STUDIA INFORMATICA 2011
Volume 32 Number 2B (97)

Joanna KLISIEWICZ, Adam PIÓRKOWSKI, Stanisława PORZYCKA
Akademia Górniczo-Hutnicza, Katedra Geoinformatyki i Informatyki Stosowanej

KONSTRUKCJA PROCESU ETL DLA DANYCH PRZESTRZENNYCH[1]

Streszczenie. Stale wzrastające zastosowanie baz danych przestrzennych we współczesnych systemach informacyjnych wiąże się z przetwarzaniem dużej ilości danych. Problemem jest także mnogość formatów takich danych. W takim kontekście warto zwrócić uwagę na hurtownie danych dedykowane danym przestrzennym. Niniejszy artykuł opisuje przykładowy problem łączenia i przetwarzania danych, które mogą pochodzić z różnych źródeł. Jako środowiska wybrano wolnodostępne rozwiązania Talend Open Studio i GeoKettle.

Słowa kluczowe: dane przestrzenne, bazy danych przestrzennych, hurtownia danych przestrzennych, proces ETL

CONSTRUCTION OF SPATIAL DATA ETL PROCESS

Summary. Steadily increasing use of spatial databases in modern information systems is associated with processing large amounts of data. Another problem is the multitude of formats for such kind of data. In this context it is worthwhile to draw attention to a dedicated data warehouse domain. This article describes an example of the problem of combining and processing data from different sources. Talend Open Studio and GeoKettle are selected as public domain solutions.

Keywords: spatial data, spatial databases, spatial data warehouse, ETL process

1. Wprowadzenie

Powszechność danych przestrzennych w dzisiejszych zastosowaniach wymusiła nowy typ baz danych – bazy danych przestrzennych [1, 2, 3]. Liczne, nowoczesne systemy pomiarowe

[1] Praca finansowana w ramach badań statutowych Katedry Geoinformatyki i Informatyki Stosowanej

pozwalają na gromadzenie ogromnych ilości danych. Przestrzenne bazy danych pozwalają na przechowywanie i efektywną analizę przestrzennych zależności pomiędzy rzeczywistymi obiektami. Wybór odpowiedniego rozwiązania, które pozwoli na obsługę bazy danych przestrzennych nie jest zadaniem prostym, głównie ze względu na mnogość różnych rozwiązań dostępnych na rynku. Wspomnieć tutaj należy o takich rozwiązaniach, jak systemy PostgreSQL z rozszerzeniem PostGIS [4], SQLite z rozszerzeniem SpatiaLite [5], MySQL Spatial [6, 7], IBM DB2 wraz z dodatkiem Spatial Extender, MS SQL Spatial Storage czy Oracle Spatial. Wszystkie te systemy zarządzania bazami danych pozwalają na zapis informacji przestrzennych będących elementami map, udostępniają funkcje przetwarzające obiekty przestrzenne, za pomocą których to funkcji można otrzymać podstawowe informacje o obiektach, czyli o ich polach powierzchni i długościach (obwodach). W wygodny sposób otrzymać można również informacje o zależnościach występujących pomiędzy obiektami, np. czy się przecinają, czy mają część wspólną, czy też wykluczają się.

Konsorcjum Open Geospatial Consortium przyjęło standardowy model zapisu danych przestrzennych o nazwie OpenGIS [8]. Model ten pozwala na opis obiektów przestrzennych przez figury geometryczne, takie jak: punkt, krzywa, linia łamana, powierzchnia, wielokąt czy też całą kolekcję obiektów. Zapis obiektów do bazy może być wykonany z użyciem jednego z dwóch formatów, jakimi są WKT i WBK. Format WKT (ang. *Well-Known Text*) pozwala na zapis danych w postaci tekstowej, np.:

- POINT(50,100),
- LINESTRING(1 1, 2 2, 3 3),
- POLYGON((0 0, 4 0, 4 4, 0 4, 0 0)).

Format WBK (ang. *Well-Known Binary*) to binarny format składowania danych, zawierający informacje o kodowaniu, rodzaju figury i jej koordynatach.

Stosowanie uniwersalnego systemu opisu danych powinno pozwolić na możliwość migracji pomiędzy różnymi systemami zarządzania bazą danych, jednak wiele środowisk (aplikacje geodezyjne, geologiczne, geofizyczne, portale internetowe) posiada inne formaty, często własne. Z tego powodu wypływa konieczność tworzenia procesów integrujących dane przestrzenne.

