**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**

1.1. Tytuł projektu

1.2. Charakterystyka dziedziny problemowej

1.3. Krótki opis obszaru analizy

1.4. Problemy i potrzeby

1.5. Cel przedsięwzięcia

1.5.1. Oczekiwania

**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. |  |  |  |  |  |
| 2. |  |  |  |  |  |
| … |  |  |  |  |  |

2. Profilowanie danych

2.1. Analiza danych

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Plik: | | | | |
| Lp. | Atrybut | Typ danych | Zakres wartości | Uwagi – ocena jakości danych |
| 1. |  |  |  |  |
| 2. |  |  |  |  |

2.2. Ocena przydatności danych w pliku do tworzenia hurtowni danych

|  |  |  |
| --- | --- | --- |
| Lp. | Plik | Ocena jakości danych |
| 1. |  |  |
| 2. |  |  |
| … |  |  |

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

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

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.**