Міністерство освіти і науки України
Національний університет «Львівська політехніка»

Болюбаш Юрій Ярославович

УДК 004.9:371.261

МЕТОДИ ТА ЗАСОБИ ОПРАЦЮВАННЯ ІНФОРМАЦІЙНИХ РЕСУРСІВ ВЕЛИКИХ ДАНИХ В СИСТЕМАХ ТЕРИТОРІАЛЬНОГО УПРАВЛІННЯ

01.05.03 – математичне та програмне забезпечення
обчислювальних машин і систем

Автореферат
дисертації на здобуття наукового ступеня
кандидата технічних наук

Львів – 2017

Дисертацією є рукопис.

Робота виконана в Національному університеті «Львівська політехніка» Міністерства освіти і науки України.

Науковий керівник: доктор технічних наук, професор
Шаховська Наталія Богданівна,
Національний університет «Львівська політехніка»,
професор кафедри інформаційних систем
та мереж

Офіційні опоненти: доктор технічних наук, доцент
Овсяк Олександр Володимирович,
Відокремлений підрозділ «Львівська філія
Київського національного університету культури і
мистецтв», професор кафедри мистецтв;

кандидат технічних наук, доцент
Угрин Дмитро Ілліч,
Чернівецький факультет Національного технічного
університету «Харківський політехнічний інститут»,
завідувач кафедри інформаційних систем.

Захист відбудеться «07» квітня 2017 р. о 14^{00} годині на засіданні спеціалізованої вченої ради Д 35.052.05 у Національному університеті «Львівська політехніка» (79013, м. Львів, вул. Професорська, 2, корп. 11, ауд. 218).

З дисертацією можна ознайомитись у Науково-технічній бібліотеці Національного університету «Львівська політехніка» (79013, м. Львів, вул. Професорська, 1).

Автореферат розісланий «03» березня 2017 р.

Учений секретар
спеціалізованої вченої ради,
доктор технічних наук, професор Р.А. Бунь

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Застосування систем аналізу та планування територіального розвитку сприяє швидкому поширенню знань, навичок та найкращих практик у певних географічних межах, таких як місто, регіон, країна тощо. Для комплексного аналізу інформації на рівні регіону необхідно:

- зберігати і керувати інформацією розміром у петабайти;
- опрацьовувати як структуровану, так і неструктуровану (у вигляді текстових звітів) інформацію, працювати з картографічними даними;
- аналізувати різнотипну інформацію, використовуючи як консолідований, так і федеративний підхід для її отримання.

Процес побудови узагальненої (комплексної) моделі регіону ускладнюється різноманітністю моделей даних, а також наявністю різних рівнів агрегації даних. Однією з популярних технологій для розроблення систем територіального управління є Великі дані.

Великі дані є терміном, який використовується для ідентифікації наборів даних, з якими не можна впоратися з використанням існуючих методологій та програмних засобів через їх великий розмір, швидкість надходження, аналіз та складність. Такі дослідники як М. Гілберт, С. Стрініваса та ін. розробили методики і програмні засоби для передачі даних та видобування інформаційних гранул з Великих даних (колекції об'єктів, які зазвичай формуються для числових атрибутів, що розташовані поряд через їх схожість, функціональну або фізичну суміжність), проте поява нових форматів даних вимагає постійного розширення та вдосконалення існуючих методів аналізу даних.

Методи машинного навчання та візуалізації даних дають змогу опрацювати та графічно подати результати аналізу даних великих обсягів (мільйони кортежів). Проте нерозв'язаною задачею залишається задача побудови відображення між моделями даних різних джерел. В роботах Alejandro Maté та Lucentia Research Group обґрунтовано використання багатовимірної моделі для представлення Великих даних та побудови відображення у реляційну модель. Проте використання бази даних «ключ-значення» як джерела даних для багатовимірної моделі є неприйнятним. Vinayak Borkar та Yingyi Bu пропонують використовувати об'єктно-орієнтований підхід, проте обмеженням є кількість зв'язків між об'єктами. Отже, єдиного підходу до опрацювання Великих даних не знайдено.

Коли аналізувати можливість використання Великих даних у системах планування та аналізу територіального розвитку, то маємо:

- сотні тисяч сутностей: особи, місця, організації (фізичні, юридичні), дати, природні ресурси (річки, ліси, озера), рекреаційний фонд (історичні пам'ятки, санаторії), законодавчі акти та звіти;
- база даних характеристик сутностей з мільйонами кортежів: документи для інтелектуального аналізу даних, онтологічні терміни, словники даних, які дають змогу відобразити зв'язки між об'єктами.

Грунтуючись на цій інформації, ми повинні вирішити, які сутності і як пов'язані між собою з метою їх подальшого аналізу.

Тому задача розроблення методів та засобів опрацювання та аналізу різнорідної інформації на основі Великих даних для процесів територіального управління є актуальною.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційна робота виконувалась у межах пріоритетного наукового напрямку, затвердженого в числі актуальних проблем Міністерством освіти і науки України, зокрема у рамках науково-дослідної теми «Комплекс інтелектуальних інформаційних технологій інтеграції даних для обліку та аналізу підвищення кваліфікації вчителів», номер держреєстрації 0113U005273 (автор розробив модель даних «сутність-характеристика» для подання та аналізу різнорідних даних).

Метою дисертаційної роботи є розробка методів та засобів опрацювання та аналізу різнорідної інформації на основі Великих даних для процесів територіального управління.

Мета дисертаційної роботи визначає необхідність розв'язання таких задач:

1) проаналізувати методи, моделі та засоби опрацювання різнотипних даних, інформаційні технології роботи з Великими даними, обґрунтування й постановка задачі;
2) розробити інформаційну модель Великих даних «сутність-характеристика»;
3) розробити метод перетворення реляційних та слабоструктурованих даних у дані, подані в моделі «сутність-характеристика»;
4) розробити метод формування відповіді на запит користувача до Великих даних;
5) розробити програмні компоненти інформаційної системи підтримки прийняття рішень для управління регіоном з використанням Великих даних.

Об'єктом дослідження є процес інтеграції слабоструктурованих даних, виділення елементів структури та збереження їх у базу даних.

Предметом дослідження є методи та засоби опрацювання Великих даних.

Методи дослідження. Для досягнення поставленої мети використано:

- методи системного аналізу – для формування концептуальної моделі Великих даних;
- методи штучного інтелекту – для виявлення закономірностей у каталозі Великих даних;
- методи об'єктно-орієнтованого аналізу та проектування – для визначення семантичних зв'язків між джерелами даних.