2. Możliwości tworzenia procesów ETL dla danych przestrzennych

Przetwarzanie dużej ilości danych pochodzących z różnych źródeł i przechowywanych w różnych formatach jest zadaniem problematycznym. Rozwiązaniem mogą być tutaj hurtownie danych, które z założenia pozwalają na łączenie heterogenicznych źródeł [9]. W szczególności hurtownie danych przestrzennych pozwalają na przetwarzanie danych przestrzennych zarówno

z map rastrowych, jak i wektorowych. Hurtownia to baza centralnie zarządzana, która posiada odpowiedni model danych i sposób ich ładowania. Proces ładowania danych to proces ETL (*ang. Extracting, Transforming, Loading*), którego poszczególnymi etapami są:

- ekstrakcja (E) – proces wyboru źródeł – wskazywane są bazy danych oraz pliki płaskie, którymi hurtownia będzie zasilana,
- transformacja (T) – czyszczenie i łączenie danych, możliwe jest także znakowanie czasowe danych,
- ładowanie (L) – załadowanie wyznaczonych w ekstrakcji i zmienionych w transformacji danych do hurtowni.

Możliwe jest stworzenie takiego procesu dedykowanego dla danych przestrzennych, gdzie dane pochodzić będą z różnych źródeł, zapisane będą w różnych systemach odniesienia, a nawet przy wykorzystaniu różnych układów współrzędnych. Hurtownia daje szansę zintegrowania tego typu danych i otrzymania efektywnych informacji o poszczególnych obiektach.

Na rynkach dominują rozwiązania komercyjne [10]. Oracle Warehouse Builder to narzędzie pozwalające na implementację procesu ETL oraz zarządzanie jego pracą. DataStage to narzędzie implementujące całą logikę tworzenia hurtowni danych i jej zasilania. Oprogramowanie Sybase IQ to serwer dedykowany dla analiz i raportowań, a podobnie rozbudowanym narzędziem w tej dziedzinie jest Cognos. Istnieją również rozwiązania innych flagowych produktów komercyjnych, takich jak SQL Server Integration Services (dla MS SQL Server) czy IBM InfoSphere Warehouse (dla IBM DB2). Możliwe jest także skorzystanie z oprogramowania dostępnego na zasadach Open Source – na uwagę zasługują Talend Open Studio [11] oraz GeoKettle [12]. Środowiska te udostępniają graficzny interfejs pozwalający na intuicyjne budowanie hurtowni, posiadają szeroką gamę sterowników, dzięki którym w wygodny sposób można połączyć się z większością dostępnych na rynku systemów zarządzania bazami danych. GeoKettle to oprogramowanie dedykowane dla przetwarzania danych przestrzennych, zaś Talend Open Studio, będące produktem generalnie przeznaczonym do tworzenia procesów ETL, dostarcza dodatek o nazwie Spatial Data Integrator, który również pozwala na dostosowanie tego oprogramowania do przetwarzania danych przestrzennych. Danymi zasilającymi hurtownie mogą być mapy wektorowe, posiadające informacje np. o drogach, rzekach, granicach państw i miejscach o szczególnym znaczeniu. Dane te mogą być przechowywane w plikach typu Shapefile, w plikach tekstowych CSV, GML, XLS czy też w tabelach zewnętrznych baz. W hurtowni dane te zostaną wczytane, odpowiednio przetworzone, a następnie wyeksportowane do odpowiedniego formatu.

3. Konstrukcja procesu ETL dla danych zawierających informacje przestrzenne

Będąc w posiadaniu map w postaci rastrowej czy też wektorowej, można skonstruować odpowiedni proces ETL, którego celem będzie poszukiwanie informacji o obiektach zawartych na tych mapach. Taki proces zostanie przedstawiony poniżej.

