STUDIA INFORMATICA 2009
Volume 30 Number 2B (84)

Adam PIÓRKOWSKI, Grzegorz GAJDA
Akademia Górniczo-Hutnicza, Katedra Geoinformatyki i Informatyki Stosowanej

KONSTRUKCJA WIELOWYMIAROWEJ BAZY DANYCH GEOLOGICZNYCH

Streszczenie. W artykule opisano tworzenie wielowymiarowej bazy danych geologicznych. Hurtownię danych utworzono na podstawie bazy danych odsłonięć geologicznych i bazy zagrożeń geologicznych. Dokonano wyróżnienia tabel wymiarów i tabel faktów. Zaproponowano schematy zasilenia tabel. Konstrukcję przeprowadzono w środowiskach IBM DataStage i Talend Open Studio.

Słowa kluczowe: hurtownie danych, wielowymiarowe bazy danych, bazy danych geologicznych

A CONSTRUCTION OF MULTIDIMENSIONAL GEOLOGICAL DATABASE

Summary. A construction of multidimensional geological database is described in this article. Data warehouse is created using databases for outcrops and database for geodynamic threats. The fact and dimension tables has been chosen. The schemas of loading are proposed. The construction is created for IBM DataStage and Talend Open Studio environments.

Keywords: data warehouse, multidimensional databases, geological databases

1. Wprowadzenie

Nauki o ziemi to dziedzina dostarczająca ogromnych ilości danych. Owe informacje wymagają dobrze zaprojektowanych systemów do składowania informacji oraz ich przetwarzania. Liczne działy geologii i geofizyki doczekały się w ciągu wielu lat implementacji takich systemów. Problemem tutaj staje się różnorodność zastosowanych technologii oraz sposób konstruowania aplikacji. Tworzenie uniwersalnej bazy danych geologicznych i

geofizycznych może być dość trudne, chociażby ze względu na trudną migrację. Dużo prostszym rozwiązaniem jest połączenie istniejących rozwiązań i stworzenie hurtowni danych geologicznych. Pierwsze przymiarki takiego pomysłu zostały przedstawione w pracy [1].

Jedną z pierwszych baz danych, jakie powstały na Wydziale Geologii, Geofizyki i Ochrony Środowiska, są GeoKarpaty – baza danych sedymentologicznych dla polskich Karpat [2]. Baza ta została stworzona w systemie MS Access. W pracy [1] przedstawiono jej schemat (silnie zdenormalizowany). GeoKarpaty doczekały się nowszych, internetowych implementacji [3, 4], wykorzystujących rozwiązania systemów zarządzania bazami danych Postgresql i MySQL. Ich schematy [1, 4, 5] są już znormalizowane.

Inna baza, która powstała przy współpracy Wydziału Geologii, Geofizyki i Ochrony Środowiska i Katedry Informatyki Akademii Górniczo-Hutniczej na zlecenie Ministerstwa Środowiska to Geozagrożenia [6]. Baza owa ma również znormalizowany schemat, jako technologie wybrano parę PHP + Postgresql.

Wstępnym celem pracy jest pokazanie możliwości połączenia wymienionych baz i stworzenie hurtowni danych geologicznych. W dalszej części badań przewidziane jest dołączanie kolejnych baz do projektu.

2. Migracja danych geologicznych

Podstawowym procesem wymiany danych między bazami danych jest migracja danych. Proces ten dokonywany jest przez oprogramowanie, które potrafi łączyć się z różnymi rozwiązaniami baz danych i umożliwia programowanie transformacji danych. Takie założenia spełniają systemy wspomagające tworzenie hurtowni danych, wśród których szczególną uwagę zwrócono na IBM DataStage [7] i Open Talend Studio [8]. Pierwszy produkt został już omówiony pod kątem tworzenia geologicznych baz danych [1]. Open Talend Studio jest rozwiązaniem bardzo podobnym, darmowym.

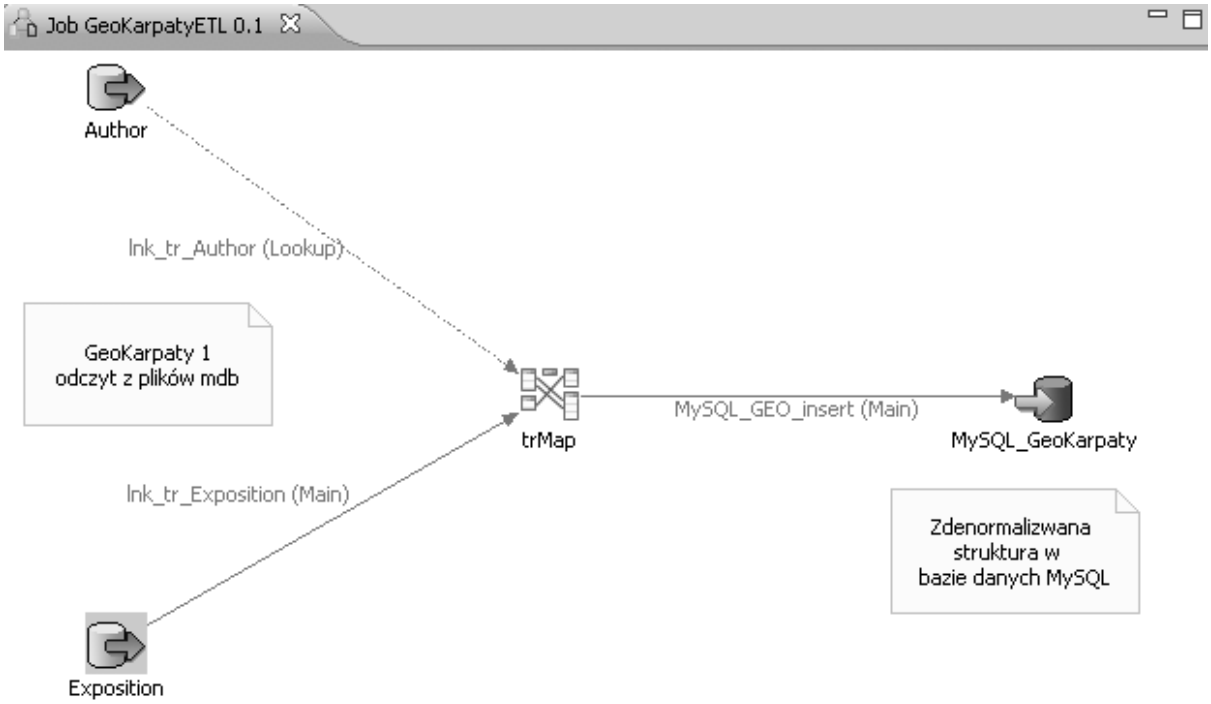
Migracja danych przeprowadzona w środowisku Talend Open Studio przebiegła analogicznie jak w IBM DataStage, w tym przypadku została zachowana zdenormalizowana struktura bazy danych GeoKarpaty, rozszerzona o dane pochodzące z tabeli *Author*.

