Optymalizacja Hurtowni Danych – raport

1. Środowisko testowe

Hardware

- Procesor: 13th Gen Intel(R) Core(TM) i5-13600KF
- Karta graficzna: NVIDIA GeForce RTX 4070 Ti
- RAM: Pamięć GoodRam IRDM PRO Deep Black, DDR4, 32 GB, 3600MHz, CL18
- Zasilacz: MSI MPG A750GF 750W
- Dysk: SSD Lexar NM620 2TB M.2 2280 PCI-E x4 Gen3 NVMe
- Chłodzenie: CPU Endorfy Fortis 5 Dual Fan (EY3A009)
- Płyta główna MSI PRO Z690-A WIFI DDR4

Software

- Microsoft Windows 11 Home PL 64 bit OEM
- Microsoft Visual Studio 2019
- Microsoft SQL Server Management Studio 2019
- Microsoft SQL Server Profiler 2019

Opis datasetu

Liczności tabel wymiarów:

- Album: 8000 wierszy
- Artist: 5000 wierszy
- Customer: 7000 wierszy
- Date: 20454 wierszy
- Junk: 3 wiersze
- Playlist: 4000 wierszy
- Song: 18000 wierszy
- Time: 86400 wierszy

Liczności tabel faktów:

- Playback: 28000 wierszy
- PlaylistCreation: 4000 wierszy
- PlaylistSong: 14000 wierszy

2. Zapytania testowe

Zapytania testowe w kolejności. Testowane zgodnie z opisami w tabeli w części Wyniki.

1. Podaj 5 najczęściej odsłuchiwanych albumów w obecnym i zeszłym miesiącu.
2. Porównaj ilość odtworzeń na różnych urządzeniach w analizowanym miesiącu relatywnie do poprzednich.
3. Podaj najbardziej odsłuchiwanych artystów na Spotify w zależności od gatunku miesiącu w stosunku do poprzednich miesięcy.

Proces tworzenia agregacji.

1. Zmiana domyślnego poziomu agregacji Music Genre na Full, a Song Title na None.
2. Zaprojektowanie przez kreator statystycznie najlepszych agregacji poprzez skorzystanie z opcji przycisku Count.
3. Wybranie opcji Performance gain reaches dostosowując procent od domyślnych 30, do 80, zwiększając o 10% co każdy pomiar.

3. Wyniki

Wynikiem są średnie z 10 prób testu dla każdej z 3 kolumn z wyłączonym cachowaniem wyników.

	MOLAP		ROLAP
	Agr.	Bez agr.	Bez agr.
Szybkość wykonywania zapytań (dla 3 różne zapytań) [ms]	13.4	19.1	62.1
	4.8	6.4	24.2
	6.2	10.9	63.4
Czas przetwarzania kostki [ms]	649	594	240
Całkowity rozmiar [Mb]	23.73	20.46	18.91

4. Wnioski

Prezentowane wyniki prezentują zależność między ilością danych przechowywanych na serwerze analitycznym, a prędkością wykonywania zapytań. Przechowywanie większej ilości danych da nam do nich szybszy dostęp, lecz poświęcimy więcej miejsca aby móc je przechowywać. Wybór każdego z tych rozwiązań zależy od dysponowanych przez nas środkami miejsca na dane oraz wymagania szybkiego wykonywania zapytań.

MOLAP bez agregacji

Rozwiązanie MOLAP bez agregacji dało wyniki znajdujące się pośrodku przedziału, zgodnie z przewidywaniami teoretycznymi. Oznacza to, że jest to metoda najbardziej uniwersalna z testowanych. Wyniki są zadowalające pod względem szybkości, ponieważ w bazie danych na serwerze analitycznym przechowywane są wszystkie dane z relacyjnej bazy danych. Duplikacja danych to konieczność posiadania większego rozmiaru dysku na przechowanie naszych danych.

MOLAP z agregacją

Rozwiązanie MOLAP z agregacją dało najlepsze wyniki pod względem szybkości wykonywania zapytań, zgodnie z przewidywaniami teoretycznymi. Odbywa się to oczywiście kosztem największego rozmiaru bazy danych na serwerze analitycznym. Agregacje pozwalają na wstępne obliczenie wartości na etapie przetwarzania kostki, co zapewni szybszy dostęp do danych w zamian za przechowywanie większej ilości zduplikowanych danych. Agregacje powodują dłuższy czas przetwarzania kostki, gdyż są tworzone podczas tego procesu.

ROLAP bez agregacji

Rozwiązanie ROLAP bez agregacji dało najlepszy wynik pod względem oszczędności miejsca na dysku - rozmiaru bazy danych na serwerze analitycznym, zgodnie z przewidywaniami teoretycznymi. Wynika to z faktu, że model nie przechowuje zduplikowanych danych na serwerze analitycznym. Koszt wychodzi przy szybkości wykonywania zapytań - najgorszy wynik z testów. Dzieje się tak, ponieważ serwer analityczny musi pobrać dane z relacyjnej bazy danych do wykonania zadanego przez nas zapytania.