.4. Problemy i potrzeby

Urząd lotnictwa cywilnego dokonuje co roku statystyk wypadków lotniczych. W tym celu urząd potrzebuje danych odnośnie wypadków w celu kategoryzowania wypadków dla różnych danych – typu faza lotu, czy napędu samolotu. Potrzebuje także statystyk na tle innych państw.

1.5. Cel przedsięwzięcia

1.5.1. Oczekiwania

Wyznaczenie zależności pomiędzy warunkami lotu, a częstotliwością wypadków i liczbą i stopniem uszkodzeń samolotów lub ubytków na zdrowiu osób uczestniczących w locie.

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

-Liczba wypadków w zależności od miesiąca i warunków pogodowych,

-Liczba wypadków w zależności od miejsca i fazy lotu,

-Śmiertelność w zależności od warunków pogodowych i kategorii samolotu,

-Śmiertelność w zależności od fazy lotu i rodzaju silnika,

-Śmiertelność w zależności od liczby silników i warunków pogodowych,

-Liczba ofiar śmiertelnych dla miesięcy i obrażeń samolotu,

-Liczba ofiar śmiertelnych dla producenta samolotu i warunków pogodowych,

-Liczba osób rannych dla linii lotniczej i uszkodzeń samolotu,

-Maksymalna śmiertelność dla warunków pogodowych i kategorii samolotu,

-Maksymalna liczba osób zmarłych dla calu lotu i budowy samolotu.

1.6. Źródła danych (lokalizacja, format, dostępność)

Wstępna analiza źródeł danych

Lp.	Plik	Typ	Liczba rekordów	Rozmiar [MB]	Opis
1.	Aviation_Data	csv	62231	13,29	Tabela zawierająca informacje o wypadkach, samolotach, czasie, miejscu i osobach rannych/zabitych
2.	USState_Codes	csv	62	0,001	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(14)		
2.	Investigation_Type	nvarchar(50)		
3.	Accident_Number	nvarchar(50)		
4.	Event_Date	date	(1948-10-24) – (2007-04-15)	
5.	Location	nvarchar(70)		
6.	Country	nvarchar(50)		
7.	Latitude	decimal(18,6)	(-77.83, 86.94)	
8.	Longitude	decimal(18,6)	(-173.24, 177.56)	
9.	Airport_Code	nvarchar(10)		
10.	Airport_Name	nvarchar(100)		
11.	Injury_Severity	nvarchar(10)		
12.	Aircraft_damage	nvarchar(15)		
13.	Aircrafr_Category	nvarchar(15)		
14.	Registration_Number	nvarchar(15)		
15.	Make	nvarchar(50)		
16.	Model	nvarchar(50)		
17.	Amateur_Built	nvarchar(4)	{Yes, No}	
18.	Number_of_Engines	int	{0, 1, 2, 3, 4}	
19.	Engine_Type	nvarchar(50)		
20.	FAR_Description	nvarchar(200)		
21.	Schedule	nvarchar(10)		
22.	Purpose_of_flight	nvarchar(15)		
23.	Air_carrier	nvarchar(100)		
24.	Total_Fatal_Injuries	int	0 - 349	
25.	Total_Serious_Injuries	int	0 - 106	
26.	Total_Minor_Injuries	int	0 - 380	
27.	Total_Uninjured	int	0 - 699	
28.	Weather_Condition	nvarchar(5)	{VMC, UNK, IMC}	
29.	Broad_phase_of_flight	nvarchar(20)		
30.	Report_Status	nvarchar(50)	{Probable Cause, Factual, Prieliminary, Foreign}	
31.	Publication_Date	date	(1980-04-16) – (2020-02-27)	

Plik: USState_Codes.csv				
Lp.	Atrybut	Typ danych	Zakres wartości	Uwagi – ocena jakości danych
1.	US_State	nvarchar(50)		
2.	Abbreviation	nvarchar(10)		

2.2. Ocena przydatności danych w pliku do tworzenia hurtowni danych

Lp.	Plik	Ocena jakości danych
1.	Aviation_Data.csv	Niezbędne do utworzenia hurtowni
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

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(15)	NOT NULL
Day	int	NOT NULL
Day In Words	nvarchar(15)	NOT NULL

DIM_PLACE:

Id	int	PK, NOT NULL
Country	nvarchar(15)	NOT NULL
Region	nvarchar(15)	NULL
Region_Code	nvarchar(5)	NULL
Latitude	decimal(18,6)	NULL
Longitude	decimal(18,6)	NULL
Airport_Code	nvarchar(10)	NULL
Airport_Name	nvarchar(100)	NULL

DIM_ACCIDENT:

