УДК: **303.436:311.42**
JEL Classification: C 19
doi: 10.31767/nasoa.1-2.2019.01

О. О. ГОРОБЕЦЬ,
аспірантка,
Національна академія статистики, обліку та аудиту;
ORCID: 0000-0003-1762-2140
Researcher ID: G-7664-2018

Великі дані – джерело статистичної інформації: на прикладі книговидавничої галузі

У статті досліджується питання використання великих даних з метою удосконалення бізнес-процесів на прикладі успішного міжнародного досвіду. Зосереджено увагу на основних складових великих даних, які формують еталонну модель управління ними. На основі опрацьованих джерел автором подано універсальну модель роботи з великими даними. Здійснено стислий огляд програми Hadoop з аналізу великих даних. На основі досвіду роботи провідних міжнародних компаній запропоновано використовувати нові джерела інформації у книговидавничій галузі України.

Ключові слова: *великі дані, еталонна модель управління великими даними, Hadoop, книговидавництво.*

Е. А. ГОРОБЕЦ,
аспирантка,
Национальная академия статистики, учета и аудита

Большие данные – источник статистической информации: на примере книгоиздательской отрасли

В статье исследуется вопрос использования больших данных с целью усовершенствования бизнес-процессов на примере успешного международного опыта. Сосредоточено внимание на основных составляющих больших данных, которые формируют эталонную модель управления ими. На основе проработанных источников автором предложена универсальная модель работы с большими данными. Выполнен краткий программы Hadoop по анализу больших данных. На основе опыта работы ведущих международных компаний предложено использовать новые источники информации в книгоиздательской отрасли Украины.

Ключевые слова: *большие данные, эталонная модель управления большими данными, Hadoop, книгоиздательство.*

O. O. HOROBETS,
Postgraduate Student,
National Academy of Statistics, Accounting and Audit

Big Data of Source of Statistical Information: On the Example of the Book Publishing Industry

The article's objective is to review domestic and foreign sources related with big data, in order to determine the potentials for their practical applications at industry level, with emphasis on book publishing. The main components of big data are highlighted, which form the reference model for big data management. An all-purpose model for big data processing is constructed by use of the analyzed information sources. A brief review of Hadoop big data analysis program is carried out. Applications of big data to improve business processes are analyzed with reference to renowned international companies, Internet platforms engaged in electronic commerce, and social networks. A brief review of Ukrainian cases of successful

electronic commerce is made. As regards big data applications in book publishing, reference is made to the practices of Jellybooks company. The algorithm of big data collection by use of electronic book is constructed and illustrated. It is concluded that any industry is capable to adapt the successful practices of electronic commerce leaders after leaning them. As far as book publishing is concerned, a well-organized system for on-line data collection will open up opportunities for quick production of statistical information, extension of the range of statistical data, monitoring of the current book publishing performance and prediction of its future developments.

Keywords: *big data, reference model for big data management, Hadoop, book publishing.*

Постановка проблеми. Великі дані, як і свого часу інтернетизація, впливають на характер суспільно-економічних відносин, модернізуючи як життя людини, так і економічну діяльність.

В Оксфордському словнику англійської мови великі дані визначаються як "надзвичайно великі сукупності даних, які можуть аналізуватися за допомогою обчислень, щоб виявити закономірності, тенденції та зв'язки, особливо ті, що мають стосунок до поведінки та взаємодії людей" [1]; в словнику "Gartner" – як "інформаційні ресурси великого обсягу, високої швидкості і/або великої різноманітності, які вимагають економічно ефективних та інноваційних форм оброблення для покращення їх розуміння, прийняття рішень та автоматизації процесів" [2]. Отже, великі дані – це передусім *великі сукупності даних* про різноманітні соціально-економічні явища і процеси, які продукуються фактично в безперервному режимі. З одного боку, це означає, що використання великих даних відкриває можливість для покращення якості процесу прийняття рішень, з іншого – специфіка великих даних ускладнює планування дій щодо їх збирання, аналізу та використання отриманих результатів. Вищезазначене обумовлює актуальність дослідження.

