Streszczenie rozprawy doktorskiej

Strumieniowe hurtownie danych zorientowane na przetwarzanie wielkich zbiorów danych kontekstowych

mgr inż. Krzysztof Pasterak
Promotor: prof. dr hab. inż. Marcin Gorawski

Tematyka niniejszej rozprawy obejmuje ogół zagadnień powiązanych ze strumieniowymi hurtowniami danych kontekstowych. Celem rozprawy jest omówienie powiązanych zagadnień teoretycznych, przedstawienie propozycji nowych metod i modeli, przeprowadzenie badań eksperymentalnych oraz analiza ich wyników.

Pierwszym zagadnieniem omawianym w rozprawie jest, stanowiący motywację do dalszych badań, system dystrybucji i składowania paliw płynnych, przy którym prowadzono badania nad problemem detekcji anomalii krytycznych. Skutkiem tych badań było opracowanie metody wykrywania wycieków wykorzystującej detekcję i interpretację trendów (algorytm TUBE), a konkluzją – spostrzeżenie o konieczności wzięcia pod uwagę kontekstu badanych zjawisk przy ich analizie. Spostrzeżenie to przybrało w rozprawie postać pierwszej tezy: „Uzyskanie w pełni jednoznacznych wyników analizy danych ukierunkowanej na wykrywanie zdarzeń anomalnych jest możliwe dopiero po uwzględnieniu kontekstu występowania poszczególnych anomalii, na który składają się dane współistniejące w czasie i przestrzeni oraz powiązane semantycznie z analizowanym zjawiskiem".

Wnioski z omówionych badań posłużyły za podstawę do sformułowania teorii danych kontekstowych, wraz z podaniem ich definicji, klasyfikacji oraz zarysu metod ich przetwarzania. Ten obszar tematyczny stanowi drugie istotne zagadnienie niniejszej rozprawy. Z owego zagadnienia wynika bezpośrednio kolejne, stanowiące próbę praktycznego ujęcia tematu przetwarzania danych kontekstowych: model strumieniowej hurtowni danych kontekstowych. Został on przedstawiony w niniejszej rozprawie jako kompletny system składowania i przetwarzania danych kontekstowych, ukierunkowany na wykrywanie oraz weryfikację anomalii krytycznych. W tym zakresie stawiana jest druga teza niniejszej rozprawy: „Możliwe jest zaprojektowanie strumieniowej hurtowni danych zorientowanej na przetwarzanie wielkich zbiorów danych kontekstowych, wykorzystującej wielotorowy model przetwarzania danych, w którym analiza danych krytycznych jest wsparta przez przeprowadzaną niezależnie wieloaspektową analizę danych kontekstowych, w celu uwiarygodnienia wyników tej pierwszej". W ramach modelu strumieniowej hurtowni danych kontekstowych zaproponowano i opisano szereg metod i modeli, przeznaczonych do wspierania przetwarzania danych kontekstowych. Są to w szczególności: silnik strumieniowej kostki CUBIT oraz

indeks przestrzenny BRI. Ta pierwsza jest odpowiednikiem kostki OLAP dla wielowymiarowych danych strumieniowych. Drugie rozwiązanie to wielowymiarowy bitowy indeks zakresowy, wspierający wykonywanie zapytań o agregaty zakresowe w wielowymiarowej przestrzeni cech.

Ostatnim zagadnieniem omawianym w niniejszej rozprawie jest problem efektywnego dostarczania agregatów wielowymiarowych. W ramach tego problemu zaprojektowano trzy nowe adaptacyjne algorytmy stronicowania, przeznaczone dla omówionego silnika CUBIT. Algorytmy te wykorzystują metody optymalizacji wielokryterialnej do zapewnienia należytej jakości usług, zarówno klienta (użytkownika), jak i źródła (bazy) danych. Rzeczone algorytmy zostały poddane analizie weryfikacyjnej oraz porównawczej – zarówno pomiędzy sobą, jak i z poprzednią generacją algorytmów. Badania przeprowadzono przy użyciu dwóch zaproponowanych metryk jakości usług dla strumieniowych hurtowni danych. Rzeczone algorytmy i metryki zostały ujęte w formie trzeciej tezy rozprawy: „Zastosowanie metod optymalizacji wielokryterialnej w procesie stronicowania w strumieniowych hurtowniach danych oraz uwzględnienie bieżących parametrów pracy i istniejących ograniczeń, pozwala na zwiększenie jakości usług, rozumianej zarówno jako poprawę efektywności i ciągłości dostarczania danych użytkownikowi, jak również zmniejszenie obciążenia źródła danych".

Weryfikację tez rozprawy przeprowadzono w sposób teoretyczny oraz empiryczny. Ten pierwszy polegał na budowie modeli i analizie powiązanych zagadnień teoretycznych. Sposób empiryczny opierał się na wynikach eksperymentów oraz wnioskach sformułowanych przy pracy z rzeczywistymi obiektami przemysłowymi. Na weryfikację pierwszej tezy rozprawy składało się scharakteryzowania sieci stacji paliw i zachodzących w niej procesów, w tym anomalii krytycznych oraz dyskusja na temat jakości wyników algorytmu wykrywania wycieków paliwa w świetle współistnienia innych zjawisk, stanowiących kontekst dla tych wycieków. Druga teza została zweryfikowana przez sformułowanie modelu strumieniowej hurtowni danych kontekstowych, wliczając w to zarówno modele cząstkowe poszczególnych baz danych, jak i modele silnika CUBIT oraz indeksu BRI. Weryfikacja trzeciej tezy została przeprowadzona w sposób empiryczny, przez wykonanie szeregu eksperymentów porównawczych dla zaproponowanych trzech nowych algorytmów wypełniania stron, przy jednoczesnym wykorzystaniu dwóch nowych metryk jakości usług.

W konkluzjach rozprawy wskazano na otwarte problemy badawcze, wliczając w to: rozwój modelu silnika CUBIT oraz indeksu BRI, a także udoskonalenie zaproponowanych algorytmów wypełniania stron. Ponadto, opisano również potencjalne drogi rozwoju poszczególnych metod i modeli, takie jak: adaptację istniejących rozwiązań do innych problemów świata rzeczywistego oraz integrację zaprezentowanego modelu strumieniowej hurtowni danych kontekstowych z modelem Spichlerza Agregatów.