

XV Konferencja PLOUG
Kościelisko
Październik 2009

„Pamięć absolutna”, czyli użycie modułu TotalRecall do archiwizacji danych historycznych

Krzysztof Mikołajczyk
Bull Polska

krzysztof.mikolajczyk@bull.com.pl

Abstrakt. W artykule zostały zamieszczone podstawowe informacje o archiwizacji danych historycznych. Przedstawione zostało wykorzystanie modułu Total Recall do zarządzania i wykonywania archiwizacji oraz różne problemy, na które można się natknąć w trakcie wykonywania pracy. Artykuł stanowi kontynuację prezentacji z konferencji PLOUG 2007.

1. Wprowadzenie

Zabezpieczanie danych historycznych jest ważne z różnych powodów. Jednym z nich są wymogi prawne – niespełnienie ich może być krytyczne. Dostęp do danych historycznych pozwala również na lepszą analizę trendów na rynku. Problem stanowi dobre zarządzanie danymi historycznymi. W grę wchodzi głównie dwa aspekty – sposób składowania danych historycznych i dostęp do nich.

Główną potrzebą związaną z tworzeniem i obsługą danych archiwalnych są wymogi prawne. Różne ustawy i zalecenia wymagają, aby został utrzymany dostęp do danych historycznych przez określony czas. Nie regulują one jednak sposobu dostępu i składowania, jest to pozostawione w kwestii organizacji.

Jednak nie tylko wymogi prawne są powodem utrzymywania danych historycznych – przydają się one także przy analizie trendów i tendencji rynkowych. W związku z tym w wielu przypadkach opłaca się zatrzymać dane historyczne. Jedyne problem związany jest ze sposobem obsługi i dostępu do tych danych.

Wiele danych pomiarowych lub obserwacyjnych jest również składanych przez długie (lub bardzo długie, nawet na zawsze) okresy czasu. Dotyczy to np. danych klimatycznych, których znajomość nawet po kilkuset latach może być bardzo przydatna.

Podstawowym problemem przy pracy z danymi archiwalnymi jest określenie, które dane są już historyczne, jak długo należy je przetrzymywać, jak ma być realizowany dostęp do tych danych, kiedy dane można definitywnie usunąć. Odpowiedź na to pytanie pozwala na efektywne zarządzanie danymi archiwalnymi.

W efektywnym zarządzaniu danymi archiwalnymi może pomóc przedstawiony w referacie moduł „Total Recall” (dostępny dla wersji 11g, opcja dodatkowo płatna). Zostaną przedstawione jego podstawowe cechy i sposób implementacji. Przedstawione zostaną również sposoby radzenia sobie z danymi historycznymi.

Moduł „Total Recall” bazuje na technologii „Flashback”, która jest dostępna również we wcześniejszych wersjach Oracle – jest to rozszerzenie funkcjonalne.

2. Moduł „Total Recall”

2.1. Podstawowe cechy

Moduł „Total Recall”, wprowadzony w wersji 11, pozwala na skuteczną i wydajną obsługę danych historycznych. Bazuje on na znanej z wcześniejszych wersji technologii „Flashback”. Technologia Flashback pozwala na składowanie danych zmodyfikowanych. Głównym powodem wykorzystywania tej technologii jest zabezpieczenie danych przed skutkami różnego typu awarii (począwszy od awarii nośników danych, a skończywszy na błędach użytkowników). Niejako efektem ubocznym jest dostęp do danych archiwalnych, jedynym problemem jest określenie czasu życia archiwum. Związane jest to z ilością danych składowanych w obszarze Flashback i miejscem potrzebnym na dyskach (składowane są wszystkie dane, nie ma sposobu na określenie wybranych tabel). Flashback Query pozwala na dostęp do danych, które znajdują się w obszarze Flashback.

Częścią modułu TotalRecall jest Flashback Data Archive, który jest rozszerzeniem technologii Flashback. Flashback Data Archive pozwala na wybór tabel, które są archiwizowane i na definicję okresu czasu, przez który dane będą one składowane.

Najważniejsze cechy modułu „Total Recall” to:

- Autoryzacja dostępu do danych historycznych.
- Zabezpieczenie danych przed zmianami
- Prosty sposób dostępu do danych
- Efektywne zarządzanie magazynem danych
- Bazuje na sprawdzonej technologii Flashback

2.2. Technologia „Flashback Data Archive”

Flashback Data Archive składa się z kilku elementów: zestaw przestrzeni tabel, w których są składowane dane archiwalne, zestaw polityk dotyczących sposobu składowania i obsługi tych danych (np. czas przechowywania danych) oraz z dodatkowego procesu FBDA (FlashBack Data Archiver), który jest odpowiedzialny za pobieranie i składowanie oryginalnych danych do przestrzeni Flashback Data Archive. Po upływie okresu składowania dane są automatycznie usuwane (można je również usunąć samodzielnie przed upływem określonego czasu).

Archiwalne tabele są rozbudowane o kilka dodatkowych kolumn, w szczególności ze znacznikami czasowymi wystąpienia zmiany. Każde usuwanie bądź modyfikacja wiersza powoduje jego archiwizowanie (wstawianie nowego wiersza nie ma wpływu na tabele archiwalne). Zapis do tabel archiwalnych jest realizowany przez proces FBDA okresowo (domyślnie co 5 minut).

Składowanie danych archiwalnych nie wymaga zmian w aplikacji (dostęp może wymagać pewnych zmian, np. użycia klauzuli „as of”), wymagane jest jedynie zdefiniowanie, które dane (z jakich tabel) i przez jaki czas mają być archiwizowane. Istnieje możliwość zdefiniowania wielu obszarów Flashback Data Archive (każdy z własnym czasem utrzymywania danych).

