Hurtownie Danych - laboratorium
Lista 3

Podstawy SQL: funkcje grupujące i okienkowe

Wstęp teoretyczny

Funkcje grupujące

Klauzula instrukcji `SELECT` dzieli wynik zapytania na grupy wierszy, zwykle w celu wykonania jednej lub więcej agregacji dla każdej grupy. Polecenie `SELECT – GROUP BY` zwraca jeden wiersz wyniku na grupę.
Stosuje się następującą składnię polecenia `GROUP BY`:

```
SELECT ...
FROM ...
WHERE ...
GROUP BY {
      column-expression
    | ROLLUP ( <group_by_expression> [ ,...n ] )
    | CUBE ( <group_by_expression> [ ,...n ] )
    | GROUPING SETS ( <grouping_set> [ ,...n ]  )
    | () --calculates the grand total
} [ ,...n ]
```

Różnice pomiędzy `ROLLUP`, `CUBE` i `GROUPING SETS`:
GROUP BY ROLLUP

GROUP BY ROLLUP (col_1, col_2, col_3, col_4, ..., col_n)
Wynik – podsumowania dla następujących zestawień kolumn:

- col_1, col_2, ..., col_n
- col_1, col_2, ..., col_{n-1}, NULL
- col_1, ..., col_{n-2}, NULL, NULL
- ...
- col_1, col_2, NULL, ..., NULL
- col_1, NULL, NULL, ..., NULL
- NULL, NULL, ..., NULL

GROUP BY CUBE

GROUP BY CUBE (col_1, col_2, col_3, col_4, ..., col_n)
Wynik – podsumowania dla wszystkich możliwych kombinacji kolumn
Ile jest możliwych podzbiorów zbioru n-elementowego?

GROUP BY GROUPING SETS

GROUP BY GROUPING SETS (...)
Wynik – podsumowania dla wszystkich wymienionych kombinacji kolumn

Funkcje okienkowe

Klauzula OVER określa podział i kolejność zestawu wierszy przed zastosowaniem powiązanej funkcji okna. Klauzula OVER definiuje okno (określony przez użytkownika zestaw wierszy), a następnie wylicza wartość dla każdego wiersza w oknie.

```
OVER (
       [ <PARTITION BY clause> ]
       [ <ORDER BY clause> ]
       [ <ROW or RANGE clause> ]
      )
```

Klauzulę OVER stosuje się do obliczania wartości zagregowanych, np. średnie ruchome, sumy bieżące, wyniki dla ostatnich *n* wierszy w każdej grupie.

PARTITION BY clause
Dzieli zbiór wyników zapytania na określone podgrupy. Funkcja okna jest stosowana do każdej podgrupy osobno i obliczenia uruchamiane są ponownie dla każdej podgrupy.

ORDER BY clause
Definiuje logiczną kolejność wierszy w każdej podgrupie wynikowego zbioru. Dla każdej podgrupy określa logiczną kolejność wykonywania obliczeń funkcji okna.

ROW or RANGE clause
Ograniczają wiersze w partycji, określając granice:

- ROWS – fizyczne ograniczenie liczby wierszy, np.
  - BETWEEN 2 PRECEEDING AND CURRENT ROW
- RANGE – logiczne ograniczenie liczby wierszy, np.
  - BETWEEN CURRENT ROW AND UNBOUNDED FOLLOWING

Obie klauzule wymagają klauzuli ORDER BY i na podstawie ustalonej kolejności determinują wynik.

Funkcje szeregujące pozwalają tworzyć ranking, przypisując odpowiednie miejsca kolejnym wynikom. Najczęściej stosowane są następujące funkcje szeregujące:

- **ROW_NUMBER** – nr pozycji
- **RANK** – ranking (to samo miejsce dla tych samych wartości)
- **DENSE_RANK** – ranking, numerowanie ciągłe
- **NTILE** – grupuje rekordy poprzez przypisanie tej samej wartości szeregującej członkom grupy

Źródła:
https://docs.microsoft.com/en-us/sql/t-sql/queries/select-group-by-transact-sql?view=sql-server-ver15
https://docs.microsoft.com/en-us/sql/t-sql/queries/select-over-clause-transact-sql?view=sql-server-ver15

Zadania do wykonania

Baza danych: **AdventureWorks**

Zad. 1. Wykorzystanie funkcji grupujących (rollup, cube, grouping sets)

1. Przygotować zestawienie przedstawiające, ile pieniędzy wydali klienci na zamówienia na przestrzeni poszczególnych lat.
   Wykonaj zestawienie przy użyciu poleceń rollup, cube, grouping sets.

