Система автоматизації процесу верифікації даних

Болдак Андрій[1], Дубінський Єгор[2]

1. Кафедра обчислювальної техніки, НТУУ "Київський політехнічний інститут", УКРАЇНА, м.Київ, просп. Перемоги 37,
E-mail: boldak.andrey@gmail.com

2. Кафедра обчислювальної техніки, НТУУ "Київський політехнічний інститут", УКРАЇНА, м.Київ, просп. Перемоги 37,
E-mail: egor.dubinskii@mail.ru

Work contains review of module for statistical data validation based on declarative domain-specific language. This module works with suite of tests in parallel-consecutive mode. Proposed module is planned to be used in ICSU World Data Center for Geoinformatics and Sustainable Development hosted by NTUU "KPI".

Ключові слова – життєвий цикл даних, препарація, актуалізація, декларативна мова описання, модуль валідації даних.

Ефективне використання даних для наукових досліджень пов'язане з реалізацією життєвого циклу даних – безперервного ітераційного процесу, що складається з етапів планування, збору, попереднього опрацювання, збереження, дослідження, інтеграції, публікації та повторного використання даних. Одним з ключових етапів життєвого циклу даних, від якості виконання якого залежить коректність результатів наукових досліджень, є етап попередньої обробки даних, пов'язаний насамперед з розв'язанням завдання валідації (перевірки на правильність) даних. Нажаль інструментарій валідації в сучасних системах забезпечення життєвого циклу даних не є достатньо розвинутим. В той же час існують певні технології, JUnit для Java, NUnit для .NET, SimpleTest для PHP, Mocha для JS, призначені для тестування програмного забезпечення. Концепція побудови згаданих технологій може бути застосована для створення засобів автоматизації процесу валідації даних.

Мета роботи полягає у зниженні трудомісткості та підвищенні якості виконання процесу валідації даних за рахунок розробки підходу, а також реалізації програмних засобів для структурної організації та паралельно-послідовного виконання системи тестів.

Досягнення мети передбачає розв'язання наступних завдань: виявлення типової структури тестів та набору засобів, необхідних для їх виконання; побудови декларативної мови для опису сценаріїв валідації даних з можливістю паралельно-послідовного виконання; проектування та реалізацію відповідних програмних засобів.

Запропонований підхід до організації процесу валідації даних передбачає дворівневу організацію системи тестів. Перший рівень складають окремі тести, які мають типову структуру і містять операції формування вибірки та препарування даних, а також операції порівняння отриманих препаратів з еталонами. Другий рівень відповідає опису графу паралельно-послідовного виконання окремих тестів.

Розроблена декларативна предметно-орієнтована мова для опису системи тестів є підмножиною предметно-орієнтованої мови аналітичної обробки даних [1] та складається з додаткових стереотипів процесів, таких як query, prepare, expect. Стереотипи parallel та synchronize призначені для опису точок паралельного запуску та синхронізації виконання тестів.

Особливості реалізації програмних засобів пов'язані з організацією паралельно-послідовного виконання тестів та полягають у використанні тестами загальнодоступного контексту, який забезпечує персистентність результатів вибірки, препарування та порівняння даних, а також у запуску окремих тестів в контейнерах, яким відповідають окремі обчислювальні процеси операційної системи.

Програмні засоби верифікації даних реалізовані з використанням серверної платформи Node.js [2] у вигляді набору мікросервісів для сервера опрацювання даних в рамках одного з проектів [3] Світового центру даних з геоінформатики та сталого розвитку [4], направлених на створення розподіленої інформаційної інфраструктури для забезпечення життєвого циклу даних для наукових досліджень.

Розроблені програмні засоби використовуються для валідації даних, що імпортуються з різноманітних офіційних джерел (World Bank, British Petroleum, Fund For Peace, World Health Organization, Heritage Foundation, Transparency International та ін.), та перевірки розрахункових показників сталого розвитку для країн світу і регіонів України, які щорічно публікує Світовий центр даних у вигляді звітів та online - презентацій [3].

Література

1. Болдак А. О. Предметно-орієнтована мова аналітичної обробки даних. / А.О. Болдак, К.В. Єфремов // Вісник НТУУ «КПІ». - К.: Век+, 2012. – № 55 «Інформатика, управління та обчислювальна техніка». – С.50-55.
2. Node.js [Електроний ресурс] – 2015. – Режим доступу: https://nodejs.org/.
3. World Data Center webapp [Електроний ресурс] – 2015. – Режим доступу: https://github.com/kpi-wdc/wdc.
4. World Data Center for Geoinformatics and Sustainable Development – 2015. – Режим доступу: http://wdc.org.ua/.