Rozdział 3

Hurtowania danych i procesy ETL

Marek Grzegorowski, Mateusz Kalisch, Michał Kozielski, Łukasz Wróbel

3.1 Wstęp

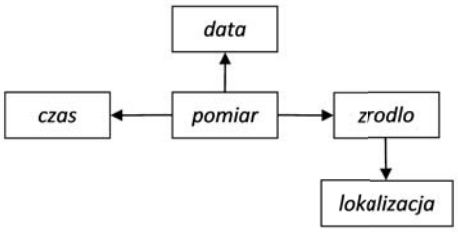
Opracowanie, z pomocą kart czujnika (omówionych w rozdziale 2), zestawu tzw. rekordów modelowych dla poszczególnych kategorii czujników stosowanych w przemyśle węglowym pozwala w jednolity sposób opisywać dane pomiarowe. Opis taki umożliwia integrację danych pomiarowych pochodzących z różnych systemów w ramach stworzonego repozytorium danych. Wypełnienie repozytorium danymi wymaga zaprojektowania procesów ETL (ang. Extract, Transform, Load) [3.7], które pobierają dane ze źródłowego systemu, przetwarzają w celu otrzymania danych ujednoliconych i ładują do stworzonego repozytorium systemu DISESOR. Repozytorium systemu DISESOR zostało zrealizowane w postaci hurtowni danych [3.3], w której centralną tabelą (faktem) jest pomiar.

W niniejszym rozdziale przedstawiona została struktura utworzonej w ramach projektu hurtowni danych oraz zagadnienia związane z ładowaniem danych (proces ETL) oraz ich integracją.

3.2 Struktura hurtowni danych

Struktura hurtowni danych stanowiącej repozytorium systemu została przygotowana w oparciu o analizę baz danych systemów monitorowania oraz informacji wchodzących w skład karty czujnika przedstawionej w rozdziale 2. Schemat repozytorium ma strukturę (ogólnie przedstawioną na rysunku 3.1), w której centralną tabelą jest tabela *pomiar* przechowująca wartości pomiarów. Każdy pomiar opisany jest wymiarami, główne z nich to: *data*, *czas* oraz *zrodlo*. Szczegółowy opis tabeli *pomiar* został zawarty w tabeli 3.1. Wymiary *data* oraz *czas* opisują czas rejestracji pomiaru, natomiast tabela *zrodlo* opisuje źródło pomiaru (czujnik lub przyrząd). Tabela *zrodlo* została przygotowana na postawie karty czujnika i zawiera m.in. takie informacje jak: wielkość mierzona, nazwa wielkości mierzonej, nazwa własna czujnika/przyrządu, typ czujnika/przyrządu, nazwa

systemu gromadzącego dane z czujnika/przyrządu, typ gromadzonych danych, zakres pomiaru. Tabela *zrodlo* opisana jest wymiarem *lokalizacja*, który pozwala na określenie miejsca zabudowy czujnika/przyrządu. Ze względu na znaczenie tabeli *zrodlo* jej szczegółowa prezentacja została dodatkowo przedstawiona w tabeli 3.2.



Rys. 3.1. Uproszczony schemat repozytorium danych pomiarowych

Tab. 3.1. Pola tabeli *pomiar*

Pole	**Opis**
wartosc	Wartość pomiaru
data_id	Klucz obcy do tabeli *data*
czas_id	Klucz obcy do tabeli *czas*
czas_kategoria_id	Klucz obcy do tabeli *czas_kategoria*
zrodlo_id	Klucz obcy do tabeli *zrodlo*
stan_id	Klucz obcy do tabeli *stan*
dyskretyzacja_id	Klucz obcy do tabeli *dyskretyzacja*

Opracowane repozytorium danych ma topologię płatka śniegu [3.5] o strukturze logicznej przedstawionej na rysunku 3.2. W tabeli 3.3 zamieszczono nazwy tabel związanych z gromadzeniem danych pomiarowych wraz z krótkim opisem ich przeznaczenia.