Наукова новизна одержаних результатів. Наукова новизна роботи полягає у розв'язанні актуального наукового завдання розроблення методів та засобів організації та інтеграції інформаційних ресурсів Великих даних. Отримано такі нові наукові результати:

- *вперше:* розроблено модель Великих даних «сутність-характеристика», яка дає змогу опрацьовувати структуровані та напівструктуровані дані та на відміну від багатовимірної моделі не містить надлишковості;

- *удосконалено:* отримання відповіді на запит користувача шляхом уніфікації елементів мов запитів до різних інформаційних джерел, що дало змогу трансформувати запит до Великих даних у запит на основі ключових слів;
- *одержав подальший розвиток:* федеративний метод інтеграції даних за рахунок визначення пари «сутність-характеристика» та узгодження семантики, що на відміну від методів інтеграції даних на рівні сховища даних дозволило інтегрувати дані з джерел з наперед невідомою структурою даних без попереднього локального завантаженняі що, в свою чергу, дало змогу підвищити ефективність подальшого аналізу Великих даних.

Практичне значення одержаних результатів:

- удосконалено алгоритми інтеграції інформаційних ресурсів за допомогою попереднього визначення структури джерел даних та їх узгодження, що дало можливість розробити алгоритм оцінювання якості Великих даних;
- розроблено алгоритми пошуку даних за запитом користувача з метою уніфікації алгоритмів опрацювання даних з різнотипних джерел даних, що дозволило збільшити релевантність відповіді користувачеві на 2%;
- спроектовано архітектуру інформаційної системи для роботи з Великими даними.

Одержані у роботі результати використано під час розроблення програмних компонент інформаційної системи для збору та опрацювання даних «Інтегратор», впровадженої в Золочівській районній раді та Регіональній агломерації «Дрогобиччина». Розроблення впроваджені також в навчальному процесі при викладанні курсу «Бази даних» Золочівського коледжу НУ «Львівська політехніка».

Особистий внесок здобувача. Усі наукові результати, подані у дисертації, одержані здобувачем особисто. У друкованих працях, опублікованих у співавторстві, внесок здобувача такий: [3, 4, 5, 14] – визначено поняття Великих даних та базові характеристики; [9, 7, 11, 13] – розроблено модель «сутність-характеристика» та визначено асоціації з неструктурованими моделями даних; [6, 10, 15] – розроблено метод перетворення даних в модель «сутність-характеристика»; [8, 12] – спроектовано архітектуру Великих даних.

Апробація результатів дисертації. Основні результати дисертаційної роботи доповідалися на семінарах та конференціях: Міжнародних конференціях «The experience of designing and application of CAD systems in microelectronics» – CADSM (Львів-Поляна, 2015); X Міжнародна конференція. «Комп'ютерні науки та інформаційні технології» – CSIT (Львів, 2015); Міжнародній конференції «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту» – ISDMCI'2013 (Євпаторія, 2013); II та III Міжнародних науково-практичних конференціях «Математика. Інформаційні технології. Освіта» (Луцьк-Світязь, 2013, 2014), III Міжнародній науково-практичній конференції «Інформаційні управляючі системи та технології» – ІУСТ (Одеса, 2014).

Публікації. Основні результати роботи відображені у 15 опублікованих працях, у тому числі 6 статей у наукових фахових виданнях України, 3 – у

наукових періодичних виданнях інших держав, що входять до наукометричних баз даних, 6 – в збірниках праць конференцій.

Структура і обсяг роботи. Дисертаційна робота складається з вступу, чотирьох розділів, висновків та додатків. Має загальний обсяг 178 сторінок, основна частина – 140 сторінок, містить 45 рисунків та 12 таблиць, 141 найменування у списку використаних літературних джерел. У додатках наведені: акти впровадження та програмний код розробленої інформаційної системи.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** обґрунтовано актуальність теми дисертаційної роботи, сформульовано мету та задачі дослідження, визначено наукову новизну та практичне значення отриманих результатів, показано зв'язок роботи з науковими темами. Подано відомості про апробацію результатів роботи, публікації та особистий внесок здобувача.

У **першому розділі** описано регіон як складну систему. Визначено задачі, які виникають під час опрацювання різнотипних даних. Розглянуто особливості аналізу соціо-еколого-економічних даних та вказано, що їх доцільно опрацьовувати з допомогою технології Великих даних. Розглянуто різні визначення Великих даних та сформовано узагальнене визначення, описано основні характеристики. Проаналізовано математичні засоби подання та опрацювання Великих даних та визначено їх обмеження. Подано програмні засоби роботи з Великими даними. Здійснено постановку задачі.

Великі дані (Big Data) в інформаційних технологіях – набір методів та засобів опрацювання структурованих і неструктурованих різнотипних динамічних даних великих обсягів з метою їх аналізу та використання для підтримки прийняття рішень. До цього класу відносять засоби паралельного опрацювання даних (NoSQL, алгоритми MapReduce, Hadoop). Визначаль-ними характеристиками для Великих даних є обсяг, швидкість, різноманіття.

Визначено, що для представлення Великих даних використовують багатовимірну та об'єктну моделі. Багатовимірне представлення даних добре використовувати для задач візуалізації даних та їх аналізу, але у зв'язку з розрідженістю гіперкуба обсяг даних у такому випадку є більший порівняно з реляційним представленням, що є неприпустими до Великих даних. Об'єктне подання дає змогу зберігати об'єкти предметної області у вигляді атрибутів, їх характеристик та зв'язків між характеристиками. За певної модифікації воно може бути використане для Великих даних. Проте залишається нерозв'язаною задача трансформації різних типів даних (текстових, напівструктурованих) в об'єктну модель. Це призводить до необхідності розроблення моделі представлення Великих даних.

Результати аналізу предметної області дали змогу сформулювати мету та задачі дослідження. Матеріали розділу опубліковано у [4, 5, 14].

У **другому розділі** здійснено формальну постановку задачі. Подано визначення різних типів даних. Визначено формальний опис Великих даних у вигляді моделі «сутність-характеристка». Подано моделі асоціацій між сутностями та характеристиками для різних категорій NoSQL баз даних.

Використано простір даних для представлення Великих даних. Подано модель федеративного сховища Великих даних.

Проаналізовано основні представлення слабоструктурованої інформації: OEM (Object Exchange Model, модель обміну об'єктами), XML (Extensible Markup Language, розширювана мова розмітки), RDF (Resource Description Framework, фреймворк опису ресурсів).

Елементи реалізації ідеї слабоструктурованої обробки й зберігання даних місятяться у безсхемних БД, що відносяться до типу NoSQL. Це лягло в основу розробленої *моделі «сутність-характеристика» для опису Великих даних*. Особливістю БД типу NoSQL, зокрема, є горизонтальне масштабування сховища даних і підтримання пошуку й індексування за будь-якими полями, а в деяких БД є можливість формування будь-яких запитів вибірки даних, що також є важливим для Великих даних.

В моделі «сутність-характеристика» усі об'єкти предметної області розділено на такі категорії:

- сутності e,
- характеристики f,
- асоціації між сутностями e та характеристиками f.

Також визначено:

- множину сутностей E;
- множину характеристик F;
- для кожних e і f зазначено номер асоціацій між e і f як $n_{e,f}$.

