Optymalizacja Hurtowni Danych - Raport

Stanisław Nieradko 193044, Bartłomiej Krawisz 193319

1. Cel raportu

Celem raportu jest analiza i optymalizacja hurtowni danych. W tym celu zostaną przeprowadzone testy wydajnościowe, a następnie zaproponowane zostaną optymalizacje.

2. Założenia

Wielkość Hurtowni

Wielkość hurtowni danych: 1360.00 MB

Liczba rekordów w zależności od tabeli faktów:

OdbycieSieEgzaminu	410400
OdpowiedzenieNaPytaniePodczasEgzaminuZeSkarga	2015
ZarezerwowanieTerminu	1342661
ZlozenieSkargi	1095

Środowisko testowe

```
//////////////////  //////////////////    KanarekLife@Laptop
//////////////////  //////////////////    ------------------
//////////////////  //////////////////    OS: Windows 11 (Home) x86_64
//////////////////  //////////////////    Host: ROG Zephyrus G14 GA402RJ_GA402RJ (1.0)
//////////////////  //////////////////    Kernel: WIN32_NT 10.0.22631.4460 (23H2)
//////////////////  //////////////////    Uptime: 17 days, 22 hours, 57 mins
//////////////////  //////////////////    Shell: PowerShell 7.4.6
//////////////////  //////////////////    Display (TL140VDXP10): 1920x1200 @ 144 Hz
                                          DE: Fluent
//////////////////  //////////////////    WM: Desktop Window Manager
//////////////////  //////////////////    WM Theme: Dark - Blue (System: Dark, Apps: Dark)
//////////////////  //////////////////    Icons: Recycle Bin
//////////////////  //////////////////    Font: Segoe UI (12pt) [Caption / Menu / Message / Status]
//////////////////  //////////////////    Cursor: Windows Default (32px)
//////////////////  //////////////////    Terminal: Windows Terminal 1.21.3231.0
//////////////////  //////////////////    Terminal Font: Cascadia Mono (12pt)
//////////////////  //////////////////    CPU: AMD Ryzen 7 6800HS (16) @ 4.75 GHz
                                          GPU 1: AMD Radeon(TM) Graphics (485.80 MiB) [Integrated]
                                          GPU 2: AMD Radeon RX 6700S (7.96 GiB) [Discrete]
                                          Memory: 13.98 GiB / 15.23 GiB (92%)
                                          Swap: 1.10 GiB / 9.50 GiB (12%)
                                          Disk (C:\): 311.79 GiB / 476.07 GiB (65%) - NTFS
                                          Disk (D:\): 6.38 GiB / 63.94 GiB (10%) - ReFS
                                          Local IP (WiFi): 10.0.0.16/24
                                          Battery: 79% [AC Connected]
                                          Locale: en-GB
```

3. Założenia teoretyczne

	MOLAP	HOLAP	ROLAP
Czas zapytania	Najkrótszy	Średni (w przypadku dobrze zaprojektowanych agregacji może być krótki)	Najdłuższy
Czas przetwarzania	Najdłuższy	Średni (w przypadku dobrze zaprojektowanych agregacji może być krótki)	Krótki
Wielkość hurtowni	Największa (wielkość miary jest zdecydowanie mniejsza jeżeli nie ma żadnych powiązanych z nią agregacji)	Średnia	Najmniejsza

4. Testowanie

Testowanie czasów wykonywania zapytań dla różnych modeli, z i bez zdefiniowanych agregacjami.

Testowanie czasów przetwarzania kostek w tych samych ustawieniach testowych

Krótki opis zapytań

Zapytanie 1: Agregacja po dacie

```
SELECT
  NON EMPTY { [Data].[Hierarchy].[Rok] } ON ROWS,
  NON EMPTY { [Measures].[Liczba rezerwacji], [Measures].[Średni czas oczekiwania na egzamin] } ON
COLUMNS
FROM
  [Data Warehouse]
```

Zapytanie 2: Agregacja po wymiarze

```
SELECT
  NON EMPTY { [Kandydat].[PKK].MEMBERS } ON ROWS,
  NON EMPTY { [Measures].[Średni czas oczekiwania na egzamin] } ON COLUMNS
FROM
  [Data Warehouse]
```