Talend Open Studio wykorzystuje gotowe komponenty realizujące połączenie ze zdefiniowaną bazą danych. Zaletą jest duża ilość komponentów wykorzystujących nie tylko ODBC i JDBC, lecz również natywne sterowniki dostępu do bazy danych (m.in.: MySQL, PostgreSQL, MS SQL Server, Oracle, Access). Inaczej niż w przypadku IBM DataStage została zorganizowana ich obsługa. Rozróżniane są komponenty odczytujące i zapisujące (aktualizujące). Przypływ danych wizualizowany jest za pomocą połączeń (linków), natomiast kie-

runek przepływu wskazują strzałki. Pojedynczy komponent zarówno odczytujący, jak i zapisujący może korzystać jednocześnie tylko z pojedynczego połączenia, przez co w celu odczytu danych z dwóch tabel konieczne jest użycie dwóch komponentów.

Zadanie realizujące migrację (rys. 1) składa się w tym przypadku z trzech komponentów programowych (dwa odczytujące – *tDBInput*, jeden zapisujący – *tMysqlOutput*). Odczyt danych z pliku *.mdb wykorzystuje ODBC, natomiast zapis – natywny sterownik do bazy danych MySQL. Parametry odczytu i zapisu mogą zostać wprowadzone na dwa sposoby:

- ręczne uzupełnienie dla danego komponentu,
- z wykorzystaniem wcześniej zdefiniowanych wartości w repozytorium projektu (może dotyczyć również zapytań).



Rys. 1. Migracja danych systemu GeoKarpaty pomiędzy różnymi silnikami baz danych z wykorzystaniem Talend Open Studio

Fig. 1. Migration of data for various database engines

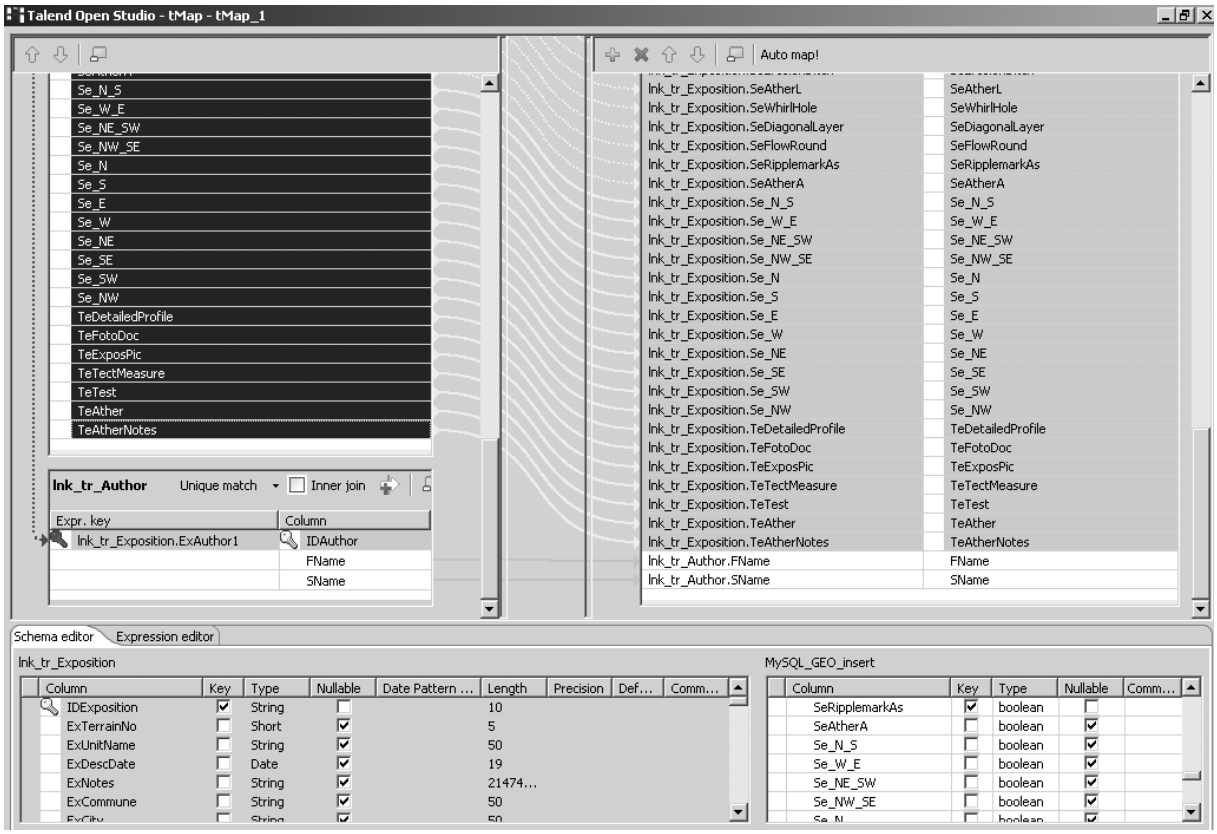
Aktualizacja danych z użyciem komponentu zapisującego może odbywać się na kilka sposobów:

- zapis (ang. insert),
- aktualizacja (ang. update),
- zapis lub aktualizacja (ang. insert or update),
- aktualizacja lub zapis (ang. update or insert),
- usunięcie (ang. delete).

