УДК 004.6

К. О. Кірей, *к.п.н., доцент, доцент кафедри інженерії програмного забезпечення*
e-mail: kirey.kea@gmail.com, kateryna.kirey@chmnu.edu.ua
Чорноморський національний університет імені Петра Могили
вул. 68 Десантників, 10, м. Миколаїв, Україна

РОЗВИТОК І ТРАНСФОРМАЦІЯ ПОНЯТТЯ BIG DATA

У статті розглядається розвиток і трансформація поняття «Великі дані» (Big Data). Нині це поняття не має чіткого трактування. Теоретичне дослідження показало, що його розвиток доцільно розглядати в історичному контексті. Спочатку вважалося, що критерієм віднесення до категорії Big Data, перш за все, є потік даних, більший 100 Гбайт на день. Однак виявилося, що цього недостатньо, щоб однозначно віднести продукт до категорії Big Data. Згодом як такий критерій було запропоновано використовувати якісні ознаки – «певну кількість V»: Volume – значне збільшення обсягу даних у корпоративних системах; Variety – різноманітність форматів і структур даних; Velocity – швидкість отримання і обробки даних для задоволення запиту тощо. Однак такий підхід не розкриває це поняття повною мірою, а торкається окремих його аспектів. В інших дослідженнях можна побачити визначення цього поняття як неможливості обробки даних традиційними способами. Однак таке визначення можна вже вважати застарілим, адже технології обробки великих обсягів даних експерти вже не відносять до новітніх. Деякі фахівці пропонують взагалі відмовитися від цього поняття. Однак, зважаючи на його значне поширення, це не видається можливим. Розвиток інформаційних технологій змінив наше ставлення до даних та інформації, а це, у свою чергу, вплинуло на сутність поняття Big Data. Нині під цим скоріше розуміється особливий підхід – ідеологія обробки великих масивів «сирих» даних. Однак не виключено, що в майбутньому це поняття може зникнути як застаріле та неактуальне, адже великі обсяги даних і сучасні підходи до їх обробки стануть звичайними засобами, і не буде необхідності робити акцент на їх інноваційній складовій.

***Ключові слова**: великі дані, Big Data, інформаційні технології.*

Вступ. Бурхливий розвиток інформаційних технологій, зокрема мережевих, технологій накопичення та зберігання даних, поява нових електронних сервісів сприяють значному зростанню потоків та обсягів даних. Це привело до зміни парадигми накопичення даних. Самі дані стали цінним ресурсом, який можна застосовувати у різних галузях. Відбувся перехід від парадигми накопичення даних – «ми знаємо, які дані нам потрібні, й збираємо саме їх» до – «ми не знаємо, які дані нам потрібні, збираємо усі дані, що можливо, а потім з'ясуємо їх цінність» [4]. З'явилися програмні й апаратні засоби, інформаційні системи, методи й технології роботи з великими обсягами даних і поняття, що описує усі ці явища – Big Data. Вже є багато ресурсів, що переймаються проблемами дослідження, використання та стандартизації великих даних. Установи, що займаються стандартизацією, також працюють над проблемами ефективного спільного використання джерел даних. Рішення з відкритими вихідними кодами дають можливість досліджувати, вивчати й експериментувати з даними великого розміру. Таким даним вже присвячено багато наукових розробок, літератури, конференцій, форумів тощо. Серед усього цього слід звернути увагу на роботи, присвячені тематиці великих даних, таких авторів, як Дж. Гурвіц, А. Ньюджент, Ф. Халпер, М. Кауфман [13]. Провідними організаціями, що займаються стандартизацією великих даних, є:

- Фонд відкритих даних (The Open Data Foundation, ODaF (www.opendatafoundation.org)) – некомерційна організація, що сприяє впровадженню міжнародних стандартів метаданих і розробці інструментів з відкритими вихідними кодами для обробки статистичних даних.
- Альянс інформаційної безпеки (Cloud Security Alliance (cloudsecurityalliance.org)) – займається впровадженням ефективних методів гарантування безпеки хмарних обчислень і навчанням відповідних IT-спеціалістів. Зокрема, альянсом організовано робочу групу

DOI: 10.24025/2306-4412.1.2019.164787

з великих даних (Big Data Working Group), метою якої є пошук масштабованих методів захисту інформації та вирішення проблем забезпечення конфіденційності великих даних [10].