Lokalizacja w kopalni posiada strukturę drzewiastą - przykład takiej struktury przedstawiony został na rysunku 3.3. Reprezentacja lokalizacji została zrealizowana w hurtowni danych jako struktura hierarchiczna (tabela *lokalizacja_hierarchia*). Dla takiej struktury lokalizacji opracowano zestaw procedur umożliwiających dostęp do dowolnego miejsca w drzewiastej strukturze lokalizacji.

Metadane dotyczące rezultatów działań i konfiguracji innych modułów systemu przechowywane są w oddzielnej bazie danych.

Zaprojektowana struktura hurtowni danych została zaimplementowana w systemie zarządzania bazą danych PostgreSQL [3.6]. System ten został wybrany

Tab. 3.2. Pola tabeli *zrodlo*

Pole	**Opis**
zrodlo_id	Klucz główny
sciezka	Identyfikator źródła postaci <miejsce>\<nazwa systemu>\[nazwa podsystemu]\<identyfikator>, np. KWK Mysłowice-Wesoła\THOR\SEMP\123456
nazwa	Nazwa szczegółowa (np. AN54, MM264, AMP1_IR)
opis	Opis źródła (np. prąd organu lewego w kombajnie)
typ	Nazwa typu źródła (np. anemometr, metanomierz, sejsmometr, wiercenie małośrednicowe - wychód zwiercin, położenie pływaka osadzarki)
typ_opis	Opis typu źródła (np. metanomierz MM-2PW)
typ_skrot	Skrót typu (np. AN dla anemometrów)
mezurand	Wielkość mierzona - ogólna (np. stężenie, ciśnienie, gęstość, temperatura)
mezurand_opis	Wielkość mierzona - sprecyzowana (np. stężenie CO_2, temperatura osłony napędu)
jednostka	Nazwa jednostki pomiarowej (np. wolt, stopień Celsjusza)
jednostka_skrot	Skrót jednostki pomiarowej (np. V, $^{\circ}$C)
grupa	Grupa (np. zagrożenia metanowe i pożarowe, zagrożenia klimatyczne i gazowe)
przeznaczenie	Przeznaczenie (np. bezpieczeństwo)
wartosc_minimalna	Wartość minimalna, NaN gdy niesprecyzowana
wartosc_maksymalna	Wartość maksymalna, NaN gdy niesprecyzowana
czestotliwosc	Częstotliwość zapisywanych pomiarów (w sekundach), 0 gdy pomiary nie są cykliczne
kryterium_rejestracji	Kryterium rejestracji: każda wartość, pomiary różniące się od poprzedniego, pomiary różniące się od poprzedniego o określony próg, jeśli zarejestrowane (np. wstrząsy), pomiary wprowadzane ręcznie
numeryczne	Flaga określająca czy źródło rejestruje pomiary numeryczne
system	Nazwa systemu, z którego pochodzi źródło (np. Thor, Hestia)
podsystem	Nazwa podsystemu (np. SEMP, UTS-2)
lokalizacja_id	Klucz obcy do tabeli *lokalizacja*

na podstawie porównania systemów MySQL oraz PostgreSQL, które są obecnie jednymi z najpopularniejszych rozwiązań typu open source. W porównaniach wykorzystano PostgreSQL w wersji 9.3 oraz MySQL w wersji 5.6. Analiza porównawcza funkcjonalności tych systemów pokazała, że PostgreSQL jest bardziej kompatybilny ze standardem SQL oraz obsługuje ten standard w szerszym zakresie niż MySQL. Przykładowo, MySQL 5.6 nie obsługuje tzw. funkcji okienkowych (ang. window functions), będących elementem standardu SQL:2003. System PostgreSQL jest również bardziej elastyczny niż MySQL pod względem możliwości rozszerzania funkcjonalności, umożliwia on m.in. tworzenie procedur składowanych w takich językach jak R, Java oraz Python. Przeprowadzone zostały również testy wydajnościowe, w których wykorzystano oprogramowanie