Aktualizacja metadanych tabeli, do której przeprowadzany jest zapis z wykorzystaniem komponentu, może przebiegać na kilka sposobów:

- usuń i utwórz tabele (ang. drop and create table),
- utwórz tabele (ang. create table),
- utwórz tabelę jeżeli nie istnieje (ang. create table if not exists),
- usuń dane z tabeli (ang. clear table),
- brak akcji.

Kolejnym komponentem Talend Open Studio, wykorzystanym w procesie migracji danych systemu GeoKarpaty, jest *tMap* (rys. 2). Za jego pomocą dokonano złączenia danych z tabeli *Exposition* i *Author* według kolumny o nazwie *IDAuthor*. Odwzorowanie danych odbyło się w stosunku 1:1 i w odróżnieniu od odniesienia wykonanego w IBM DataStage zachowana został pierwotna zdenormalizowana struktura.



Rys. 2. Schemat migracji danych w Talend Open Studio
Fig. 2. Migration data schema for Talend Open Studio

Komponent *tMAP* jest stricte związany z przetwarzaniem i operuje na już pobranych danych, które mogą pochodzić z jednego lub kilku źródeł. Posiada możliwość przeprowadzania różnego rodzaju złączeń (wewnętrzne i zewnętrzne), dokonywania filtracji według wybranych wierszy lub kolumn, a także ich odniesienia do odrębnie zdefiniowanych struktur wyjściowych. Oferuje on także wygodny i intuicyjny interfejs graficzny, za pomocą którego dużą część pracy można zaprojektować myszką. Komponent tMAP umożliwia dokonywanie transformacji pośredniej do zmiennych przechowywanych w pamięci, które następnie są

odnoszone do właściwych struktur wyjściowych. Takie podejście stanowi warstwę pośrednią i daje dodatkowe możliwości transformacji danych wejściowych.

3. Tworzenie schematu hurtowni danych geologicznych

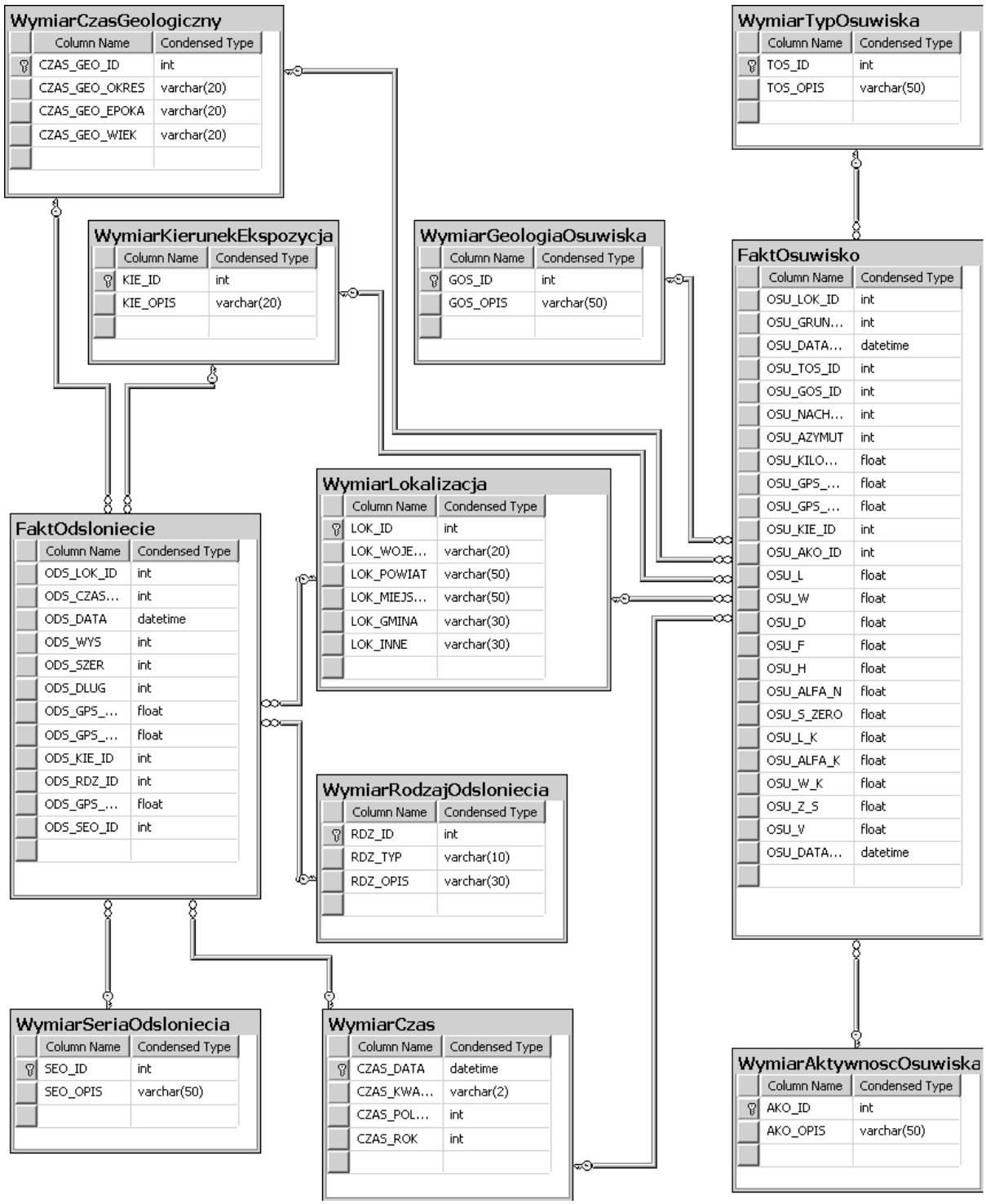
Drugi etap przeprowadzonych prac konstrukcyjnych to przygotowanie proponowanego schematu hurtowni danych geologicznych (rys. 3) na podstawie baz danych GeoKarpaty II i Geozagrożenia. Na tym etapie prac jako silnik docelowej hurtowni danych mogła zostać wykorzystana dowolna relacyjna baza danych *open source*, zdecydowano się jednak na MS SQL Server w darmowej wersji Express. W chwili obecnej przechowuje ona jedynie schemat, jednak patrząc przyszłościowo w przypadku wdrożenia hurtowni geologicznej i osiągnięcia przez nią dużych rozmiarów rozwiązanie typu open source może nie zapewnić dostatecznej wydajności. MS SQL Server zapewnia dobrą skalowalność, gdyż w łatwy sposób można dokonać jego aktualizacji do wyższej wersji (Standard, Enterprise), przystosowanej do przechowywania i indeksowania dużych zbiorów danych i tworzenia klastrów danych.