Przykładowe dane fragmentu terenu stanu Północnej Karoliny (USA) [13] (rys. 1) zawierają informacje zgromadzone w następujących tabelach:

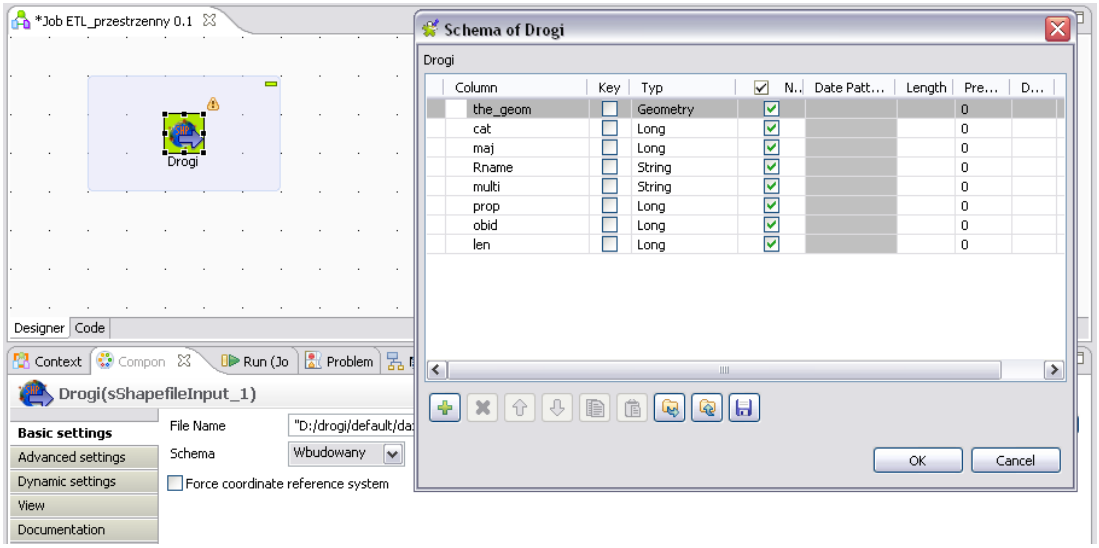
- drogi (koordynaty oraz atrybuty dróg),
- kolej (dane dotyczące linii kolejowych),
- miejsca (dane strategicznych miejsc, takich jak: tamy, mosty, parki, cmentarze, kościoły, lotniska, strumienie),
- straż (jednostki straży pożarnej),
- szkoły,
- ulice,
- zaludnienie (miejsca o największym zaludnieniu),
- szpitale.

Rys. 1. Mapa części stanu Północna Karolina [13]
Fig. 1. A part of North Carolina state map [13]

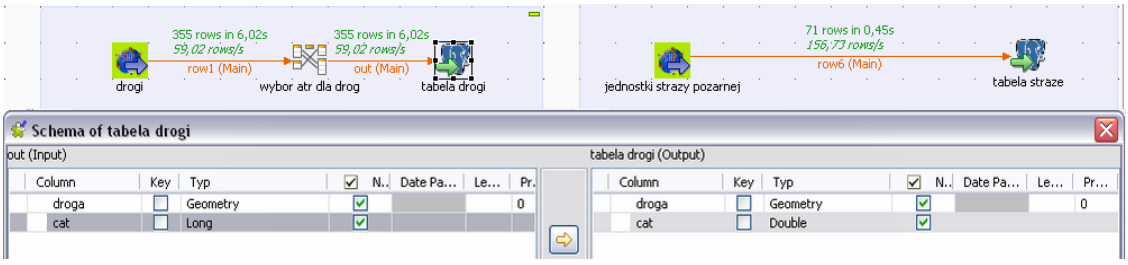
Zadaniem tworzonego procesu ETL mogą być zapytania, takie jak wyznaczenie odległości szkół od rzek (wyznaczenie potencjalnych miejsc ewakuacji ludności w razie powodzi) lub sprawdzenie, czy zachowana jest bezpieczna odległość 500 metrów pomiędzy zbiornikami wodnymi a cmentarzami (w innym przypadku istnieje ryzyko skażenia wody; w Polsce o takiej regulacji mówi Dziennik Ustaw 00.23.295).

3.1. Proces ETL dla danych przestrzennych w Talend Open Studio (Spatial Data Integrator)

Implementacja procesu ETL w środowisku Talend Open Studio zaczyna się od wybrania danych, którymi hurtownia będzie zasilana. W tym celu wykorzystywany jest komponent o nazwie „sShapeFile" (przedrostek „s" oznacza, iż komponent pochodzi z grupy komponentów „spatial"). W komponencie tym należy wskazać plik wejściowy oraz zadać schemat pliku. Niestety środowisko to w obecnej wersji nie potrafi samodzielnie określić schematu danych źródła. Wszelkie dodatkowe informacje zawarte w pliku z danymi muszą być podane. W ręcznym tworzeniu schematu istotne jest dobranie prawidłowych typów zapisanych danych, gdyż środowisko nie potrafi dokonywać konwersji (rys. 2).