- Національний інститут стандартів та технологій (National Institute of Standards and Technology, NIST (www.nist.gov)) – американська державна установа, що розробляє нові стандарти. Зокрема, у березні 2012 р. інститут заснував проект з великих даних, метою якого є навчання установ щодо використання великих даних у наукових дослідженнях, охорони навколишнього середовища, розвитку біомедицини, системи навчання та гарантування безпеки країни.
- Фонд програмного забезпечення Apache (Apache Software Foundation, ASF (www.apache.org)) – забезпечує організаційну, юридичну та фінансову підтримку широкого кола проектів, пов'язаних з розробкою ПЗ з відкритим вихідним кодом. Зокрема, одним із ключових проектів фонду є керування платформою Hadoop, що забезпечує стандартні методи кластерної обробки наборів великих даних [7].
- Організація щодо розвитку стандартів структурованої інформації OASIS (Organization for the Advancement of Structured Information Standards (www.oasis-open.org)) – здійснює некомерційну діяльність стосовно розробки стандартів великих даних.

Нині вже є багато рішень, пов'язаних з великими даними, серед яких слід звернути увагу на розробки таких компаній, як Google (Google AI (ai.google/research)), Amazon (Сервіси хмарних технологій (aws.amazon.com/ru/?nc2=h_lg)), IBM (www.ibm.com/analytics/products, cognitiveclass.ai/learn/big-data/), Oracle (Big Data, Big Data Appliance, Big Data Cloud Service, Big Data Connectors, Big Data Discovery, Big Data Platform, Big Data Preparation Cloud Service, Big Data SQL, Big Data Spatial and Graph (www.oracle.com/ru/products/ oracle-a-z.html#b), Microsoft (www.micro-soft.com/en-us/sql-server/big-data) тощо. Аналітичні компанії також не обійшли стороною це явище. Так, аналітична компанія Gartner вперше згадує про великі дані у циклі зрілості технологій (Hype Cycle Emerging Technologies) за 2011 р. під назвою "Big Data and Extreme Information Processing and Management" [14]. За даними аналітичної компанії Forrester, у майбутньому буде спостерігатися значне зростання ринку рішень класу Big Data. У їхньому звіті прогнозується, що ринок таких рішень буде зростати, при цьому «нереляційні» платформи, такі як сегменти Hadoop і NoSQL, зростатимуть майже вдвічі швидше (рис. 1).

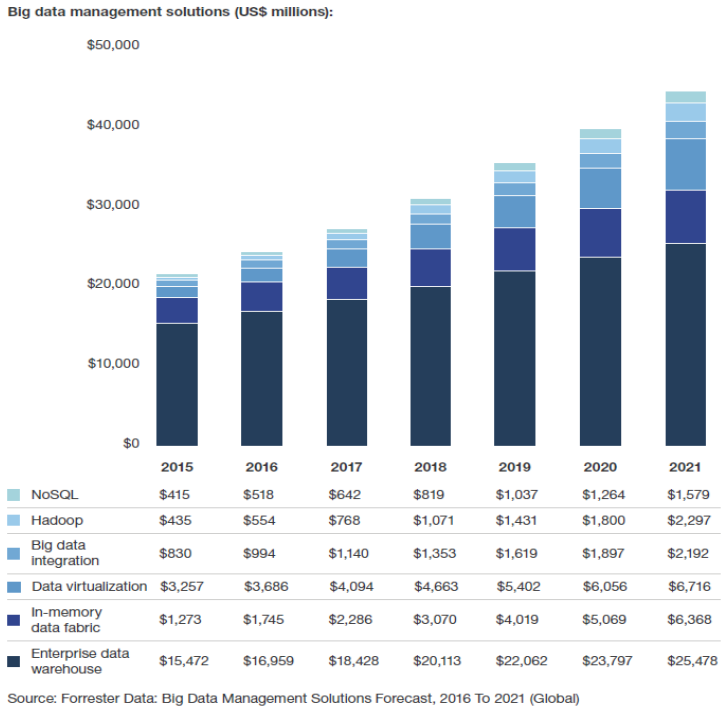

Рисунок 1 – Діаграма за період 2015–2021 рр. ринку рішень управління великими даними від компанії Forrester

DOI: 10.24025/2306-4412.1.2019.164787

Google Trends (trends.google.com) показує активне зростання вживання терміна Big Data у пошукових запитах, починаючи саме з 2011 р. (рис. 2). Така ж тенденція притаманна і нашій країні (рис. 3).

