УДК 528.852

О.В. Зайцев, к.т.н., доцент
Воєнно-дипломатична академія імені Євгенія Березняка, м.Київ

СТРУКТУРНО-ФУНКЦІОНАЛЬНА МОДЕЛЬ СИСТЕМИ ІНТЕГРУВАННЯ ІНФОРМАЦІЇ НА ОСНОВІ ТЕХНОЛОГІЙ BIG DATA

Досліджено пропозиції ринку сучасних технологій Big Data для оброблення та аналізу даних великих обсягів. Запропоновано архітектуру системи оброблення та аналізу інформації на основі технологій Big Data. Запропоновано варіант структурно-функціональної моделі системи інтегрування інформації для визначення її оптимального складу залежно від потреби споживачів та пропозицій на основі технологій Big Data.

Ключові слова: інформаційні потоки, інтеграція інформації, технології Big Data, структурно-функціональний аналіз.

Постановка проблеми в загальному вигляді. Сучасний стан інформаційної діяльності більшості крупних організацій визначається постійним збільшенням обсягів інформаційних потоків, енергетичного та частотно-часового простору та залученням різнорідних джерел інформації [1-3]. Серед проблем комплексного використання даних в єдиному інформаційному просторі слід відзначити неможливість або економічну недоцільність зберігання та реєстрації всього обсягу потоків даних, низьку інформативна ємність окремих потоків даних, складність зберігання великого обсягу не структурованих даних, довгу тривалість їх зберігання, проблеми пошуку, розповсюдження, передачі та візуалізації даних. За допомогою традиційних технологій оброблення та управління базами даних вже не можна вирішити ці проблеми.

Прикладами великих обсягів даних, які можуть дуже швидко накопичуватися, але при цьому інформаційна щільність яких низька [3, 4], можуть бути протоколи роботи користувачів інформаційно-телекомунікаційних систем, дані телеметрії, датчиків, слабо структуровані і неструктуровані дані, записи в соціальних мережах, веб-сайти, тощо.

Разом з тим зростають можливості цифрових мереж з доставки зафіксованої інформації до місць обробки, зростають обчислювальні можливості комп'ютерів для аналізу доступної інформації. Також останнім часом зростають можливості комп'ютерів щодо синтезу остаточних рішень. Прикладом такої системи є *IBM Watson (www.ibm.com/uk/WatsonAnalytics)*. Сукупність технологій, що стоять за цим, зараз умовно називають *великі дані* (англ. *Big Data*) [5].

Таким чином, постає протиріччя між об’єктивним збільшенням обсягів інформаційних потоків, зростанням вимог користувачів щодо якості

інформаційного забезпечення з одного боку та обмеженими можливостями наявних засобів зберігання, обробки та аналізу даних – з другого.

Аналіз останніх досліджень і публікацій. Питання інформаційного забезпечення розглянуто в [6, 7], в яких визначено підходи до підвищення якості функціонування інформаційних систем. Серед наукових досліджень зарубіжних вчених, слід відзначити публікації [8, 9], в яких висвітлено питання щодо типів, принципів та методів оброблення та використання великих обсягів даних. Незважаючи на певну методологічну розробленість зарубіжними вченими цієї проблеми, ряд аспектів залишається малодослідженим. Так, надзвичайно важливим є вирішення питання визначення раціонального складу системи на основі технологій *Big Data* для комплексної обробки та аналізу інформації.

Виходячи із зазначеного метою статті є аналіз концепції *великих даних* як нового підходу до збору, обробки та аналізу інформації, дослідження можливостей її практичного застосування під час її комплексної обробки та аналізу з різних джерел, розроблення структурно-функціональної моделі системи обробки інформації на основі технологій *Big Data*.

Виклад основного матеріалу. *Архітектура системи обробки інформації на основі технологій Big Data.* Базовою архітектурою для обробки великого масиву даних вважають в *SN*-архітектуру (англ. *Shared Nothing Architecture*) [10], яка забезпечує масивно-паралельну обробку даних на основі технології *NoSQL*. В основі архітектури лежать три основні принципи. По-перше, дані рівномірно розподіляються на внутрішніх дисках великої кількості серверів, об'єднаних єдиною файловою системою. По-друге, не дані передаються програмам для обробки, а програма передається до даних. Третій принцип - дані обробляються паралельно, причому цю можливість закладено архітектурно в програмному інтерфейсі.

Сьогодні програмні та апаратні рішення *Big Data* [12] дозволяють забезпечити підтримку інформаційної роботи підрозділів на всіх етапах. Для реалізації запропонованої архітектури на практиці вже розроблено досить багато апаратних та програмних рішень (рис. 1) як на відкритій, так і на комерційній основі. Розглянемо деякі з них докладніше.

Рішення з відкритим кодом. Сьогодні дедалі популярною стає модель роботи з *Big Data*, яку реалізовано в проекті *Apache Software Foundation – Apache Hadoop (hadoop.apache.org)*.

Apache Hadoop складається з двох компонентів: розподіленої кластерної системи *Hadoop Distributed File System (HDFS)* і програмного інтерфейсу *Map Reduce*. Також існують і інші технології з відкритим вихідним кодом (*open source*), які розробляються та підтримуються *Apache Software Foundation* і доповнюють *Hadoop*: мова програмування *R,* мова опису програм *Pig*, що аналізує великі набори даних; рішення для організації сховищ даних *Hive*; середовище зберігання даних *HBase*; служба переносу даних *Flume*;

бібліотека пошукової системи *Lucene*; технологія послідовного впорядкування даних *Avro*; служба синхронізації *ZooKeeper*; система планування потокової обробки завдань *Oozie*.

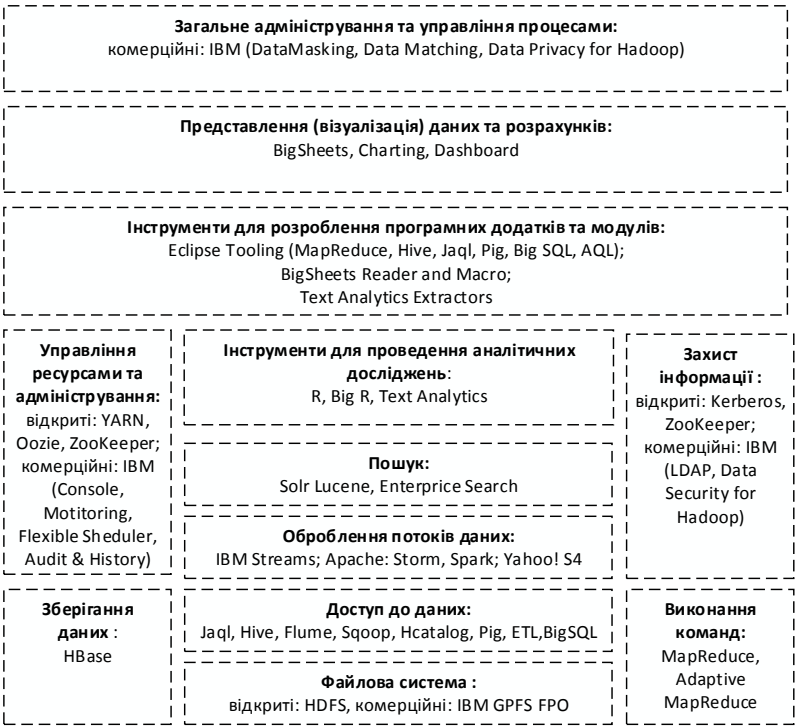

Рис. 1. Програмне забезпечення технології *Big Data*

Перевагами використання рішень з відкритим кодом є такі: можливість подальшого вдосконалення та адаптації програмних засобів; безкоштовність програмного забезпечення; наявність у вільному доступі додаткових програмних модулів для нарощування можливостей системи. Серед недоліків слід зазначити: обмежений базовий функціонал проекту; обмеженість технічної підтримки зі сторони розробника; необхідність самостійного вдосконалення та адаптації платформи під свої потреби.

Комерційні рішення Big Data представлено програмно-апаратними комплексами світових лідерів у сфері систем управління базами даних, систем статистичного аналізу, систем отримання та оброблення даних: *Cloudera, EMC, Oracle, IBM, Informatica, QlikView, Teradata*. Такі рішення поставляються як програмно-апаратні комплекси та містять кластери серверів і програмне забезпечення для масово-паралельної обробки. Використання

комерційних рішень звичайно прискорює процеси розгортання та впровадження технологій, проте потребує значних фінансових ресурсів.

У результаті аналізу наявних пропозицій ринку програмних та апаратних засобів *Big Data*, зважаючи на інформацію про джерела інформації, інформаційні потоки та етапи її обробки, пропонується варіант програмно-апаратної платформи на основі технологій *Big Data* (рис. 2), який має забезпечувати масивно-паралельну обробку та аналіз потоків великих за обсягами неструктурованих, частково структурованих та структурованих даних.

Рис. 2. Варіант програмно-апаратної платформи на основі технології *Big Data*

Структурно-функціональна модель системи обробки PI на основі технологій Big Data. Зважаючи на велику вартість створення систем *Big Data*, в яких використовуються розподілені в просторі різнорідні дані та потужні обчислювальні ресурси, під час розробки та експлуатації цієї системи доцільно використовувати методи структурно-функціонального аналізу [13]. Змістовне формулювання задачі структурно-функціонального аналізу зводиться до такого: необхідно визначити призначення, загальні характеристики та властивості системи, а також вимоги до структурних, функціональних, експлуатаційних та економічних показників.

Залежно від переліку завдань, що вирішуються в системі, можна визначити конкретні вимоги або властивості системи. Визначимо структуру системи на основі технологій *Big Data* як складну ієрархічну систему та

обґрунтуємо вимоги до кожного елемента інфраструктури та функціональних елементів на всіх ієрархічних рівнях.

Множину вимог до системи подамо у вигляді впорядкованої структури класів (1)

$$V_0 = \{V_i \mid i = 1, \ldots, k\}, \tag{1}$$

де V_0 — множина вимог до системи, V_i — *i*-й клас вимог до системи, k — кількість вимог.

Для системи побудованої на основі технологій *Big Data* визначимо таку множину класів:

V_1 — клас структурних властивостей (кількісні характеристики: кількість вузлів у системі, кількість процесорів, пропускна здатність каналів зв'язку, обсяг пам'яті для зберігання даних, обсяг оперативної пам'яті тощо; якісні характеристики: гетерогенність, топологія, інтерфейси, відповідність стандартам).

V_2 — клас функціональних властивостей, що визначає перелік функцій системи і кількісні показники їх виконання: підтримка запитів користувачів на виконання завдань, доступ, передачу і реплікацію даних; їх єдина реєстрація; балансування навантаження; збір статистики, прозорість, керованість, адаптованість.

V_3 — клас експлуатаційних властивостей, що визначає зручність експлуатації системи та відповідність системи зовнішнім умовам: наявність певного програмного забезпечення, зручний для користувача інтерфейс, вимоги до якості роботи сервісів (доступність, надійність, оперативність), можливість нарощувати елементи системи.

V_4 — клас властивостей безпеки: конфіденційність, цілісність та доступність інформації; можливість спостерігати за роботою системи, відповідність системи визначеним рівням гарантій якості.

V_5 — клас економічних властивостей визначає вимоги до вартості створення та експлуатації системи: вартість обладнання, вартість експлуатації, вартість послуг, що надаються системою.

Кожна властивість v_{ij} класу V_i визначається набором показників:

$$W_{ij} = \{w_{ijm} \mid m = 1, \ldots, k_{ij}\}, \tag{2}$$

де w_{ijm} — кількість показників для властивості v_{ij}, які можуть мати як кількісний, так і якісний вигляд. Вимоги до кількісних показників задаються за допомогою інтервальних оцінок у вигляді

$$\underline{w_{ijm}} < w_{ijm} < \overline{w_{ijm}} \tag{3}$$

Вимоги до якісних показників зручно задати за допомогою нечітких термів [14].

Формалізуємо описання системи на основі технологій *Big Data*, як

складної ієрархічної системи. Результати структурної декомпозиції системи подано на рис. 3.

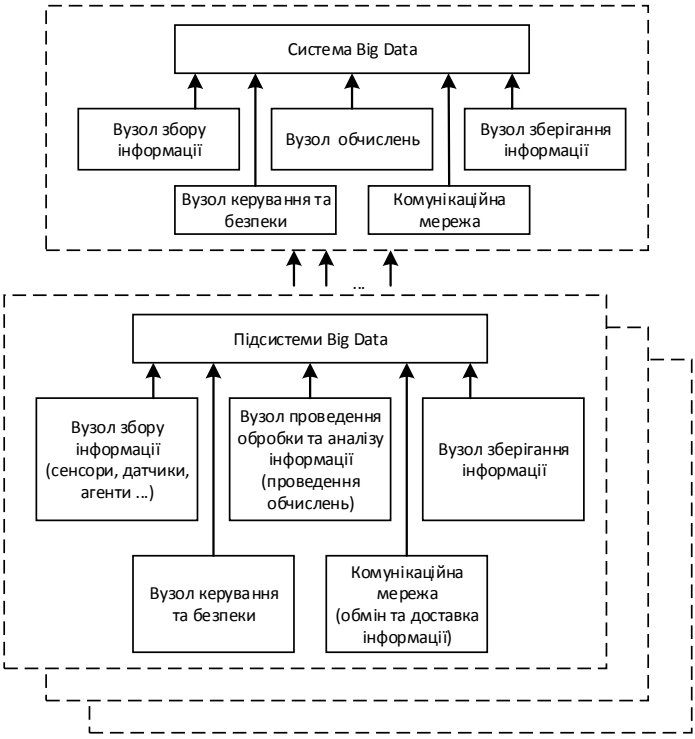

Рис. 3. Структурна модель системи на основі технологій *Big Data*.

Інтеграція підсистем Big Data в єдину систему може реалізовуватися за такими двома принципами.

1. Кожна підсистема є незалежним компонентом загальної системи, при цьому оброблення даних і підготовка інформаційного продукту в основному здійснюється безпосередньо власними обчислювальними ресурсами. Такий підхід доречно використовувати, якщо немає потреби обробляти великі обсяги даних.

2. Гібридний підхід, за якого для оброблення і зберігання даних спільно використовуються ресурси інших підсистем. При цьому використовується підхід до оброблення даних на основі сервісів.

Такий підхід вимагає реалізації системи узгодженого розподіленого керування системою в цілому, з використанням механізмів балансування навантаження на компоненти системи та розподіленого управління безпекою

інформації.

При цьому кожен *r*-й рівень ієрархії системи Big Data складається з F_r функціональних елементів. Тоді множину всіх функціональних елементів системи можна описати наступним чином:

$$L = \left\{ L_{rf} \mid r = \overline{1,R}, f = \overline{1,F_r} \right\} \tag{4}$$

де *R* — загальне число рівнів ієрархії, L_{rf} — *f-й* функціональний елемент *r*-ого ієрархічного рівня.

Кожний функціональний елемент L_{rf} системи характеризується наступним вектором показників:

$$p_{rf} = \left\{ p_{rfj} \mid j = 1, \ldots, n_{rf} \right\}, \tag{5}$$

де p_{rf} — кількість показників функціонального елементу L_{rf}, який виконує набір функцій

$$\Phi_{rf} = \left\{ u_{rfk} \mid k = 1, \ldots, m_{rf} \right\}. \tag{6}$$

Кожна з функцій u_{rfk} у (6) залежить від значень показників вектора p_{rf}:

$$u_{rfk} = u_{rfk}(p_{rf}) \tag{7}$$

і впливає на реалізацію вимог до системи в цілому. Згідно з [13] склад і вид функцій (6, 7) визначається в процесі системного аналізу.

Для простоти викладу відмовимося від подвійної індексації. Тоді узагальнений вектор показників можна подати у вигляді:

$$\mathrm{p} = (p_1, p_2, ..., p_{N_0},)^{\mathrm{T}}, \mathrm{p} \in \mathrm{R}^{N_0}, \tag{8}$$

де $P = \left\{ p_{rf}, r = 1, \ldots, R, f = 1, \ldots, F_r \right\}$ — множина показників функціональних елементів на всіх ієрархічних рівнях системи.