Аналіз досліджень та публікацій. Серед досліджень статистичного аспекту великих даних варто згадати роботи українських учених О. Васєчко, О. Журавльова, О. Корепанова, О. Осауленка, В. Саріогло та інших.

О. Васєчко вказує, що великі дані класифікуються за такими групами: адміністративні дані; транзакції або бізнес-інформація; дані сенсорних уловлювачів; дані мобільних сенсорних пристроїв; поведінкові дані; інформація щодо індивідуальної та громадської думки [3, с. 9]. О. Журавльов підкреслює, що робота з великими даними – це, в першу чергу, дієвий інструмент ефективного перерозподілу ресурсів, який дозволить отримати швидкі та ефективні "перемоги" тут і зараз. Це та аналітика, яка сприятиме абсолютно іншій культурі політикотворення – коли політика формується на основі фактів, а не теоретичних міркувань. Доступ до відкритих даних та широке використання аналізу даних покращить процес прийняття рішень у публічному секторі, в тому числі в питаннях законодавства та державного управління на національному та місцевому рівнях [4, с. 10]. О. Корепанов, досліджуючи питання розвитку "розумних" міст в Україні, уточнює, що "великі" дані із багатьох джерел є напівструктурованими або мультиструктурованими, а не неструктурованими. Такі дані передбачають наявність логічної схеми, яка дозволяє отримати інформацію для аналізу [5]. О. Осауленко у монографії "Офіційна статистика в системі національної інформаційної безпеки" стверджує, що великі дані потенційно є джерелом більш релевантної і вчасної статистичної інформації порівняно з традиційними її джерелами [6]. В. Саріогло зауважує, що зростання інтересу до питань імплементації "великих даних" у статистику пов'язано, насамперед, зі значним комерційним успіхом цього підходу в США. Однак реальна корисність "великих даних" для офіційної статистики та оптимальні шляхи їх упровадження ще не з'ясовані, і з приводу цих питань тривають наукові дискусії [7, с. 19].

Серед іноземних дослідників варто згадати роботи Д. Бойда, В. Майер-Шенбергера, М. Кауфмана, К. Кроуфорда, К. О'Ніл, Б. Френкса та інших.

Незважаючи на значний науковий доробок з питань великих даних, дуже невелика кількість робіт присвячена практичним аспектам використання великих даних на галузевому рівні, а дослідження з питань застосування великих даних для потреб книговидавничої галузі загалом відсутні.

Метою статті є огляд вітчизняних і зарубіжних джерел, пов'язаних із великими даними, з метою визначення можливостей їх практичного застосування в окремих видах економічної діяльності, зокрема книговидавничій.

Результати дослідження. Інформація про шість загальних ознак великих даних – обсяг, швидкість, різноманітність, достовірність, життєздатність, цінність – наразі є загальновідомою. Говорячи про обсяг великих даних, варто акцентувати увагу на те, що потужність пам'яті людського мозку дорівнює 2,5 петабайта (1 петабайт = 1000 терабайт), і цього об'єму вистачає на повноцінне життя людини. В. К. Джейн у своїй книзі "Великі дані та Hadoop" стверджує, що до 2020 р. інформація, яку виробляє населення, зросте до 44 зетабайт, тобто до 44 трлн гігабайт. Але "…і 50 Мб можна назвати великими даними якщо вони мають занадто складну структуру для звичайної СУБД" [8].

Вивчаючи питання збирання та використання великих даних, варто звернутися до еталонної моделі управління великими даними ("BDMcube"), яку в 2016 р. презентував М. Кауфман [9]. При роботі з великими даними М. Кауфман пропонує опрацьовувати їх за складовими кубу, покладаючись при цьому на людську інтелектуальність (рис.1).