Проте виявилося, що це поняття вельми розпливчате, неоднозначне і викликає багато суперечностей серед фахівців. Отже, **метою** статті є дослідження сутності поняття Big Data (Великі дані) та окреслення подальшого напряму розвитку цієї дефініції.

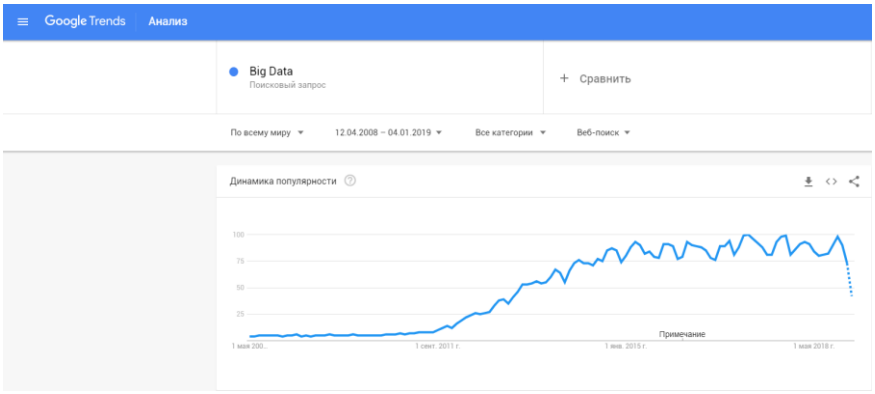

Рисунок 2 – Статистика використання пошукового запиту Big Data в світі за даними аналітичної платформи Google Trends

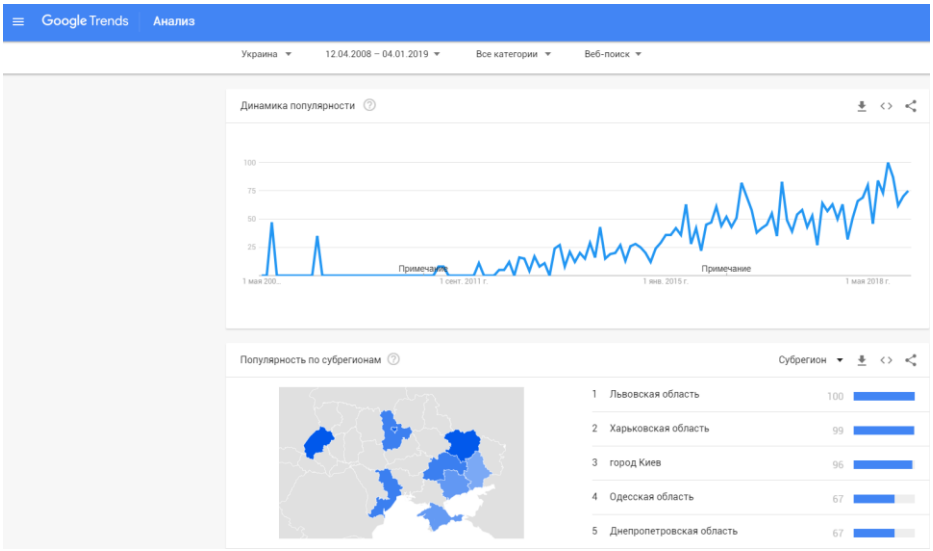

Рисунок 3 – Статистика використання пошукового запиту Big Data в регіонах України за даними аналітичної платформи Google Trends

Основна частина. Історично склалося, що поняття Big Data має дату появи – 4 вересня 2008 р., коли вийшов спеціальний номер британського журналу Nature, присвячений проблематиці бурхливого зростання глобальних даних та їхньої ролі у науці й суспільстві [8]. І саме тут уперше звернено увагу на зміну відношення до даних, дані розглядаються як ресурс на рівні природних копалин.