Загальна кількість сутностей визначається як $|E|$, загальна кількість характеристик є потужністю множини F: $|F|$. Також описано:

- для кожної характеристики f множину $e(f) = \{e \in E : n_{e,f} > 0\}$ усіх асоційованих з f сутностей;
- для кожної сутності e множину $f(e) = \{f \in F : n_{e,f} > 0\}$ усіх асоційованих з e характеристик.

Коли є кілька сутностей, пов'язаних з характеристикою, використано кількісне представлення інформації, тобто кількість звернень, які необхідно, щоб знайти потрібний об'єкт. Загалом, якщо ми знаємо, що невідомий об'єкт належить множині, яка складається з N елементів, то можемо розділити цей набір на дві половини і, задаючи питання, з'ясувати, до якої половини належить шуканий об'єкт. Отже, тоді кількість об'єктів становитиме $N/2$. Загалом, після відповіді на q бінарних запитань (відповідь на які становить так або ні) матимемо множину з $N \cdot 2^{-q}$ елементів, що містить необхідний об'єкт.

Коли ми досягнемо q^{-1}, то це означатиме, що множина складатиметься з одного елемента, і ми точно визначимо необхідну нам альтернативу. Таким чином, для N альтернатив відповідна інформація (кількість бінарних запитань) складатиме $N \cdot 2^{-q} = 1$, отже, буде рівна $q = \log_2(N)$.

Так само здійснено пошук сутності. Маємо $|E|$ сутностей з кількістю інформації $\log_2(|E|)$. Коли ми знаємо, що якась сутність асоційована з характеристикою (маємо $|e(f)|$ сутностей), то кількість питань рівна $\log_2(|e(f)|)$. Таким чином, той факт, що сутність пов'язана з характеристикою f, дозволяє зменшити кількість питань до

$$\log_2(|E|) - \log_2(|e(f)|) = \log_2\left(\frac{|E|}{|e(f)|}\right). \quad (1)$$

Загальна важливість характеристики f для сутності e визначена як $\log_2\left(\frac{|E|}{|e(f)|}\right)$ з фактором важливості $1 + \log_2(n_{e,f})$. Результуюча кількість питань (або важливість характеристики для сутності) визначена як

$$I(e,f) = (1 + \log_2(n_{e,f})) \cdot \log_2\left(\frac{|E|}{|e(f)|}\right). \quad (2)$$

Формула (2) є так званою зворотною частотою документа tf-idf, отже дає змогу описувати кількість інформації у неструктурованому документі. За аналогією до tf-idf оцінки частоти терміну t в документі d маємо $wf_{t,d} = 1 + \log tf_{t,d}$. Для кожної сутності e маємо важливість $I(e,f)$ для різних характеристик f. Значення важливості необхідно нормалізувати, оскільки різні сутності мають різну кількість характеристик:

$$V(e,f) = \frac{(1 + \log_2(n_{e,f})) \cdot \log_2\left(\frac{|E|}{|e(f)|}\right)}{\sqrt{\sum\left((1 + \log_2(n_{e,f})) \cdot \log_2\left(\frac{|E|}{|e(f)|}\right)\right)^2}}. \quad (3)$$

Оцінено важливість характеристики f для сутності e як відстань між терміном і документом, і, на відміну від реляційної моделі, ця відстань не завжди дорівнює одиниці.

Для кожної сутності e є вага $V(e,f)$. Таким чином, мірою близькості між двома об'єктами E_1 і E_2 є відстань між відповідними векторами $(V(e_1,f), V(e_2,f), ...)$:

$$d(e_1,e_2) = \sum_{f \in F} |V(e_1,f) - V(e_2,f)|. \quad (4)$$

Формула (4) репрезентує відстань між сутністю і характеристикою в певному документі. Відстань залежить від кількості характеристик: наприклад, якщо на додаток до документів, ми зберігаємо їх копії, відстань збільшується вдвічі. Щоб уникнути цієї залежності, відстань $d(e_1,e_2)$ нормалізується в інтервалі [0,1] шляхом ділення на максимально можливе значення цієї відстані.

Далі у розділі охарактеризовано властивості метрики *d* та введено поняття моделі асоціацій між об'єктами та характеристиками для різних категорій NoSQL баз даних.

Носій даних у моделі «ключ-значення» (інша назва – колонкова БД) описується кортежем вигляду:

$$KV = \{< k, v >\}, \tag{5}$$

де k – ключ, який приймає унікальні значення у кожній парі, v– значення, що відповідає цьому ключу; $e \leftrightarrow k; f \leftrightarrow v$. Сигнатура моделі є підмножиною сигнатури реляційної моделі і визначена як $O = \langle \pi, \sigma \rangle$, де π – операція проекції за атрибутами (ключ або значення), σ – операція селекції атрибутів (вибір значення за ключем, ключів за значенням, ключів за значенням предків). Вага $d(e, f) = 0$.

Для версійного розподіленого зберігання великих обсягів даних в Google була спроектована модель BigTable. Це неповна реляційна модель даних з підтримкою динамічного контролю над розміщенням даних. Її основа – рядки $r \in E$, стовпці $c \in F$ і тимчасові мітки t:

$$\begin{gathered} BigTable = \{< r, c, t >\}, \\ f \leftrightarrow r(t), \\ e \leftrightarrow c. \end{gathered} \tag{6}$$

Якщо в кількох стовпцях зберігаються дані одного типу, такі стовпці, згідно моделі Bigtable, утворюють сімейство:

$$colF = \{c_i, c_j \mid dom(c_i) \in T \wedge dom(c_j) \in T\}, \tag{7}$$

що дає змогу стиснути однорідні дані, тим самим зменшивши обсяг. Саме сімейства стовпців є одиницею доступу до даних. Якщо $\pi_c(r) \neq \varnothing$, то $d(\pi_c(r), c) \geq 0$, в іншому випадку $d(\pi_c(r), c) = 1$.

Носій моделі «об'єкт-документ» описується кортежами вигляду:

$$OD = \{< f_0, < f_1 : e_1, f_2 : e_2, f_n : e_n, f_{n+1} : d_1, f_2 : d_2, f_{n+l} : d_l >\}, \tag{8}$$

де f_0 – ідентифікатор документу, $f_1..f_m$ – характеристики (атрибути) документу, $e_1..e_m$ – атомарні значення характеристик $f_1..f_m$, $d_1..d_l$ – посилання на інші документи, $d_i = e(f_i)$.

Операції цієї моделі визначено як модифікацію операцій об'єктної моделі представлення даних.

Операція визначення вузлів елемента:

$$v(f_i) = \{C\} \cup \{f_0 \mid i = \overline{1, n}\} \cup \{e(f_i) \mid i = \overline{0, n+l}\}, \tag{9}$$

де C – колекція документів f_0.

Операція визначення значень вузлів:

$$v(f_i) = \{e_{ij} \mid i = \overline{1, n}, j = \overline{0, m+l}\}, \tag{10}$$

де e_{ij} – значення атрибутів f_i.

