ZESZYTY NAUKOWE POLITECHNIKI ŚLĄSKIEJ 1996

Seria: INFORMATYKA z. 30 Nr kol. 1315

Wojciech KADŁUBOWSKI

ODTWARZANIE SYSTEMU REPLIKOWANEGO PO WYSTĄPIENIU AWARII

Streszczenie. W publikacji omówiono działanie systemu replikacji baz danych. Opisane zostały sposoby doprowadzenia systemu replikowanego do spójności po wystąpieniu awarii. Przedstawiono metody odtwarzania bazy pierwotnej przy użyciu zarówno skoordynowanych kopii archiwalnych, jak również tylko kopii bazy pierwotnej. Zaprezentowane zostały możliwości odtworzenia kolejek replikatora oraz sposoby wykrywania błędów w systemie replikacyjnym. We wnioskach podano wskazówki dotyczące zabezpieczeń najważniejszych komponentów systemu.

RECOVERY OF REPLICATION SYSTEM AFTER FAILURE.

Summary. In this article the work of database replication systems was discused. The methods of bringing the replication system to consistency state have been demonstrated. The pimary database recovery methods by usage of coordinated dumps, as well as primary database dump have been shown. Among others, the possibility of stable queue recovery and methods of error detection in replication system were presented. The summary presented suggestions for safety precautions for main system components.

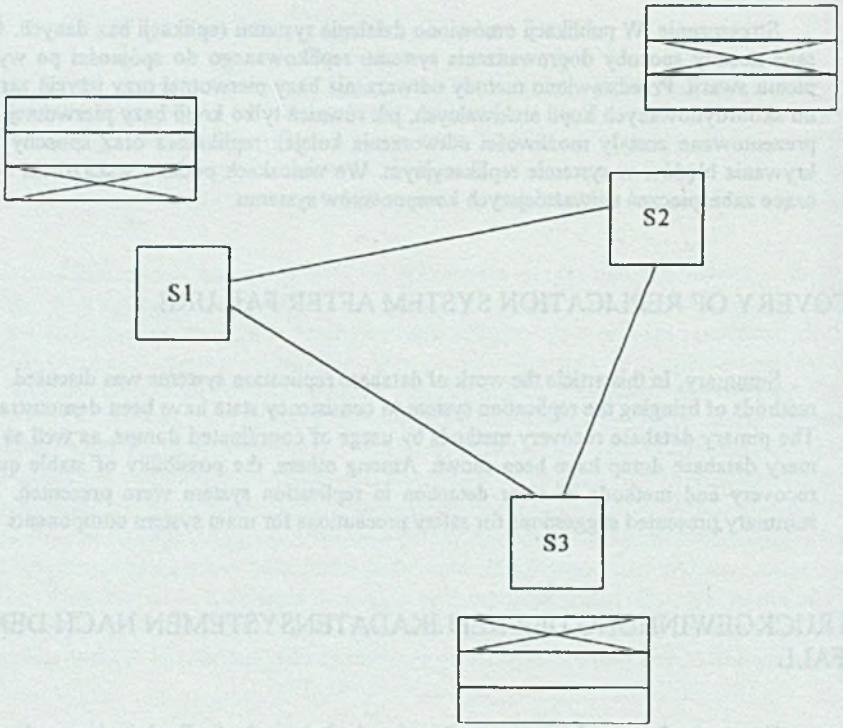
DIE RUCKGEWINNUNG DER REPLIKADATENSYSTEMEN NACH DEM UNFALL

Zusammenfassung. In diesem vorligenden Artikel wurde die Funktionierung der Replikadatenbanken dargestellt. Es wurde die Methoden, die diese Systeme fur den vorhergehende Stand nach dem Unfall fuhren geschrieben. Der Aritkel stellt also die Methoden der Ruckgewinnung der Hauptkopie durch die Benutzung koordinierte urkundlichen Kopien oder nur urkundliche Hauptkopie dar. Im schluss wurde die Hinweisungen betreffende der Vorsichten der wichtigsten Systemkomponenten gegeben.

