Igor Józefowicz 193257, Patryk Sowiński-Toczek 191711

Optymalizacja hurtowni danych - raport

1. Cel laboratorium

Celem laboratorium jest zbadanie wydajności różnych modeli hurtowni danych i ukazanie problemów i różnic dotyczących fizycznych modeli kostki oraz agregatów.

2. Założenia wstępne

Wielkość bazy danych (hurtowni danych): 4’048 MB

Liczba wierszy w tabeli faktów, względem której przeprowadzane są testy (fact_attendance): 1’068’000

Środowisko testowe

- Urządzenie - laptop ASUS TUF Gaming F15
- Procesor
    - 12th Gen Intel(R) Core(TM) i5-12500H
    - Szybkość podstawowa: 2,50 GHz (4,5 GHz w trybie Turbo)
    - Gniazda: 1
    - Rdzenie: 12
    - Procesory logiczne: 16
    - Pamięć podręczna poziomu 1: 1,1 MB
    - Pamięć podręczna poziomu 2: 9,0 MB
    - Pamięć podręczna poziomu 3: 18,0 MB
- RAM - 32 GB
    - średnie zajęcie pamięci operacyjnej podczas testów: 67%
- Dysk C (na której pracuje hurtownia) - 512 GB
- System operacyjny: Windows 11 Pro, 23H2
- Uruchomione aplikacje podczas testów:
    - Visual Studio Community, tytani data warehouse project
    - SQL Server Management Studio 20
    - SQL Server Profiler
    - Discord
    - Google Chrome

3. Przeprowadzenie testów

Przeprowadzenie testów czasu wykonania różnych zapytań dla różnych modeli hurtowni danych, z uwzględnieniem i bez uwzględnienia zdefiniowanych agregacji, oraz przeprowadzenie testów czasu procesowania kostek.

Krótki opis zapytań wykonywanych w ramach przeprowadzania testów:

Igor Józefowicz 193257, Patryk Sowiński-Toczek 191711

1. Jakie oceny wystawili uczniowie; TOPCOUNT 10

```
SELECT
  NON EMPTY { [Measures].[AVG_Rating] } ON COLUMNS,
  NON EMPTY
    TOPCOUNT(
      {
        [Students].[Id].[Id].ALLMEMBERS
        * [Students].[First Name].[First Name].ALLMEMBERS
        * [Students].[Last Name].[Last Name].ALLMEMBERS
      },

      10,
      [Measures].[AVG_Rating]
    )
    DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME
ON ROWS
FROM [Tytani]
CELL PROPERTIES VALUE, BACK_COLOR, FORE_COLOR, FORMATTED_VALUE, FORMAT_STRING,
FONT_NAME, FONT_SIZE, FONT_FLAGS
```

2. Liczba obecnych, nieobecnych i spóźnionych uczniów w danym roku i miesiącu w danej lokalizacji

```
SELECT NON EMPTY { [Measures].[Fact Attendance Count] } ON COLUMNS, NON EMPTY {
([Junk].[Status].[Status].ALLMEMBERS * [Locations].[City].[City].ALLMEMBERS *
[Locations].[Name].[Name].ALLMEMBERS * [Date].[Month].[Month].ALLMEMBERS * [Date].
[Day].[Day].ALLMEMBERS ) } DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME ON
ROWS FROM [Tytani] CELL PROPERTIES VALUE, BACK_COLOR, FORE_COLOR, FORMATTED_VALUE,
FORMAT_STRING, FONT_NAME, FONT_SIZE, FONT_FLAGS
```

3. Zestawienie tematów i ich średnich ocen wystawianych przez uczniów

```
 SELECT NON EMPTY { [Measures].[AVG_Rating] } ON COLUMNS, NON EMPTY { ([Subjects].
[Name].[Name].ALLMEMBERS ) } DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME ON
ROWS FROM [Tytani] CELL PROPERTIES VALUE, BACK_COLOR, FORE_COLOR, FORMATTED_VALUE,
FORMAT_STRING, FONT_NAME, FONT_SIZE, FONT_FLAGS
```

Procesowanie kostki [ms]		**Zapytanie 1. [ms]**		**Zapytanie 2. [ms]**		**Zapytanie 3. [ms]**	
MOLAP	**ROLAP**	**MOLAP**	**ROLAP**	**MOLAP**	**ROLAP**	**MOLAP**	**ROLAP**
3595	1726	20	82	49	114	15	73
3920	1739	23	93	51	166	13	82
4281	1744	16	88	50	119	11	71
3489	1816	19	70	50	160	15	85
3827	1527	15	82	52	120	16	76
3649	1847	17	71	51	178	19	87
3792	1793	15	87	62	165	12	93
3651	1860	15	79	55	145	26	92
3633	1729	19	72	60	128	14	92

3818	1624	14	71	45	111	16	75

Procesowanie kostki [ms]		**Zapytanie 1. [ms]**		**Zapytanie 2. [ms]**		**Zapytanie 3. [ms]**	
MOLAP	**ROLAP**	**MOLAP**	**ROLAP**	**MOLAP**	**ROLAP**	**MOLAP**	**ROLAP**
3832	x	6	x	56	x	5	x
3897	x	5	x	54	x	5	x
4447	x	7	x	57	x	4	x
3870	x	7	x	55	x	5	x
4065	x	6	x	56	x	4	x
3974	x	7	x	58	x	3	x
4104	x	5	x	51	x	5	x
4376	x	7	x	52	x	5	x
4304	x	8	x	50	x	4	x
3746	x	6	x	56	x	6	x

