УДК 004.

Н.Б. Шаховська (докт. техн. наук), Ю.Я. Болюбаш, О.М. Верес (канд. техн. наук)
Національний університет «Львівська політехніка», м. Львів,
кафедра інформаційних систем та мереж,
E-mail: natalya233@gmail.com

ОРГАНІЗАЦІЯ ВЕЛИКИХ ДАНИХ У РОЗПОДІЛЕНОМУ СЕРЕДОВИЩІ

У статті уведено поняття терміну Великі дані та проаналізовано причину їх появи. Визначено причини використання NoSQL та інших нереляційних засобів зберігання даних. Описано простір даних як технологію роботи з Великими даними.
Ключові слова: *великі дані, інформаційний продукт, простір даних.*

Постановка проблеми

Опрацювання інформаційних ресурсів, що використовують різні моделі даних, схеми керування тощо вимагає розроблення уніфікованого методу доступу до них для того, щоб надати можливість користувачу вибирати адекватний інструментарій для вивчення та використання різних засобів опрацювання даних. Необхідність у цьому виникає в організацій, робота яких полягає в опрацюванні великої кількості різнотипних, взаємонезалежних джерел даних, для яких не всі семантичні взаємозв'язки відомі і вказані. У деяких випадках семантичні зв'язки невідомі через невизначену кількість початкових джерел або через брак кваліфікованих людей у визначенні таких зв'язків. У інших випадках, не всі семантичні зв'язки необхідні для класифікації послуг користувачам. Тому в користувачів немає єдиної схеми, за якою вони можуть створювати запити відносно цільових задач.

Внаслідок керування різнотипними даними з метою розв'язання аналітичних задач стратегічного рівня виникає задача якості даних – відповідності вимогам користувачів. На рівні задач, для яких використовується точкове джерело, якість даних цього джерела є достатньою, і задовольняє (повністю чи частково) потреби осіб, що приймають рішення на їх основі. Проте використання даних з декількох джерел, наперед неузгоджених та з невідомими структурами, призводить до того, що якість даних різко знижується і вже не може задовольняти потреб користувача через неузгодженість форматів, різне подання, необхідне для вирішення проблеми.

За усієї важливості відомих результатів, теоретичні та експериментальні дослідження повинні розвиватися в напрямку: розроблення ефективних засобів опрацювання даних з різнотипних інформаційних ресурсів та вироблення засад і критеріїв оцінювання якості інтегрованих даних, які би підвищували ефективність прийнятих рішень.

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. Є альтернативою традиційним системам управління базами даних і рішеннями класу Business Intelligence. До цього класу відносять засоби паралельного опрацювання даних (NoSQL, алгоритми MapReduce, Hadoop) [1, 2].

Визначальними характеристиками для Великих даних є обсяг (volume, в сенсі величини фізичного обсягу), швидкість (velocity в сенсах як швидкості приросту, так і необхідності високошвидкісної обробки та отримання результатів), різноманіття (variety, в сенсі можливості одночасної обробки різних типів структурованих і напівструктурованих даних).

З одного боку, через свою неоднорідність і постійного зростання Big Data вимагають до себе нестандартних підходів у зберіганні та опрацюванні. Для ефективної роботи необхідні комплексні рішення моніторингу, фільтрації, структурування та пошуку ієрархічних зв'язків. З іншого - використовуючи Big Data, можна спостерігати за величезною множиною змінних, і на основі наданої інформації виявляти глобальні тренди і висновки, розглядаючи певну ситуацію в перспективі.

Модель федеративного сховища Великих даних

Для технології Великі дані необхідних є опрацювання інформації з різних за виразною потужністю типів джерел інформації: структурованих, напів-структурованих, 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).