Źródło finansowania: Komitet Badań Naukowych - 108/E-364/BW/95

1. Wstęp

Systemy rozproszone są dziś powszechnie wykorzystywane. Typowy system rozproszony posiada zasoby zgromadzone w różnych węzłach sieci. W przypadku bazy danych fragmenty tablic rozmieszczone są na różnych serwerach, połączonych ze sobą poprzez sieć. Jeśli aplikacja klienta uaktualnia informacje zawarte w bazie, to system rozproszony dokonuje zmian w odpowiednich tablicach (mogą się one znajdować na jednym bądź na kilku serwerach). Wymaga się przy tym, by wszystkie fragmenty tablic, które mają być uaktualniane, były dostępne w momencie wykonywania transakcji rozproszonej (patrz rys. 1).



Przy serwerach zaznaczono tablice bazy rozproszonej
S1-S2 - Serwery bazy rozproszonej przechowujące fragmenty tablic

Rys. 1. Przykład bazy rozproszonej
Fig. 1. Example of distributed database

Pewnym przykładem systemu rozproszonego może być także replikacja (częściowa lub całkowita) bazy danych. W przypadku replikacji baza danych (lub jej część) zostaje skopiowana i umieszczona na kilku serwerach w sieci (lokalnej lub rozległej). Aplikacje klientów uaktualniają bazę pierwotną, natomiast system sam rozsyła zmiany dokonane w bazie do jej replik. Rozsyłanie uaktualnień może odbywać się zależnie od przyjętych założeń natychmiast po dokonaniu zmian, co jakiś określony czas, po dokonaniu określonej liczby zmian itd. Systemy rozproszone ze względu na swoją specyfikę są złożone, a to z kolei powoduje różnorakie komplikacje w przypadku awarii. W prezentowanej publikacji przedstawiono metody odtworzenia spójnego systemu replikowanego po wystąpieniu awarii.

2. Działanie systemu replikacyjnego

System replikacji składa się z następujących elementów:
- bazy danych,
- agenta wykrywającego zmiany zachodzące w bazie,
- replikatora bazy danych.

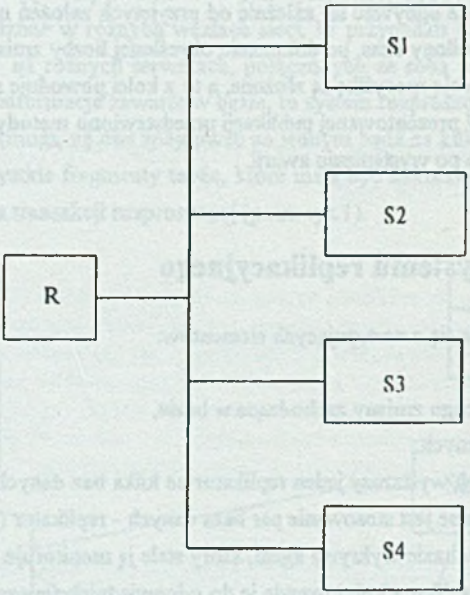
W sieciach lokalnych wystarczy jeden replikator na kilka baz danych (patrz rys. 2.). W sieciach rozległych korzystne jest stosowanie par baza danych - replikator (patrz rys. 3.).

Zmianę dokonaną w bazie wykrywa agent, który stale ją monitoruje. Przekazuje on wykryte zmiany do replikatora. Ten z kolei rozsyła je do odpowiednich serwerów lub innych replikatorów. Replikator przekazuje także wszelkie uaktualnienia otrzymane z innych replikatorów lub z innych baz do bazy, którą sam obsługuje. W przypadku gdy połączenie z którymś z serwerów ulegnie awarii, replikator przechowuje w odpowiednich kolejkach na dysku informacje dla danej repliki bazy. Po wznowieniu połączenia wszystkie zmiany odnotowane w kolejkach zostaną przesłane do odpowiednich serwerów. Podobnie przedstawia się sprawa z awariami bądź wyłączeniami serwerów i replikatorów.

3. Awarie systemu replikacji

Każdy system komputerowy może ulec awarii. Uszkodzenia występujące w systemie rozproszonym są szczególnie dotkliwe, ponieważ mogą doprowadzić do niespójności bazy. Dlatego też ważne jest, by system taki zapewniał sprawne narzędzia usuwające wszelkie usterki i przywracające spójność bazy. Opisane poniżej procedury awaryjnego zachowania się systemu replikacyjnego zostały przedstawione na przykładzie bazy Sybase. W Sybase replikator nazywa się Replication Server, a agent wychwytujący zmiany w bazie Log Transfer Manager.