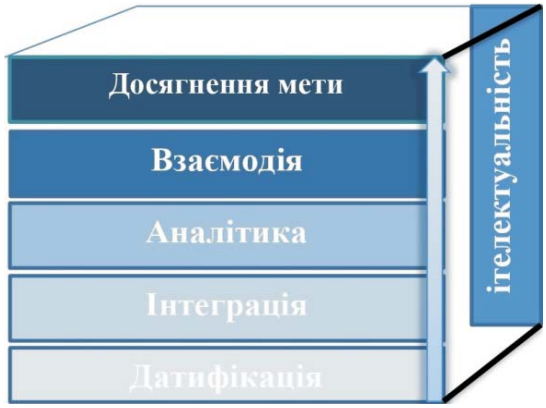

Рис 1. Куб управління великими даними ("BDMcube")

Джерело: [9]

Тобто, враховуючи вищезгадані ознаки великих даних при їх значному обсязі, передусім потрібно здійснити *датифікацію*, тобто ідентифікувати ці дані – дати їм назву, визначити їх призначення та мету збирання, для подальшого їх *інтеграції* в певний процес (або галузь), щоб удосконалити його роботу. Після цього необхідно провести *аналітичне дослідження* великих даних та отриманих результатів, вивчаючи їх *взаємодію* як з об'єктом дослідження (під час їх уведення в нього), так і між собою, а *після досягнення мети* рекомендується пройти тими ж кроками, але у зворотному порядку, задля уникнення помилок та більш глибокого аналізу, покладаючись на *інтелектуальність* дослідницької групи.

Уже стала відомою вільна програмна платформа для організації розподіленого зберігання і оброблення сукупностей великих даних з використанням моделі програмування MapReduce – Hadoop. Всі модулі в Hadoop спроектовані з урахуванням припущення, що апаратне забезпечення часто виходить із ладу і такі ситуації повинні автоматично опрацьовуватись фреймворком [10]. У своїй роботі Hadoop використовує так звані "озера даних", що дозволяє вважати програмну платформу універсальною з точки зору її архітектури. Адже така будова дозволяє об'єднувати "озера даних", гармонізуючи різні типи та форми даних та одночасно працюючи з ними без необхідності їх уніфікації. Зараз архітектура Hadoop використовується у Facebook, Netflix та ін. Адміністратори останніх відмічають, що Hadoop значно скоротила їх витрати на сховища даних. З допомогою налаштувань Hadoop може активізувати процес "пізнього використання" даних, тобто самостійно визначати, коли і які дані будуть необхідні, та використовувати конкретне сховище даних. Hadoop розрізняє такі види великих даних:

1. структуровані дані – відносні дані (таблиці);
2. напівструктуровані дані (дані CSV, XML);
3. неструктуровані дані (Word, PDF, Text, Media Logs) [8].

Підсумувавши все вищезазначене, ми розширили та деталізували "BDMcube" М. Кауфмана і створили універсальну модель роботи з великими даними, яка забезпечує наочність усіх процесів оброблення великих даних та опис можливих результатів у разі успішного виконання кожного критерію (рис. 2).