Zapytanie 3: Zapytanie ogólne

```
SELECT
    NON EMPTY { [Measures].[Liczba pytań] } ON COLUMNS,
    NON EMPTY {
        TopCount(
            ([Pytanie].[Tresc].[Tresc].ALLMEMBERS),
            50,
            [Measures].[Liczba pytań]
        )
    }
    DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME ON ROWS
FROM
    (
        SELECT
            ( { [Skarga].[Typ Skargi].&[Treść Pytań] } ) ON COLUMNS
        FROM
            [Data Warehouse]
    )
WHERE
    ( [Skarga].[Typ Skargi].&[Treść Pytań] )
```

Cache i Ustawienia Agregacji

- Podczas testów cache usuwany był przed każdym zapytaniem z wykorzystaniem poniższego polecania:

```
<ClearCache xmlns="http://schemas.microsoft.com/analysisservices/2003/engine">
  <Object>
    <DatabaseID>DataWarehouse</DatabaseID>
  </Object>
</ClearCache>
```

- Testy czasu przetwarzania przeprowadzane były natomiast z wykorzystaniem kolejnego polecenia:

```
<Batch xmlns="http://schemas.microsoft.com/analysisservices/2003/engine">
  <Parallel>
    <Process xmlns:xsd="http://www.w3.org/2001/XMLSchema" xmlns:xsi="http://www.w3.org/2001/
XMLSchema-instance" xmlns:ddl2="http://schemas.microsoft.com/analysisservices/2003/engine/2"
xmlns:ddl2_2="http://schemas.microsoft.com/analysisservices/2003/engine/2/2"
xmlns:ddl100_100="http://schemas.microsoft.com/analysisservices/2008/engine/100/100"
xmlns:ddl200="http://schemas.microsoft.com/analysisservices/2010/engine/200"
xmlns:ddl200_200="http://schemas.microsoft.com/analysisservices/2010/engine/200/200"
xmlns:ddl300="http://schemas.microsoft.com/analysisservices/2011/engine/300"
xmlns:ddl300_300="http://schemas.microsoft.com/analysisservices/2011/engine/300/300"
xmlns:ddl400="http://schemas.microsoft.com/analysisservices/2012/engine/400"
xmlns:ddl400_400="http://schemas.microsoft.com/analysisservices/2012/engine/400/400"
xmlns:ddl500="http://schemas.microsoft.com/analysisservices/2013/engine/500"
xmlns:ddl500_500="http://schemas.microsoft.com/analysisservices/2013/engine/500/500">
      <Object>
        <DatabaseID>DataWarehouse</DatabaseID>
      </Object>
      <Type>ProcessFull</Type>
      <WriteBackTableCreation>UseExisting</WriteBackTableCreation>
    </Process>
  </Parallel>
</Batch>
```

- Utworzone strategie agregacji zostały zdefiniowane na podstawie domyślnych ustawień dla każdego wymiaru.
- Pomiary czasowe zostały przeprowadzone z wykorzystaniem narzędzia `Microsoft SQL Server Management Studio.`

Wyniki

Wyniki w milisekundach (ms), w nawiasach wartości odchylenia standardowego.

Agregacje utworzone z domyślnymi parametrami.

średnia (odch. std.)	**ROLAP**		**HOLAP**		**MOLAP**	
	Bez Agregacji	Agregacja	Bez Agregacji	Agregacja	Bez Agregacji	Agregacja
Czas zapytania dot. Daty [ms]	103,2 (24,21)	97,2 (4,71)	99,3 (9,73)	8,2 (10,49)	16,7 (5,17)	8,2 (10,16)
Czas zapytania dot. wymiaru [ms]	131,5 (29,51)	128,9 (3,31)	128,2 (3,39)	48,4 (10,09)	60,4 (5,58)	46,3 (4,06)
Czas zapytania ogólnego [ms]	41,1 (4,18)	42,3 (2,87)	41,9 (3,48)	42 (2,91)	7,8 (4,83)	8 (4,97)
Czas przetwarzania [ms]	1773	1744	2524	6245	4595	6947
Rozmiar hurtowni [MB]	9,95	9,95	9,95	11,2	24,6	25,92

Wnioski

Czas przetwarzania, zgodnie z oczekiwaniami, jest najkrótszy w modelu ROLAP. Dane nie są w tym typie przeliczane, zapisywane są tylko metadane kostki oraz mapowania danych, dlatego jest to działanie szybkie. Z kolei w modelu MOLAP czas przetwarzania jest najdłuższy, ponieważ dane są kopiowane oraz wykonywane są na nich wstępne obliczenia, co wymaga więcej czasu. HOLAP jest pośrednim rozwiązaniem, które łączy zalety ROLAP i MOLAP, dlatego czas przetwarzania jest dłuższy niż w ROLAP, ale krótszy niż w MOLAP.