Tab. 3.3. Spis tabel hurtowni danych wraz z krótkim opisem ich przeznaczenia

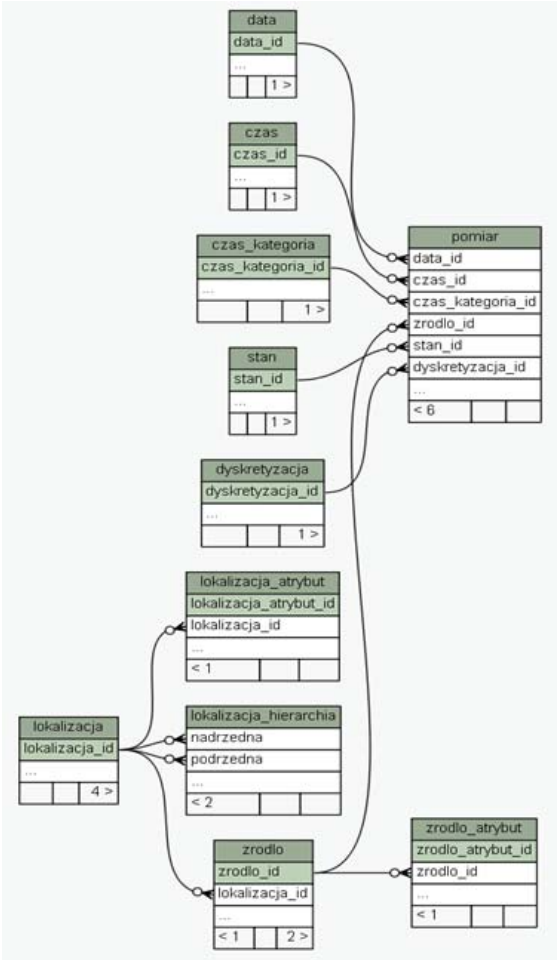
Tabela	**Opis**
czas	Czas pomiaru, od 00:00:00 do 23:59:59 (h:min:sek) co 1 sekundę.
czas_kategoria	Kategoria czasu (np. okres wydobywczy, okres niewydobywczy).
data	Data rejestracji pomiaru, wstępnie wypełniona od 2000-01-01 do 2016-12-31.
dyskretyzacja	Umożliwia zamianę wartości pomiarowych na dyskretne jeżeli źródło rejestruje pomiary typu nominalnego.
lokalizacja	Lokalizacja źródła pomiaru. Może to być zarówno miejsce w strukturze kopalni (np. wyrobisko), jak również monitorowane urządzenie (np. kombajn chodnikowy). Do tabeli pomiar wpisywany jest identyfikator lokalizacji najbardziej specyficznej (lokalizacja jest tym bardziej specyficzna im niżej położona jest w hierarchii).
lokalizacja_atrybut	Cechy specyficzne dla danej lokalizacji (np. rozeznanie górnicze lub wysokość wyrobiska). Celem tabeli jest przechowywanie wszystkich dodatkowych informacji, które występują tylko dla określonych lokalizacji lub które nie są jeszcze nam znane.
lokalizacja_hierarchia	Hierarchiczna struktura lokalizacji. Zawiera wszystkie możliwe ścieżki w drzewie hierarchii.
pomiar	Wartości pomiarów.
stan	Stan pomiaru, np. do określenia czy wartość jest powyżej progu alarmowego, kalibracja, awaria.
zrodlo	Informacja o źródle pomiaru (najczęściej jest to czujnik).
zrodlo_atrybut	Tabela analogiczna do lokalizacja_atrybut, przechowuje cechy specyficzne dla określonego źródła.

HammerDB w wersji 2.15 umożliwiające przeprowadzenie testów opracowanych przez TPC (Transaction Processing Performance Council http://www.tpc.org/).

3.3 Procesy ETL

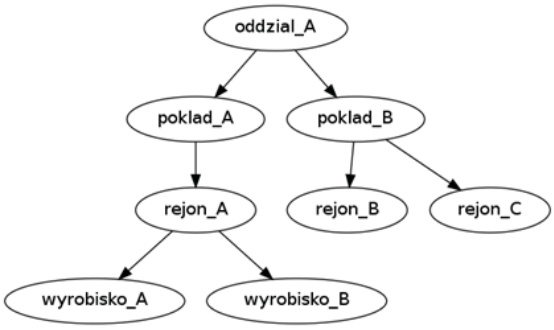
Zaprojektowana hurtownia danych zasilana jest przez procesy ETL (ang. Extract, Transform, Load) danymi pochodzącymi z systemów dyspozytorskich i/lub monitorowania takich jak np. THOR lub Hestia. Ze względu na różnice istniejące w bazach danych różnych systemów monitorowania i dyspozytorskich proces ETL jest realizowany poprzez schematy zasilania danymi opracowane odrębnie dla każdego systemu.