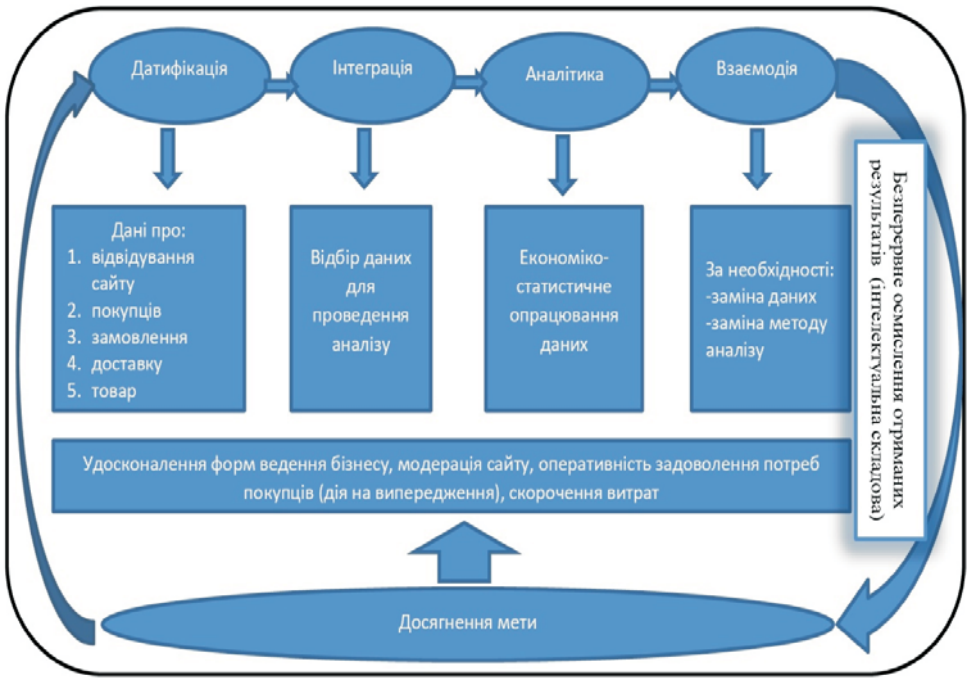

Рис. 2. Універсальна модель роботи з великими даними

Джерело: авторська розробка на основі [9]

Зрозуміло, що відстежувати, зберігати та аналізувати абсолютно усю доступну інформацію не потрібно. Займаючись обробленням будь-якої інформації, необхідно розуміти, для чого вона нам, тобто який вона може надати кінцевий результат. Саме задля такого постійного моніторингу і необхідний так званий "інтелектуальний контроль" за збиранням, аналізом та кінцевим результатом великих даних.

Найвідомішим піонером успішної електронної комерції є Інтернет-платформа Amazon яка у 1994 р. розпочала діяльність із продажу книг. Зараз Amazon – це найбільша у світі Інтернет-платформа з різноманітними товарами: від книг до їжі, від технічних засобів до корму для тварин. З метою кращого задоволення потреб покупців Amazon використовує великі дані, які вона збирає за допомогою заповнених анкет реєстрації (анкета включає відомості про склад родини, домашніх тварин, хобі, харчові уподобання та ін.), моніторингу попередніх покупок, адрес доставки замовлень, теплової карти сайту. Отже, Amazon застосовує "360-градусний погляд" на окремого покупця. Уявити собі обсяг великих даних на серверах Amazon можна на основі інформації про ціну цієї компанії на фондовому ринку, яка у січні 2019 р. сягнула 737 млрд дол.: Amazon стала найдорожчою компанією світу, обійшовши Microsoft [11].

Яскравим прикладом цілеспрямованого використання великих даних та логічної побудови алгоритмів є славнозвісна історія з магазином Target та школяркою, адже за допомогою даних про зміну вподобань школярки алгоритм дізнався про вагітність дівчини раніше, ніж вона, та повідомив про це їй та її батькам за допомогою рекламної продукції для вагітних [12].

Компанія Apple, за допомогою додатку Siri зберігає у хмарному середовищі зразки голосу, запити та місце знаходження клієнта, однак, як стверджують її працівники, ці

дані є конфіденційними та зашифрованими і використовуються з метою визначення найпопулярніших запитів у різних регіонах світу.

Соціальні мережі (Instagram, Twitter, Facebook), різного роду месенджери, електронна пошта збирають дані за схожими алгоритмами, аналізуючи при цьому поведінку, вподобання, трендові захоплення, місце перебування людей.

Успішним прикладом електронної комерції з використанням великих даних на вітчизняному ринку є такі Інтернет-магазини як Rozetka, Makeup, Yakaboo та інші. Варто зазначити, що три згадані магазини мають різну спеціалізацію: Rozetka займається продажем переважно технічних засобів, Makeup – парфумерії та косметичних засобів, Yakaboo – книг та канцелярії. Однак усі вони працюють за єдиним принципом – моніторингом купівельних уподобань, адже якщо певна людина є зареєстрованим клієнтом і відвідує один із вищезазначених сайтів, їй відразу ж на пошту прийде пропозиція купівлі щойно переглянутого нею товару.

Звичайно, говорити про таке використання великих даних як у Target в Україні ще зарано, однак вищезазначені приклади вітчизняної електронної комерції свідчать про її потенціал.

