Michał Sadkowski 197776

Dawid Wesołowski 197943

Optymalizacja hurtowni danych – sprawozdanie

1. Cel laboratorium

Celem zadania jest pokazanie problemów związanych z różnymi fizycznymi modelami kostki oraz projektowaniem agregacji.

2. Założenia początkowe

Rozmiar bazy danych (hurtowni danych): 464MB

Ilość wierszy w Interwencje_Fakty: 371350

Ilość wierszy w Patrole_Fakty: 50160

Ilość wierszy w Kary_Fakty: 165150

Środowisko testowe:

Pomiary zostały wykonane na maszynie wirtualnej korzystającej z zasobów komputera osobistego opartego o procesor Intel Core i5-10400F (8 wątków i 90% jego użycia do dyspozycji VM), 32GB RAM (16384MB do dyspozycji VM) i 2TB dysk SSD (80GB do dyspozycji VM). Wykorzystano system Windows 10, do oceny czasu przetwarzania kostki użyliśmy SQL Server Managment Studio z rozszerzeniem SQL Server Profiler. Podczas wykonywania pomiarów jedynymi aktywnymi aplikacjami na maszynie były SSMS oraz Visual Studio.

3. Testowanie

Testowanie czasów wykonywania zapytań dla różnych modeli, z określonymi i bez określonych agregacji. Testowanie czasów przetwarzania kostki w tych samych warunkach testowych.

Krótki opis zapytań:

1) Porównaj liczbę patroli i wypadków drogowych w poszczególnych dzielnicach w analizowanym miesiącu względem poprzednich.

```
WITH
MEMBER [Measures].[Wypadki Poprzedni Miesiac] AS
   ([Measures].[Interwencje Fakty Count],
[Data].[Kalendarz].CurrentMember.PrevMember)
```

```
SELECT
  {
    [Measures].[Interwencje Fakty Count],
    [Measures].[Wypadki Poprzedni Miesiac]
  } ON COLUMNS,
  NON EMPTY [Data].[Kalendarz].[Miesiac].Members ON ROWS
FROM [policjaHDview]
WHERE ([Zdarzenia].[Kategoria].&[Drogowe])
```

2) Zidentyfikuj okresy (dni tygodnia, pory dnia) o największej liczbie wypadków.

```
SELECT
  {[Measures].[Interwencje Fakty Count]} ON COLUMNS,
  TopCount(
    ([Data].[Dzien Tygodnia].[Dzien Tygodnia].Members * [Czas].[Pora
Dnia].[Pora Dnia].Members),
    10,
    [Measures].[Interwencje Fakty Count]
  ) ON ROWS
FROM [policjaHDview]
WHERE ([Zdarzenia].[Kategoria].&[Drogowe])
```

3) Przeanalizuj rozkład rodzajów kar dla typów przewinień.

```
SELECT
  [Opis Kary].[Rodzaj].[Rodzaj].Members ON COLUMNS,
  [Zdarzenia].[Rodzaj].[Rodzaj].Members ON ROWS
FROM [policjaHDview]
WHERE ([Measures].[Kary Fakty Count])
```

Wszystkie pomiary zostały wykonane dziesięciokrotnie. Wyniki przedstawia poniższa tabela:

Tabela 3.1. Czas procesowania (jednostka ms) kostki i zapytań dla modeli MOLAP oraz ROLAP z i bez agregacji

Bez agregacji							
Procesowanie kostki		Zapytanie 1		Zapytanie 2		Zapytanie 3	
MOLAP	ROLAP	MOLAP	ROLAP	MOLAP	ROLAP	MOLAP	ROLAP
4270	1030	11	317	75	107	141	113
4698	1394	9	99	9	93	9	129
4725	1264	10	93	23	82	9	143
4349	1288	10	91	10	77	9	112
4499	1239	11	146	10	82	9	176
4901	1377	11	98	10	76	9	133
4317	1154	10	145	12	80	9	143
5084	1393	10	85	10	110	9	132
4443	1184	9	86	17	77	8	117
5022	1158	10	91	9	100	8	150
Z agregacjami							
Procesowanie kostki		Zapytanie 1		Zapytanie 2		Zapytanie 3	
MOLAP	ROLAP	MOLAP	ROLAP	MOLAP	ROLAP	MOLAP	ROLAP
4810	1168	6	113	10	81	50	121
4670	1273	5	85	9	100	10	115
4932	1247	7	92	10	78	9	122

5068	1162	7	90	9	76	8	114
5662	1287	6	101	8	78	10	128
4670	1155	6	80	10	84	8	120
4618	1251	6	102	11	113	9	121
4509	1205	6	124	8	86	10	134
4667	1154	6	91	10	84	9	112
5171	1273	8	86	13	75	9	128

Po wykonaniu pomiarów, zdecydowalismy się odrzucić wartości odstające i wyliczyć średnią oraz odchylenie standardowe dla każdej kolumny. Podsumowane wyniki przedstawiaja tabele poniżej:

Tabela 3.2. i 3.3. Średnia i odchylenie standardowe czasu procesowania (jednostka ms) kostki i zapytań dla modeli MOLAP i ROLAP z i bez agregacji.

