**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 (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)**

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 | int | NOT NULL |
| Region | int | 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 |

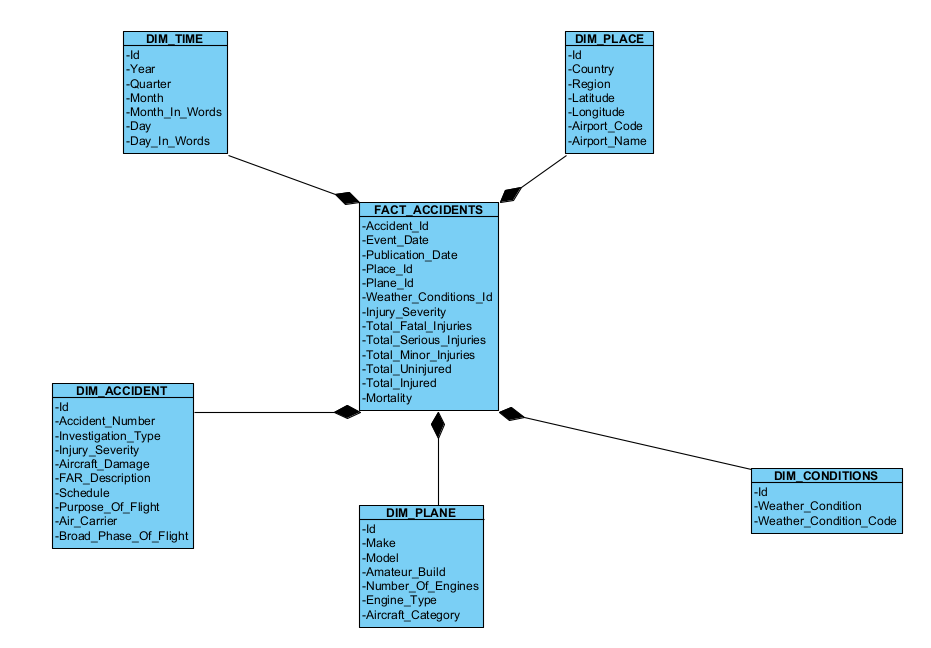
DIM\_CONDITIONS:

|  |  |  |
| --- | --- | --- |
| Id | int | PK, NOT NULL |
| Weather\_Condition | nvarchar | NOT NULL |
| Weather\_Condition\_Code | nvarchar | NOT NULL |

FACT\_ACCIDENTS:

|  |  |  |
| --- | --- | --- |
| Accident\_Id | int | NOT NULL |
| Event\_Date | int | NOT NULL |
| Publication\_Date | int | NOT NULL |
| Place\_Id | int | NOT NULL |
| Plane\_Id | int | NOT NULL |
| Weather\_Conditions\_Id | int | NOT NULL |
| Injury\_Severity | nvarchar | 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 | 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.)

**Wnioski:**

**Projekt – etap II (12.05./24.05.2022 r.)**

**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 (26.05./07.06.2022 r.)**

**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 (02.06./14.06.2022 r.)**

**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.**