Говорячи про книговидавництво як вид економічної діяльності, варто зауважити, що це єдина галузь, в якій важко спрогнозувати попит на товар (книги). Книговидавничий бізнес будується здебільшого на інтуїтивних відчуттях: видавця – стосовно популярності автора (адже якщо письменник популярний у Бразилії, це аж ніяк не означає, що й в Україні він користуватиметься популярністю); читача – стосовно того, чи сподобається йому книга. Однак, вивчаючи досвід компанії Jellybooks, можна говорити про інше. Провідною ідеєю цієї компанії є те, що вона безкоштовно, за згодою видавців та авторів, надає своїм зареєстрованим читачам доступ до електронних оригінальних версій книг, навіть до тих, які ще не вийшли друком, отримуючи взамін, за згодою читачів, інформацію про процес читання книг за допомогою спеціальних алгоритмів [13].

Цей досвід є унікальним і застосовується для аналізу інформації, невідомої нікому, окрім самого читача (рис. 3).

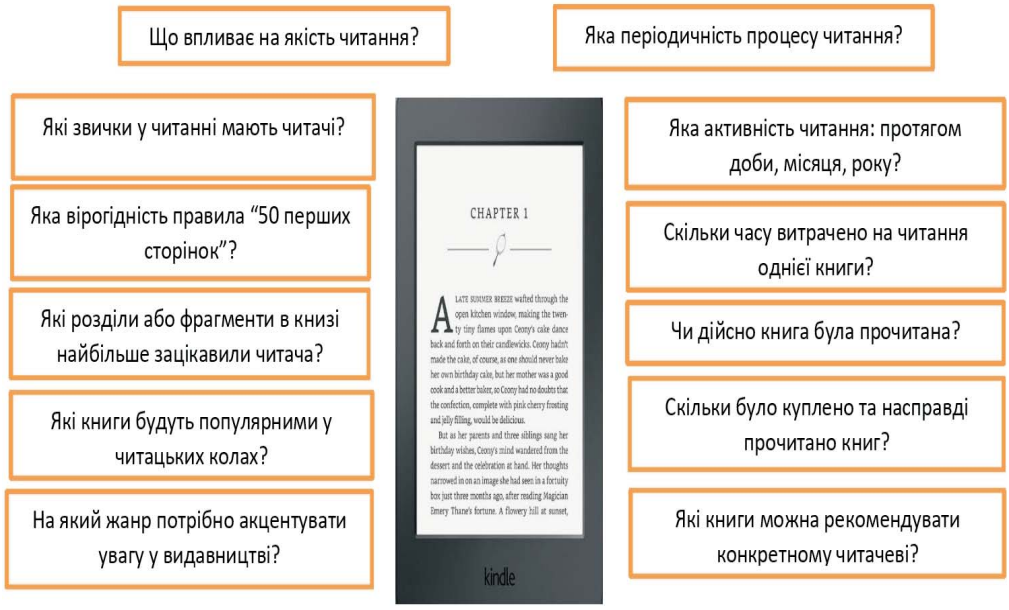

Рис. 3. Збирання даних за допомогою електронної книги

Джерело: авторська розробка на основі [13]

Усі питання, зазначені на рис. 3, є надважливими як для видавця та автора, так і для дослідника проблем читаності та книговидавництва. Наприклад, видавця перш за все зацікавлять відповіді на запитання, які стосуються звичок читачів, популярності

літературного жанру, достовірності читання; автора насамперед зацікавить те, який фрагмент книги читачеві сподобався найбільше, чи була прочитана його книга, чи зміг читач прочитати перші п'ятдесят сторінок і не покинути книгу і чи була його книга взагалі прочитана. Що ж стосується дослідника, то такий вид збирання інформації є універсальним, адже під час звичного анкетування респондент може надавати неточні або взагалі неправдиві відповіді. За допомогою такого збирання даних дослідник точно знатиме, скільки часу читач витрачає на прочитання однієї книги, яка загальна активність та якість читання протягом усього року та чи всі ті книги, які купуються, були дійсно прочитані, що, в свою чергу, дозволить зібрати нові статистичні дані для нових досліджень.

Висновки. Отже, великі дані – це одна з головних ознак XXI ст. Однак окрім переваг великих даних потрібно пам'ятати і про їх недоліки, а саме: неточність алгоритмів, недостовірність машинних розрахунків, ризик втратити дані, порушення конфіденційності та ін. Вивчивши позитивний досвід гігантів електронної комерції, будь-яка галузь може застосувати його у своїй діяльності. Що стосується книговидавничої галузі, то за допомогою налагодженої системи збирання електронних даних з'явиться можливість своєчасно отримувати статистичну інформацію, розширювати її діапазон, здійснювати поточний моніторинг стану галузі, а також прогнозувати майбутній розвиток.