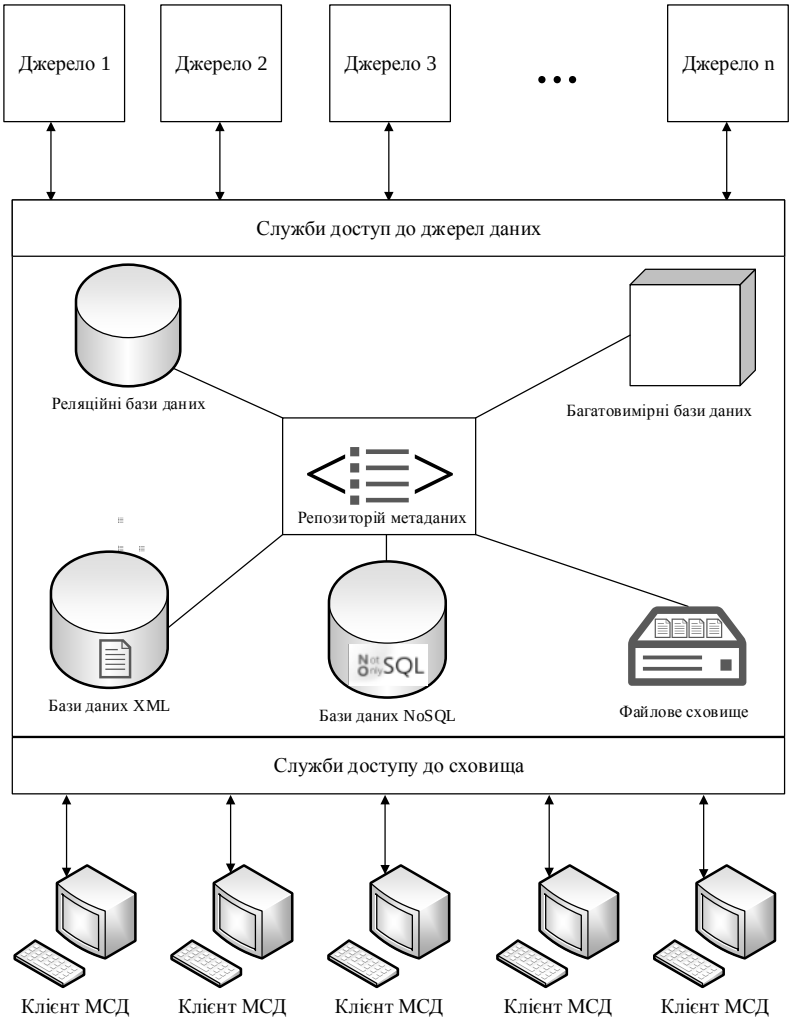


Рисунок 1 - Архітектура федеративного сховища даних

Основними характерними властивостями, які відрізняють федеративні сховища даних для Великих даних від інших сховищ даних, є наступні:

- наявність своєї системи керування сховищем даних, за допомогою якої здійснюється робота з сховищем (виконання запитів до сховища);
- наявність у сховищі реляційної БД, основними призначенням якої є зберігання структурованих даних і даних, до яких здійснюється частий доступ;

- наявність у сховищі багатовимірної БД, яка може містити як атомарні, так і узагальнені дані, основним призначенням багатовимірної бази даних є зберігання даних, до яких виконуються складні запити;

- наявність у сховищі бази даних XML та баз даних NoSQL, основним призначенням якої є зберігання слабко-структурованих даних і слабко-структурованої частини частково-структурованих даних;

- збереження неструктурованих даних у вигляді файлів, що зберігаються безпосередньо у файловій системі;

- взаємодія з джерелами даних, що здійснюється за допомогою інтегратора, та 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 сховища даних, який дає змогу користувачам звертатися до даних за допомогою єдиного інтерфейсу, незалежно від фізичного та логічного розташування цих даних у сховищі.

Необхідність наявності баз даних NoSQL виникає через:

- проблеми з розширення РБД, коли набір даних занадто великий;

- СУБД не були призначені для дистрибуції;

- поширення набули бази даних багатовузлових рішень;

- необхідність у масштабуванні або горизонтальному масштабуванні.

Іншими шляхами скалювання РБД є:

- мульти-майстер реплікації - система реплікації з декількома майстрами відповідає за поширення зміни даних, зроблені кожним учасником в решті частини групи, і вирішення конфліктів, які можуть виникнути через одночасні зміни, зроблені різними членами ;

- дані тільки додаються, а не оновлюються та знищуються;

- зменшення використання операції з'єднання (Join), тим самим зменшуючи час запиту (включаючи в тому числі денормалізовані дані, які призводять до появи ще більших баз даних);

- баз даних в пам'яті (in-memory).