Zaprojektowany schemat hurtowni danych geologicznych ma postać konstelacji faktów. Wyróżniono w nim dwie centralne tabele faktów oraz osiem tabel wymiarów, opisujących dane fakty. Do faktów zaliczamy:

- FaktOdsloniecie – dane związane z odsłonięciami w Karpatach Fliszowych,
- FaktOsuwisko – dane na temat osuwisk z rejonu całej Polski.

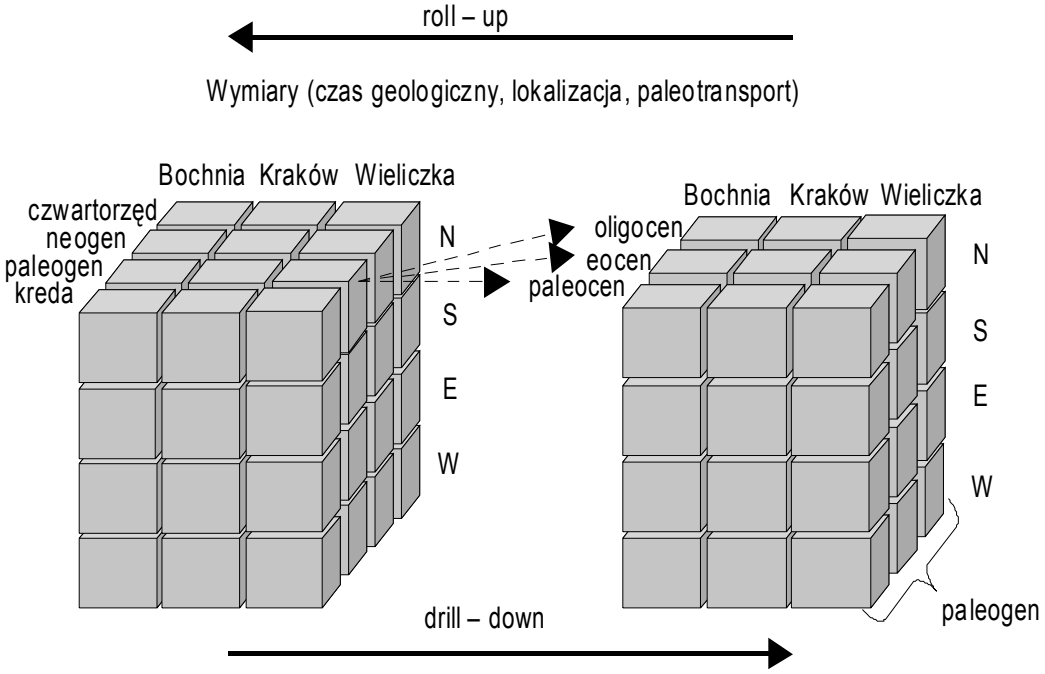
Do tabel wymiarów zaliczamy:

- WymiarCzasGeologiczny – reprezentuje tabelę stratygraficzną z podziałem na okres, epokę i wiek,
- WymiarTypOsuwiska – kwalifikacja typu osuwiska,
- WymiarKierunekEkspozycja – orientacyjny kierunek paleotransportu dla odsłonięcia lub ekspozycji zbocza skarpy w przypadku osuwiska,
- WymiarGeologiaOsuwiska – kwalifikacja rodzaju geologicznego osuwiska,
- WymiarLokalizacja – reprezentuje lokalizację danego zjawiska geologicznego,
- WymiarRodzajOdsloniecia – kwalifikacja rodzaju odsłonięcia,
- WymiarSeriaOdsloniecia – może posłużyć do kwalifikacji odsłonięć z innych serii poza Śląską, np. pienińskiego pasa skałkowego, czy niecki podhalańskiej,
- WymiarCzas – kwalifikuje okres wykonania badań,
- WymiarAktywnoscOsuwiska – kwalifikacja aktywności osuwiska.