Rys. 2. Atrybuty źródła danych przestrzennych
Fig. 2. Spatial data attributes



Rys. 3. Weryfikacja schematów danych
Fig. 3. Schema verification

Następnym krokiem jest wskazanie bazy, do której nastąpi zapis. W tym celu wykorzystywany jest komponent „sPostgisOutput", pozwalający na bezpośredni zapis do bazy danych PostgreSQL. Konieczna jest konfiguracja połączenia, czyli podanie lokalizacji bazy, jej instancji, portu, nazwy użytkownika i hasła. Dodatkowo należy wskazać nazwę tabeli, do której dane zostaną migrowane. Tabela, do której zapis będzie wykonany, nie musi zostać wcześniej utworzona w bazie. Środowisko Talend Open Studio samo tworzy tabele, dobierając typy atrybutów

według własnego algorytmu. W takim przypadku istotne jest prawidłowe zdefiniowanie wszystkich typów atrybutów – w prezentowanym przypadku należy zmienić typ „LongInt" na „Double" (rys. 3) ze względu na wymagania systemu PostGIS, połączenie obiektów geometrycznych (Geometry) nie stwarza problemów. Pomocny jest tutaj komponent „tMap" (przedrostek „t" jest charakterystyczny dla rodzimych komponentów Talend Open Studio), pozwalający na selekcję atrybutów, a także operacje na nich (np. arytmetyczne).

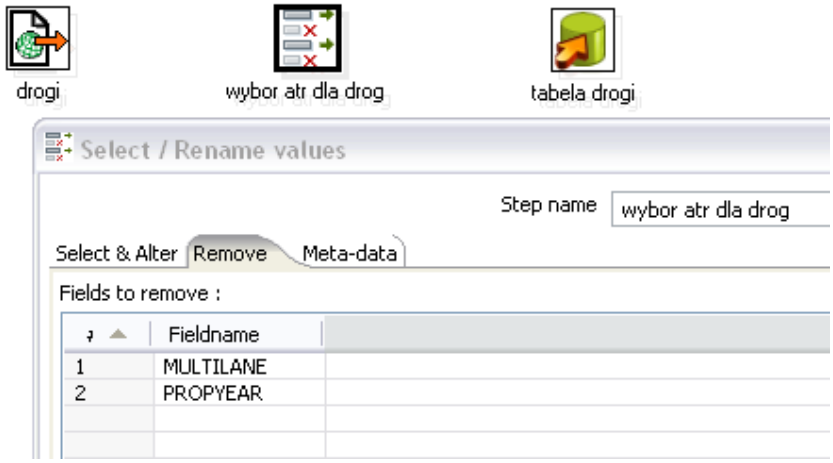
Na rys. 4 przedstawiono gotowy proces ETL dla wszystkich użytecznych połączeń źródeł danych. Po pozytywnym zakończeniu procesu migracji do bazy w oknie Run (Job), otrzymuje się statystyki wykonanego zadania.

Rys. 4. Proces ETL dla danych przestrzennych w Talend Open Studio
Fig. 4. ETL process for spatial data in Talend Open Studio

3.2. Proces ETL dla danych przestrzennych w GeoKettle

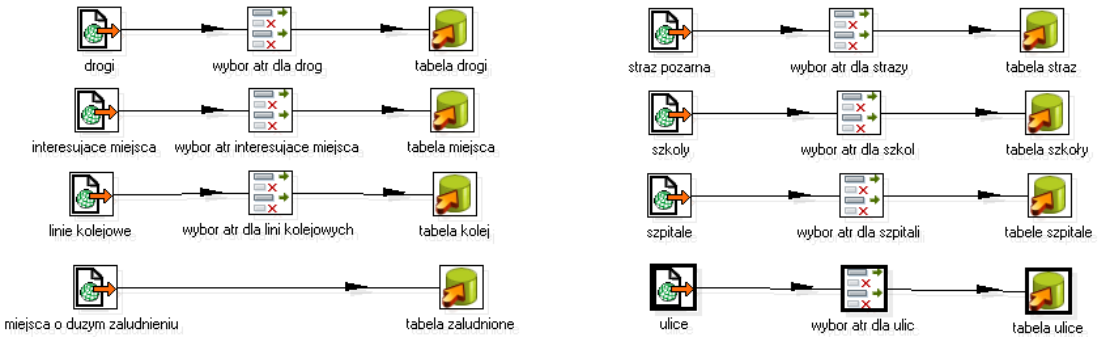
Konstrukcja procesu ETL w środowisku GeoKettle przebiega podobnie jak w przypadku Talend Open Studio. Pierwszym krokiem jest ustalenie danych wejściowych – w tym wypadku wykorzystywany jest komponent „GIS File Input". W komponencie tym można wskazać pliki wykorzystujące różne formaty danych. W prezentowanym przypadku wybrano te same pliki wektorowe (np. drogi.shp) co dla oprogramowania Talend Open Studio. Środowisko GeoKettle, w odróżnieniu od konkurenta, potrafi rozpoznać schemat danych.