Серез причин появи NoSQL треба виділити:

- вибух соціальних мереж (Facebook, Twitter) з великими потребами в даних;

- появу хмарних рішень, таких як Amazon S3 (рішення простого зберігання);

- використання динамічно типізованих мов (Ruby / Groovy), зрушення в динамічно типізованих даних з частими змінами схеми;

- формування спільнот, що користуються та розробляють програмне забезпечення з відкритим вихідним кодом.

При проектуванні федеративного сховища даних Великих даних для забезпечення його оптимального функціонування пропонується поєднати підходи проектування на базі структурованості джерел даних і запитів до сховища даних. В основі цього лежить CAP-теорема. За цією теоремою є три властивості системи: узгодженість, доступність і подільність. Можна мати не більше двох з цих трьох властивостей для будь-якої системи з розділюваними даними. Проте, щоб масштабувати, необхідно розбити на розділи, що у свою чергу, призводить до втрати узгодженості та доступності.

Нехай N - кількість вузлів з реплікою даних, W - кількість вузлів, які мають підтвердити оновлення, R - мінімальна кількість вузлів, які встигають здійснювати операцію читання.
Тоді для $W + R > N$ виконується вимога суворої узгодженості.

Для задоволення вимог CAP-теореми спроектуємо інформаційну структура Великих даних (рис. 2).

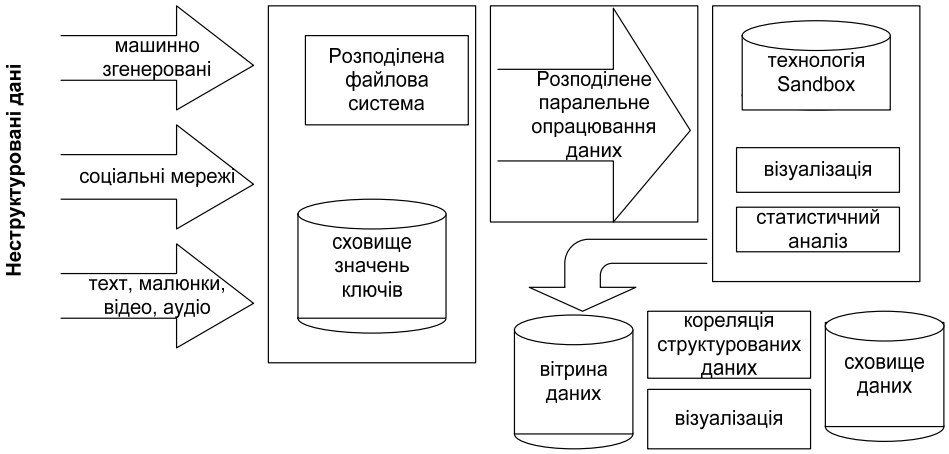


Рисунок 2 - Інформаційна структура Великих даних

Для роботи з Великими даними передбачається чотири фази перетворення даних:

1. Набуття – робота з фіксованими або придбаними з використанням розподілених файлових систем даними (Hadoop Distributed File Systems [HDFSs]) і базами даних NoSQL (Oracle NoSQL баз даних).

2. Організація - парадигма програмування MapReduce використовується для інтерпретації й уточнення даних.

3. Аналіз - очищені і організовані дані подаються в реляційну базу даних (бази даних SQL), щоб здійснити належний аналіз.

4.Підтримка рішень – використання методів підтримки прийняття рішень для подальшого аналізу даних.

Ці чотири фази обробки даних не можуть здійснюватись з використанням однієї машини.