Також визначено відношення над елементами носія:

- відношення «елемент-елемент» між документами та колекцією $OD \times C \rightarrow EE$ (11)
- відношення «елемент-атрибут»: $f_i \times OD \rightarrow EA$. (12)

- відношення «елемент-посилання»: $f_i \times d_j \to ER$. (13)
- відношення «елемент-дані»: $f_i \times e_j \to ED$. (14)

Графова модель даних подана як:

$$O = \langle ID, A, z, r \rangle, \quad (15)$$

де *ID* – множина ідентифікаторів, вузлів графа; *A* – множина позначених спрямованих дуг (*p*,*l*,*c*), $p,c \in ID$, *l* – «рядок-мітка», запис (*p*,*l*,*c*) означає, що між вузлами *p* та *c* є зв'язок *l*, $\forall A : d(p,c) \geq 0$; z – функція, що відображає кожний вузол $n \in ID$ в конкретне значення складеного або атомарного типу, $z : n \to v$; *V* – особливий кореневий вузол графа.

Відмінність структури XML-документа, що складається з вкладених елементів-тегів, від графової моделі полягає, в основному, у трактуванні тегів і міток: в графах мітки використовуються як позначення зв'язків між елементами схем даних, і мітки не потрібні для позначення елемента, а в XML документно-орієнтованій моделі потрібно, щоб кожний (нетекстовий) елемент даних мав ідентифікуючу ознаку. Також XML транслюється в структуру даних «дерево», що є частковим випадком графової моделі.

Опис ресурсів у вигляді RDF-набору даних – це трійка «суб'єкт»-«предикат»-«об'єкт», тобто множина *U* (Universal Resource Identifier, URI, уніфікований ідентифікатор ресурсів) складається з елементів *f*, множина *B* (Black nodes) з порожніх вузлів, множина *L* (Literal) з *RDF*-літералів, $B \in e, L \in e$, визначається набір (*f*, *e*(*f*), e), де *f* – «суб'єкт»; *e*(*f*) – «предикат»; *e* – «об'єкт».

RDF-графова модель даних: нехай *t* = (*f*, *e*(*f*), *e*) є *RDF*-набором даних, де $(f,e(f),e) \in (UB) \times U \times (UBL)$, причому *t* називається основним, якщо він не містить вузлів, які не мають ідентифікаторів. RDF-граф *G* є множиною $T \supseteq t$.

Далі у розділі описано Великі дані як елемент простору даних. Побудовано архітектуру федеративного сховища даних, яке не вимагає проміжного збереження даних, отриманих з інших джерел, що дає змогу зменшити ємнісну складність запиту.

Для технології Великих даних необхідним є опрацювання інформації з різних за виразною потужністю типів джерел інформації: структурованих, слабоструктурованих, неструктурованих. Відповідно федеративне сховище даних, побудоване на їх основі, містить реляційні бази даних, багатовимірні бази даних, бази даних XML, бази даних NoSQL, файлове сховище, репозиторій метаданих, інтегратор джерел даних і подання для доступу до сховища (рис. 1).

Основними характерними властивостями, які відрізняють федеративні сховища даних для Великих даних від інших сховищ даних, є такі.

- Наявність реляційної БД, основним призначенням якої є зберігання структурованих даних і даних, до яких здійснюється частий доступ.
- Наявність багатовимірної БД, яка може містити як атомарні, так і узагальнені дані. Основним призначенням багатовимірної бази даних є зберігання даних, з якими виконуються складні запити.

- Наявність бази даних XML та баз даних NoSQL, основним призначенням якої є зберігання слабоструктурованих даних.
- Збереження неструктурованих даних у вигляді файлів, що зберігаються безпосередньо у файловій системі.
- Взаємодія з джерелами даних здійснюється за допомогою інтегратора, полягає у відслідковуванні змін даних і метаданих, які відбуваються у джерелах, і застосуванні цих змін відповідно до налаштувань сховища даних.
- Уніфікований доступ користувачів до сховища даних через подання (view), який дає змогу користувачам звертатися до даних за допомогою єдиного інтерфейсу, незалежно від фізичного та логічного розташування цих даних у сховищі.

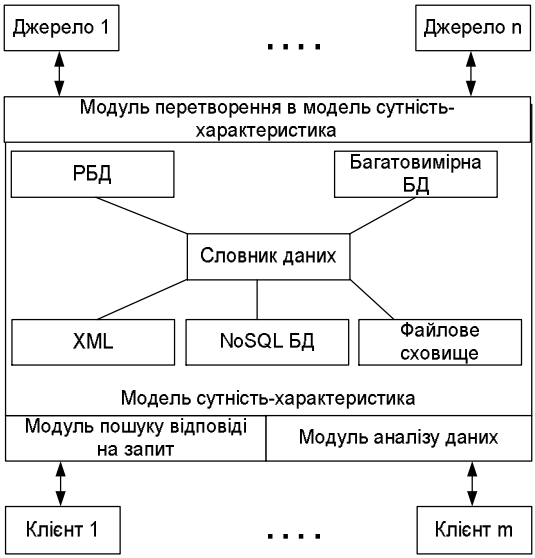

Рис. 1. Архітектура федеративного сховища даних

Результатом розділу є побудова інформаційної моделі даних «сутність-характеристика» для Великих даних, що дало змогу опрацьовувати дані різних форматів.

Матеріали розділу опубліковано у [1-3, 9, 7, 11, 13].

У **третьому розділі** розроблено метод обміну різнотипними даними та приведення реляційних та XML-даних до моделі «сутність-характеристика». Описано метод формування відповіді на запит користувача з врахуванням типу даних джерела.

Основною метою перетворення даних у модель «сутність-характеристика» є забезпечення можливості опрацювання даних будь-якої структури.

Метод перетворення даних у модель сутність характеристика полягає у перерахунку важливості характеристики для сутності, а також фізичному пертворенні схеми даних у пару «сутність-характеристика» (рис. 2).

Для розширення XML використано архітектуру MVVM («View-Model», «Вигляд-Модель»). Цей шаблон застосовується при проектуванні архітектури програмного додатка й використовує поділ моделі в частині логіки її функціонування та її подання на користувацькому інтерфейсі. Основною її відмінністю від відомої архітектури програмного додатка MCV (Model-View-Controller, Модель-Вигляд-Контролер) є відсутність вимоги прив'язки даних до їхнього подання.