Projekt procesu ETL został zrealizowany z wykorzystaniem oprogramowania Talend Open Studio [3.1]. Aplikacja umożliwia użytkownikowi zbudowanie



Rys. 3.2. Struktura logiczna hurtowni danych systemu DISESOR

procesu w postaci diagramu, którego zadaniem jest pobranie informacji ze źródła, wprowadzenie modyfikacji, jeżeli jest to konieczne, i zapisanie w lokalizacji docelowej. Oprogramowanie Talend Open Studio pozwala na pracę z wieloma rodzajami baz danych, w tym PostgreSQL wybranym, jak wspomniano w poprzednim punkcie, do realizacji repozytorium danych. Środowisko aplikacji zostało oparte o interfejs wykorzystywany między innymi przez oprogramowanie Eclipse i wymagane jest zainstalowane środowisko Java 7 w celu jego poprawnego funkcjonowania. Działanie aplikacji polega na generowaniu kodu w języku Java na podstawie zbudowanego diagramu, dzięki takiemu rozwiązaniu możliwe jest bezpośrednie wykorzystanie języka Java w niektórych elementach grafu, co w znacznym stopniu może ułatwić pracę i zredukować liczbę pojedynczych elementów w diagramie. W chwili uruchomienia przygotowanego procesu, wygenerowa-



Rys. 3.3. Hierarchiczna struktura lokalizacji

ny kod jest wykonywany. W przypadku problemów z prawidłowym działaniem procesu, wyświetlane są odpowiednie komunikaty o błędach lub wyjątkach. Istnieje możliwość podziału projektu na pod-procesy, które można uruchamiać np. sekwencyjnie lub równolegle.

Talend Open Studio umożliwia projektowanie schematów ładowania danych, które zaprojektowane jednokrotnie mogą być realizowane cyklicznie w celu zasilania repozytorium (tutaj hurtowni danych) nowymi danymi. Na rysunku 3.4 zaprezentowano przykładowy schemat procesu ETL dla systemu THOR. Proces ładowania danych składa się z m.in. następujących kroków:

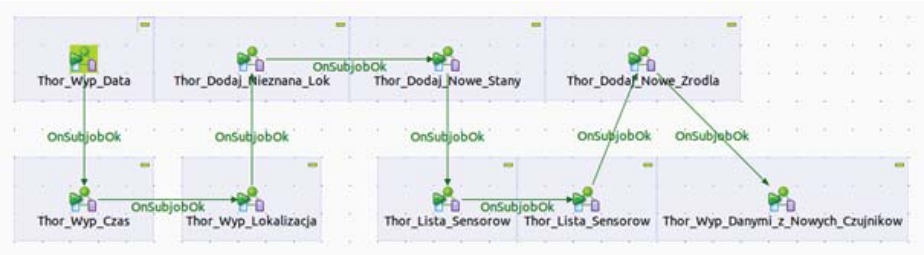
- wypełnienie tabeli *data* datami z określonego zakresu,
- wypełnienie tabeli *czas* z sekundową częstotliwością,
- dodanie do tabeli *lokalizacja* nowych lokalizacji z tabeli systemu zasilającego (tutaj THOR),
- dodanie do tabeli *lokalizacja* identyfikatora nieznanej lokalizacji (w przypadku nieznanej lub nieopisanej w systemie monitorowania lokalizacji źródła pomiaru),
- dodanie do tabeli *stan* nowych stanów,
- dodanie nowych źródeł do tabeli *zrodlo* w systemie DISESOR,
- kopiowanie wartości pomiarowych z systemu źródłowego (tutaj THOR) do systemu DISESOR.

