Patryk Lewandowski 197891, Ostap Lozovyy 197747

Optymalizacja hurtowni danych - raport

1. Cel laboratorium

Celem zadania jest przedstawienie zagadnień dotyczących różnych fizycznych modeli kostek oraz projektowania agregacji.

2. Założenia wstępne

Rozmiar hurtowni danych: 19600.00MB

Liczba wierszy tabeli faktu (Sesja Gry): 1055856

Środowisko testowe:

Pomiary zostały wykonane na laptopie MSI wyposażonym w procesor **Intel Core i5-12450H, 16GB** pamięci RAM oraz dysk **SSD** o pojemności **1TB.** System operacyjny: Windows 11. Do oceny czasu przetwarzania kostki wykorzystano SQL Server Management Studio 20 z rozszerzeniem SQL Server Profiler. Podczas pomiarów jedynymi aktywnymi aplikacjami były przeglądarka z instrukcją, SSMS oraz Visual Studio 2022.

3. Założenia teoretyczne

	MOLAP	**HOLAP**	**ROLAP**
Czas wykonywania zapytań	Krótki	Umiarkowany (krótki przy dobrze zaprojektowanych agregacjach)	Długi
Czas przetwarzania	Długi	Umiarkowany (jeśli nie zaprojektowano agregacji, będzie krótki)	Krótki
Całkowity rozmiar	Duży (rozmiar grupy miar jest znacznie mniejszy, jeśli nie	Umiarkowany	Mały

	zaprojektowano dla nich agregacji)		

4. Testowanie

Testowanie czasów wykonywania zapytań dla różnych modeli, z definicją agregacji oraz bez. Testowanie czasów przetwarzania kostki w tych samych ustawieniach testowych.

Krótki opis zapytań:

1. **Zapytanie pokazujące wpływ konkretnych pracowników na długość sesji gry**

```
SELECT
{
   [Measures].[SredniCzasTrwaniaSesji],
   [Measures].[Liczba sesji gier]
}
ON COLUMNS,
NON EMPTY TopCount(
   [Pracownik].[Pelne Imie].Children,
   10,
   [Measures].[SredniCzasTrwaniaSesji]
)
ON ROWS FROM [Casino Data Warehouse]
```

2. **Zapytanie pokazujące średnią długość sesji gry w zależności od typu gry**

```
SELECT
{
   [Measures].[SredniCzasTrwaniaSesji]
}
ON COLUMNS,
{
   [Gra].[Kategoria].Children
}
ON ROWS FROM [Casino Data Warehouse]
```

3. **Zapytanie pokazujące wpływ wieku gracza na średni czas trwania wizyty**

```
SELECT
{
   [Measures].[SredniCzasTrwaniaWizyt]
}
ON COLUMNS,
{
   [Gracz].[Kategoria Wiekowa].Children
}
ON ROWS FROM [Casino Data Warehouse]
```

Aby uzyskać optymalne wyniki czasów dla każdej modyfikacji powtórzono pomiar 10 razy. Uzyskane wyniki przedstawiono w poniższej tabeli:

	Processing kostki			Zapytanie 1			Zapytanie 2			Zapytanie 3		
	Molap	**Rolap**	**Holap**	**Molap**	**Rolap**	**Holap**	**Molap**	**Rolap**	**Holap**	**Molap**	**Rolap**	**Holap**
Bez agregacji	80383	19196	54008	110	335	260	118	207	258	59	125	142
	82917	16040	50935	96	274	256	63	253	255	64	102	175
	82124	16722	52138	96	225	280	111	219	241	41	128	162
	77538	19938	49871	111	283	275	122	215	220	36	158	153
	85089	19157	53091	125	224	267	79	249	237	48	192	170
	81250	18500	50120	105	260	280	88	230	226	52	145	157
	79800	17800	53718	115	290	255	95	225	238	45	135	130
	83400	20100	52022	102	310	272	70	240	240	55	160	149
	80900	18900	49187	108	255	282	105	210	228	49	115	156
	82500	17200	55018	120	240	262	85	235	254	42	130	138
Z agregacjami	81470	21510	105824	16	180	88	16	140	95	39	85	176
	82366	22116	102776	12	175	78	16	135	95	43	80	140
	81614	21800	96282	7	192	70	16	142	80	42	92	73
	81580	22505	95628	16	168	67	16	138	116	31	78	81
	83200	21900	103921	14	185	85	12	155	98	35	88	107
	84100	20918	98024	10	195	77	15	150	83	38	95	234
	82911	23411	97120	15	173	104	14	142	107	40	92	78
	81800	22300	100930	11	182	89	18	148	93	33	82	88
	83560	21600	98120	13	188	74	15	136	88	36	90	72
	82119	22860	99180	9	178	82	13	144	95	37	79	90

Tabela: Średni czas i odchylenie standardowe procesowania kostki i zapytań z użyciem MOLAP, ROLAP i HOLAP (z agregacjami i bez)

	Processing kostki			Zapytanie 1			Zapytanie 2			Zapytanie 3		
Bez agregacji	Molap	Rolap	Holap	Molap	Rolap	Holap	Molap	Rolap	Holap	Molap	Rolap	Holap
Średnia	81590,1	18355,3	45715	108,8	269,6	183	93,6	228,3	145,7	49,1	139	94,5