Одним із головних завдань структурно-функціонального аналізу є визначення перетворення

$$F: X \rightarrow Y, \tag{9}$$

з множини *X* допустимих показників системи в простір *Y* необхідних властивостей за набором кількісних вимог, заданих у вигляді (3), і якісних вимог, заданих у вигляді нечітких термів.

При цьому потрібно вирішити задачу структурно-параметричної ідентифікації, що дає змогу визначити структуру ієрархічної системи в цілому, структуру функціональних елементів усіх ієрархічних рівнів, а також вид перетворення (9).

Висновки

Для вирішення проблем збору, обробки та аналізу потоків даних великих обсягів, запропоновано використовувати технології *Big Data*. Запропоновано загальну архітектуру побудови системи збору, обробки та аналізу інформації та варіант програмно-апаратної платформи на основі технологій *Big Data.* Для оптимізації складу апаратного та програмного забезпечення системи відповідно до задач, що вирішуються, пропонується використати методи структурно-функціонального аналізу. У подальших дослідженнях планується всебічно дослідити запропоновану структурно-функціональну модель та порівняти отримані результати з аналогами.

1. *Кірпічніков Ю.А. Петрушен М.В., Андрощук О.В.* Аналіз поняття інтеграційної платформи та методів інтеграції даних інформаційних систем управління // Збірник наукових праць ЦВСД НУОУ ім. І.Черняховського, стаття. – 2017. – №2(60).
2. *Ziegler, P. and Dittrich, K.* 2007. Data Integration-Problems, Approaches and Perspectives. Database Technology Research Group, Department of Informatics, University of Zurich. [Електронний ресурс]. – Режим доступу: https://pdfs.semanticscholar.org/ac7c/ed257ae5598d4a6048ea9c182773d317126c.pdf.
3. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим / *В. М. Шенбергер, К. Кукьер;* пер. с англ. Инны Гайдюк. – М. : Манн, Иванов и Фербер, 2014. – 240 с.
4. *Lynch C.* How do your data grow? / C. Lynch // Nature. – 2008. – V. 455. № 7209. – P. 28–29.
5. *Dijcks Jean-Pierre.* Big Data for the Enterprise / Jean-Pierre Dijcks. // Oracle. – October 2011. [Електронний ресурс] – Режим доступу: http:// bigdatawithoracle– 521307.pdf.
6. *Берко А.Ю., Висоцька В.А.* Моделі та методи проектування інформаційних систем електронної контент-комерції / А. Ю. Берко, // Вісн. Нац. ун-ту "Львівська політехніка". Серія : Інформаційні системи та мережі. – 2008. – № 621. – С. 29–48.
7. *Берко А. Ю.* Структурно-семантична інтеграція даних на основі фактологічної реляційної моделі / А. Ю. Берко // Вісн. Нац. ун-ту "Львівська політехніка". Серія : Комп'ютерні науки та інформаційні технології. – 2010.– № 663. – С. 60–69.
8. *Arputhamary, B., Arockiam, L.*: Data integration in Big Data environment. Bonfring Int. J. Data Mining 5(1), 1–5 (2015). doi: 10.9756/BIJDM.8001 [Електронний ресурс]. – Режим доступу: http://www.journal.bonfring.org/papers/dm/volume5/BIJ-8001.pdf
9. *Jenn Webb, Tim O'Brien* Big Data Now // O'Reilly Media, Inc., 2014. – [Електронний ресурс]. – Режим доступу: http://www.oreilly.com/data/free/files/bigdatanow2013.pdf.
10. *Solutions Big Data* // IBM Inc., 2015. – [Електронний ресурс]. – Режим доступу: http://www-05.ibm.com/fr/events/netezzaDM_2012/ Solutions_Big_Data.pdf.
12. The Big Data Landscape // Creative Commons Inc., 2015. – [Електронний ресурс]. – Режим доступу: http://www.mammothdb.com/ the-big-data-landscape-by-mammothdb/
13. *Згуровский М.З., Панкратова Н.Д.* Системный анализ: проблемы, методология, приложения. – К.: Наукова думка, 2005. – 744 с.
14. *Zadeh L.* Fuzzy sets // Inform. Control. – 1965. – V.8. – 338 p.

Поступила 17.09.2018р.