Спочатку фахівці стверджували, що критерієм відношення до категорії Big Data, перш за все, є потік даних, більший за 100 Гбайт на день. Проте виявилося, що цього недостатньо, щоб однозначно віднести продукт до категорії Big Data. Згодом аналітична компанія Gartner вводить визначення поняття Big Data через «три V» [11]:

- Volume (об'єм) – це збільшення обсягів даних у корпоративних системах за рахунок збільшення обсягів трансакцій та використання традиційних і нових типів даних. Спеціалісти компанії Gartner наголошують на тому, що занадто великий обсяг даних породжує проблему їхнього зберігання, проте це також породжує проблему аналізу даних.
- Variety (різноманіття та неструктурованість даних) полягає у різноманітному фор-

DOI: 10.24025/2306-4412.1.2019.164787

маті наявних даних, це можуть бути табличні дані (бази даних), ієрархічні дані, документи, дані електронної пошти, дані вимірювання, відео, нерухомі зображення, аудіо, дані біржових цінних паперів, дані фінансових операцій тощо. Це породжує проблему переведення великих обсягів транзакційної інформації в рішення.

- Velocity (швидкість обробки) означає як швидкість отримання даних, так і швидкість їх обробки для задоволення запиту.

Такий підхід підхопили інші фахівці й розширили ознаки. Так, компанія Forrester визначає поняття Big Data як технологію в галузі апаратного і програмного забезпечення, яка об'єднує, організовує, керує й аналізує дані, що характеризуються «чотирма V»: об'ємом (Volume), різноманітністю (Variety), мінливістю (Variability) та швидкістю (Velocity) [12, с. 4–5]. Тут під об'ємом розуміється дуже великий обсяг інформації, накопичений у базах даних, що складно обробляти та зберігати традиційними засобами СУБД. Аналітики компанії Forrester роблять акцент на необхідності використання нових підходів і вдосконалення наявних інструментів обробки та збереження даних. Розуміння ознаки різноманітності даних аналітиками компанії Forrester і Gartner є схожим. Швидкість тут розуміється як швидкість не тільки обробки даних, а й їх накопичення. Аналітики звертають увагу на те, що стали більш затребувані технології обробки даних у реальному часі. Також було додано четверту ознаку Variability – мінливість інформації. Наприклад, такою є інформація, що безперервно надходить з датчиків пристроїв або з Інтернету і має важливе значення для аналізу, прогнозування та прий-няття рішень. На рис. 4 зображено діаграму співвідношення цих чотирьох ознак Великих даних.

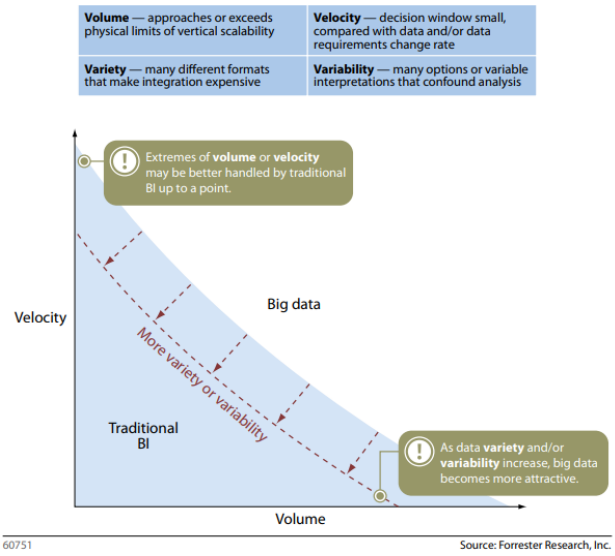

Рисунок 4 – Діаграми чотирьох ознак Великих даних за даними компанії Forrester [12, с. 5]

Згодом цей список ознак було поповнено: «п'ять V» (додано Viability – життєздатність, Value – цінність), «сім V» (додано Veracity – достовірність, Visualization – візуалізація) тощо. Так, наприклад, спеціалістами вітчизняного освітнього ресурсу про новітні технології «The Future» було виокремлено «п'ять V» [3]: Volume (об'єм), Velocity (швидкість), Variety (різноманітність), Variability (мінливість) та Veracity (достовірність). Під достовірністю тут розуміється саме виокремлення достовірних даних. Необхідність п'ятої ознаки пояснюється тим, що якість зафіксованих даних може дуже відрізнятися, що має істотний вплив на результати аналізу.