Rys. 3. Proponowany schemat hurtowni danych geologicznych
Fig. 3. Proposed schema for geological warehouse

Do danych wymiarów, których charakterystyka pozwala na stworzenie hierarchii atrybutu, możemy zaliczyć *WymiarCzasGeologiczny*, *WymiarCzas*, *WymiarRodzajOdsloniecia*. Takie uszeregowanie pozwala np. na zaawansowane rozwijanie i zwijanie wzdłuż hierarchii danego atrybutu (rys. 4).



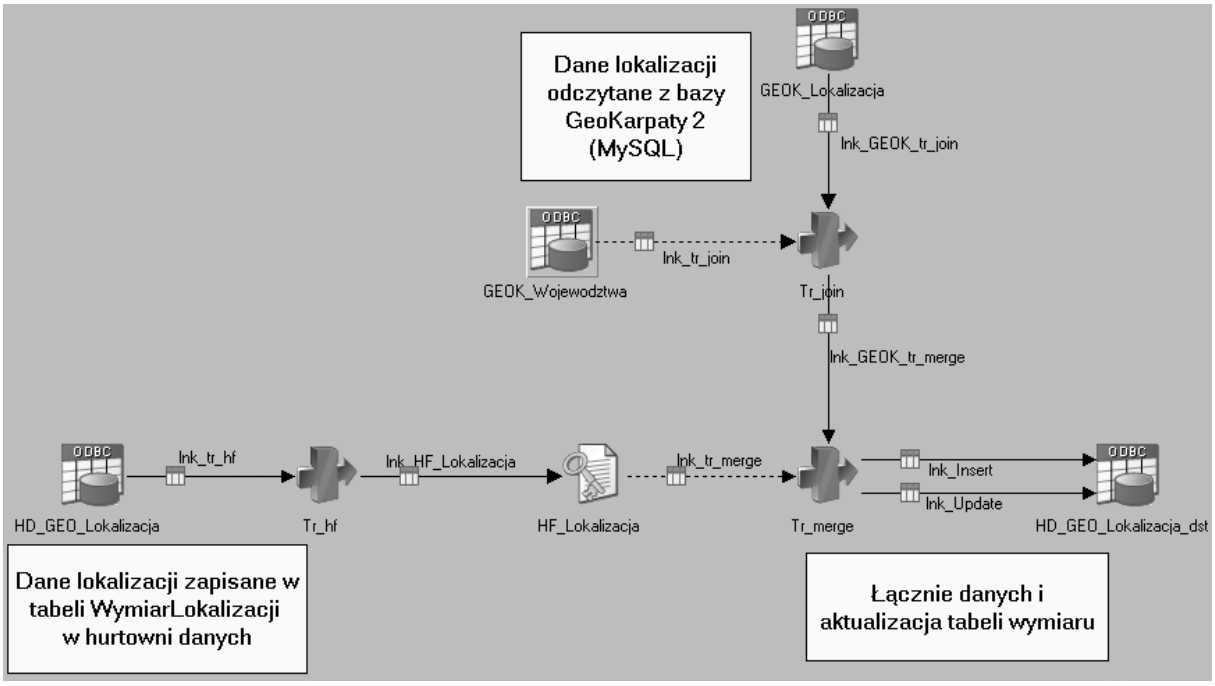
Rys. 4. Hierarchia wymiaru hurtowni danych geologicznych na podstawie czasu geologicznego
Fig. 4. Dimension hierarchy for geological warehouse

4. Implementacja zadań zasilających hurtownię

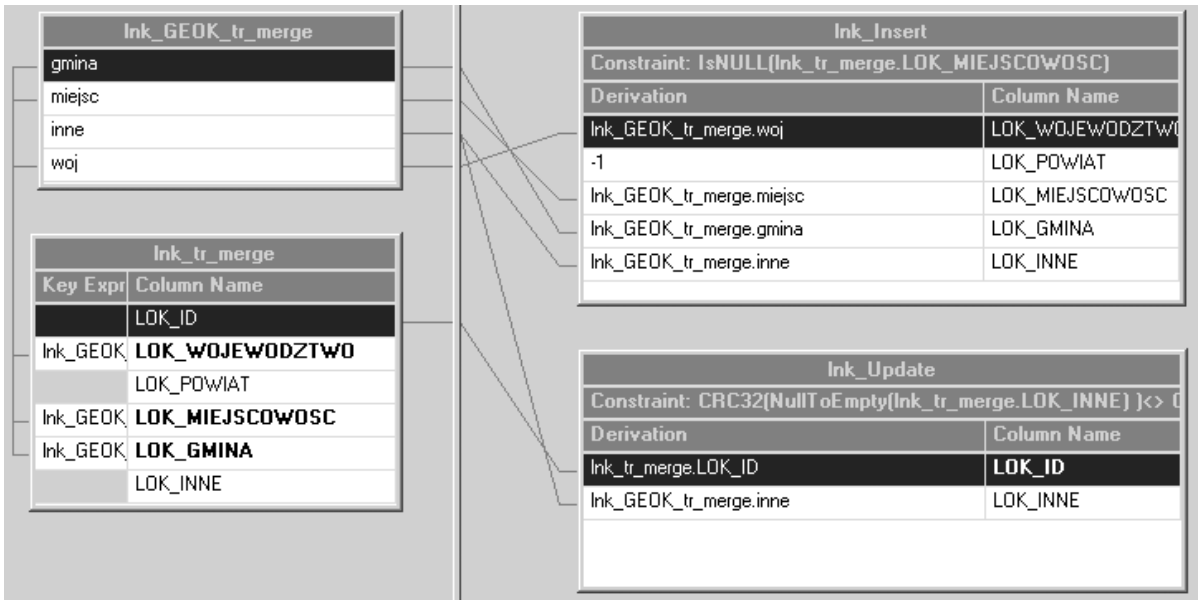
Końcowy etap prac konstrukcyjnych to stworzenie zadań zasilających wybrany wymiar i fakt geologiczny. Implementacji został poddany *WymiarLokalizacja* i *FaktOsuwisko*. Prace podobnie jak w przypadku migracji zostały wykonane w dwóch środowiskach IBM DataStage i Talend Open Studio.

