

Kolumnowa Baza Danych na przykładzie SAP Hana

10.05.2023

Elbląg

Kuba Swiśtocki 19423

1. Przechowywanie

Sposób przechowywania danych znacząco różni się między tabelami kolumnowymi a tabelami rzędownymi. Baza kolumnowa przechowuje swoje dane w sposób kolumnowy, co oznacza, że biorąc za przykład tabelę posiadającą pola takie jak "Imię", "Nazwisko" i "Kraj", baza będzie przechowywała w jednym miejscu wszystkie imiona, w drugim nazwiska, a w trzecim kraje. Przechowywać będzie dane tego samego typu. Przykład:

Niech dane wyglądają następująco:

Nazwisko	Imię	Kraj
SMITH	Gerl	Poland
BROWN	Matthew	Brazil
WILSON	John	Switzerland
CAMPBELL	Casper	Germany
THOMSON	Nikola	Germany
ROBERTSON	Wayne	Poland
STEWART	Little	Poland
ANDERSON	Silva	Ukrain

Teraz sposób przechowywania danych przez tabele rzędowną

Rząd 1	SMITH	Gerl	Poland
Rząd 2	BROWN	Matthew	Brazil
Rząd 3	WILSON	John	Switzerland
Rząd 4	CAMPBELL	Casper	Germany
Rząd 5	THOMSON	Nikola	Germany
Rząd 6	ROBERTSON	Wayne	Poland
Rząd 7	STEWART	Little	Poland
Rząd 8	ANDERSON	Silva	Ukrain

A teraz jak takie przechowywanie wygląda w tabeli kolumnowej

Imię	Nazwisko	Kraj
Gerl	SMITH	Poland
Matthew	BROWN	Brazil
John	WILSON	Switzerland
Casper	CAMPBELL	Germany
Nikola	THOMSON	Ukrain
Wayne	ROBERTSON	
Little	STEWART	
Silva	ANDERSON	

Podczas przechowywania danych w tabelach kolumnowych utworzona zostaje również tabela łącząca, w tym przypadku będzie ona wyglądała tak:

Imię	Nazwisko	Kraj
1	1	1
2	2	2
3	3	3
4	4	4
5	5	4
6	6	1
7	7	1
8	8	5

Przydaje się to podczas wywoływania zapytań, na przykład prostego SELECT, które ma pokazać Imię oraz kraj wszystkich osób z konkretnej tabeli. Baza nie musi w ogóle korzystać z zbędnych kolumn, ponieważ nazwisko nie było uwzględnione w zapytaniu. W przypadku bazy danych rzędowych, nazwiska nie brałyby udziału w operacji ze względu na zapis pełnych rzędów w jednym miejscu. Zostałyby pominięte podczas wyświetlania wyniku zapytania. Podczas operacji i zapytań na tabeli opisanej powyżej, osiągnięcia znacząco nie wyróżniałyby się podczas korzystania z kolumnowego lub rzędowego przechowywania danych. Jednakże, jeśli dodamy wiele nowych kolumn przechowujących cechy dokładniej opisujące osobę, wtedy zauważymy, że baza kolumnowa szybciej wyświetla wyniki w porównaniu do bazy rzędowej, ponieważ nie musi brać pod uwagę wszystkich zbędnych kolumn. Baza kolumnowa została zoptymalizowana pod względem przetwarzania danych, dlatego podczas potrzeby przeprowadzenia szybkich analiz liczbowych czy bankowych, baza kolumnowa przeważa nad bazą rzędową, ponieważ ładowane i wykorzystywane są tylko te informacje, które są potrzebne jako wynik zapytania lub do dalszych obliczeń, omijając niepotrzebne wartości. Ze względu na odmienny sposób przechowywania danych (pionowo i poziomo), indeksowanie również odbywa się w inny sposób. Indeksy są tworzone na poziomie kolumn, a nie wierszy, co umożliwia filtrowanie i przeszukiwanie tylko tych danych, które są brane pod uwagę, unikając przeszukiwania całych wierszy, jak ma to miejsce w bazie rzędowej. Podczas zapisu danych do tabel automatycznie uruchamiana jest kompresja. Dlatego też, przedstawione wcześniej tabele spowodują również powstanie odpowiadających im tabel słownikowych, aby zaoszczędzić jak najwięcej miejsca.

2. Kompresja

Kolumnowa baza danych, ze względu na swoją strukturę i sposób przechowywania danych, jest wyjątkowo podatna na kompresję danych. Wynika to z faktu, że dane z jednej kolumny są przechowywane w jednym miejscu, co oznacza powtarzalność typu danych, czyli taką samą charakterystykę, na przykład liczby zmiennoprzecinkowe lub ciągi znaków. Poniżej przedstawiam kompresję słownikową :

Kraj	
Poland	1
Brazil	2
Switzerland	3
Germany	4
Germany	4

Poland	1
Poland	1
Ukrain	5

Polega na wyeliminowaniu wartości powtarzających się i zastąpieniu ich wartościami liczbowymi co oczywiście powoduje zmniejszenie się rozmiaru danych. Taki rodzaj kompresji jest bardzo powszechny w hurtowniach danych, np. w systemach SAP BW.

3. Ładowanie danych

Klasyczne bazy danych oparte są na przechowywaniu danych w wierszach. W tym podejściu dane są ładowane do bazy poprzez przenoszenie wierszy, na przykład z pliku .csv, do tabeli w bazie danych. Każdy wiersz w pliku odpowiada pojedynczemu wpisowi w tabeli.