Rozmiar hurtowni danych jest najmniejszy w ROLAP, ponieważ zapisywane są tylko metadane kostki oraz mapowania. HOLAP ma identyczny rozmiar hurtowni jak ROLAP. MOLAP z kolei jest ponad dwukrotnie większy. Zawiera on, oprócz metadanych i mapowań, kopie wszystkich danych i wstępnie obliczone agregacje.

Czasy zapytań są za to najkrótsze w MOLAP, ponieważ dane są już przetworzone i gotowe do zapytań. W ROLAP czas zapytań jest znacznie dłuższy, gdyż dane są pobierane z relacyjnej bazy danych i przetwarzane na bieżąco. HOLAP, pomimo że powinien mieć krótszy czas zapytań niż ROLAP, ma czas zapytań do niego zbliżony. Jest to spowodowane brakiem zdefiniowanych agregacji.

Z agregacjami czas przetwarzania dla ROLAP się praktycznie nie zmienił, rozmiar hurtowni też pozostaje taki sam. Dla HOLAP czas przetwarzania jest znacznie dłuższy, a rozmiar hurtowni trochę większy. Dla MOLAP czas przetwarzania też jest dłuższy, ale stosunkowo nie zmienił się on tak bardzo jak dla HOLAP. Rozmiar hurtowni dla MOLAP też się zwiększył, podobnie jak dla HOLAP.

Czas zapytań dla ROLAP z agregacjami nieznacznie się skrócił. ROLAP nie oblicza wstępnie agregacji, dlatego ich zdefiniowanie nie wpływa znacznie ani na wielkość, ani na czas zapytań. Dla HOLAP i MOLAP z agregacjami, czas zapytań jest znacznie krótszy, ale tylko dla zapytań dotyczących daty i wymiaru. Zapytanie ogólne we wszystkich trzech modelach jest za to nawet trochę dłuższe, co może być spowodowane brakiem kompatybilności tego zapytania ze zdefiniowanymi agregacjami.

Podsumowując, dobór typu bazy danych zależy od potrzeb użytkownika: ROLAP jest najlepszy, gdy ważna jest elastyczność i oszczędność miejsca, ale z dłuższym czasem zapytań. MOLAP sprawdza się, gdy kluczowa jest szybkość zapytań, zwłaszcza z agregacjami, kosztem większego rozmiaru hurtowni i dłuższego przetwarzania. HOLAP łączy zalety obu rozwiązań, oferując kompromis między czasem przetwarzania, przestrzenią i szybkością zapytań.

Pełna lista pomiarów

Typ Hurtowni	Zapytanie	Agregacje	Czasy przetwarzania [ms]									
ROLAP	Zapytanie dot. daty	Bez Agregacji	172	95	96	96	95	94	98	95	97	94
		Z Agregacjami	108	94	96	93	94	100	94	94	98	101
	Zapytanie dot. wymiaru	Bez Agregacji	215	120	122	119	124	130	120	122	120	123
		Z Agregacjami	133	133	125	124	132	127	130	131	127	127
	Zapytanie ogólne	Bez Agregacji	52	39	38	39	40	38	42	41	43	39
		Z Agregacjami	40	41	50	42	42	42	40	43	41	42
HOLAP	Zapytanie dot. daty	Bez Agregacji	126	99	93	100	93	93	98	98	97	96
		Z Agregacjami	38	5	5	4	4	5	6	5	5	5
	Zapytanie dot. wymiaru	Bez Agregacji	132	130	132	119	128	129	128	125	127	132
		Z Agregacjami	58	43	74	43	43	44	47	44	43	45
	Zapytanie ogólne	Bez Agregacji	51	41	39	40	42	40	40	41	41	44
		Z Agregacjami	49	42	40	41	42	39	44	42	39	42
MOLAP	Zapytanie dot. daty	Bez Agregacji	31	14	14	17	14	16	14	15	17	15
		Z Agregacjami	37	4	4	5	6	4	5	7	5	5
	Zapytanie dot. wymiaru	Bez Agregacji	37	4	4	5	6	4	5	7	5	5
		Z Agregacjami	54	44	51	44	43	44	45	43	51	44
	Zapytanie ogólne	Bez Agregacji	21	7	6	5	9	6	5	8	5	6
		Z Agregacjami	22	7	6	6	6	6	7	8	6	6