Zasilenie wymiaru lokalizacja (rys. 5) odbywa się na podstawie źródłowych tabel *lokalizacja* i *wojewodztwa* z bazy danych GeoKarpaty II. Dane otrzymane w złączeniu tych tabel porównywane są ze stanem faktycznym zapisanym w docelowej tabeli *WymiarLokalizacja*, który jest odczytywany z pliku mieszającego (*HF_Lokalizacja*).

Porównanie zawartości tabel źródłowych z aktualnym stanem zapisanym w hurtowni odbywa się z udziałem komponentu Transformer Stage (rys. 6). W przypadku wystąpienia nowych danych lokalizacji dokonywane jest wstawienie danych do docelowej tabeli *WymiarLokalizacja*, natomiast w przypadku zmiany opisu danej lokalizacji następuje aktualizacja.

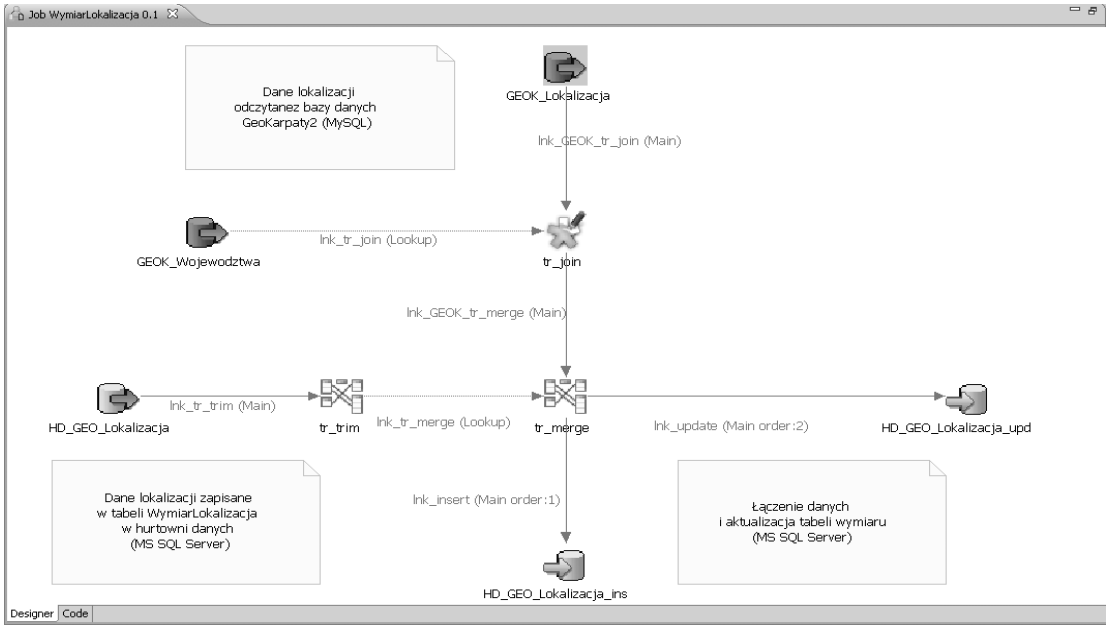


Rys. 5. Proponowany schemat zasilenia tabeli WymiarLokalizacja w IBM DataStage
Fig. 5. Schema of loading the table WymiarLokalizacja in IBM DataStage



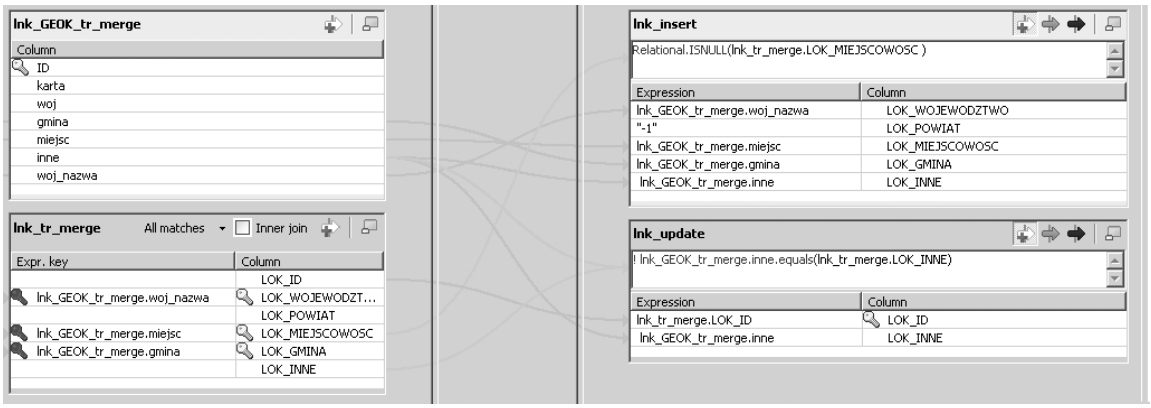
Rys. 6. Porównanie danych źródłowych i docelowych tabeli WymiarLokalizacja w IBM DataStage
Fig. 6. A comparison of data for table WymiarLokalizacja in IBM DataStage

Analogiczne zadanie zasilenia danymi tabeli *WymiarLokalizacja,* realizujące opisaną powyżej funkcjonalność, zostało przeprowadzone w Talend Open Studio (rys. 7). Do połączenia z bazami danych zostały wykorzystane sterowniki natywne w przypadku GeoKarpaty II (MySQL) i w przypadku hurtowni danych geologicznych sterownik ODBC (MS SQL Server).