Вважається, що Великі дані є технологією Hadoop. Проте Великі дані є дещо більше, ніж Hadoop. Одною з ключових вимог є розуміння і навігація по федеративних джерелах великих даних - щоб виявити дані на місці. Нова технологія підтримує також індекси, пошук та навігацію різних джерел Великих даних. Hadoop являє собою набір можливостей з відкритим вихідним кодом. Два з найбільш відомих з них є: Hadoop FS для зберігання різноманітної інформації, MapReduce - двигун паралельної обробки. Сховища даних також керують Великими даними, оскільки обсяг структурованих даних швидко зростає. Можливість запуску глибоких аналітичних запитів на величезних обсягах структурованих даних є великою проблемою даних. Це вимагає наявності величезних сховищ даних паралельної обробки і спеціально побудованих засобів для аналізу даних. Великий дані містять не тільки статичні, але й динамічні дані. Потокові дані представляють абсолютно іншу проблему Великих даних - здатність швидко аналізувати і діяти у відповідності з даними в той час як вони ще прибувають.

Однією з технологій, що доцільно використовувати для роботи з Великими даними, є простір даних.

Простір даних – це блоковий вектор, що містить множину інформаційних продуктів предметної області, поділену на три блоки: структуровані дані (бази, сховища даних), напівструктуровані дані (XML, електронні таблиці) та неструктуровані дані (текст). Над цим вектором та його окремими елементами визначено операції та предикати, які забезпечують:

- перетворення різних елементів вектора один в одного;
- об'єднання елементів одного типу;
- пошук в елементах за ключовим словом.

Моделі даних, що підтримуються у просторі даних, утворюватимуть ієрархію відповідно до їх виразної потужності:

- реляційна;
- багатовимірна;
- об'єктно-реляційна моделі;
- об'єктно-орієнтована модель;
- розширена мова розмітки інформації (XML) зі схемою;
- середовище описання ресурсів (Resource Description Framework – RDF);
- стандартний засіб описання зв'язків між об'єктами даних – онтології, описані за допомогою Web Ontology Language – OWL;
- структурований текст (у тому числі HTML);
- напівструктурований текст.

Придатність моделей даних до підтримання мов запитів та до використання в глобальній мережі подано на рис. 3.

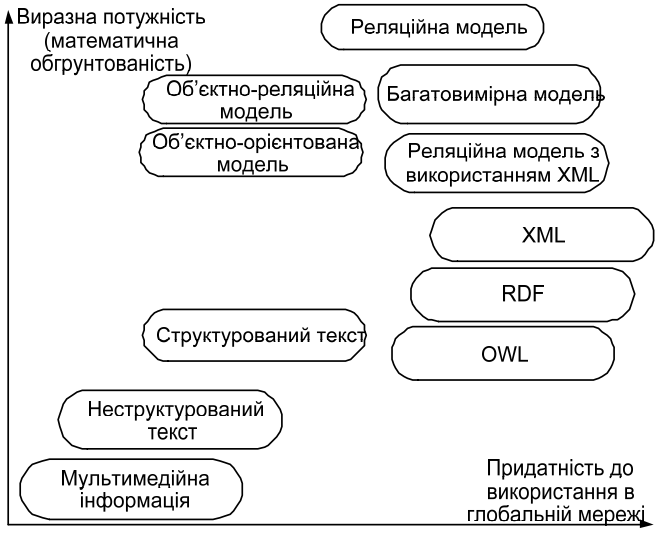


Рисунок 3 - Придатність моделей даних до підтримання мов запитів та до використання в глобальній мережі

Кожен учасник простору даних підтримує деяку модель даних і деяку мову запитів, відповідну цій моделі. Запит до такого програмного засобу відповідає тому, що зазвичай підтримується у файлових системах стосовно до їх директорій: зіставлення імен, пошук в діапазоні дат, сортування за розміром файлу та ін. На наступному рівні простору даних модель даних повинна підтримувати мультимножини слів з метою здійснення ефективного пошуку необхідної інформації за ключовими словами, внаслідок чого отримаємо певну можливість бачення вмісту учасників простору даних. Нижче рівня моделі мультимножини слів в ієрархії може розташовуватися модель напівструктурованих даних, заснована на позначених графах. Оскільки джерела даних є різнотипні, то необхідно визначити платформу та архітектуру ПД.

