

Data Warehouse Optimization

Cel laboratorium

Celem zadania jest przedstawienie zagadnień związanych z różnymi modelami fizycznymi kostki i projektowaniem agregatów

Założenia wstępne

Baza danych (hurtownia) zawiera bardzo małe ilości krotek w wymiarach, w ilości od kilkudziesięciu do kilkuset krotek. Wymiar czasu zawiera jeden rekord na każdą minutę w dobie, a wymiar dat zawiera ponad 13 tys wierszy. W hurtowni są dwie tabele faktów. Jedna jest mniej liczna → około 5 tys wierszy. Druga, jest główną tabelą faktów i zawiera 125 tysięcy wierszy.

Środowisko testowe:

- Windows 10 Home
- Intel Core I5-5200U
- RAM 8GB
- Microsoft Visual Studio 2019
- Microsoft SQL Server Management Studio 2018
- Microsoft Excel - Office 360
- Microsoft SQL Server Profiler 2018

Testowanie

Testowanie polegało na wykonywaniu konkretnych zapytań z wykorzystaniem różnych modeli wielowymiarowych zarówno bez, jak i z zdefiniowanymi agregacjami. Jako pytania testowe wybrano:

- liczba przejazdów wedle dni tygodnia i kategorii wiekowych kierowców
- suma opóźnień poszczególnych kierowców w poszczególnych latach
- suma opóźnień i liczba przejazdów danych modeli tramwajów w poszczególnych latach

	MOLAP		ROLAP		HOLAP	
	Aggr.	Brak	Aggr.	Brak	Aggr.	Brak
Zapytanie 1 [ms]	14.8	21.5		416.2	188.3	176.4
Zapytanie 2 [ms]	7.8	19.9		971.6	12.8	150.0
Zapytanie 3 [ms]	7.9	22.1		838.8	4.4	250.9
Przetwarzanie kostki [s]	0.03	0.04		0.01	0.01	0.01
Rozmiar bazy [Mb]	5.41	4.65		3.44	4.20	3.44

Wnioski

Na podstawie wykonanych testów można wywnioskować więc, że im więcej danych znajduje się w analitycznej bazie danych, tym szybciej będą wykonywane zapytania. Dodatkowe przechowywanie agregacji może znacznie przyspieszyć wykonywanie zapytań, lecz takie zdefiniowane agregaty muszą być wówczas przechowywane na serwerze OLAP. Pojawia się problem, w którym należy podjąć decyzję, czy zależy nam na szybkości wykonywanych zapytań, czy na rozmiarze zajmowanej pamięci.

Model MOLAP zarówno w teorii, jak i w praktyce okazał się najszybszym spośród dostępnych modeli. W tym modelu wszystkie dane znajdują się w analitycznej bazie danych, co znacznie przyspiesza wykonywanie zapytań, ale jednocześnie powoduje przechowywanie zduplikowanych danych. Objawia się to dużo większym rozmiarem w stosunku do pozostałych rozwiązań.

Model ROLAP w teorii powinien być najwolniejszy ze wszystkich testowanych modeli i tak jest w praktyce. Wynika to z faktu, że na serwerze OLAP przechowywane są tylko dane dotyczące kostki, a reszta informacji w relacyjnej bazie danych. Zaletą tego modelu jest bardzo mały rozmiar bazy analitycznej, więc może okazać się dobrym rozwiązaniem, gdy będzie nam zależeć nie na szybkości, a na zajmowanej pamięci.

Model HOLAP jest rozwiązaniem pośrednim między MOLAP i ROLAP. Przy zdefiniowanych agregacjach szybkością dorównuje modelowi MOLAP równocześnie zajmując o wiele mniej pamięci na serwerze. Jednak gdy ich nie zdefiniujemy, to prędkość wykonywania zapytań drastycznie spada.