Надалі цей список «V» можна продовжувати, конкретизуючи його для кожної задачі або бізнес-процесу, проте такий підхід не розкриває це поняття повною мірою, а лише торкається окремих його аспектів. Як справедливо зазначає Джудіт Гурвіц: «Вищезазначені ознаки не завжди виражені у великих даних рівною мірою. Наприклад, методи роботи з великими даними можна застосувати для управління порівняно невеликим об'ємом різноманітних і складних даних або при обробці дуже простих даних, але у великій кількості» [2, с. 27]. Усі ці ознаки є якісними. Постають питання, коли, наприклад, обсяг даних великий, а коли він ще недостатньо великий, чим

DOI: 10.24025/2306-4412.1.2019.164787

це виміряти, що взяти за точку відліку, яку швидкість обробки вважати великою тощо.

Розглядаючи Big Data як технологію, тут можна виокремити три напрями за завданнями, що вирішуються:

- збір, первісне оброблення та зберігання даних для їхнього подальшого використання;
- структурування розрізненого контенту: текстового, графічного, відео, аудіо, звукового тощо;
- бізнес-аналітика на великих обсягах даних.

Для кожного напряму існують свої вимоги та ознаки технологій цього класу. Отже, неможливо і немає потреби поєднувати в одному понятті усі ці напрями. Так, у статті про Великі дані з Вікіпедії надаються альтернативні визначення поняття Великих даних через зазначені ознаки і звертається увага на різний контекст розгляду цього поняття: «Вели́кі да́ні (англ. Big Data) в інформаційних технологіях – набори інформації (як структурованої, так і неструктурованої) настільки великих розмірів, що традиційні способи та підходи (які здебільшого базуються на рішеннях класу бізнесової аналітики та системах управління базами даних) не можуть бути застосовані до них. Альтернативне визначення називає великими даними феноменальне прискорення нагромадження даних та їх ускладнення. Важливо також відзначити те, що часто під цим поняттям у різних контекстах можуть мати на увазі як дані великого об'єму, так і набір інструментів та методів (наприклад, засоби масово-паралельної обробки даних системами категорії NoSQL, алгоритмами MapReduce чи програмними каркасами проекту Hadoop)» [1].

Фахівці сходяться на думці, що Big Data – це зонтичний термін, синергія величезної кількості технологій, багато з яких самі по собі гідні найпильнішого розгляду. Проте з'явилася нова галузь і сформовані підходи щодо роботи з даними. З'явилися засоби, що дають змогу працювати з конкретними бізнес-проблемами і знаходити розумні рішення задач нового рівня [9]. Так, аналітична компанія Forrester у своєму дослідженні TechRadar: Big Data, Q1 2016 описала поточний стан та перспективи розвитку 22 провідних технологій класу Big Data (рис. 5). Кількість і розмір стрілок у кружечках означають швидкість просування цієї технології по кривій розвитку TechRadar. Квадратик або вертикальні лінії показують те, що технологія практично не розвивається або досягла максимуму своїх можливостей.

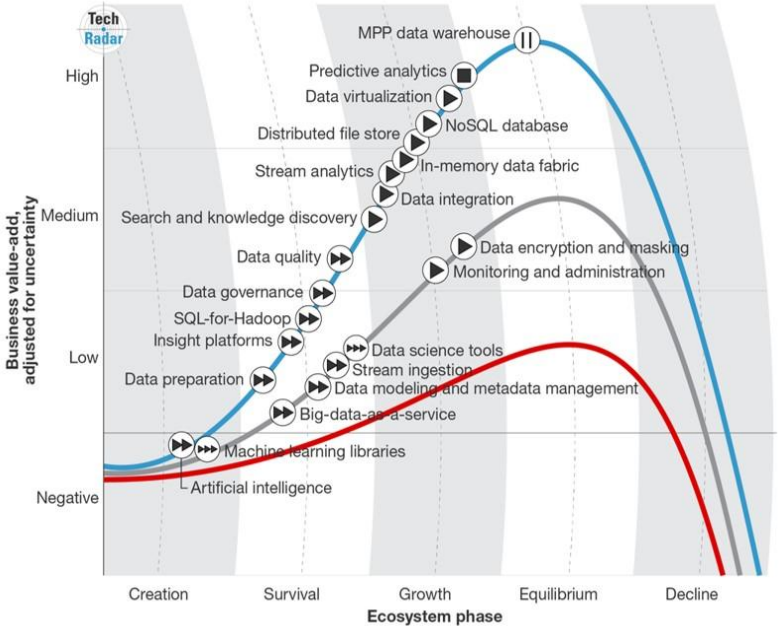

Рисунок 5 – Поточний стан та перспективи розвитку 22 провідних технологій Big Data від компанії Forrester