Id	nvarchar(14)	PK, NOT NULL
Accident_Number	nvarchar(50)	NOT NULL
Investigation_Type	nvarchar(50)	NOT NULL
Injury_Severity	nvarchar(10)	NULL
Aircraft_damage	nvarchar(15)	NULL
FAR_Description	nvarchar(200)	NULL
Schedule	nvarchar(10)	NULL
Purpose_of_flight	nvarchar(15)	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_Build	nvarchar(3)	NULL
Number_of_Engines	int	NULL
Engine_Type	nvarchar(50)	NULL
Aircraft_Category	nvarchar(15)	NULL

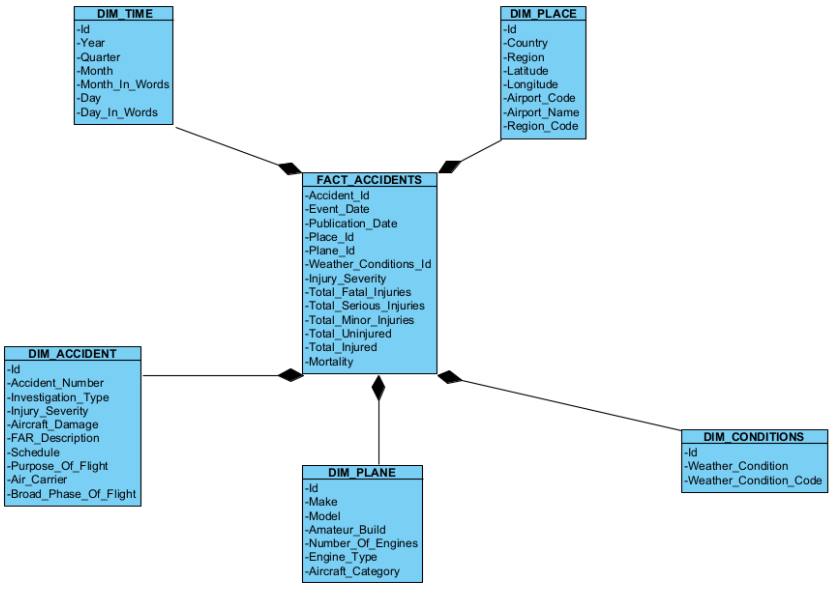
DIM_CONDITIONS:

Id	int	PK, NOT NULL
Weather_Condition	nvarchar(16)	NOT NULL
Weather_Condition_Code	nvarchar(10)	NOT NULL

FACT_ACCIDENTS:

Accident_Id	int	NOT NULL
Event_Date	int	NOT NULL
Publication_Date	int	NULL
Place_Id	int	NOT NULL
Plane_Id	int	NOT NULL
Weather_Conditions_Id	int	NOT NULL
Injury_Severity	nvarchar(16)	NOT NULL
Total_Fatal_Injuries	int	NOT NULL
Total_Serious_injuries	int	NOT NULL
Total_Minor_Injuries	int	NOT NULL
Total_Uninjured	int	NOT NULL
Total_Injured	int	NOT NULL
Mortality	decimal(18,6)	NOT NULL

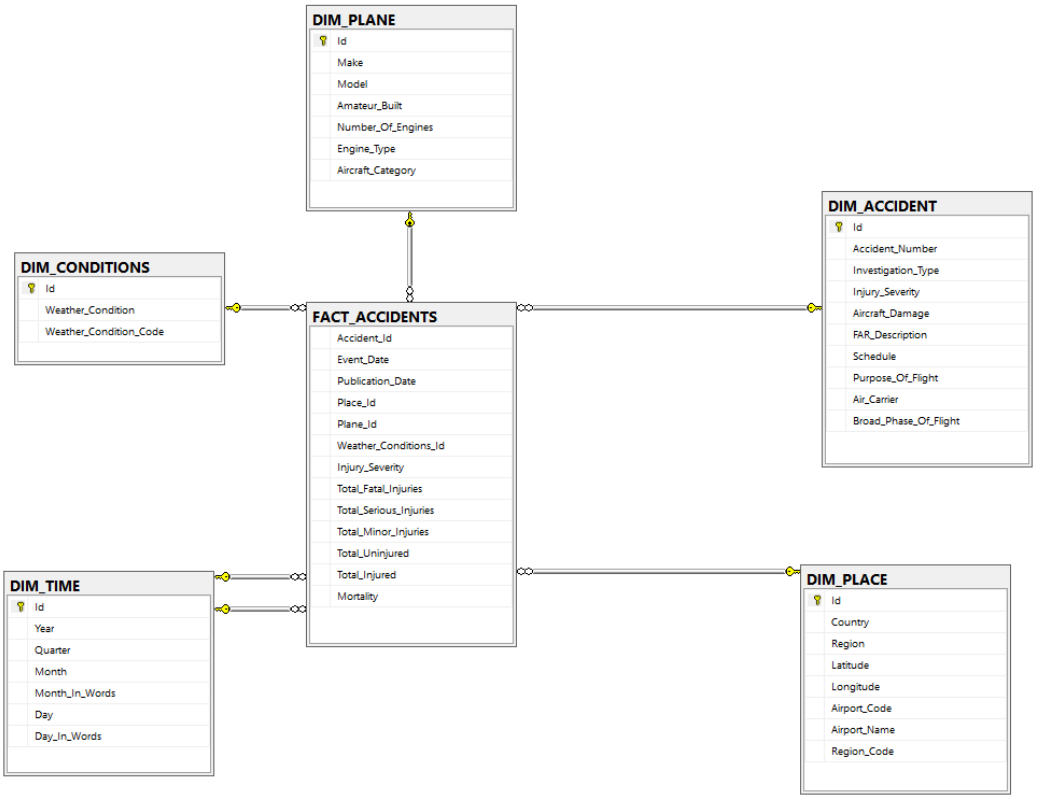
2.5. Diagram klas – model danych utworzony na podstawie danych zgromadzonych w plikach