Do połączenia z bazą docelową służy komponent „Table output", który należy skonfigurować. Tabelę docelową należy utworzyć ręcznie. Komponent o nazwie „Select values" jest wykorzystywany w celu wybrania atrybutów, jakie będzie posiadała tabela przechowująca dane w bazie. Możliwy jest tutaj wybór tylko tych kolumn, które będą niezbędne w dalszych analizach (rys. 5). Jest to szansa na odrzucenie danych, które w dalszych etapach będą zbędne. W ten sposób oszczędza się przestrzeń dyskową, jaką zajmie ostateczna baza, oraz skraca się czasy operacji.



Rys. 5. Proces ETL dla danych przestrzennych w GeoKettle
Fig. 5. ETL process for spatial data in GeoKettle

Ostatnią czynnością jest połączenie komponentów. Tak przygotowany proces jest gotowy do wykonania (rys. 6).



Rys. 6. Proces ETL dla danych przestrzennych w GeoKettle
Fig. 6. ETL process for spatial data in GeoKettle

3.3. Przykładowe zapytania do bazy danych

Po wykonaniu procesu dane z plików przeniesione do baz danych przestrzennych mogą być przetwarzane za pomocą zapytań SQL. Poniżej przedstawione są dwa przykładowe zapytania.

3.3.1. Szkoły zagrożone powodzią

Polecenie ma na celu wyszukanie szkół znajdujących się w odległości mniejszej niż 300 metrów od strumieni.

```
SELECT ST_Distance(szkoly.the_geom, miejsca.the_geom) AS "odleglosc", szko-
ly.namelong FROM szkoly, miejsca WHERE ST_Distance(szkoly.the_geom, miejs-
ca.the_geom)<300 AND miejsca.class LIKE 'Stream';
```

W wyniku system zwrócił tylko jedną krotkę spełniającą warunek:

```
    odleglosc     |    namelong
------------------+-----------------
 246.248068673285 | ROOT ELEMENTARY
(1 row)
```

w związku z czym tylko jedna szkoła o nazwie „Root Elementary” jest najbardziej zagrożona w wypadku powodzi, gdyż znajduję się poniżej zadanej odległości od strumieni.

3.3.2. Cmentarze ulokowane blisko zbiorników wodnych

W tym przypadku poszukiwane są cmentarze, które leżą w zbyt bliskiej odległości od zbiorników wody. Zbiorniki te mogą zostać skażone.

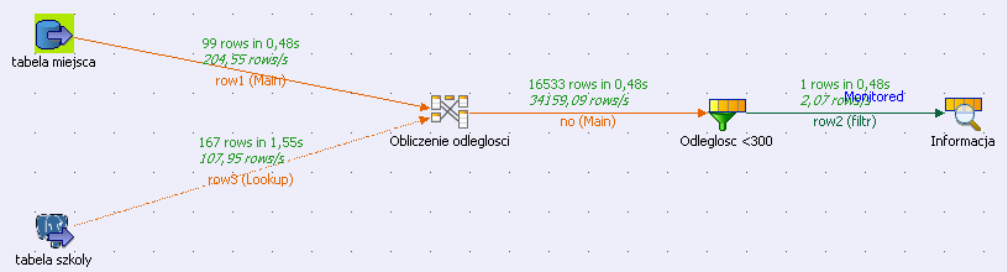
```
SELECT ST_DISTANCE(miejsca.the_geom, zbi.the_geom) AS "odleglosc", miej-
sca.featurenam FROM miejsca, zbi WHERE miejsca.class LIKE 'Cemetery' AND
ST_DISTANCE(miejsca.the_geom, zbi.the_geom)<500;
```

W wyniku otrzymano również tylko jedną krotkę:

```
    odleglosc      |   featurenam
-------------------+----------------
 302.877420969247 | Rand Cemetery
(1 row)
```