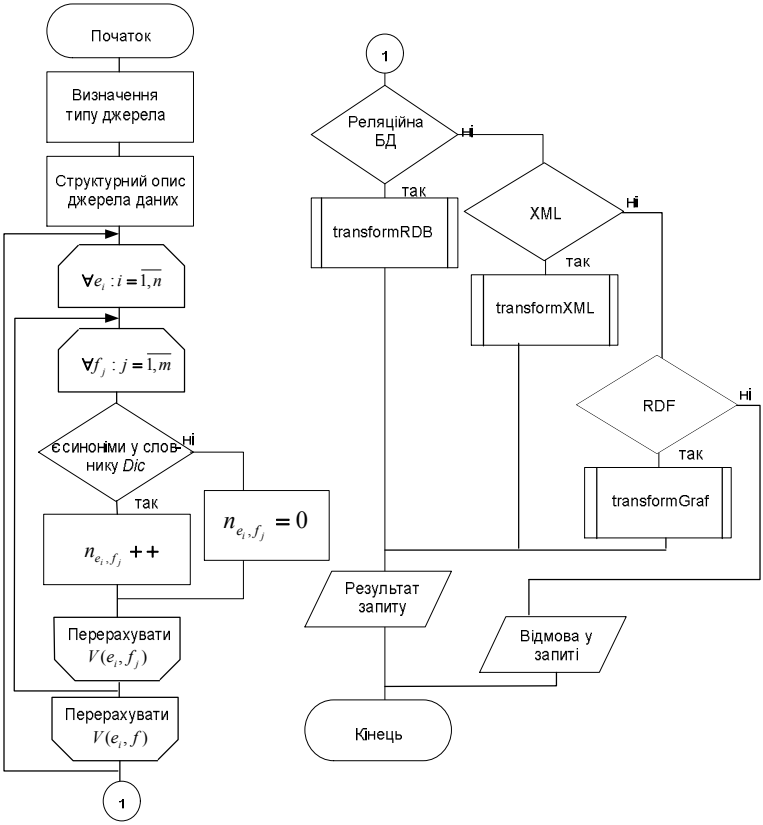

Рис. 2. Схема перетворення реляційних даних у модель «сутність- характеристика»

Відповідно до цього підходу розроблені компоненти для автоматизованої конвертації з реляційної БД в модель XML з врахуванням особливостей моделі «сутність-характеристика» (рис. 3).

Далі розроблено метод перетворення реляційних даних у модель «сутність-характеристика» transformRDB, який містить такі кроки.

Крок 1. Створити кореневий елемент схеми RDB для моделі даних Big Data.

Крок 2. Для кожної сутності *e* створити окремий елемент Entity і розмістити його під кореневим елементом.

Крок 3. Для кожної характеристики *f* сутності *e* з важливістю більшою за порогове значення створити атрибут *Feature*, розмістити його усередині відповідного опису сутності й задати його тип.

Крок 4. В атрибуті необхідно вибрати й задати тип даних та визначити граничні значення.

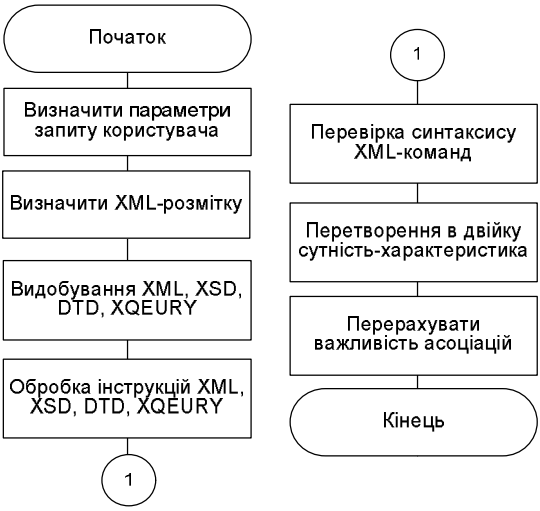

Рис. 3. Схема конвертації реляційних таблиць у XML-БД з врахуванням моделі «сутність-характеристика»

Для перетворення графової моделі в модель «сутність-харакетристика» важливим є визначення ваги зв'язку між елементами. Оскільки першим параметром моделі є характеристика, другим – зв'язок, а третім – сутність, то перетворення між моделями полягатиме у числовому вираженні асоціації між елементами RDF-моделі (рис. 4).

Далі у розділі побудовано *метод формування відповіді на запит користувача* до Великих даних:

Крок 1. Ініціалізація параметрів пошуку.

Крок 2. В циклі виконується потокове читання вузла за вузлом з документа джерела.

Необхідна сутність присутня в елементі-джерелі? Так. Оброблюваний вузол є шуканим?

Так: пошук характеристики вузла.

Ні: обробка наступного вузла.

Крок 3. Перевірка вузла на виконання умов відповідності шуканому значенню характеристики.

Приєднання як дочірнього елемента до оброблюваного цільового вузла елемента батьківського документа.

Крок 4. Очищення приєднаного дочірнього елемента від внутрішніх елементів. Перехід на крок 2.

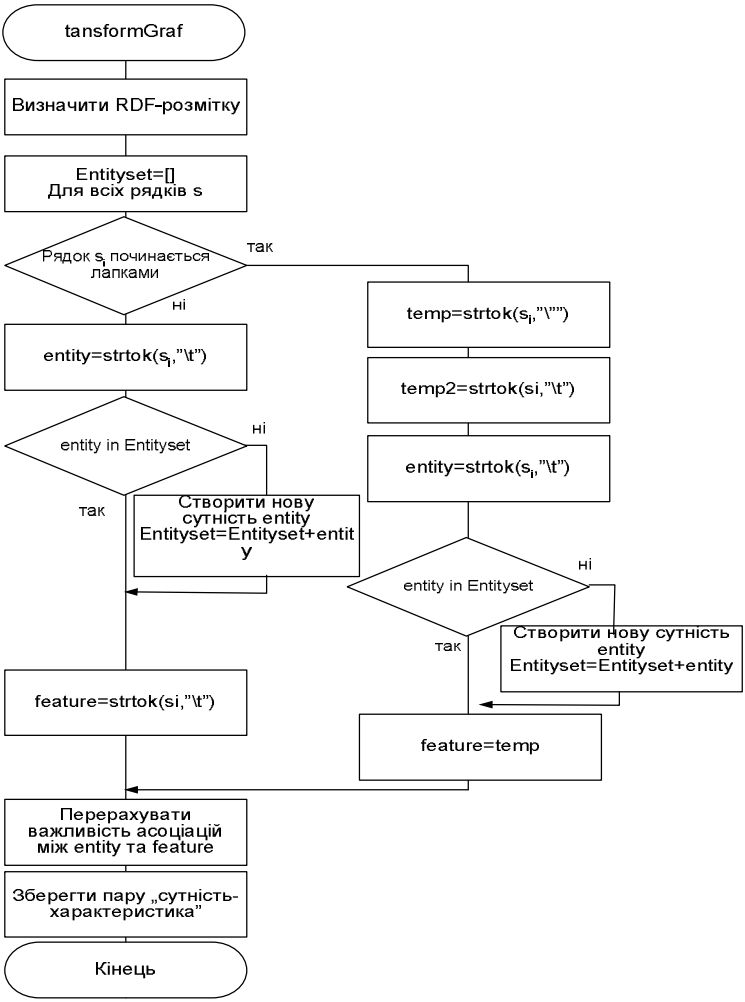

Рис. 4. Схема конвертації графової БД з врахуванням моделі «сутність-характеристика»

У розділі розроблено метод перетворення реляційних даних у модель «сутність-характеристика», що дає змогу здійснювати федеративне опрацювання даних, а також метод відповіді на запит користувача до Великих даних.