Rys. 7. Proponowany schemat zasilenia tabeli WymiarLokalizacja w Talend Open Studio
Fig. 7. Schema for loading WymiarLokalizacja table in Talend Open Studio

Złączenie źródłowych danych z tabel *lokalizacja* i *wojewodztwa* w bazie danych GeoKarpaty II z użyciem Talend Open Studio zostało wykonane za pomocą komponentu tJoin, który daje możliwość przeprowadzenia złączenia wewnętrznego i zewnętrznego. Właściwe odniesienie danych do docelowych struktur oraz filtracja dokonywane są za pomocą komponentu tMap (rys. 8), który został już wcześniej opisany.

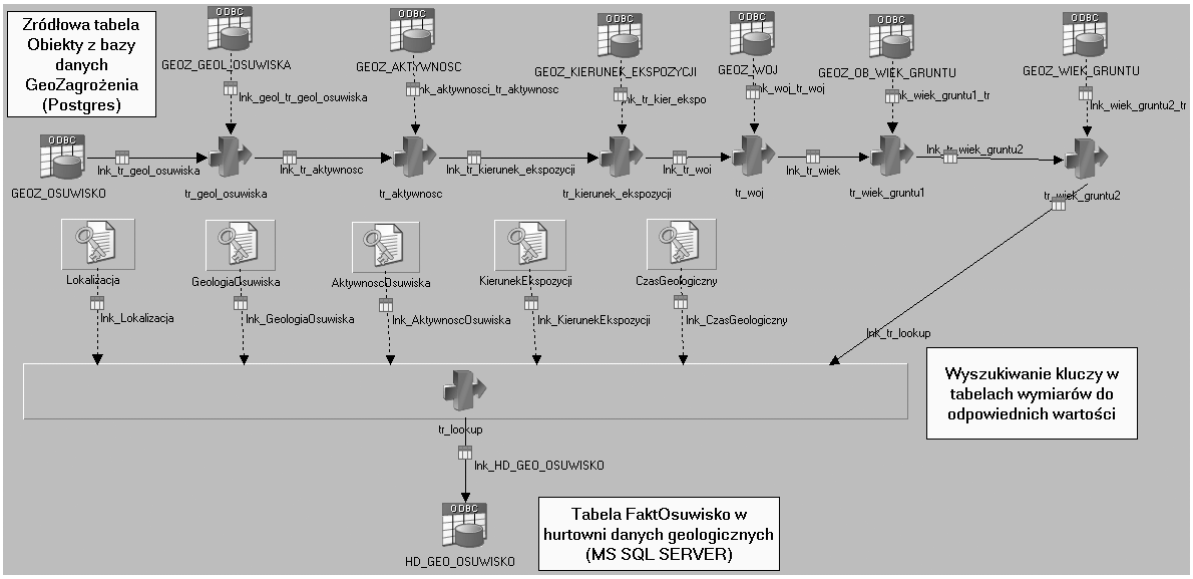


Rys. 8. Porównanie danych źródłowych i docelowych tabeli WymiarLokalizacja w Talend Open Studio
Fig. 8. A comparison of data for table WymiarLokalizacja in Talend Studio

Atutem rozwiązania zaprojektowanego w Talend Open Studio w porównaniu do IBM DataStage jest elastyczność w sposobie wykonania, gdyż na jego podstawie generowany jest kod źródłowy w zależności od wybranego rodzaju projektu w języku Java lub Perl, który w praktyce z dołączonymi odpowiednimi bibliotekami może być uruchamiany dowolnie poza

środowiskiem programistycznym. Zadanie stworzone w IBM DataStage może być obsługiwane jedynie za pomocą dedykowanego programu uruchomieniowego DataStage Director.

Zadanie realizujące zasilenie faktu zostało zrealizowane dla tabeli *FaktOsuwisko* (rys. 9). Źródłową bazą danych w tym przypadku stanowią Geozagrożenia i tabela Obiekty, do której w celu uzyskania całości danych dokonano wielokrotnych złączeń z tabelami: *rodz_geol_osuwisko*, *aktywnosc*, *ekspozycja_zbocza_skarpy*, *wojewodztwa*, *wiek_gruntow*, *obiekty_wiek_gruntow*.