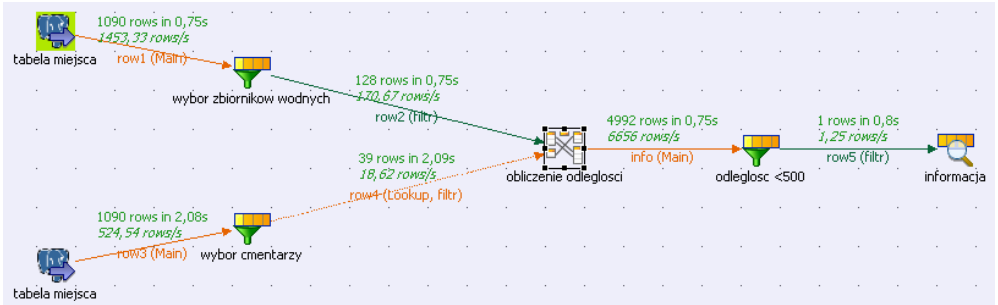
Zgodnie z wynikami zwróconymi przez bazę, jeden cmentarz jest zagrożeniem dla zbiorników wodnych.

3.4. Zapytania realizowane przez proces ETL



Rys. 7. Proces ETL dla danych przestrzennych w GeoKettle
Fig. 7. ETL process for spatial data in GeoKettle

Migracja danych do bazy danych ułatwia wykonywanie zapytań SQL, co ma miejsce w przypadku prostych zapytań. W praktyce zapytania mogą wiązać różne źródła danych, wówczas konieczne jest wykonanie takiego zapytania w środowisku ETL. Realizacja opisanych w punkcie 3.1 kwerend dla środowiska Talend Open Studio przedstawiona została na rysunkach 7 i 8.



Rys. 8. Proces ETL dla danych przestrzennych w GeoKettle
Fig. 8. ETL process for spatial data in GeoKettle

Otrzymane dane dla pierwszego zapytania prezentują się następująco:

```
.-----------------+---------------+--------------------.
|                      Informacja                      |
|=----------------+---------------+-------------------=|
|Odleglosc        |Zagrozona_szkola|Zagrazajacy_Strumien|
|=----------------+---------------+-------------------=|
|246.2480686731176|ROOT ELEMENTARY |Beaverdam Creek     |
'-----------------+---------------+--------------------'
```

natomiast dla drugiego:

```
.-----------------+--------------+-----------------.
|                     informacja                    |
|=----------------+--------------+----------------=|
|odleglosc        |Nazwa_cmenatrza|Zagrozony_zbiornik|
|=----------------+--------------+----------------=|
|302.8774209688703|Rand Cemetery  |Lake Benson       |
'-----------------+--------------+-----------------'
```

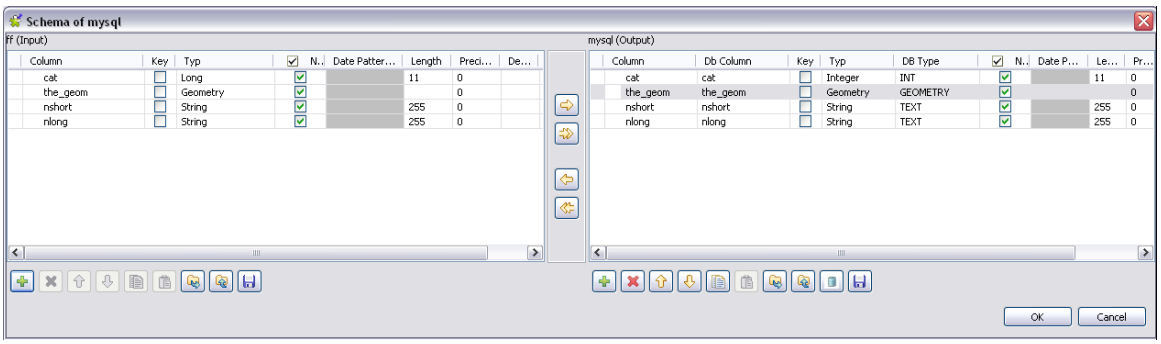
Środowisko GeoKettle w chwili obecnej nie udostępnia pełnej możliwości tworzenia takich kwerend (brak implementacji wszystkich zapytań analitycznych, w szczególności wyliczania odległości), aczkolwiek prace nad tym trwają.