Результати розділу опубліковано у [6, 10, 15].

У **четвертому розділі** спроектовано архітектуру інформаційної системи для роботи з Великими даних для опрацювання даних регіону. Апробовано розроблені методи та алгоритми.

Архітектура інформаційної системи опрацювання Великих даних подана на рис. 5. Виділено рівні сервісів, їх підтримки, платформ та даних.

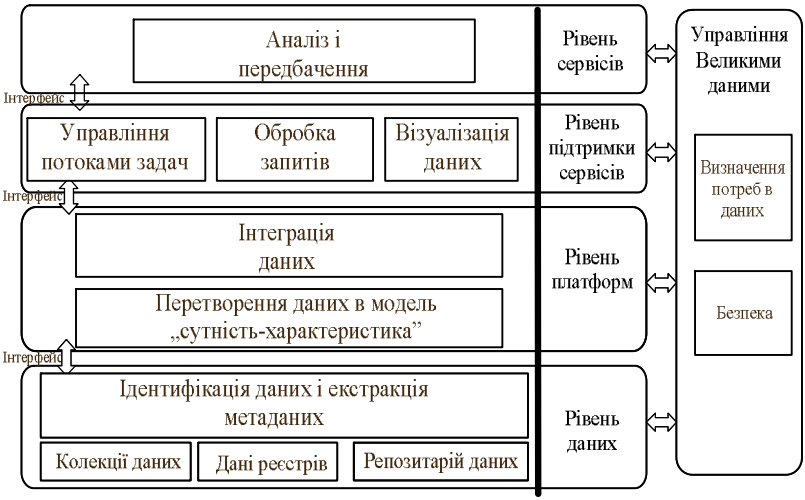

Рис. 5. Архітектура інформаційної системи для роботи з Великими даними

Розроблено діаграму послідовності опрацювання запиту користувача (рис. 6).

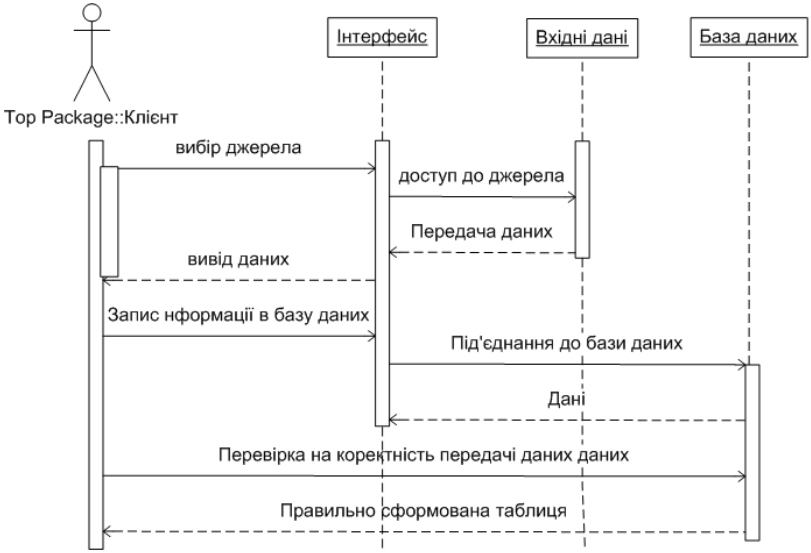

Рис. 6. Діаграма послідовності

У першу чергу здійснюється федеративне опрацювання даних з джерел. Аналізується відносна кількість об’єктів або документів, наявних у джерелах даних, до загальної кількості об’єктів, які потрапили у федеративне сховище (рис. 7).

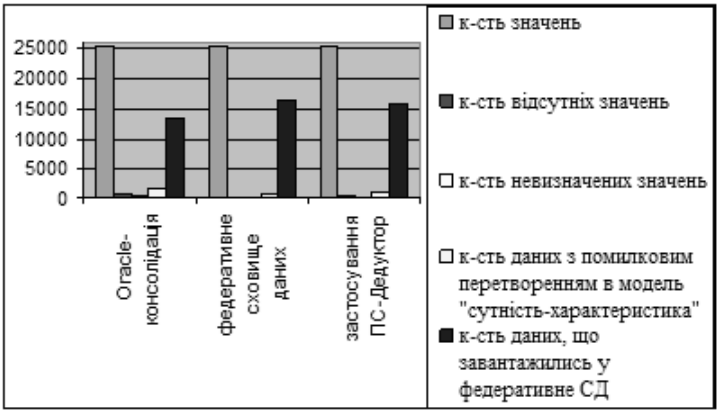

Рис. 7. Аналіз повноти накопичених описів об'єктів

На поданій діаграмі проаналізовано роботу алгоритму федеративного запиту. Роботу алгоритму порівняно з роботою алгоритму немодифікованої інтеграції, застосованому в Oracle Data Integrator. Дані у федеративне сховище потрапляють з баз даних різних установ, структури даних яких наперед невідомі. Кількість записів вхідних баз даних, що мають потрапити у федеративне сховище даних – 15000.

Розроблено діаграму інтерфейсу (рис. 8) та проаналізовано правильність перетворення запитів до різних моделей даних (Таблиця 1).

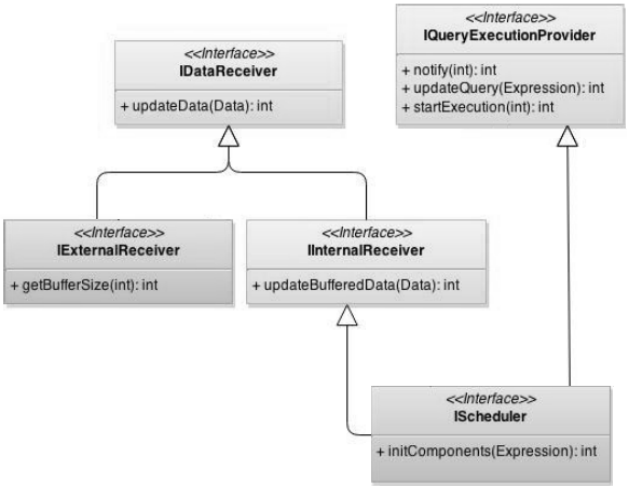

Рис. 8. Діаграма інтерфейсу взаємодії автоматичного перетворення запитів

З цією метою порівнювалися розроблені засоби в системі Інтегратор з середовищем DocumentDB, у якому є можливість формування запитів на мові SQL та NoSQL. Результати порівняння подано в таблиці у відсотковому значенні

правильно поданих запитів. Загальна кількість запитів, що тестувались – по 50 запитів кожної категорії. Правильність формування запитів перевірялася експертно.