2.3. Określenie wymagań archiwizacji

Przy obsłudze danych historycznych ważne jest przede wszystkim określenie, które dane i jak długo mają być dostępne (składowane). Czas dostępności danych archiwalnych jest definiowany dla obszarów Flashback Data Archive, a więc dla różnych czasów potrzebne będą różne obszary (a co za tym idzie dodatkowe przestrzenie tabel). Ponieważ dane archiwalne składowane są w innych przestrzeniach tabel (dedykowanych dla obszarów Flashback Data Archive), można je umieścić na innych dyskach (np. na innej macierzy, która oferuje zasoby dyskowe o dużej pojemności i mniejszej szybkości, dzięki czemu koszt utrzymywania dużej ilości danych archiwalnych jest odpowiednio niski).

Trzeba również pamiętać, że część danych powinna być okresowo usuwana z bazy podstawowej (dane te zostaną automatycznie umieszczone w archiwum). Przeniesienie danych historycznych w inne miejsce (usunięcie ich z bazy podstawowej) pozwoli na zmniejszenie ilości danych bieżących, a co za tym idzie poprawienie wydajności. Należy tylko dobrze określić, kiedy dane mogą być przeniesione do archiwum (usunięcie ich z bazy nie powinno powodować zakłóceń w dostępie do innych danych).

Ważne jest również, kto ma dostęp do tych danych i w jaki sposób się do nich dostaje.

2.4. Implementacja

Włączenie składowania danych jest stosunkowo proste – należy zdefiniować nowy obszar Flashback Data Archive (i przypisać doń wcześniej zdefiniowane przestrzenie tabel), zdefiniować jego charakterystykę oraz dla określonej tabeli włączyć (bądź wyłączyć) logowanie. Może to być wykonane podczas tworzenia tabeli, jak również poprzez modyfikację (jeśli obszar Flashback Data Archive nie jest domyślny, to konieczne jest wskazanie jego nazwy). Wszystkie tabele związane z tym samym obszarem Flashback Data Archive mają ten sam czas życia. Do tworzenia obszaru

wymagane jest dodatkowe uprawnienie „FLASHBACK ARCHIVE ADMINISTER” (lub rola DBA).

2.5. Dostęp do danych

Dostęp do danych na ogół będzie wymagał modyfikacji w aplikacji lub też napisania nowych fragmentów aplikacji (w szczególności, gdy dane bieżące i historyczne są ze sobą łączone). Najprostszą metodą dostępu do danych archiwalnych jest użycie klauzuli „as of” lub „versions between” i wskazanie znacznika czasowego. W tym momencie dostajemy określone dane historyczne wyciągnięte z archiwum.

2.6. Monitorowanie

Informacje o archiwizowanych danych można znaleźć w perspektywach słownika danych – DBA_FLASHBACK_ARCHIVE (obszary archiwizowania i związany z nimi czas życia danych), DBA_FLASHBACK_ARCHIVE_TABLES (lista tabel, dla których archiwizacja jest włączona, i ich przypisanie do obszarów archiwizacji) i DBA_FLASHBACK_ARCHIVE_TS (lista przestrzeni tabel przeznaczonych do archiwizacji i ich przypisanie do obszarów archiwizacji).

2.7. Ograniczenia

Technologia ta ma pewne ograniczenia – podstawowe to brak możliwości usunięcia kolumny z tabeli. Można to wykonać poprzez wyłączenie archiwizowania, ale to powoduje usunięcie wszystkich danych archiwalnych związanych z daną tabelą.

Dane do archiwum trafiają po każdej modyfikacji (lub usunięciu) danych. Czasem jednak nie jest konieczna znajomość, jak się zmieniały dane np. w przeciągu miesiąca, tylko jaką miały wartość w określonych momentach czasowych (np. 1 każdego miesiąca o godzinie 0:00). TotalRecall zapamiętuje wszystkie zmiany, a nie tylko te niezbędne. Może to powodować nadmierne zużycie przestrzeni dyskowej. Dlatego w tym przypadku należy zadbać o usuwanie z archiwum części danych.

3. Podsumowanie

Moduł TotalRecall jest dużym krokiem w kierunku ułatwienia obsługi danych archiwalnych. Dzięki niemu dane archiwalne są na bieżąco przenoszone (utrzymywane) do archiwum. Archiwizacja na ogół wymaga tylko niewielkiego nakładu pracy, dostęp do danych w archiwum nieco więcej, jednak jest to zdecydowanie mniej, niż w przypadku całkowitej obsługi archiwizacji na poziomie aplikacji.

Drugą rzeczą, o której warto pamiętać, jest zadbanie o poprawę wydajności bazy produkcyjnej. Trzeba pamiętać i dobrze zdefiniować, kiedy i jakie dane z bazy podstawowej mogą być usunięte. Usunięcie części danych jest dopiero początkiem pracy. Należy wykonać zestaw dodatkowych czynności, aby osiągnąć pełny sukces w optymalizacji wydajności bazy produkcyjnej.

Bibliografia

- Dokumentacja produktowa Oracle
- [Ala04] Alapati, Sam R., Kim, Charles: Oracle Database 11g, (books.google.com)
- [Mik07] Mikołajczyk, K.: Archiwizacja danych historycznych, Materiały konferencyjne, PLOUG, Zakopane 2007
- [Foo08] Foote, R.: Constraints With Oracle Total Recall (Remember A Day), Richard Foote's Oracle Blog (richardfoote.wordpress.com)