Proces ładowania danych do bazy kolumnowej nazywany jest ETL (ang. Extract, Transform, Load), co oznacza ekstrakcję, transformację, a następnie ładowanie. W pierwszym kroku, czyli ekstrakcji danych, dane są pobierane z żądanych źródeł, takich jak inne systemy bazodanowe, aplikacje lub serwery. Na tym etapie dane mogą być wyświetlane, ale nie są jeszcze przenoszone ani ładowane.

Po ekstrakcji danych przechodzi się do drugiego etapu, czyli transformacji. W tym etapie dane są formatowane, aby miały odpowiednią strukturę dla bazy danych kolumnowej. Oprócz przygotowania danych do załadowania, w tym etapie można również przeprowadzić operacje filtrowania danych, na przykład usuwanie niechcianych rekordów. Konieczność filtrowania zależy od użytkownika i specyfikacji systemu.

Trzeci etap to samo ładowanie, gdzie przygotowane wcześniej dane są załadowane do odpowiednich tabel. W przypadku bazy danych SAP HANA istnieje możliwość bezpośredniego ładowania plików płaskich i tworzenia tabeli o strukturze odpowiadającej plikowi, który ma zostać załadowany. To znacznie ułatwia projektowanie tabel do późniejszego wykorzystania.

4. Optymalizacja

Kolumnowe bazy danych różnią się od bazy danych rzędowych pod względem optymalizacji, co wynika z różnych kierunków, w jakich poszły ich rozwój. Bazy kolumnowe są zoptymalizowane pod kątem odczytu (ang. Read-optimized), co oznacza, że są szybsze w udzielaniu odpowiedzi na zapytania, nawet te skomplikowane z wieloma warunkami. Jednakże, wynika z tego, że takie bazy danych mają słabszą wydajność w zakresie zapisywania danych, ponieważ wymagają procesu ETL przed załadowaniem danych. Z tego powodu tradycyjne bazy danych rzędowe są uważane za zoptymalizowane pod kątem zapisu (ang. Write-optimized).

Wybór między tymi dwoma typami baz danych zależy od tego, do jakich celów będą wykorzystywane przez przedsiębiorstwo lub firmę. Jeśli firma planuje korzystać z bazy danych do zadań analitycznych, analizy biznesowej lub tworzenia raportów, powinna rozważyć użycie bazy kolumnowej, ponieważ będzie ona bardziej efektywna w obsłudze zapytań dotyczących dużej ilości informacji w krótkim czasie, co jest charakterystyczne dla wspomnianych zadań. Natomiast jeśli firma używa bazy danych w innym celu, na przykład do zapisywania dużych ilości transakcji bankowych lub obsługi ogromnych zamówień, powinna rozważyć użycie bazy danych rzędowej ze względu na zoptymalizowane właściwości zapisu.

5. Skalowalność

Ważną cechą i jednocześnie różnicą między tymi dwoma typami baz danych jest skalowalność bazy kolumnowej. Baza kolumnowa jest wyjątkowo łatwa do skalowania, co oznacza, że dodawanie nowych kolumn do istniejącej tabeli nie wymaga, aby dane z tych kolumn były przechowywane w tym samym węźle, co dane z pozostałych kolumn. Dzięki temu kolumny mogą być przechowywane osobno. W przypadku tradycyjnych baz danych rzędowych byłoby to znaczącym problemem, ponieważ wiersze danych nie mogą być rozdzielane i muszą być przechowywane w ciągłym formacie.

Skalowalność bazy kolumnowej oznacza, że można łatwo dodawać nowe kolumny do istniejącej tabeli, bez konieczności przenoszenia danych w innych kolumnach. Jest to korzystne w przypadku, gdy struktura danych ewoluuje i wymaga dodania nowych informacji. Dzięki temu rozwiązaniu można elastycznie dostosowywać bazę danych do zmieniających się wymagań, bez konieczności modyfikowania istniejących danych.

W przypadku tradycyjnych baz danych rzędowych, gdybyśmy chcieli dodać nową kolumnę do tabeli, musielibyśmy przenieść wszystkie dane w innych kolumnach, aby utrzymać ciągłość wierszy. To może być skomplikowane i czasochłonne, szczególnie w przypadku dużych zbiorów danych.

Skalowalność bazy kolumnowej pozwala na efektywne zarządzanie danymi i dostosowywanie struktury bazy danych do aktualnych potrzeb, co jest istotne w dynamicznym środowisku biznesowym.

6. Wady

Baza kolumnowa nie jest jednak pozbawiona wad. Poza faktem, że baza rzędowa lepiej sprawdza się przy zapisywaniu danych, rozwiązanie kolumnowe jest bardziej skomplikowane pod względem projektowania, implementacji i utrzymania. Przed wdrożeniem bazy kolumnowej konieczne jest poświęcenie więcej czasu i wysiłku w porównaniu do bazy rzędowej. Już na etapie wdrożenia należy przemyśleć, jakie dane będą przechowywane w poszczególnych kolumnach oraz jakie algorytmy kompresji zostaną zastosowane, aby zminimalizować koszty.

Bazy kolumnowe często wymagają zaawansowanego i złożonego sprzętu, aby osiągnąć oczekiwaną wydajność, co wiąże się z większymi wydatkami na infrastrukturę. Na przykład baza danych SAP HANA, która oprócz bycia bazą kolumnową, jest również bazą in-memory, przechowuje dane w pamięci. Oczywiście jest, że pamięć o takiej samej pojemności jak dysk SSD lub HDD jest znacznie droższa.