В інших дослідженнях можна натрапити на визначення цього поняття як неможливості обробки даних традиційними способами. Проте таке визначення є тимчасовим,

DOI: 10.24025/2306-4412.1.2019.164787

адже способи, що нині є новітніми, завтра вже можуть стати традиційними. Так, зокрема, технології обробки великих обсягів даних експерти вже не відносять до новітніх. Компанія Gartner у циклі зрілості технологій за 2015 р. прибрала технологію Big Data (рис. 6). Аналітики компанії пояснюють це розмиванням терміна – технології, що входять у поняття великих даних, стали повсякденною реальністю бізнесу. Тобто до складу поняття Big Data входить велика кількість технологій, що вже частково відносяться до інших популярних сфер і стали повсякденним робочим інструментом.

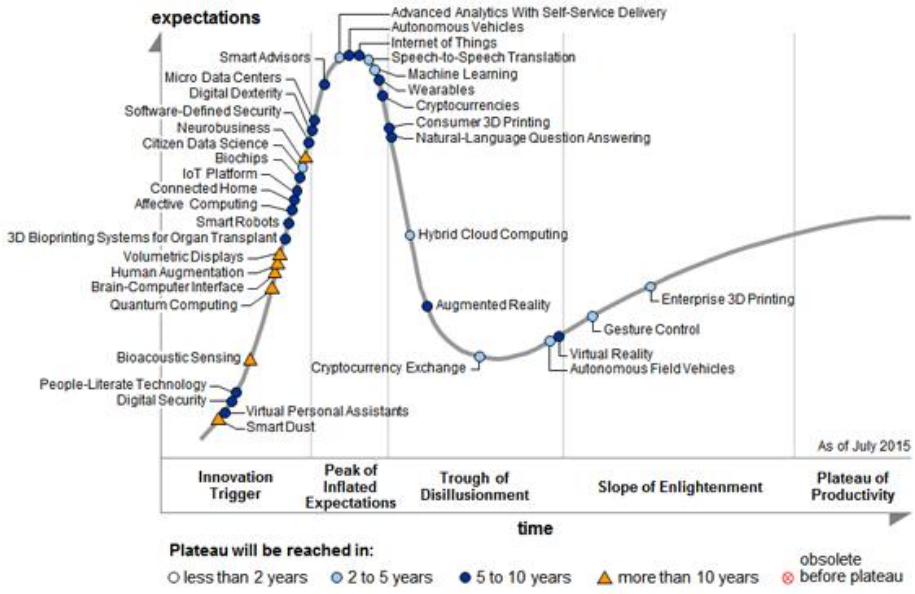

Рисунок 6 – Цикл зрілості технологій за 2015 р., складений компанією Gartner

Висновки. Отже, визначення поняття Big Data потребує інших підходів. Як протилежність цьому існують точки зору, коли взагалі пропонується відмовитися від цього поняття [5]. На думку багатьох експертів, настала епоха, коли важливо не просто вміти акумулювати інформацію, а витягувати з неї бізнес-вигоду. Першими цього висновку дійшли індустрії, які безпосередньо працюють зі споживачем: телекомунікаційна і банківська. Тепер процеси взаємодії виходять на новий рівень, даючи змогу налагодити зв'язок між різними пристроями з використанням інструментів доповненої реальності, та відкривають нові можливості оптимізації бізнес-процесів компаній. Нині спостерігається втрата інтересу з боку реального бізнесу до поняття великих даних. На діаграмі Gartner його місце зайняли інші більш вузько спрямовані технології. Це, у першу чергу, машинне навчання (Machine Learning) – засоби пошуку правил і зв'язків у дуже великих обсягах інформації. Такі технології дають можливість не просто перевіряти гіпотези, але й шукати не відомі раніше фактори впливу. Також, крім машинного навчання, це технології для зберігання даних і паралельного доступу до них (NoSQL Database), попередньої обробки потоків інформації (Marshalling); рішення для візуалізації та самостійного аналізу (Advanced Analytics with Self-Service Delivery). Крім того, зберігають своє значення засоби інтелектуального аналізу даних (Business Intelligence і Data Mining), що виходять на новий технологічний рівень. Отже, в літературі починають вживатися інші поняття, що, на думку фахівців, є більш влучними, наприклад Data Mining. Проте відмовитися від поняття Big Data вже неможливо, адже воно глибоко увійшло у життя. Цей процес змінив наше відношення до даних та інформації, а це, у свою чергу, вплинуло і на сутність поняття Big Data. Якщо спочатку Big Data розумілося як певний стик технологій, то нині, як влучно зазначають IT-фахівці, це скоріше особливий підхід – ідеологія процесінгу інформації, що застосовується для обробки великих масивів «сирих» даних [6]. На нашу думку, такий підхід найбільш влучно відображає сутність поняття Big Data на сучасному етапі розвитку. Проте