	Klient	Rok	Kwota
1		2013	48965887,9632
2		2014	22419498,3157
3		2011	14155699,525
4		2012	37675700,312
5			123216786,1159
6	A. Leonetti	2013	1814,1819
7	A. Leonetti	2014	1586,6583
8	A. Leonetti		3400,8402
9	Aaron Adams	2013	130,3458
10	Aaron Adams		130,3458

Przykładowe rozwiązanie

2. Przygotować zestawienie przedstawiające łączną kwotę zniżek z podziałem na kategorie, produkty oraz lata.

	Kategoria	Produkt	Rok	Kwota
1	Accessories	All-Purpose Bike Stand	2013	0.00
2	Accessories	All-Purpose Bike Stand	2014	0.00
3	Accessories	All-Purpose Bike Stand		0.00
4	Accessories	Bike Wash - Dissolver	2013	79.92
5	Accessories	Bike Wash - Dissolver	2014	26.94
6	Accessories	Bike Wash - Dissolver		106.87
7	Accessories	Cable Lock	2012	19.89
8	Accessories	Cable Lock	2013	3.41
9	Accessories	Cable Lock		23.30

Przykładowe rozwiązanie

Zad. 2. Wykorzystanie funkcji okienkowych (over, over partition by, row_number, rank, dense_rank, ntile)

1. Dla kategorii ‘Bikes’ przygotuj zestawienie prezentujące procentowy udział kwot sprzedaży produktów tej kategorii w poszczególnych latach w stosunku do łącznej kwoty sprzedaży dla tej kategorii. W zadaniu wykorzystaj funkcje okna.

	Nazwa	Rok	Procent
1	Bikes	2011	7.94
2	Bikes	2012	34.51
3	Bikes	2013	42.57
4	Bikes	2014	14.98
5			100.00

Przykładowe rozwiązanie

wykonaj podobne zestawienia dla pozostałych kategorii.

2. Przygotuj zestawienie dla sprzedawców z podziałem na lata i miesiące prezentujące liczbę obsłużonych przez nich zamówień w ciągu roku, w ciągu roku narastająco oraz sumarycznie w obecnym i poprzednim miesiącu. W zadaniu wykorzystaj funkcje okna.

	Imię i nazwisko	Rok	Miesiąc	W miesiącu	W roku	W roku narastająco	Obecny i poprzedni miesiąc
1	Amy Alberts	2012	6	3	7	3	3
2	Amy Alberts	2012	9	2	7	5	5
3	Amy Alberts	2012	12	2	7	7	4
4	Amy Alberts	2013	1	1	29	1	1
5	Amy Alberts	2013	2	1	29	2	2
6	Amy Alberts	2013	3	1	29	3	2

Przykładowe rozwiązanie

3. Przygotuj ranking klientów w zależności od liczby zakupionych produktów. Porównaj rozwiązania uzyskane przez funkcje `rank` i `dense_rank`.

4. Przygotuj ranking produktów w zależności od średniej liczby sprzedanych sztuk. Wyróżnij 3 (prawie równoliczne) grupy produktów: sprzedających się najlepiej, średnio i najsłabiej.

Zad. 3. Ocena jakości danych – profilowanie danych

1. Przeanalizować, scharakteryzować i ocenić dane znajdujące się w pliku *dane_lista3.csv* wykorzystując wybrane oprogramowanie, np. Python, R, Matlab, Integration Services Project w Visual Studio, itp.

2. W przypadku wyboru Visual Studio należy utworzyć projekt *Integration Services Project*. Z menu bocznego *SSIS Toolbox* należy:
   a. wybrać bloczek *Data Profiling Task* (przeciągnąć na kanwę projektu),
   b. określić dane w sekcji *Destination*,
   c. skonfigurować *Quick Profile*,
   d. uruchomić pakiet (Run);
   e. aby obejrzeć wynik należy użyć *Data Profile Viewer* lub ponownie edytując bloczek *Data Profiling Task* użyć opcji *Open Profile Viewer*.

Rozwiązania:

Wnioski: