Висновки

У цій роботі було розглянуто підхід перетворення RDB на базу даних NoSQL, що складається з двох модулів: модуля перетворення та модуля очищення даних.

Запропонований підхід придатний для перетворення баз даних та очищення даних. Як майбутній напрямок цей підхід може бути розширений, щоб забезпечити можливість багаторазового введення даних у разі використання великого обсягу даних. Час трансформації може бути додатково скорочено за допомогою найсучасніших технологій. Модуль перетворення даних може бути вдосконалений в якості майбутньої роботи для підтримки інших RDB та NoSQL. Модуль очищення даних може бути додатково вдосконалено з використанням статистичних методи для ідентифікації дублікатів.

Список літератури

1. *S. Ramzan, I. Bajwa, and R. Kazmi, ‘‘An intelligent approach for handling complexity by migrating from conventional databases to big data,’’ Sym- metry, vol.10, no. 12, p. 698, Nov. 2018.*
2. *MongoDB Production Deployments [Електроний ресурс]. URL:http://www.mongodb.org/about*
3. *A.Davoudian,L.Chen,andM.Liu,‘‘AsurveyonNoSQLstores,’’ACM Comput.Surv., vol. 51, no. 2, p. 40, Apr. 2018.*

УДК 004.855:519.216

Дужак А. О., студент
Зелінська О.В., к.т.н., доцент, доцент кафедри інформаційних технологій

ПРИКЛАДНІ АСПЕКТИ ВИКОРИСТАННЯ ВЕЛИКИХ ДАНИХ ДЛЯ РОЗВ’ЯЗАННЯ ОКРЕМИХ ПРОБЛЕМ СТАТИСТИКИ

Донецький національний університет імені Василя Стуса, м. Вінниця

«Великі дані (Big Data) – позначення структурованих и неструктурованих даних величезних обсягів і значного розмаїття, що піддаються ефективній обробці програмних інструментів, які горизонтально масштабуються та з’явились у кінці 2000-х років, і альтернативних традиційних систем управління базами даних і рішенням класу рішень Business Intelligence».

Основні принципи роботи з великими даними такі:

- **Горизонтальна масштабованість**. Це — базовий принцип обробки великих даних.
- **Відмовостійкість**. Цей принцип витікає з попереднього.

- **Локальність даних**. Оскільки дані розподілені по великій кількості обчислювальних вузлів, то, якщо вони фізично знаходяться на одному сервері, а обробляються на іншому, витрати на передачу даних можуть бути невиправдано великими.

Актуальність пошуку нових джерел інформації для сучасної офіційної (державної) статистики обумовлена низкою факторів. Серед них як найбільш важливі, на наш погляд, слід виділити насамперед такі: зростання потреби користувачів у більш своєчасних (оперативних) даних; значне погіршення відкритості респондентів та зменшення їх бажання брати участь у традиційних обстеженнях; стрімкий розвиток IT технологій, що призводить до експоненційного зростання обсягів даних у державних реєстрах та у приватних компаній і надає можливість фактично у режимі реального часу обробляти значні масиви інформації; покращення кваліфікації працівників органів статистики та їх спроможності швидко засвоювати й упроваджувати новий інструментарій у статистичну практику. Представляється доцільним навести окремі приклади проблем, які можуть розв'язуватись органами державної статистики України за умови 68 забезпечення доступу останніх до даних адміністративних реєстрів, даних мобільних операторів, інформації з сайтів тощо. В Україні існує серйозна проблема (причому це проблема, насамперед, політичного характеру) з проведенням перепису населення – перший і останній всеукраїнський перепис було проведено у 2001 році. Через це все більшої популярності набувають спекулятивні міркування щодо можливості отримання даних, аналогічних даним перепису, з інших джерел: із державних реєстрів, від мобільних операторів щодо кількості абонентів тощо. Констатуючи, що в Україні у теперішній час абсолютно відсутні умови для заміни такого спостереження, як перепис населення інструментами перетворення даних адміністративних реєстрів у статистичні дані та їх об'єднання і комплексного використання, слід тем не менше зазначити таке. Процедури оцінки чисельності населення за адміністративнотериторіальними одиницями, або за територіальними комірками площею 1 км² і дрібнішими з використанням геолокаційних даних мобільних операторів стають все вживанішими [1; 2]. Загально відомо, що дані мобільних операторів (достатньо великих, які надійно покривають усю територію країни) дуже добре відображають розподіл населення за адміністративно-територіальними одиницями (далі – АТО), або за будь-якими територіальними комірками. Для ілюстрації на рис. 1 представлено порівняння розподілів населення за даними демографічної статистики на 01.01.2019 р. та даними одного з трьох найбільших мобільних операторів по АТО Вінницької області.