Każdy z wymienionych i przedstawionych na rysunku 3.4 elementów jest osobnym diagramem realizującym wybrany podproces.

3.4 Czyszczenie i ujednolicanie danych

Dane zebrane w hurtowni danych przeznaczone są do dalszej analizy i przetwarzania przez kolejne moduły systemu DISESOR. Z tego powodu wymagane są dalsze operacje czyszczenia i ujednolicania danych, które są realizowane przez moduł nazwany ETL2 [3.4].

Zadaniem ETL2 jest zintegrowanie danych pochodzących z zadanego przez użytkownika okresu czasu i z zadanych przez niego źródeł (w szczególności czujników) oraz przygotowanie ujednoliconego zbioru wynikowego. Dane pomiarowe



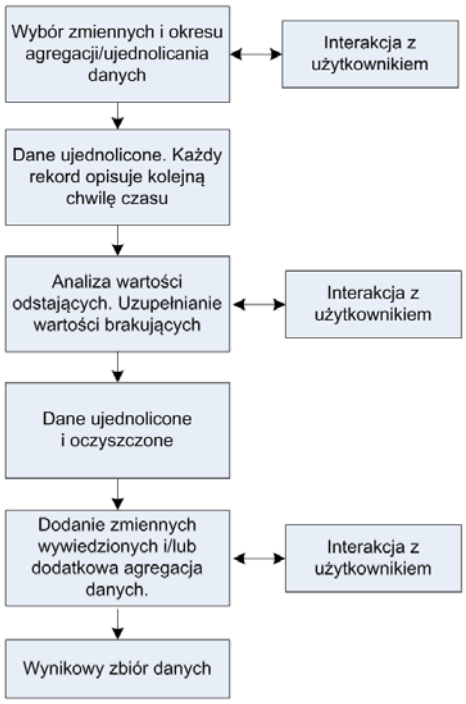
Rys. 3.4. Ogólny widok procesu ETL dla systemu THOR

gromadzone są z różnymi częstotliwościami, ponadto w niektórych systemach zapisuje się dane pomiarowe jedynie po istotnej (zdefiniowanej w bazie danych systemu pomiarowego lub systemu monitorowania) zmianie. Moduł ETL2 ujednolica dane w taki sposób, że każdy rekord w wynikowym zbiorze danych opisuje pewien zdefiniowany przez użytkownika okres czasu np. 30 sekund. W module ETL2 następuje także wywołanie procedur oczyszczania danych (identyfikacji wartości odstających i uzupełniania wartości brakujących), a także agregacji danych (np. 10 pomiarów zostaje zastąpionych 1 pomiarem) oraz manualnego definiowania tzw. zmiennych wywiedzionych (np. nowa zmienna powstaje jako suma wartości dwóch innych zmiennych). Do definiowania zmiennych wywiedzionych użytkownik otrzymuje zestaw pewnych predefiniowanych funkcji matematycznych i logicznych (np. koniunkcje i/lub alternatywy warunków, reguły logiczne). Agregacja danych - a więc zastąpienie grupy następujących po sobie rekordów (okna czasowego) jednym rekordem - odbywa się również za pośrednictwem predefiniowanego zestawu funkcji agregujących (m.in. maksimum, minimum, dominanta, średnia). Ogólny schemat przetwarzania danych w module ETL2 zaprezentowano na rysunku 3.5, a w dalszej części rozdziału wymienione są metody stosowane przy realizacji poszczególnych operacji czyszczenia i ujednolicania danych.

Procedury oczyszczania danych pomiarowych uruchamiane są na każde żądanie użytkownika systemu związane z wyborem pewnego podzbioru dostępnych danych w celu wykonania na ich podstawie analizy. Oznacza to, że jeśli w systemie nie zdefiniowano żadnych zadań (modeli) analitycznych, nie jest również uruchamiana żadna procedura oczyszczania danych.