DOI: 10.24025/2306-4412.1.2019.164787

у майбутньому це поняття може зникнути як застаріле та неактуальне, адже великі обсяги даних і сучасні підходи до їхньої обробки можуть стати звичайними засобами, і не буде потреби робити наголос на їх інноваційній складовій.

Список літератури

1. Великі дані. Матеріал з Вікіпедії – вільної енциклопедії. 2019. URL: https://uk.wikipedia.org/wiki/Великі_дані (дата звернення: 11.03.2019). Назва з екрану.
2. Гурвиц Дж., Ньюджент А., Халпер Ф., Кауфман М. Просто о больших данных. Москва: Эксмо, 2015. 400 с.
3. Кравчук С. Що таке Big Data? The Future. URL: http://thefuture.news/bigdata (дата звернення: 11.03.2019). Назва з екрану.
4. Найдич А. Big Data: проблема, технология, рынок. КомпьютерПресс. 01.2012. URL: http://compress.ru/ article.aspx?id=22725 (дата звернення: 11.03.2019). Назва з екрану.
5. Реймер Д. Gartner: Big Data больше не существует! URL: http://denreymer.com/gartner-end-of-big-data (дата звернення: 11.03.2019). Назва з екрану.
6. Савчук И. Big Data – технология, рождающая новый тип бизнеса. *БИТ*. 2014. № 3 (36). URL: http://bit.samag.ru/archive/article/1352 (дата звернення: 11.03.2019). Назва з екрану.
7. Apache Hadoop. The Apache Software Foundation. URL: http://hadoop.apache.org (дата звернення: 11.03.2019). Назва з екрану.
8. Big Data. *Nature*. 2008. Vol. 455 (7209). P. 1–136. URL: http://www.nature.com/nature/journal/v455/n7209/index.html (дата звернення: 11.03.2019). Назва з екрану.
9. BIG DATA 2016: Форум умных решений. Издательство «Открытые системы», 07.04.2016. URL: https://www.osp.ru/news/articles/2016/14/13048945 (дата звернення: 11.03.2019). Назва з екрану.
10. Big Data Working Group. Cloud Security Alliance. URL: https://cloudsecurityalliance.org/working-groups/big-data/#_overview (дата звернення: 11.03.2019). Назва з екрану.
11. Gartner says solving 'Big Data' challenge involves more than just managing volumes of data. Gatner, STAMFORD, Conn., June 27, 2011. URL: http://www.gartner.com/newsroom/id/1731916 (дата звернення: 11.03.2019). Назва з екрану.
12. Hopkins B., Evelson B. Expand your digital horizon with Big Data. Forrester Research, Inc. Reproduction Prohibited, September 30, 2011, 16 p.
13. Hurwitz & Associates. URL: https://hurwitz.com (дата звернення: 11.03.2019). Назва з екрану.
14. Smith D. Hype cycle for cloud computing. Gatner, 27 July 2011. URL: https://www.gartner.com/doc/1753115/hype-cycle-cloud-computing- (дата звернення: 11.03.2019). Назва з екрану.