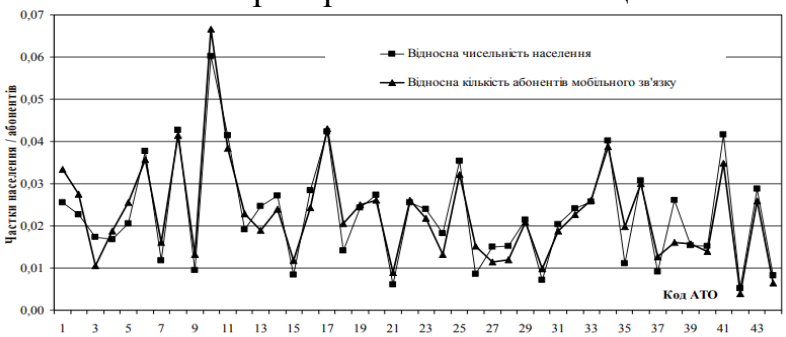

Рис. 1. Розподіл часток населення та кількості абонентів мобільного зв'язку за адміністративно-територіальними одиницями Вінницької області

Візуальна схожість наведених розподілів підтверджується коефіцієнтом кореляції Пірсона, який у цьому випадку дорівнює r = 0,9384.

Зазначимо, що загалом аналогічні значення коефіцієнта кореляції характеризують статистичний взаємозв'язок цих розподілів для інших регіонів України та інших операторів. Відповідно, виникає можливість достатньо адекватного моделювання чисельності населення за АТО за даними мобільних операторів. Для побудови таких моделей не вистачає лише точних даних щодо чисельності населення, які, в свою чергу, можуть бути отримані за даними перепису населення України. І все ж існує певна можливість побудувати такі моделі, хоча і менш надійні, використовуючи актуальні дані реєстрів, які достатньо повно відстежують чисельність окремих категорій населення. У такому випадку мова йде, насамперед, про реєстр осіб пенсійного віку. Іншим прикладом, що ілюструє потенціал великих даних для практиці статистичних спостережень, є оцінка та прогнозування динаміки трудової міграції. В Україні обстеження трудової міграції органами статистики здійснюються один раз на 5 років. Враховуючи актуальність і гостроту цього питання, масштаби й динаміку трудової міграції і необхідність її урахування при оцінці та прогнозуванні попиту, і пропозиції робочої сили, представляється доцільним приділити увагу можливості використання даних з сайтів, щодо пошуку та пропозиції роботи. Такі дослідження проведені в Інституті демографії та соціальних досліджень імені М. В. Птухи НАН України. Результати цих досліджень частково опубліковані в роботі [3].

Список літератури

1. *Khodabandelou G., Gauthier V., El-Yacoubi M., Fiore M. Population Estimation from Mobile Network Traffic Metadata. 2016. Doi: 10.1109/TMC. 2018.2871156*
2. *Handbook on the use of Mobile Phone data for Official Statistics UN Global Working Group on Big Data for Official Statistics. UN Global Working Group on Big Data for Official Statistics. Draft. 2017. URL: https://unstats.un.org/bigdata/taskteams/mobilephone/Handbook%20on%20Mobile%20Phone%20Data%20for%20official%20statistics%20-%20Draft%20Nov%202017.pdf*
3. *Веремчук А. В., Розбицький М. А. Оцінка потенціалу «великих даних» для досліджень трудової міграції // Демографія та соціальна економіка. 2019. 1 (35). С. 196–208.*
4. *https://www.it.ua/knowledge-base/technology-innovation/big-data-bolshie-dannye*

УДК 004.6

Ковальчук С.Ю., студент 2 курсу спеціальності 122 «Комп'ютерні науки»
Потапова Н. А., к.е.н., доцент,