W przypadku zdefiniowania zadania analitycznego użytkownik wybiera pewien zestaw zmiennych (pomiarów), który podlegał będzie dalszej analizie. Użytkownik określa również pewien przedział czasu, który objęty będzie analizą. Przed przystąpieniem do analizy użytkownik ma możliwość podglądu wybranych danych, dodatkowo prezentowane są pewne informacje zbiorcze dotyczące pobranego zbioru danych. Informacjami tymi są:

– częstotliwość zapisu poszczególnych pomiarów (wartości te mogą być regularne np. co 2 sekundy lub nieregularne jeśli w bazie danych systemu monitorowania odzwierciedlane są jedynie sytuacje, w których pomiar się zmienił);



Rys. 3.5. Ogólna charakterystyka procesu przetwarzania danych w module ETL2

– dla zmiennych typu symbolicznego i porządkowego prezentowany jest histogram częstości występowania danej wartości;
– dla zmiennych typu numerycznego prezentowany jest rozkład wartości pomiarowych (procedura doboru szerokości kolumn histogramu będącego podstawą do utworzenia rozkładu jest automatyczna);
– jeśli zmienna jest typu numerycznego prezentowane są również wartości: średnia, mediana, kwartyle, wartość minimalna i maksymalna;
– jeśli dla danej zmiennej określono symbol brakującej wartości i/lub wartości niepoprawnej prezentowana jest liczba tych wartości (dla zmiennych symbolicznych i porządkowych) lub sumaryczna długość czasu dla którego brakuje wartości pomiarowych lub są one niepoprawne; podawany jest również najdłuższy okres czasu w jakim brakowało wartości pomiarowych;
– jeśli dla danej zmiennej zdefiniowano nazwy przedziałów wartości (np. stany w jakich może znajdować się czujnik – alarm, ostrzeżenie, normalny etc.) użytkownik otrzymuje również informacje o liczbie poszczególnych stanów w jakich, w interesującym go okresie czasu, znajdowały się wartości pomiarowe;
– dla każdej zmiennej prezentowana jest również liczba wartości pomiarowych, które zostały uznane za odstające; do identyfikacji wartości odstających wykorzystany zostanie m.in. rozstęp międzykwartylowy.

Na podstawie powyższych informacji użytkownik decyduje o kolejnych krokach przetwarzania danych. Krokami tymi są:

- ujednolicenie częstotliwości pomiarów;
- wybór procedury uzupełniania brakujących i pustych wartości dla każdej zmiennej;
- wybór sposobu wygładzania danych dla każdej zmiennej.

Ujednolicenie częstotliwości pomiarów oznacza, że użytkownik podaje pewien przedział czasu (interwał), zgodnie z którym generowane będą przetworzone rekordy. Przykładowo, jeśli użytkownik poda interwał równy 10 sekund, to w wynikowym zbiorze danych każdy kolejny wiersz będzie zawierał dane pomiarowe z kolejnych 10 sekund. Jeśli dla wybranych zmiennych akwizycja danych pomiarowych następuje częściej (np. co 2 sekundy) to pomiary pośrednie zostaną pominięte lub zastąpione wybraną statystyką (np. minimum, maksimum, medianą) opisującą dane w zadanym oknie. Jeśli dla wybranych zmiennych akwizycja danych następuje rzadziej niż 10 sekund (np. co 20 sekund) to pomiary zostaną odpowiednio zduplikowane. W przypadku, gdy częstotliwość akwizycji danych nie jest wielokrotnością interwału podanego przez użytkownika, wybierana jest wartość najbliższa temu interwałowi. Procedury uzupełniania brakujących i/lub nieważnych wartości są następujące:

- brak uzupełniania (wszystkie brakujące wartości zastępowane są znakiem „?”);
- zastępowanie znakami specjalnymi (brakujące wartość zastępowane są znakiem „?”, wartości poza zakresem oraz wartości nieważne zastępowane są znakiem „N”);
- ostatnia poprawna wartość (użytkownik definiuje wartość parametru MaxU określającą jaki maksymalnie okres brakujących wartości może być w ten sposób uzupełniony);
- średnia z k ostatnich prawidłowych wartości (wartość parametru k definiowana jest przez użytkownika; UWAGA: średnia kroczy więc w kolejnych krokach będziemy wykorzystywać już uśrednione dane; użytkownik definiuje również wartość parametru MaxU określający jaki maksymalnie ciąg brakujących wartości może być w ten sposób uzupełniony);
- uzupełnianie brakujących wartości na podstawie interpolacji liniowej; wyznaczane jest równanie prostej pomiędzy ostatnią prawidłową wartością pomiarową zarejestrowaną przed brakiem danych i pierwszą prawidłową wartością pomiarową zarejestrowaną po braku danych; na podstawie wyznaczonego równania prostej uzupełniane są braki (użytkownik definiuje również wartość parametru MaxU określającą jaki maksymalnie ciąg brakujących wartości może być w ten sposób uzupełniony);
- uzupełnianie brakujących wartości wartością: najczęściej występującą, średnią, medianą.

