**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. 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 |

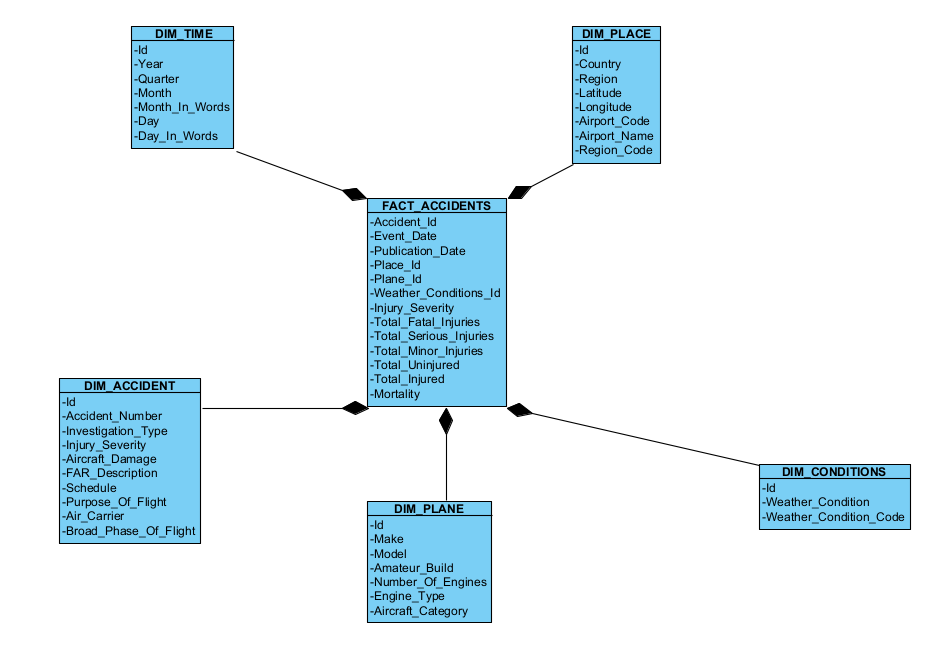
DIM\_CONDITIONS:

|  |  |  |
| --- | --- | --- |
| 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

);

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),

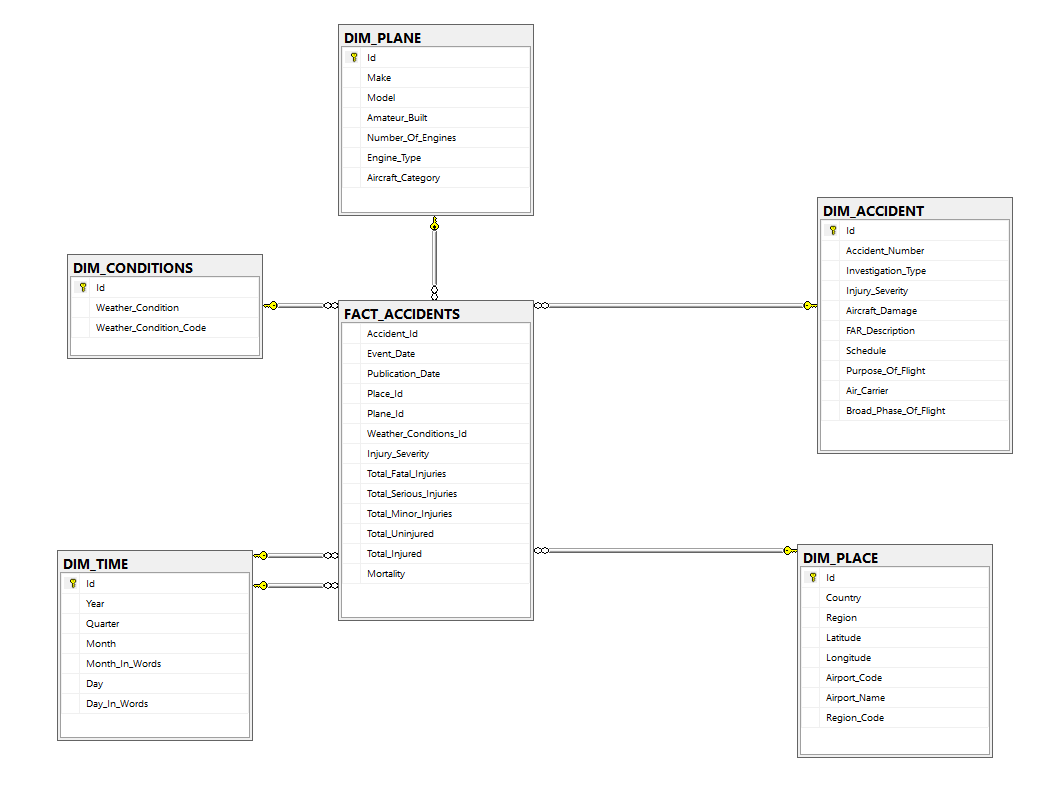
CONSTRAINT ACCIDENT\_FOREIGN\_KEY FOREIGN KEY(Accident\_Id) REFERENCES DIM\_ACCIDENT(Id),

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.