3.5. Zapytania wykorzystujące rozproszone źródła danych

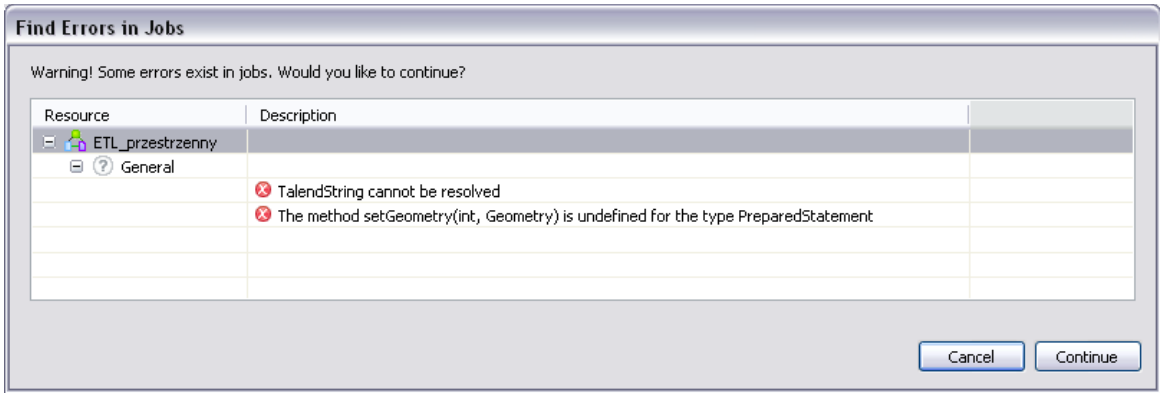
Cechą charakterystyczną środowisk oferujących konstrukcję procesów ETL jest łączenie rozproszonych źródeł danych. W przypadku omawianego przykładu rozdzielono dane między systemy MySQL i PostgreSQL. Następnie konstruowano proces ETL łączący te źródła.

Środowisko GeoKettle pozwoliło na poprawną konstrukcję takiego procesu. Połączenie danych dokonywane jest intuicyjnie.

Inaczej jest w przypadku Talend Open Studio. Obecna wersja środowiska nie posiada implementacji funkcji związanych z przetwarzaniem danych przestrzennych dla wszystkich systemów zarządzania bazami danych. Prawidłowo zaimplementowany jest interfejs do rozszerzenia PostGIS. Samo związanie z bazą danych MySQL (w tym podsystemem MySQL Spatial) i złączenie na podstawie danych typu GEOMETRY jest możliwe (rys. 9), jednakże system sygnalizuje błąd w niepełnej implementacji wykonania złączenia dla tego typu (rys. 10).



Rys. 9. Złączenie danych na podstawie atrybutów przestrzennych
Fig. 9. Data join using spatial attributes



Rys. 10. Błąd podczas łączenia danych przestrzennych w TOS
Fig. 10. Joining spatial data error in TOS

4. Wnioski

Dzięki oprogramowaniom Talend Open Studio oraz GeoKettle możliwe jest tworzenie procesów ETL dla baz danych przestrzennych. Pozwala to na przetwarzanie informacji o obiektach zawartych w mapach przy użyciu operatorów definiowanych przez standard OGC OpenGIS (lub SQL/MM).

Oprogramowanie Talend Open Studio posiada przyjazny interfejs, jednak zdecydowanym jego minusem jest brak możliwości podglądu danych na bieżąco oraz konieczność ręcznego ustawiania schematu pliku wejściowego. Dostępny w środowisku komponent „tMap" jest bardzo wygodnym narzędziem służącym zarówno do definiowania atrybutów tabeli wejściowej, jak i do wykonywania bezpośrednich operacji (dodawania, mnożenia, filtrowania) bez użycia dodatkowych komponentów, jak to ma miejsce w konkurencyjnym GeoKettle. Niewątpliwie problemem w obecnej wersji Talend Open Studio jest niepełna implementacja operacji na danych przestrzennych dla części systemów zarządzania bazami danych, takich jak MySQL.

GeoKettle, pomimo iż wymaga ręcznego stworzenia tabeli wyjściowej, to jednak posiada lepszą obsługę zarówno plików wejściowych, jak i połączeń z bazami. Bardzo przydatną opcją jest możliwość podglądu danych oraz to, iż program jest w stanie automatycznie ustalić schemat danych w plikach wejściowych.

W artykule przedstawiono zagadnienia na prostych przykładach w celu zachowania przejrzystości. Pominięto np. aspekty integracji map wektorowych z rastrowymi, co będzie tematem dalszych badań.