Таблиця 1

Порівняння правильності виконання запитів до різних джерел даних

	Інтегратор				DocumentDB			
вид запиту	select	select...join	insert	delete	select	select...join	insert	delete
РБД (Microsoft SQL Server Database Services)	98	95	93	93	98	92	94	94
XML (XBase)	86	82	-	-	-	-	-	-
NoSQL (mongoDB)	91	86	84	84	89	82	81	81

Отримані результати лягли в основу розробленого програмного забезпечення для аналізу соціо-еколого-економічних показників розвитку регіону (рис. 9).

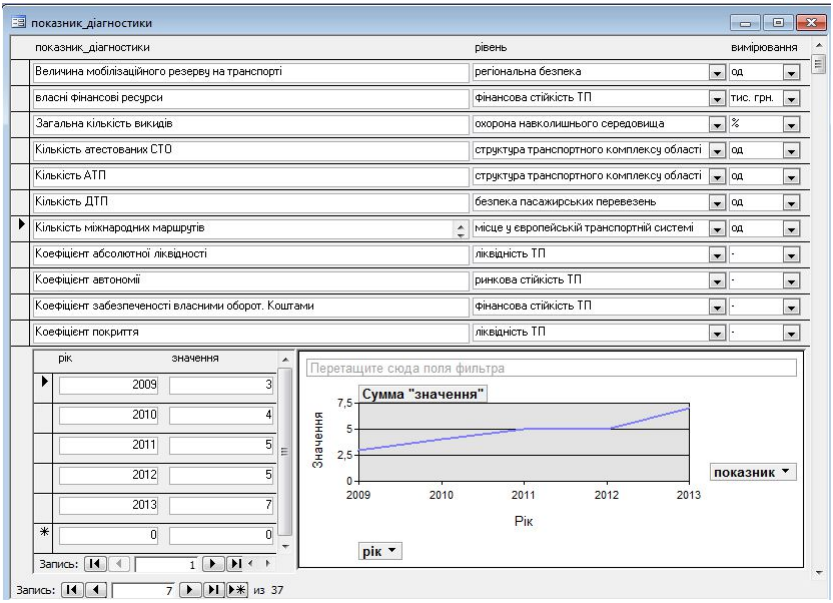

Рис. 9. Вигляд форми для аналізу показників розвитку регіону

У розділі побудовано архітектуру інформаційної системи опрацювання Великих даних, схему сховища даних регіону, алгоритм отримання даних з

різних джерел та їх аналізу, розроблено засоби для розрахунку та аналізу основних показників розвитку регіону.

Результати розділу опубліковано у [8, 12].

У **додатках** наведено акти впровадження результатів дисертаційної роботи та програмний код розробленої інформаційної системи.

ВИСНОВКИ

У дисертаційній роботі розв'язано важливе наукове завдання розроблення методів та засобів організації та інтеграції інформаційних ресурсів Великих даних для процесів управління регіоном. У результаті виконання цієї роботи одержано такі результати.

1. Проаналізовано методи, моделі та засоби опрацювання різнотипних даних, інформаційних технологій роботи з Великими даними. Обгрунтовано актуальність розв'язання завдання організації та інтеграції інформаційних ресурсів Великих даних у системах територіального управління.
2. Розроблено інформаційну модель Великих даних «сутність-характеристика», яка дає змогу організовувати структуровані та слабоструктуровані дані і на відміну від багатовимірної моделі не містить надлишковості.
3. Розроблено методи перетворення реляційних та слабоструктурованих даних у дані, подані в моделі «сутність-характеристика», що дало змогу уніфікувати форму запитів до різних моделей даних.
4. Розроблено федеративний метод інтеграції даних через визначення пари «сутність-характеристика» та узгодження семантики, що на відміну від методів інтеграції даних на рівні сховища даних дозволило інтегрувати дані з джерел з наперед невідомою структурою даних без попереднього локального завантаження, і що, в свою чергу, дало змогу підвищити ефективність подальшого аналізу Великих даних. Цей метод став основою для розроблення методу формування відповіді на запит користувача.
5. Розроблено метод формування відповіді на запит користувача до Великих даних шляхом уніфікації елементів мов запитів до різних інформаційних джерел, що дало змогу трансформувати запит до Великих даних у запит на основі ключових слів.
6. Розроблено програмні компоненти інформаційної системи підтримки прийняття рішень «Інтегратор» для управління регіоном з використанням Великих даних.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Болюбаш Ю. Я. Методи та засоби опрацювання великих даних у системах територіального управління / Ю. Я. Болюбаш // Науковий вісник Національного лісотехнічного університету: збірник наукових праць. – Львів : РВВ НЛТУ України. – 2016. – Вип. 26.4. – С. 341–354.
2. Болюбаш Ю. Я. Методи опрацювання Великих даних у федеративному сховищі даних / Ю. Я. Болюбаш // Вісник Національного університету «Львівська політехніка». – 2016. – № 843: Комп'ютерні науки та інформаційні технології. – С. 356–365.
3. Big Data information technology and data space architecture / N. Shakhovska, O. Veres, Y. Bolubash, L. Bychkovska // Sensors & Transducers. – 2015. – Vol. 195, Is.12. – P. 69–77.
4. Шаховська Н. Б. Опрацювання невизначеності у великих даних / Н. Б. Шаховська, Ю. Я. Болюбаш // Радіоелектроніка, інформатика, управління. – 2014. – № 1. – С. 96–105.
5. Шаховська Н. Б. Організація великих даних у розподіленому середовищі / Н. Б. Шаховська, Ю. Я. Болюбаш, О. М. Верес // Наукові праці Донецького національного технічного університету. Серія: Обчислювальна техніка та автоматизація. – 2014. –№ 2. – С. 147–155.
6. Шаховська Н. Б. Робота з великими даними – показниками соціо-еколого-економічного розвитку регіону / Н. Б. Шаховська, Ю. Я. Болюбаш // Восточно-Европейский журнал передовых технологий. – 2013. – № 5(2). – С. 4–8.
7. Шаховська Н. Б. Модель Великих даних «сутність-характеристика» / Шаховська Н. Б., Болюбаш Ю. Я. // Вісник Національного університету «Львівська політехніка». – 2015. – № 814 : Інформаційні системи та мережі. – С. 186–196.
8. Shakhovska N. Dataspace architecture and manage its components class projection / N. Shakhovska, Y. Bolubash // Econtechmod / Polish Academy of Sciences, Branch in Lublin. – 2015. – Vol. 4, №. 1. – P. 89–97.
9. Shakhovska N. Big Data model «entity and characteristics» / N. Shakhovska, Y. Bolubash // Econtechmod / Polish Academy of Sciences, Branch in Lublin. – 2015. – Vol. 4, No. 2. – P. 51–58.
10. Шаховська Н. Б. Методи прогнозування соціо-еколого-економічного розвитку регіону Західного Бугу / Н. Б. Шаховська, Ю. Я. Болюбаш // Матеріали IX Міжнародної науково-практичної конференції «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту» (ISDMCI), 20-24 травня 2013 р., Євпаторія. – Херсон, 2013. – С. 324–326.
11. Шаховська Н. Б. Федеративне сховище даних для Big Data / Н. Б. Шаховська, Ю. Я. Болюбаш // Математика. Інформаційні технології. Освіта: матеріали Міжнародної науково-практичної конференції, 6-8 червня 2014 р., Луцьк; Світязь. – Луцьк, 2014. – С. 87–88.
12. Шаховська Н. Б. Проектування федеративного сховища даних на основі Великих даних / Н. Б. Шаховська, Ю. Я. Болюбаш // Матеріали III Міжна-