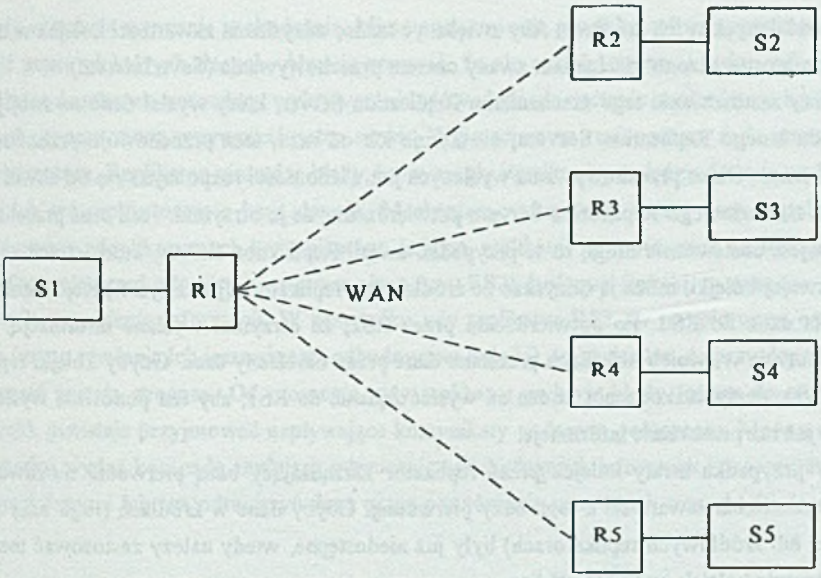
Nieniszczące stany awaryjne mogą być zazwyczaj obsługiwane automatycznie przez system bez udziału użytkowników. Replication Server pracuje w przypadku awarii sieci, awarii



R - Replikator
S1-S2 - Serwery bazy replikowanej obsługiwane przez replikator sieci LAN

Rys. 2. Zastosowanie replikatora w sieci LAN
Fig. 2. Usage of replictor in LAN network

różnych urządzeń czy też wyłączenia innych replikatorów i serwerów baz. Uszkodzenia powodujące zniszczenie pierwotnej bazy, replikatora lub repliki bazy muszą być usunięte przez administratora systemu. Wykonuje on pewne procedury dzięki, którym cały system replikacyjny jest ponownie spójny.



S1 - Serwer pierwotnej bazy danych
R1 - Replikator pierwotnej bazy danych
S2-S5 - Serwery replik bazy danych
R2-R5 - Replikatory wtórnych baz danych

Rys. 3. Zastosowanie replikatora w sieci WAN
Fig. 3. Usage of replicator in WAN network

3.1. Uszkodzenie kolejek replikatora

Zdarzają się awarie, w których zniszczone zostają kolejki dyskowe zarządzane przez Replication Server, a które zawierają informacje przesyłane do innych Replication Serverów lub serwerów bazy. Istnieją dwie procedury odbudowy kolejek:

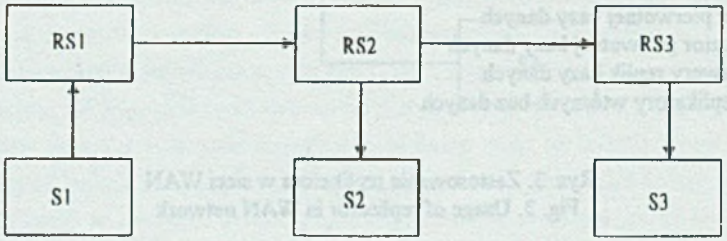
- odbudowa w trybie on-line (przy odbudowie kolejek korzysta się z informacji dostępnych jeszcze w systemie),
- odbudowa w trybie off-line (w trybie tym konieczne jest sięgnięcie do kopii archiwalnej).

W procesie odbudowywania kolejki w trybie on-line najpierw kasowane są dane z kolejek (jeśli jeszcze się tam jakieś znajdują). Następnie Replication Server próbuje odzyskać zawartość utraconej kolejki w źródłach, tzn. bądź z Replication Server, z którym posiada połączenie, bądź z logu bazy pierwotnej którą zarządza. Warunkiem powodzenia tej procedury jest do-

stępność danych w ich źródłach. Aby zwiększyć szansę odzyskania zawartości kolejek w trybie on-line, wprowadzono mechanizm zwany czasem przechowywania (Save Interval).