	Procesowanie kostki		Zapytanie 1		Zapytanie 2		Zapytanie 3	
Bez agregacji	MOLAP	ROLAP	MOLAP	ROLAP	MOLAP	ROLAP	MOLAP	ROLAP
Średnia	4630,8	1248,1	10,1	125,1	18,5	88,4	22	134,8
Odch. std.	283,8	113,8	0,7	67,5	19,3	12,4	39,7	18,5

	Procesowanie kostki		Zapytanie 1		Zapytanie 2		Zapytanie 3	
Z agregacjami	MOLAP	ROLAP	MOLAP	ROLAP	MOLAP	ROLAP	MOLAP	ROLAP
Średnia	4877,7	1217,5	6,3	96,4	9,8	85,5	13,2	121,5
Odch. std.	328,5	51,6	0,8	13	1,4	11,4	12,3	6,6

3.4. Średni czas procesowania kostki i zapytań dla modeli MOLAP i ROLAP z i bez agregacji.

	MOLAP		ROLAP	
	Bez agregacji	Z agregacjami	Bez agregacji	Z agregacjami
Prędkość zapytań (dla 3 różnych)	10,1	6,3	125,1	96,4
	18,5	9,8	88,4	85,5
	22	13,2	134,8	121,5
Czas procesowania	4630,8	4877,7	1248,1	1217,5
Rozmiar	172,27MB	174,57MB	167,24MB	167,24MB

4. Wnioski

Przeprowadzone testy wydajnościowe wykazały różnice między modelami kostki MOLAP i ROLAP. Model MOLAP okazał się być znacznie szybszy pod względem czasu wykonywania zapytań analitycznych (np. dla zapytania 1: 10,1 ms vs 125,1 ms bez agregacji) w stosunku do modelu ROLAP. Jest to efekt przechowywania danych w zopytmalizowanej strukturze na serwerze OLAP.

Zastosowanie agregacji wpłynęło pozytywnie na wydajność obliczeń w obu modelach. W przypadku MOLAP, pozwoliło to na redukcję czasu o 30-40%, co spowodowało jednak wydłużenie czasu procesowania kostki. Wynika to z tego, że agregacje to gotowe sumy – serwer bierze tylko gotowy wynik. Wpływa to jednakże na zwiększenie rozmiaru na dysku.

Z kolei model ROLAP, mimo znacznie gorszych czasów zapytań, wykazał ogromną przewagę w szybkości przetwarzania kostki (blisko 4-krotnie szybszy). Wynika to z faktu, iż ROLAP nie kopiuje danych, a jedynie odwołuje się do źródłowej bazy relacyjnej – powoduje to również mniejszy rozmiar na dysku.

W naszych pomiarach różnica w zajmowanym rozmiarze jest mała, mimo że model ROLAP powinien znacznie zredukować rozmiar danych zajmujących miejsce na dysku. Jest to spowodowane tym, że wymiary nadal są przechowywane w trybie MOLAP (domyślne ustawienie SSAS), natomiast tabela faktów przeszła na tryb ROLAP. Ze względu na małą tabelę faktów, różnica w rozmiarze jest niewielka.

Warto nadmienić, że nierówności w zysku z agregacji w modelu ROLAP mogą wynikać z optymalizacji ze strony silnika SQL. Jeżeli serwer uznaje, że szybciej przeczytać małą tabelę faktów (jak w naszym przykładzie) niż korzystać z widoku, to optymalizator powinien zignorować agregację. Ponadto różnica ok. 30ms w czasie procesowania kostki w modelu ROLAP bez oraz z agregacją nie zgadza się z teorią. Czas procesowania z agregacją powinien być dłuższy. Może to wynikać z błędu pomiarowego lub chwilowego obciążenia serwera.

Podsumowując, model MOLAP jest optymalnym wyborem w scenariuszach, gdzie priorytetem jest szybkość raportowania dla użytkownika końcowego – znacznie szybsze wykonywanie zapytań. Natomiast model ROLAP sprawdzi się lepiej w środowiskach o bardzo dużej ilości danych, gdzie czas procesowania kostki MOLAP byłby nieakceptowalny.