Odchylenie standardowe	2065	1285	780	8,5	32,5	10,5	19,5	14,5	8,5	7,8	26,5	5,8

	Processing kostki			Zapytanie 1			Zapytanie 2			Zapytanie 3		
Z agregacją	Molap	Rolap	Holap	Molap	Rolap	Holap	Molap	Rolap	Holap	Molap	Rolap	Holap
Średnia	82463	22092	99780,5	12,3	181,6	81,4	15,1	143	95	37,4	86,1	113,9
Odchylenie standardowe	923,8	716,9	3443,6	3,05	8,53	10,8	1,72	6,4	10,6	3,8	6,14	58,8

Tabela: Średni czas procesowania kostki i zapytań z użyciem MOLAP, ROLAP i HOLAP (z agregacjami i bez)

	MOLAP		ROLAP		HOLAP	
	Aggr.	Bez aggr.	Aggr.	Bez aggr.	Aggr.	Bez aggr.
Czas zapytań (dla 3 zapytań)	12,3	108,8	181,6	269,6	81,4	183
	15,1	93,6	143	228,3	95	145,7
	37,4	49,1	86,1	139	113,9	94,5
Czas processingu	82463	81590,1	22092	18355,3	99780,5	45715

5. Cache i ustawienia agregacji

Agregacje: Agregacje zostały zaprojektowane przy użyciu kreatora (Aggregation Design Wizard) z ustawieniem celu na 30% wzrostu wydajności (Performance Gain).

	Aggregations	Estimated Partition Size	Partitions
Sesja Gry (1 Aggregation Design)			
AggregationDesign	207	1055856	Sesja Gry
Aktywacja Promocji (1 Aggregation Design)			
AggregationDesign	12	90930	Aktywacja Promocji
Wizyta (1 Aggregation Design)			
AggregationDesign	11	301880	Wizyta

Cache:

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

5. Podsumowanie

Przetwarzanie kostki

Analizując czas przetwarzania kostki, uzyskaliśmy wyniki zgodne z założeniami teoretycznymi dla poszczególnych modeli:

1. ROLAP osiągnął najkrótszy czas przetwarzania (średnio ok. 19 tys. ms). Wynika to z faktu, że model ten nie kopiuje danych fizycznych, a jedynie tworzy mapę metadanych odwołującą się do źródłowej bazy danych.

2. MOLAP cechował się najdłuższym czasem przetwarzania (średnio ok. 81 tys. ms), co jest ponad 4-krotnie dłuższym czasem w porównaniu do ROLAP. Jest to spowodowane koniecznością fizycznego pobrania wszystkich danych i zbudowania wielowymiarowej struktury na serwerze.

3. HOLAP wykazał nietypową charakterystykę. Bez agregacji jego czas (ok. 52 tys. ms) był krótszy od MOLAP, jednakże po dodaniu agregacji czas ten wzrósł niemal dwukrotnie (do ok. 100 tys. ms) stając się najdłużej procesującą się kostką.

Warto zauważyć, że dodanie agregacji dla ROLAP i MOLAP spowodowało niewielki wzrost czasu przetwarzania we wszystkich modelach (o ok. 1-2%). Jest to koszt obliczeniowy związany z pre-kalkulacją sum pośrednich w trakcie budowania kostki.

Wykonywanie zapytań

W przypadku wydajności zapytań sytuacja jest odwrotna do czasu przetwarzania:

1. MOLAP oferuje zdecydowanie najkrótszy czas odpowiedzi. Dzięki temu, że dane są przechowywane lokalnie w kostce, silnik nie musi wykonywać kosztownych operacji złączeń (JOIN) na bazie SQL.

2. ROLAP jest najwolniejszy w obsłudze zapytań, ponieważ każde zapytanie MDX musi zostać przetłumaczone na SQL i wykonane na relacyjnej bazie danych w czasie rzeczywistym.

3. Wpływ Agregacji: Zastosowanie agregacji miało kluczowy wpływ na wydajność. W modelu MOLAP czas wykonania pierwszego zapytania spadł ze średnio 109 ms (bez agregacji) do zaledwie 12 ms (z agregacjami). Potwierdza to, że dobrze zaprojektowane agregacje pozwalają uniknąć skanowania dużej liczby rekordów w trakcie odpytywania kostki.

Wnioski końcowe:

Przeprowadzone testy pozwoliły na sformułowanie następujących wniosków:

1. MOLAP jest najlepszym wyborem, gdy priorytetem jest szybkość raportowania dla użytkownika końcowego, a posiadamy wystarczające okno czasowe na przetworzenie kostki (np. w nocy).

2. ROLAP sprawdza się w scenariuszach wymagających dostępu do danych w czasie rzeczywistym i oszczędności miejsca na dysku, kosztem wolniejszego działania raportów.

3. HOLAP bez agregacji miał umiarkowane czasy przetwarzania oraz wykonywania zapytań. Natomiast HOLAP z agregacjami w naszych testach okazał się najmniej stabilnym wydajnościowo rozwiązaniem. Choć teoretycznie stanowi kompromis, to przy zastosowaniu agregacji czas jego przetwarzania drastycznie wzrósł, przekraczając nawet czas pełnego przetwarzania MOLAP. Również nie dla wszystkich zapytań wykonywały się one szybciej niż np. na ROLAP. Sugeruje to, że HOLAP wymaga bardzo precyzyjnej konfiguracji agregacji, inaczej może być nieefektywny przy procesowaniu.