Przy zastosowaniu tego mechanizmu Replication Server, który wysłał dane ze swojej kolejki do innego Replication Servera, nie kasuje ich od razu, lecz przechowuje przez zadany okres czasu. Okres przechowywania wysłanych już wiadomości rozpoczyna się od chwili uzyskania z docelowego Replication Servera potwierdzenia, że je otrzymał. Jeśli czas przechowywania jest dostatecznie długi, to w przypadku awarii Replication Server, który stracił zawartość swojej kolejki, może ją odzyskać ze źródłowego replikatora (patrz rys.4.).Replikator RS1 wysyła dane do RS2. Po potwierdzeniu przez RS2, że otrzymał wysłane informacje, RS1 przechowuje w swoich kolejkach przesłane dane przez określony czas. Gdyby kolejki replikatora RS2 uległy uszkodzeniu, może on wysłać żądanie do RS1, aby ten ponownie wysłał do niego już raz przekazane informacje.

W przypadku utraty kolejek przez replikator zarządzający bazą pierwotną możliwe jest odtworzenie ich zawartości z logu bazy pierwotnej. Gdyby dane w źródłach (logu bazy pierwotnej lub źródłowych replikatorach) były już niedostępne, wtedy należy zastosować metodę odtwarzania kolejek w trybie off-line.

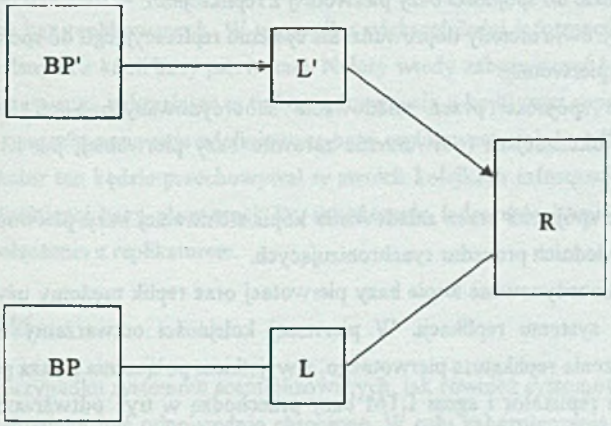


S1 - Baza pierwotna
S2-S3 - Bazy replikowane
RS1 - Pierwotny replikator
RS2-RS3 - Replikatory baz wtórnych (replikowanych)

Rys. 4. Przykład zastosowania mechanizmu Save Interval
Fig. 4. Example of usage Save Interval

Po zakończeniu odbudowy kolejek replikator wysyła komunikat do pozostałych , mających z nim łączność replikatorów, o możliwości przyjmowania informacji gromadzonych przez nie w trakcie awarii. W przypadku replikatora zarządzającego bazą pierwotną Replikation Server wysyła komunikat do agenta LTM o możliwości przyjęcia uaktualnień, które po-

wstały w trakcie usuwania uszkodzenia. Mimo zastosowania procedur odtwarzania może wystąpić utrata niektórych danych. Aby się upewnić, że nie zostały stracone żadne informacje, replikator uruchamia procedury wykrywania błędów. Jeśli nie zostanie stwierdzony brak jakiś danych, system może przywrócić pełen serwis. W przeciwnym razie system żąda interwencji administratora. Replikator sprawdza błędy, które mogły wyniknąć pomiędzy dwoma replikatorami lub też replikatorem a bazą danych. Mechanizm wykrywania utraty danych polega na znajdowaniu zduplikowanych komunikatów. Jeśli np. replikator RS3 otrzymał od RS2 już daną informacje przed odtworzeniem przez niego (tzn. RS2) dyskowej kolejki, to oznacza, że nie nastąpiło zagubienie informacji. W przypadku gdy replikator RS3 po raz pierwszy otrzyma dane (mimo wysłania ich jeszcze raz z odbudowanej kolejki), to stwierdza, że prawdopodobnie ich część została stracona. Od momentu, gdy replikator wykryje błędy (utratę przesyłanych danych), przestaje przyjmować napływające komunikaty w danym połączeniu. Można w tym przypadku wydać komendę anulującą odrzucanie nadchodzących informacji i pracować z niepełnymi danymi lub też odtworzyć dane przez powtórzenie wszystkich transakcji z logu off-line.