родної науково-практичної конференції Інформаційні управляючі системи та технології (ІУСТ), 23-25 вересня 2014 р., Одеса. – Одеса, 2014. – С. 293–296.
13. Шаховська Н. Б. Робота з великими даними – показниками соціо-еколого-економічного розвитку регіону / Н. Б. Шаховська, Ю. Я. Болюбаш // Математика. Інформаційні технології. Освіта: матеріали ІІ Міжнародної науково-практичної конференції, 3-5 червня 2013 р., Луцьк-Світязь. – Луцьк, 2013. – С. 41–42.
14. Data Space architecture for Big Data managering / Shakhovska Natalya, Veres Oleh, Bolubash Yuri, Liliana Bychkovska // Proceedings of the International Conference on Computer Sciences and Information Technologies, 14-17 September 2015, Lviv. – Lviv, 2015. – P. 184–187.
15. Shakhovska N. B. Big Data federated repository model / N. B. Shakhovska, Y. J. Bolubash, O. M. Veres // Proceedings of 13th International Conference: The Experience of Designing and Application of CAD Systems in Microelectronics, CADSM, 24-27 February 2015, Lviv. – Lviv, 2015. – P. 382–384.

АНОТАЦІЇ

Болюбаш Ю. Я. Методи та засоби опрацювання інформаційних ресурсів Великих даних в системах територіального управління. – На правах рукопису.

Дисертаційна робота на здобуття наукового ступеня кандидата технічних наук за спеціальністю 01.05.03 – математичне та програмне забезпечення обчислювальних машин і систем. – Львів: Національний університет «Львівська політехніка» Міністерства освіти і науки України, 2017.

У дисертаційній роботі розв'язано важливе наукове завдання організації та інтеграції інформаційних ресурсів Великих даних у системах територіального управління.

Подано визначення Великих даних та описано основні характеристики. Проаналізовано математичні засоби подання та опрацювання Великих даних та визначено їх обмеження. Подано програмні засоби роботи з Великими даними. Визначено формальний опис Великих даних. Подано моделі асоціацій між сутностями та характеристиками для різних категорій NoSQL баз даних. Використано простір даних для представлення Великих даних. Подано модель федеративного сховища Великих даних. Для представлення Великих даних використано простір даних, який дає змогу працювати з різнотипними даними. Проте основною операцією інтеграції є не консолідація, а федералізація, що дає змогу зменшувати ємнісну складність запитів. Розроблено метод обміну різнотипними даними та приведення реляційних даних до моделі «сутність-характеристика». Розроблено метод відповіді на запит користувача з урахуванням типу даних джерела. Спроектовано схему даних регіону. Апробовано розроблені методи і алгоритми.

Ключові слова: інформаційні ресурси, Великі дані, простір даних, прогнозування методом ковзного середнього, система територіального управління.

Болюбаш Ю. Я. Методы и средства обработки информационных ресурсов Больших данных в системах территориального управления. – На правах рукописи.

Диссертация на соискание ученой степени кандидата технических наук по специальности 01.05.03 – математическое и программное обеспечение вычислительных машин и систем. – Львов: Национальный университет «Львивська политэхника» Министерства образования и науки Украины, 2017.

В диссертационной работе решено важное научное задание организации и интеграции информационных ресурсов Больших данных для выявления и исследования основных трендов прогнозирования экологических и экономических процессов региона.

Дано определение Больших данных и описаны основные характеристики. Проанализированы математические средства представления и обработки Больших данных и определены их ограничения. Приведены программные средства работы с Большими данными. Определено формальное описание Больших данных. Представлено модели ассоциаций между сущностями и характеристиками для различных категорий NoSQL баз данных. Использовано пространство данных для представления Больших данных. Представлена модель федеративного хранилища Больших данных. Для представления Больших данных использовано пространство данных, которое позволяет работать с разнотипными данными. Однако основной операцией интеграции является не консолидация, а федерализация, что позволяет уменьшать емкостную сложность запросов. Разработан метод обмена разнотипными данными и приведение реляционных данных в модели «сущность-характеристика». Разработан метод обмена разнотипными данными и приведения реляционных данных к модели «сущность-характеристика». Разработан метод ответа на запрос пользователя с учетом типа данных источника. Спроектирована схема данных региона. Апробированы разработанные методы и алгоритмы.

Ключевые слова: информационные ресурсы, Большие данные, пространство данных, прогнозирования методом скользящего среднего, система территориального управления.

Boliubash Y. Methods and tools of Bid data information resources processing in territorial administration systems. – On the rights of manuscript.

The thesis for the degree of candidate of technical sciences, specialty 01.05.03 – mathematical and software of computers and systems. – Lviv Polytechnic National University, Ministry of Education and Science of Ukraine, Lviv, 2017.

The dissertation solved an important scientific task of organizing and integrating Big data information resources and identifying the main trends of forecasting environmental and economic processes in the region.

In the first chapter the main definitions of Big data are given and the main characteristics are described. There are analyzed mathematical means of submission and processing of Big data. Their limitations are defined. There are posted software for Big data processing and analysis.

In the second chapter the formal description of Big data is defined. The model «entity-characteristics» for Big data representation is given. The method for data transformation from relation model to model of «entity-characteristics» is developed. There are posted patterns of associations between entities and characteristics for various categories of NoSQL databases. The data space is chosen as model for Big data representation. The federated repository Big model data is built. To represent the data there is used data space, which allows you to work with heterogeneous data. However, the main operation is federalisation that, allowing capacitive reduce the complexity of requests.

In the third chapter the method of heterogeneous data sharing and bringing to the relational data model «entity-characterization» is developed. The method of heterogeneous data sharing and bringing to relational data model «entity-characterization» is given.

In the fourth chapter there is designed the scheme of data region and architecture of Big data. The software for the development of methods and algorithms testing is created.

To achieve the desired goal, the development of a formal model of the Big Data information technology is made and its structural elements are described.

There is projected Big data architecture and instrumentation tools for practical realization. There are chased program tools for variant data integration realization. The realization specification is described. There are described language tools and user interface realization.

The annexes are presented acts of implementing the results of the thesis and the code of developing of information system.

Keywords: information resources, Big data, data space, forecasting by the moving average, system of territory management.

Здано у видавництво 01.03.2017 р.
Формат 60x84/16. Папір офсетний. Друк цифровий
Умовн. друк. арк. 0,9.
Наклад 120 прим.

Надруковано ТзОВ «Графік Стар»
м. Львів вул. В. Великого, 2
email: soroka@soroka.lviv.ua
тел. 244 28 37