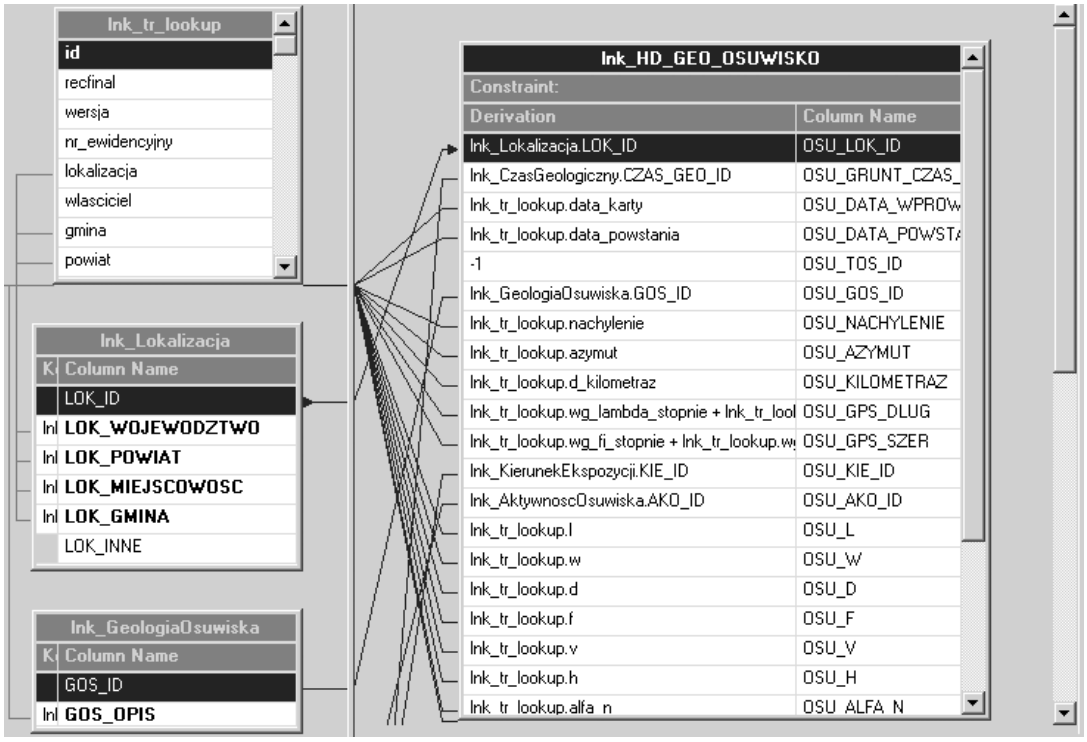


Rys. 9. Schemat zasilenia tabeli FaktOsuwisko w IBM DataStage
Fig. 9. Schema for loading FaktOsuwisko table in IBM DataStage

Dane ze skorelowanych wymiarów odczytywane są z plików mieszających, natomiast całość transformacji (rys. 10) polegającej na wyszukaniu kluczy w odpowiadających tabelach wymiarów i zapisaniu ich do tabeli faktu przeprowadzona została z użyciem komponentu Transformer Stage (*tr_lookup*).

5. Podsumowanie

W pracy przedstawiono możliwości łączenia baz danych geologicznych w hurtownię. Wydzielono tablice wymiarów i faktów. Naświetlono aspekty zasilenia tych tabel. Do konstrukcji wykorzystano produkt komercyjny i produkt darmowy (open source). Dalsze badania powinny objąć większą liczbę baz danych geologicznych oraz powinny podążyć w stronę wizualizacji i przetwarzania danych, w szczególności interesujące jest zastosowanie zbiorów przybliżonych do opisu zjawisk geofizycznych [9].



Rys .10. Łączenie i wyszukiwanie kluczy w tabelach wymiarów w IBM DataStage
Fig. 10. Linking keys for dimension tables in IBM DataStage

Praca finansowana w ramach badań statutowych KGIS nr 11.11.140.561.

BIBLIOGRAFIA

1. Gajda G., Piórkowski A.: Możliwości konstrukcji hurtowni danych geologicznych. Bazy danych – rozwój metod i technologii – bezpieczeństwo, wybrane technologie i zastosowania. Praca zbiorowa pod red. Stanisława Kozielskiego [i in.]. WKŁ, Warszawa 2008.
2. Kotlarczyk J., Krawczyk A., Leśniak T., Słomka T.: Geologiczna baza danych GeoKarpaty dla polskich Karpat fliszowych. Wydawnictwo własne WGGiOŚ AGH, Kraków 1997.
3. Onderka Z., Piórkowski A.: Projekt i implementacja geologicznej bazy danych w sieci Internet. W: Nowe technologie sieci komputerowych - praca zbiorowa. WKŁ, Warszawa 2006.
4. Malec O., Kyc M., Piórkowski A.: Internetowa baza danych odsłonięć GeoKarpaty II. Materiały kongresowe Pierwszego Polskiego Kongresu Geologicznego, Kraków 26–28 czerwca 2008. Polskie Towarzystwo Geologiczne, 2008.
5. Gajda G., Piórkowski A.: Konstrukcja rozproszonej bazy danych dla danych odsłonięć geologicznych w Karpatach. Materiały kongresowe Pierwszego Polskiego Kongresu Geologicznego, Kraków 26–28 czerwca 2008. Polskie Towarzystwo Geologiczne, 2008.

6. Valenta M., Siwik L.: Geo-zagrożenia – komputerowy system ewidencji zagrożeń geodynamicznych w Polsce. Bazy danych – struktury, algorytmy, metody – wybrane technologie i zastosowania. Praca zbiorowa pod red. Stanisława Kozielskiego [i in.]. WKŁ, Warszawa 2006.
7. IBM InfoSphere DataStage. http://www.ibm.com/software/data/integration/datastage/.
8. Talend: First provider of Open Source ETL and Data Integration Software. http://www.talend.com/.
9. Stepaniuk J.: Rough – Granular Computing in Knowledge Discovery and Data Mining. Springer-Verlag, Berlin 2008.

Recenzent: Dr inż. Tomasz Traczyk

Wpłynęło do Redakcji 27 stycznia 2009 r.

Abstract

The data warehouses for geological data is the main topic of this article. Some of simple geological databases are included. The first one – Geocarpathian – is a database for outcrops. The second is a newer version of Geocarpathian – its normalized and has a thin-client implemented. Geo-threats is the third database, collecting geodynamic threats. The first step is to define a construction of multidimensional geological database. The schema for this is shown on fig. 3.. The fact and dimension tables has been chosen. The schemas of loading are proposed. The construction is created for IBM DataStage and Talend Open Studio environments.

Adres

Adam PIÓRKOWSKI: Akademia Górniczo-Hutnicza, Katedra Geoinformatyki i Informatyki Stosowanej, al. Mickiewicza 30, 30-059 Kraków, Polska, pioro@agh.edu.pl.
Grzegorz GAJDA: Akademia Górniczo-Hutnicza, Katedra Geoinformatyki i Informatyki Stosowanej, al. Mickiewicza 30, 30-059 Kraków, Polska.