Платформа підтримання ПД (ППД) – це набір програмного забезпечення, що керує організацією, зберіганням і пошуком даних у просторі. Також здійснює контроль безпеки та цілісності (рис. 4.).

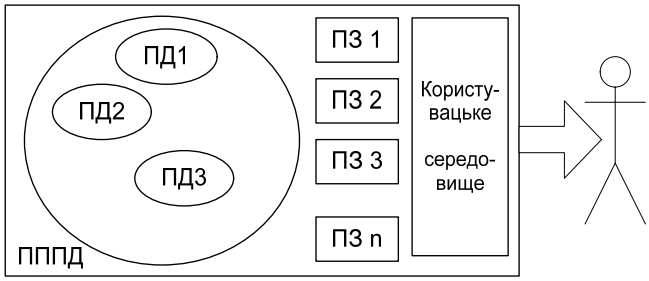


Рисунок 4 - Платформа підтримання простору даних

Архітектуру простору даних спроектуємо за рівнями (рис. 5). Рівень застосувань призначений для реалізації операцій над даними у просторі даних. Рівень онтологій використовується для встановлення зв'язку між джерелами. Останній рівень містить джерела даних та забезпечує доступ до даних та виконання операцій рівня застосувань безпосередньо в джерелі (наприклад, операція вибірки на рівні реалізації виконується як запит в конкретній базі даних).

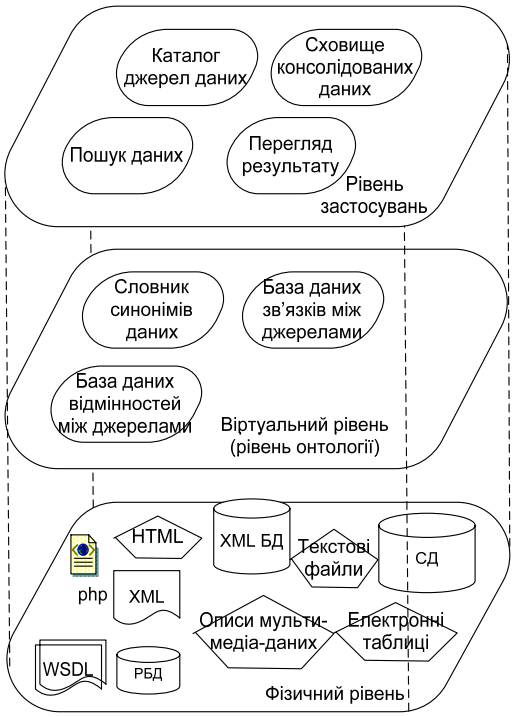


Рисунок 5 - Рівні реалізації фізичної моделі простору даних

Формалізація каталогу простору даних

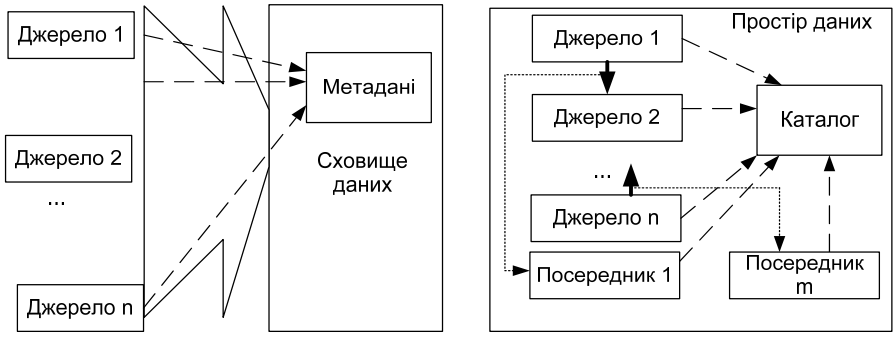
Каталог *Cg* – це реєстр ресурсів даних, що містить базову інформацію про кожного з них: джерело, ім'я, місцезнаходження в джерелі, розмір, дату створення і власника, технологічну платформу, протоколи та режим доступу, частоту оновлення та ін. Формується на основі метаописань джерел даних:

$$\text{Metadata}(\mathbf{DB}, \mathbf{DW}, \mathbf{Wb}, \mathbf{Nd}, \mathbf{Gr}) \Rightarrow Cg.$$

Він не лише містить описову інформацію (тобто виконує роль метаданих), але й зберігає для кожного учасника схему джерела, статистичні дані, швидкість зміни, точність, можливості відповідей на запити, інформацію про власника і дані, про політику доступу і підтримання конфіденційності.