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(15) NOT NULL,
	"Day" INT NOT NULL,
	"Day_In_Words" NVARCHAR(15) NOT NULL
);

CREATE TABLE DIM_PLACE
(
	Id INT PRIMARY KEY,
	Country NVARCHAR(15) NOT NULL,
	Region 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(16) NOT NULL,
	Weather_Condition_Code NVARCHAR(10) NOT NULL
);

CREATE TABLE DIM_PLANE
(
	Id INT PRIMARY KEY,
	Make NVARCHAR(50) NULL,
	Model NVARCHAR(50) NULL,
	Amateur_Built NVARCHAR(50) NULL,
	Number_Of_Engines INT NULL,
	Engine_Type NVARCHAR(50),
	Aircraft_Category NVARCHAR(15)
);

CREATE TABLE DIM_ACCIDENT
(
	Id INT PRIMARY KEY,
	Accident_Number NVARCHAR(50) NOT NULL,
	Investigation_Type NVARCHAR(50) NOT NULL,
	Injury_Severity NVARCHAR(10) NULL,
	Aircraft_Damage NVARCHAR(15) NULL,
	FAR_Description NVARCHAR(200) NULL,
	Schedule NVARCHAR(10) NULL,
	Purpose_Of_Flight NVARCHAR(15) NULL,
	Air_Carrier NVARCHAR(100) NULL,
	Broad_Phase_Of_Flight NVARCHAR(20) NULL
);
```

```sql
CREATE TABLE FACT_ACCIDENTS
(
	Accident_Id INT NOT NULL,
	Event_Date INT NOT NULL,
	Publication_Date INT NULL,
	Place_Id INT NOT NULL,
	Plane_Id INT NOT NULL,
	Weather_Conditions_Id INT NOT NULL,
	Injury_Severity NVARCHAR(10) NOT NULL,
	Total_Fatal_Injuries INT NOT NULL,
	Total_Serious_Injuries INT NOT NULL,
	Total_Minor_Injuries INT NOT NULL,
	Total_Uninjured INT NOT NULL,
	Total_Injured INT NOT NULL,
	Mortality DECIMAL(18,6) NOT NULL
);

ALTER TABLE DIM_PLACE
ADD Region_Code NVARCHAR(5) 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(Accident_Id) REFERENCES
DIM_ACCIDENT(Id),
		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(Event_Date) REFERENCES
DIM_TIME(Id),
		CONSTRAINT PUBLICATION_DATE_FOREIGN_KEY FOREIGN KEY(Publication_Date)
REFERENCES DIM_TIME(Id);
```



Wnioski:

Pierwszy etap projektu jest najprawdopodobniej najtrudniejszym etapem, ze względu na wybranie zbioru danych, w którym moglibyśmy znaleźć przynajmniej 5 wymiarów i przynajmniej 3 miary, z czego 1 nieaddytywną. W przypadku zbioru „Aviation Accident Database & Synopses" duża część danych okazała się uszkodzona/problematyczna, ponieważ pola były oddzielone przecinkami, które występowały także w wartościach jednej z kolumn. Innym problemem były także dodatkowe znaki ukryte w pliku, których nie dostrzegłem wcześniej. Na szczęście błędy te występowały relatywnie pod koniec, więc większość zbioru została zachowana.

Jeśli chodzi o dobór wymiarów – myślę, że wyjaśnienia może wymagać jedynie tabela DIM_ACCIDENT – powstała ona, ponieważ wiele z cech wypadku było ściśle związane z wypadkiem i atrybuty z tabeli DIM_ACCIDENT pozwalają na filtrowanie po zbiorze danych. Ewentualną zamianą mogłoby być utworzenie tabeli DIM_FLIGHT, w której znalazłoby się większość atrybutów z DIM_ACCIDENT.