Użytkownik decyduje również czy zidentyfikowane w procesie wstępnej analizy danych wartości odstające mają być traktowane jak wartości brakujące. Jeśli wybrana zostanie ta opcja wartości odstające zastępowane są zgodnie z metodami wymienionymi powyżej. Użytkownik posiada również możliwość manualnego

zastępowania (zamiany) dowolnych wartości pomiarowych (niezależnie od tego czy są one brakujące, odstające czy poprawne).

Oczyszczanie danych obejmuje również procedury zamiany wartości oraz wygładzania danych pomiarowych. Procedury te będą następujące:

- wygładzanie za pomocą k punktowej średniej ruchomej;
- wygładzanie danych za pomocą znanego z pakietu Statistica filtru 4253H; przekształcenie to polega na kilkakrotnym wygładzaniu średnią lub medianą ruchomą i jest silnym narzędziem wygładzającym szereg; wybranie tej opcji powoduje, że wykonane będą następujące przekształcenia: (1) wygładzanie 4-punktową medianą ruchomą centrowaną za pomocą dwupunktowej mediany, (2) 5-punktową medianą ruchomą, (3) 3-punktową medianą ruchomą i (4) 3-punktową średnią ważoną, z wagami Hanninga (0,25, 0,5, 0,25), (5) następnie obliczane są reszty, przez odjęcie przekształconego szeregu od oryginalnego, (6) kroki 1 - 4 powtarza się dla reszt, (7) przekształcone reszty dodaje się do przekształconego szeregu;
- zmiana oryginalnych wartości pomiarowych na wartości lingwistyczne zdefiniowane w słowniku zakresów (np. jeśli dla pomiaru zdefiniowano zakresy – normalny, ostrzeżenie, alarm – to wszystkie numeryczne wartości pomiarowe zostają zastąpione odpowiednimi nazwami zakresów).

Funkcje przetwarzania danych opisane powyżej zostały zaimplementowane jako operatory w środowisku RapidMiner [3.2]. Dzięki temu użytkownik może definiować bardzo zaawansowane schematy przetwarzania łącząc operatory dostępne standardowo w środowisku RapidMiner z tymi opracowanymi w ramach projektu. Przykładowo, do identyfikacji wartości odstających lub uzupełniania wartości brakujących może zostać wykorzystana cała grupa zaawansowanych metod, które zostały opisane w rozdziale 9.

Literatura

[3.1] R. Barton. *Talend Open Studio Cookbook*. Packt Publishing Ltd, 2013.

[3.2] M. Hofmann, R. Klinkenberg. *RapidMiner: Data mining use cases and business analytics applications*. CRC Press, 2013.

[3.3] R. Kimball, M. Ross. *The data warehouse toolkit: the complete guide to dimensional modeling*. John Wiley & Sons, 2011.

[3.4] M. Kozielski, Ł. Wróbel. Decision support and maintenance system for natural hazards, processes and equipment monitoring. *Eksploatacja i Niezawodnosc — Maintenance and Reliability*, 18(2):218–228, 2016.

[3.5] A. Pelikant. *Hurtownie danych: od przetwarzania analitycznego do raportowania*. Helion, 2011.

[3.6] G. Smith. *PostgreSQL 9.0: High Performance*. Packt Publishing Ltd, 2010.

[3.7] P. Vassiliadis. A survey of extract–transform–load technology. *International Journal of Data Warehousing and Mining (IJDWM)*, 5(3):1–27, 2009.