Оскільки джерела простору даних фізично не переносять у нього інформацію та не можуть обмінюватись між собою інформацією, то у каталозі необхідно зберігати дані і про зв'язки між джерелами.

Зазначимо, що поняття каталогу простору даних ширше, ніж поняття метаданих простору даних. Окрім традиційних за Захманом 6-ти типів метаданих, каталог простору даних зберігає ще інформацію про зв'язки між джерелами і протоколи обміну інформацією між ними. Відмінності між поданням джерел даних у метаданих та каталозі схематично подані на рис. 6.



а) б)

Рисунок 6 - Подання даних у а) метаданих у сховищі даних; б) каталозі простору даних

Зв'язки у каталозі можуть зберігатися у вигляді метаданих, перетворень запитів, графів залежності або текстових описань тощо. Зв'язок між елементами простору даних поданий на рис. 7.

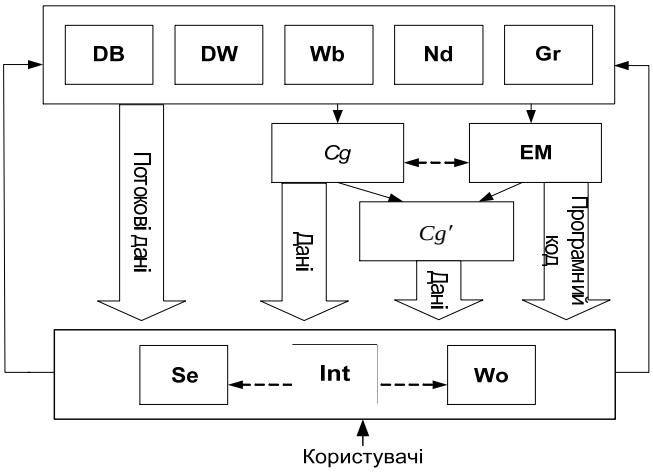


Рисунок 7 - Схема елементів простору даних

Над каталогом розміщене середовище керування моделями **EM**, яке дає змогу створювати нові зв'язки і маніпулювати наявними зв'язками (наприклад, об'єднувати або інвертувати відображення, зливати схеми і створювати єдині подання декількох джерел).

Важливою компонентою простору даних є федералізоване сховище даних (Cg'), яке слугує для досягнення наступних цілей:

- створення асоціацій між об'єктами даних від різних учасників;
- вдосконалення доступу до джерел з обмеженими власними засобами доступу;
- забезпечення можливості виконання деяких запитів без доступу до реального джерела даних;

- консолідації даних як результат запиту користувача;
- підтримання високого рівня доступності і відновлення.

Отже, зв'язок між каталогом *Cg*, середовищем керування моделями **EM** і федералізованим сховищем даних *Cg′* можна подати як відображення:

$$\mathbf{EM}(Cg) \Rightarrow Cg'.$$

Що більше моделей здатне «розрізнити» середовище керування, то точнішою буде інформація в *Cg′* і ефективніше можна буде здійснювати процедури інтеграції, пошуку та опрацювання даних у просторі даних **DS**.

Висновки

1. Введено інформаційну модель Великих даних та визначено складові федералізованого сховища даних.
2. Показано, що простір даних доцільно використовувати для роботи з Великими даними.
3. Описано структуру простору даних.

Список використаної літератури

1. Шаховська Н.Б. Формальне подання простору даних у вигляді алгебраїчної системи / Шаховська Н.Б. // Системні дослідження та інформаційні технології / Національна академія наук України, Інститут прикладного системного аналізу. – Київ, 2011. – № 2. – С.128 – 140.
2. Шаховська Н.Б., Болюбаш Ю.Я. Аналіз методів опрацювання показників соціо-еколого-економічного розвитку регіону// Східно-європейський журнал передових технологій, Том 5, № 2(65), 2013. – С. 4-8.
3. Згуровський М.З. Основи системного аналізу / Згуровський М.З., Панкратова Н.Д.. - К.: Видавнича група BHV, 2007. - 544 с.