References

1. Big Data (2019). Wikipedia – the free encyclopedia. URL: https://uk.wikipedia.org/wiki/Великі_дані (accessed 11 March 2019).
2. Hurwitz, J., Nugent, A., Halper, F., Kaufman, M. (2015). Big Data for Dummies. Moscow: Eksmo, 400 p. [in Russian].
3. Kravchuk, S. Is this Big Data? The Future. URL: http://thefuture.news/bigdata (accessed 11 March 2019).
4. Naidych, A. (2012). Big Data: problem, technology, market. ComPress. E-journal URL: http://compress.ru/ article.aspx?id=22725 (accessed 11 March 2019).
5. Reymer, D. Gartner: Big Data no longer exists! URL: http://denreymer.com/gartner-end-of-big-data (accessed 11 March 2019).
6. Savchuk, I. (2014). Big Data is a technology that creates a new type of business. *BIT*, No. 3 (36). URL: http://bit.samag.ru/archive/article/1352 (accessed 11 March 2019).
7. Apache Hadoop. The Apache Software Foundation. URL: http://hadoop.apache.org (accessed 11 March 2019).
8. Big Data (2008). *Nature*, vol. 455 (7209), pp. 1–136. URL: http://www.nature.com/nature/journal/v455/n7209/index.html (accessed 11 March 2019).
9. BIG DATA 2016: Smart Solutions Forum (2016). Open Systems, 07 April. URL: https://www.osp.ru/news/articles/2016/14/13048945 (accessed 11 March 2019).
10. Big Data Working Group. Cloud Security Alliance. URL: https://cloudsecurityalliance.org/working-groups/big-data/#_overview (accessed 11 March 2019).
11. Gartner says solving 'Big Data' challenge involves more than just managing volumes of

DOI: 10.24025/2306-4412.1.2019.164787

data (2011). Gatner, STAMFORD, Conn., 27 June. URL: http://www.gartner.com/newsroom/id/1731916 (accessed 11 March 2019).
12. Hopkins, B., Evelson, B. (2011). Expand your digital horizon with Big Data. Forrester Research, Inc. Reproduction Prohibited, 30 September, 16 p.
13. Hurwitz & Associates. URL: https://hurwitz.com (accessed 11 March 2019).
14. Smith, D. (2011). Hype cycle for cloud computing. Gatner, 27 July. URL: https://www.gartner.com/doc/1753115/hype-cycle-cloud-computing- (accessed 11 March 2019).

E. A. Kirey, *Ph.D., associate professor*
e-mail: kirey.kea@gmail.com, kateryna.kirey@chmnu.edu.ua
Petro Mohyla Black Sea National University
68 Desantnykiv str., 10, Mykolaiv, Ukraine

DEVELOPMENT AND TRANSFORMATION OF BIG DATA CONCEPT

The article discusses the development and transformation of Big Data concept (Big Data). This concept does not have a clear interpretation today. Theoretical research has shown that it is reasonable to consider its development in a historical context. Initially it has been considered that the data flow, bigger than 100 Gb a day, is, first of all, a criterion of reference to category Big Data. However, it has turned out that this is not enough to clearly categorize the product as Big Data. Subsequently as such a criterion qualitative signs, that is "a certain number of V": Volume – a significant increase in the amount of data in corporate systems; Variety – a variety of formats and structures of available data; Velocity – the speed of receiving and processing data to satisfy the request, etc., have been used. However, this approach does not fully disclose this concept, but deals with its individual aspects. In other studies, you can find the definition of this concept as the inability to process data in traditional ways. However, such a definition can already be considered obsolete, since experts no longer refer technologies for processing large volumes of data to the latest ones. Some experts suggest to abandon this concept altogether. However, due to its wide spread, this is not possible. The development of information technologies has changed our attitude to data and information, and this, in its turn, has influenced the essence of Big Data concept. Now this rather means a special approach – the ideology of processing large amounts of "raw" data. However, it is possible that in future this concept may disappear as outdated and irrelevant one, because large amounts of data and modern approaches to their processing may become common tools and there will be no need to focus on their innovative component.

Keywords: *Big Data, information technologies.*

DOI: 10.24025/2306-4412.1.2019.164787