BP - Baza pierwotna
BP' - Baza pierwotna odtwarzana z kopii
L - Agent LTM wyłączony na czas odtwarzania
L' - Agent LTM włączony na czas odtwarzania kopii
R - Replikator

Rys. 5. Odtwarzanie kolejek replikatora w trybie off-line
Fig. 5. Recovery of stable queue of replicator in off-line mode

Na podobnej zasadzie replikator wykrywa błędy powstałe w czasie awarii pomiędzy Replication Serverem a bazą danych. Jeżeli po odtworzeniu kolejek dyskowych (lub innych elementów systemu) w trybie on-line wystąpiły błędy lub brak jest informacji w źródłach (np. logu bazy pierwotnej), to niezbędne staje się zastosowanie procedury odtwarzania w trybie off-line. Metoda off-line polega na załadowaniu starszej wersji bazy danych, a także archiwalnych logów transakcji w oddzielnej bazie oraz uruchomieniu dla nich agenta LTM. Mimo że LTM pobiera dane z archiwalnej bazy, to przesyła je do Replication Servera, tak jakby przekazywał je ze zwykłej bazy pierwotnej (patrz rys.5.). W czasie odtwarzania kolejek agent LTM bazy pierwotnej musi być wyłączony.

3.2. Awarie bazy pierwotnej

Często po niegroźnych usterkach bazy danych można odtworzyć system bez większych problemów. Najczęściej po restarcie bazy danych replikator synchronizuje się z nią, sprawdzając jednocześnie czy nie wystąpiły jakieś błędy. Jeśli uszkodzenie bazy jest poważniejsze, należy odtworzyć bazę pierwotną z kopii archiwalnych i za pomocą odpowiednich procedur replikatora doprowadzić do spójności bazy pierwotnej z replikacjami.

- Wyróżniamy dwie metody doprowadzenia systemu replikacyjnego do spójności po odtworzeniu bazy pierwotnej:
- odtworzenie spójności przez załadowanie skoordynowanych kopii archiwalnych we wszystkich lokalizacjach (odtworzenie zarówno bazy pierwotnej, jak i replik jednocześnie),
- odtworzenie spójności przez załadowanie kopii archiwalnej bazy pierwotnej oraz wykonanie odpowiednich procedur synchronizujących.

Posiadając skoordynowane kopie bazy pierwotnej oraz replik możemy użyć je do odtworzenia spójnego systemu replikacji. W pierwszej kolejności odtwarzamy bazę pierwotną. Wszystkie połączenia replikatora pierwotnego, z wyjątkiem połączenia z bazą pierwotną, zostają zawieszone, a replikator i agent LTM bazy przechodzą w tryb odtwarzania. Agent LTM przebiega w tym trybie odtworzony log bazy. Następnie zwiększany jest numer generacyjny bazy, aby nadchodzące w normalnym trybie pracy komunikaty nie były odrzucane. Podobnie postępujemy w miejscach replikacji, wykorzystując kopie archiwalne skoordynowane z kopią bazy pierwotnej. Po zakończeniu odtwarzania replik odnawiamy połączenia pomiędzy wszystkimi elementami systemu w normalnym trybie pracy. System replikacyjny przywraca pełny serwis.

Każda baza pierwotna w systemie replikacji posiada numer generacyjny bazy. Numer ten jest zapisany w bazie danych, jak również w tablicach systemowych replikatora zarządzającego bazą. W przypadku odtwarzania bazy z taśmy, odtwarzany jest również jej numer generacyjny.

Aby replikator poprawnie działał (np. mógł sprawdzać czy nastąpiła utrata lub duplikacja danych), należy dopasować jego numer generacyjny do numeru odtwarzanej bazy. Po zakończeniu odtwarzania należy powrócić do starego numeru, gdyż w przeciwnym razie replikator (a także baza) może ignorować napływające komunikaty.

Odtworzenie spójnego systemu po awarii bazy pierwotnej bez użycia skoordynowanych kopii archiwalnych nieco się różni. W tym przypadku po odzyskaniu danych pierwotnych z kopii (stosujemy procedurę podobną do poprzedniej) uruchamiamy w miejscach replikacji odpowiednie procedury zapewniające synchronizację replik z odtworzoną bazą pierwotną. Testują one wiersze w replikowanych tablicach i korygują niezgodności w stosunku do bazy pierwotnej.