References

1. Shakhovska, N.B. (2011) "A formal representation of the data space in the form of algebraic system", *System Research and Information Technologies*, no. 2, pp. 128–140.
2. Shakhovska, N.B. and Bolubash, Yu.Ja. (2013) "Analis metodiv opratsuvannia pokaznykiv sotsio-ekologo-ekonomichnogo rozvytku regionu", *Shidno-yevropeyskij zhurnal peredovyh tehnologij*, vol. 5, no. 2(65), pp. 4-8.
3. Zgurovskyj, M.Z. and Pankratova N.D. (2007), *Osnivy systemnogo analizu*, BHV, Kiev, Ukraine.

Надійшла до редакції:
20.03.2014 р.

Рецензент:
канд. техн. наук, проф. Турупалов В.В.

Н.Б. Шаховська , Ю.Я.Болюбаш, О.М. Верес
Национальный университет «Львовская политехника»

Организация Больших данных в распределенной среде. *В статье введено понятие срока Большие данные и проанализированы причину их появления. Подано информационную модель федеративного хранилища данных и описаны его составные элементы. Определены особенности использования NoSQL и других нереляционных средств хранения данных. Описаны пространство данных как технологию работы с Большими данными. Описаны уровни физической модели пространства данных.*

Ключевые слова*: большие данные, информационный продукт, пространство данных.*

N.B.Shakhovska, Yu.Ja.Bolubash, O.M. Veres
Lviv Polytechnic National University

Big data organizing in a distributed environment. *This paper introduced the concept of the term Big Data and analyzes the cause of their appearance. Big Data is a set of methods and tools for processing different types of structured and unstructured data dynamic large amounts for their analysis and use of decision support . There is an alternative to traditional database management systems and solutions class Business Intelligence. To this class belong the parallel data processing means (NoSQL, algorithms MapReduce, Hadoop). Defining characteristic for Big data is the amount (volume, in terms of volume size), speed (velocity in terms of both growth rate and the need for high-speed processing and the results), diversity (variety, in terms of the possibility of simultaneous processing of different types of structured and semi-structured data). One of the technologies that should be used for large data region is the data space available. Data space is a block vector containing a set of information products subject divided into three categories: structured data (databases, data warehouses), semi-structured data (XML, spreadsheets) and unstructured data (text). Above this vector and its individual elements there are defined operations and predicates. There is posted the federated information model describes the data warehouse and its components. The features use non-relational NoSQL and other means of storage are described. We describe the data space as the technology of working with large data. The levels of physical model data space are given.*

Keywords*: big data, information products, data space.*

Шаховська Наталія Богданівна, Україна, закінчила Національний університет «Львівська політехніка», докт. техн. наук, доцент, професор кафедри інформаційних систем та мереж, декан базової вищої освіти Інституту комп'ютерних наук та інформаційних технологій Національного університету «Львівська політехніка» (вул. С.Бандери, 12, м. Львів, 79013, Україна). Основні напрями наукової діяльності – бази, сховища даних, моделювання інформаційних систем, розподілені системи, хмарні технології.

Болюбаш Юрій Ярославович, Україна, закінчив Тернопільський державний педагогічний університет, здобувач Національного університету «Львівська політехніка», заступник директора з навчальної роботи Золочівського коледжу. Основні напрями наукової діяльності – бази даних, Великі дані, методи прогнозування.

Верес Олег Михайлович, Україна, канд. техн. наук, доцент кафедри «Інформаційні системи та мережі» Інституту комп'ютерних наук та інформаційних технологій Національного університету «Львівська політехніка». Основні напрями наукової діяльності – інформаційне моделювання, системи баз даних та знань, інтелектуальні системи підтримання прийняття рішень, методи прийняття рішень в слабко структурованому середовищі, системний аналіз, моделювання складних систем.