BIBLIOGRAFIA

1. Cichociński P., Dębińska E.: Baza danych przestrzennych wspomagająca samorządy lokalne w prowadzeniu polityki rozwoju przedsiębiorczości. ZN Pol. Śl., s. Informatyka, Studia Informatica, Vol. 31, No. 2B, Gliwice 2010, s. 371÷380.
2. Lenart A.: Zastosowanie GIS w systemach ERP, [w:] Kozielski S., Małysiak B., Kasprowski P., Mrozek D. (red.): Bazy Danych: Rozwój metod i technologii. WKiŁ, Warszawa 2008.
3. Gorawski M., Płuciennik E.: Analiza danych w systemie telemetrycznych hurtowni danych z wykorzystaniem rozszerzeń SQL/Oracle10g, [w:] Kozielski S., Małysiak B., Kasprowski P., Mrozek D. (red.): Bazy Danych: Modele, Technologie, Narzędzia. WKiŁ, Warszawa 2005.
4. SpatiaLite Home Page, http://www.gaia-gis.it/spatialite/.
5. PostGIS Home Page, http://postgis.refractions.net/.
6. MySQL Spatial, http://dev.mysql.com/doc/refman/5.1/en/spatial-extensions.html.
7. Piórkowski A.: Mysql Spatial and Postgis – Implementations of Spatial Data Standards. EJPAU 14(1), Vol. 14, No. 03, 2011.
8. OGC – The Open Geospatial Consortium, http://www.opengeospatial.org/.
9. Piórkowski A., Gajda G.: Konstrukcja wielowymiarowej bazy danych geologicznych. ZN Pol. Śl., s. Informatyka, Studia Informatica, Vol. 30, No. 2B, Gliwice 2009, s. 179÷190.
10. Gajda G., Piórkowski A.: Możliwości konstrukcji hurtowni danych geologicznych, [w:] Kozielski S., Małysiak B., Kasprowski P., Mrozek D. (red.): Bazy Danych: Rozwój metod i technologii. WKiŁ, Warszawa 2008.
11. Talend Home Page, http://www.talend.com.
12. GeoKettle Home Page, http://www.spatialytics.org/projects/geokettle/.
13. GRASS Tutor Data, http://grass.fbk.eu/download/data6.php.

Recenzenci: Dr inż. Jacek Frączek
Dr hab. inż. Zygmunt Mazur, prof. Pol. Wrocławskiej

Wpłynęło do Redakcji 18 stycznia 2011 r.

Abstract

The universality of spatial data in modern applications, forced a new type of database – spatial databases [1, 2, 3]. Many modern measurement systems allow for collection of huge amounts of data. Spatial databases allow the storage and efficient analysis of the spatial relationships between real objects. Collection and processing of such large amounts of data from heterogeneous sources is a problematic task. The solution may be a spatial database as well as more specialized structures such as data warehouses.

These systems allows to record the spatial information of the elements of maps, provide the processing functions of spatial objects, these function can obtain basic information about the objects, as their surface areas and lengths (circuits). It is possible to get information about the dependencies that occur between objects, if they intersect or are part of a joint or exclusive. This article describes an example of the problem of combining and processing data from different sources. Two based on open sources solutions Talend Open Studio and GeoKettle, were selected to studies. Conducted in both environments as ETL (Extraction, Transformation, Loading) for vector maps of the state of North Carolina (fig. 1).

Talend Open Studio was used for migrating data to a PostGIS spatial database. To write to the database only significant attributes, they were selected to limit the excess of unnecessary data. A similar process was created in an environment GeoKettle, where it served in a simpler way is to connect to database it is also possible to view data on the fly.

That database was used to extract information about the schools at risk of flooding and the water reservoirs that may be contaminated by being too close to cemeteries. (fig. 7). These queries were made in a graphical environment Talend Open Studio, GeoKettle in its current form does not provide some necessary functions (like distance function) to achieve that.

Adresy

Joanna KLISIEWICZ: Akademia Górniczo-Hutnicza, Katedra Geoinformatyki i Informatyki Stosowanej, al. Mickiewicza 30, 30-059, Kraków, Polska.

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

Stanisława PORZYCKA: Akademia Górniczo-Hutnicza, Katedra Geoinformatyki i Informatyki Stosowanej, al. Mickiewicza 30, 30-059, Kraków, Polska, , porzycka@agh.edu.pl.