3.3. Awarie baz replikowanych

Najbardziej wrażliwym elementem w systemie replikacyjnym jest baza pierwotna. Jeśli jest ona dostatecznie dobrze chroniona, to istnieje możliwość odzyskania danych w miejscach replikacji. Jedną z takich możliwości odtworzenia repliki bazy danych jest ponowne jej zdefiniowanie oraz przesłanie danych z bazy pierwotnej poprzez sieć. Sposób ten może mieć zastosowanie do małych baz replikowanych. W przypadku większej ilości informacji korzystne staje się odtworzenie danych z kopii bazy pierwotnej. Należy wtedy zabezpieczyć bazę replikowaną przed próbą dokonywania uaktualnień w trakcie odtwarzania z kopii oraz zapewnienia spójności systemu. W tym celu ponownie zdefiniowaną bazę replikowaną należy odłączyć od jej replikatora. Replikator ten będzie przechowywał w swoich kolejkach informację napływającą z na bieżąco uaktualnianej bazy pierwotnej. Po zakończeniu ładowania kopii bazy pierwotnej przywraca się połączenie z replikatorem.

3.4. Wnioski

Zarówno w przypadku systemów scentralizowanych, jak również systemów replikacyjnych dane pierwotne powinny być odpowiednio chronione. W celu zabezpieczenia najbardziej krytycznych danych można użyć duplexnigu. Jeżeli dysk komputera ulegnie uszkodzeniu, system przełączy się automatycznie na zapasowy. Do najbardziej krytycznych elementów systemu replikacyjnego , do których można zastosować duplexing, należą: pierwotna baza danych, log transakcji bazy pierwotnej oraz kolejki replikatora. Odtworzenie tych komponentów w trybie on-line nie zawsze jest możliwe, natomiast odzyskanie ich z kopii archiwalnych może być czasochłonne i złożone.

LITERATURA

[1] Butler B., Canter S.: Bazy danych SQL, PC Magazine po polsku, 1994, No.2.

[2] Edelstein H.: The Challenge of replication, DBMS, 1995, Vol.8, No. 3, 46-52.

[3] Edelstein H.: The Challenge of replication, DBMS, 1995, Vol.8, No. 4, 62-70.

[4] Kadłubowski W.: Zwiększenie bezpieczeństwa systemu rozproszonego poprzez zastosowanie replikacji bazy daych. II Conference on Real-Time Systems '95, Szklarska Poręba, September 20-23, 1995.

[5] Kadłubowski W.: Wykorzystanie replikacji baz danych w systemach rozproszonych, II Krajowa konferencja "Komputerowe wspomaganie badań naukowych", Wroclaw, 14-16 grudzień 1995r.

[6] Łakomy M.: Sybase i produkty dla środowiska klient/serwer, ComputerWorld, 1994, No. 7, 24-27.

[7] Łakomy M.: Replikacja danych w bazach, ComputerWorld, 1994, No. 20, 41.

[8] Dokumentacja, Replication server overview, Sybase, Inc. oct 20,1993.

[9] Dokumentacja, Replication server administration guide, Sybase, Inc. oct 21,1993.

[10] Dokumentacja, Replication server design guide, Sybase, Inc. oct 21,1993.

Rezenzent: Dr inż. Katarzyna Stąpor

Wpłynęło do Redakcji 20 grudnia 1995 r.

Abstract

In this article the work of replication system on Sybase database has been shown. Author presents the elements of such systems and its work in replication environment. The replicator in LAN (see Fig.2.) and WAN (see Fig. 3.) networks has been discussed. The automated procedures usage in case of non destroying failures were shown (see chapter 3.1.). Examples of recovery from different destroying failures, when they occure, were discussed as well. Among others, the recovery of stable queue of replicator in an on-line and off-line mode was presented. Also the recovery of primary database and its replication was demonstrated. The mechanism of automated data retrieval in on-line mode, called *Save Interval* was shown(see Fig.4.,chapter 3.1.). The primary database recovery methods by usage of coordinated dumps as well as primary database dumps were also presented. The summary presented suggestions for safety precautions for main system components.