	Procesowanie kostki [ms]		**Zapytanie 1. [ms]**		**Zapytanie 2. [ms]**		**Zapytanie 3. [ms]**	
Bez agregatów	**MOLAP**	**ROLAP**	**MOLAP**	**ROLAP**	**MOLAP**	**ROLAP**	**MOLAP**	**ROLAP**
Średnia	3765,50	1740,50	17,30	79,50	52,50	140,60	15,70	82,60
Odchylenie standardowe	222,76	101,98	2,87	8,26	5,15	25,10	4,27	8,42
Z agregatami	**MOLAP**	**ROLAP**	**MOLAP**	**ROLAP**	**MOLAP**	**ROLAP**	**MOLAP**	**ROLAP**
Średnia	4061,50	x	6,40	x	54,50	x	4,60	x
Odchylenie standardowe	243,07	x	0,97	x	2,68	x	0,84	x

27% agregacja fact_attendance

Czyszczenie cache - uruchamiane przed każdym zapytaniem:

```
<ClearCache xmlns=http://schemas.microsoft.com/analysisservices/2003/engine>
   <Object>
      <DatabaseID>your analytical database ID</DatabaseID>
   </Object>
</ClearCache>
```

4. Dyskusja

Odchylenie standardowe

Duże wartości odchylenia standardowego najprawdopodobniej wynikają z lekko różniącego się stanu obciążenia komputera podczas wykonywania pomiarów. Bez laboratoryjnych warunków, testowanie na zwykłym laptopie pracującym na systemie Windows 11 prowadzi do mniej dokładnych pomiarów.

Agregaty

Wykorzystanie agregatów jest kluczowe dla optymalizacji wydajności - agregacja w MOLAP znacząco skraca czas przetwarzania, co potwierdzają wyniki testów.

Czas procesowania kostki

Średni czas przetwarzania kostki w **MOLAP** był wyraźnie **dłuższy** niż w ROLAP. Jest to związane z kompromisem między dłuższym czasem procesowania kostki, ale za to późniejszym wzrostem wydajności przy zapytaniach.

Wynika to z samej natury podejścia MOLAP. MOLAP przetwarza dane podczas procesowania kostki, tworząc wielowymiarowe kostki, które zawierają zagregowane informacje gotowe do szybkiej analizy. Inwestujemy zasoby na początku, aby uzyskać znaczną poprawę wydajności w późniejszych etapach użytkowania.

Należy również zauważyć, że dłuższy czas procesowania kostki w MOLAP może być odczuwalny przy bardzo dużych zbiorach danych lub przy częstych zmianach w danych źródłowych. W takich przypadkach konieczne jest ponowne procesowanie kostek, co może prowadzić do przerw w dostępie do zaktualizowanych danych.

Czas wykonywania zapytań

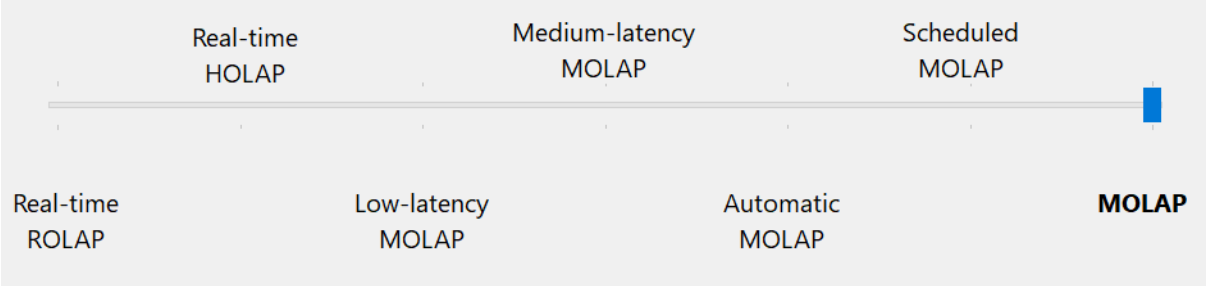
Czas wykonywania poszczególnych zapytań był w **MOLAP** był znacznie **szybszy** od ROLAP. *Dzieje się tak, ponieważ w MOLAP dane są już wstępnie przetworzone i przechowywane w formie zoptymalizowanych struktur wielowymiarowych. To pozwala na natychmiastowy dostęp do zagregowanych wyników bez potrzeby dynamicznego wykonywania operacji na dużych tabelach źródłowych.

Z kolei ROLAP, działając bezpośrednio na relacyjnych bazach danych, musi za każdym razem przetwarzać dane na żywo, co prowadzi do znacznie dłuższego czasu wykonywania zapytań. Taka architektura sprawdza się lepiej w sytuacjach, gdzie dane zmieniają się często i nie ma możliwości/potrzeby wykonywania wstępnego przetwarzania.

Możliwości konfiguracji MOLAP i ROLAP

Podczas wykonywania zadania, zauważyliśmy że możliwych do zaznaczenia modeli lub wariantów modeli jest więcej niż tylko klasyczne MOLAP i ROLAP.

Igor Józefowicz 193257, Patryk Sowiński-Toczek 191711



Wnioski

MOLAP wyraźnie wyróżnia się pod względem wydajności wykonywania zapytań, co czyni go świetnym rozwiązaniem dla hurtowni danych z dużym obciążeniem analitycznym. Natomiast ROLAP, mimo gorszej wydajności przy zapytaniach, oferuje większą elastyczność i niższe wymagania wstępne, co może być korzystne w dynamicznych środowiskach o mniej przewidywalnych schematach dostępu do danych.

Decyzja między MOLAP a ROLAP powinna zatem być oparta na specyfice zastosowania:

- Jeśli kluczowa jest szybkość analizy i przewidywalne obciążenie – MOLAP jest lepszym wyborem.
- Jeśli środowisko wymaga częstych aktualizacji danych i bardziej elastycznego dostępu – ROLAP może być bardziej praktyczny, mimo większych opóźnień w wykonywaniu zapytań.