В подальших наукових дослідженнях планується глибше розглянути питання великих даних, зосередивши особливу увагу на ризиках, які можуть виникнути під час збирання та використання інформації.

Список використаних джерел

1. Big Data. English Oxford Living Dictionaries.URL: https://en.oxforddictionaries.com/definition/big_data
2. Big Data. Gartner IT Glossary. URL:https://www.gartner.com/it-glossary/big-data/
3. Васєчко О. О. Сучасні виклики статистичної вищої освіти і науки // Статистика України. 2014.№ 4. С. 4–10.
4. Журавльов О. В. Імплементація концепції "великих даних" у державну статистику // Науковий вісник Національної академії статистики, обліку та аудиту. 2017. № 3. С. 7–15.
5. Корепанов О. С.Методологічні засади статистичного забезпечення управління розвитком "розумних" сталих міст в Україні: моногр. К.: ДП "Інформ.-аналіт. агентство", 2018. С. 95.
6. Осауленко О. Г. Офіційна статистика в системі національної інформаційної безпеки: моногр. К.: ТОВ "Август Трейд", 2017. С. 295.
7. Саріогло В. Г. "Великі дані" як джерело інформації та інструментарій для офіційної статистики: потенціал, проблеми, перспективи // Статистика України. 2016. № 4. С. 12–19.
8. Jain V. K. Big Data and Hadoop. Khanna Publishing, 2017. URL: https://books.google.com.ua/books?id=i6NODQAAQBAJ&printsec=frontcover&hl=ru&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false
9. KaufmannM. (2016). Towards areference modelfor big data management. Research report, Faculty of Mathematics and Computer Science, University of Hagen, retrieved July 15, 2016. URL: https://ub-deposit.fernuni-hagen.de/receive/mir_mods_00000583
10. Wikipedia.ApacheHadoop. URL: https://uk.wikipedia.org/wiki/Apache_Hadoop
11. Bernard Marr & Co. Amazon: Using Big Data to understand customers. URL: https://www.bernardmarr.com/default.asp?contentID=712
12. Hill K. How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did. Forbes. February 16, 2012. URL: https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/
13. Jellybooks. URL: https://www.jellybooks.com/

References

1. Big Data. English Oxford Living Dictionaries. Retrieved from https://en.oxforddictionaries.com/definition/big_data

2. Big Data. Gartner IT Glossary. Retrieved from https://www.gartner.com/it-glossary/big-data/
3. Vasiechko O. O. (2014). Suchasni vyklyky statystychnoi vyshchoi osvity i nauky [Modern challenges to the statistical higher education and science]. Statystyka Ukrainy – Statistics of Ukraine, 4, 4–10 [in Ukrainian].
4. Zhuravlov O. V. (2017). Implementatsiia kontseptsii "velykykh danykh" u derzhavnu statystyku [Implementing the concept of "big data" in the official statistics]. Naukovyi visnyk Natsionalnoi akademii statystyky, obliku ta audytu – Scientific bulletin of National Academy of Statistics, Accounting and Audit, 3, 7–15 [in Ukrainian].
5. Korepanov O. S. (2018). Metodolohichni zasady statystychnoho zabezpechennia upravlinnia rozvytkom "rozumnykh" stalykh mist v Ukraini [The methodological framework for statistical support to management of the "smart" sustainable cities development in Ukraine]. Kyiv: Publishing house "Inform.-analit. ahentstvo", p. 95 [in Ukrainian].
6. Osaulenko O. H. (2017). Ofitsiina statystyka v systemi natsionalnoi informatsiinoi bezpeky [The Official statistics in the national information security system]. Kyiv: "Avhust Treid" Ltd, p. 295 [in Ukrainian].
7. Sariohlo V. H. (2016). "Velyki dani" yak dzherelo informatsii ta instrumentarii dlia ofitsiinoi statystyky: potentsial, problemy, perspektyvy ["Big data" as a source for information and a tool for the official statistics: potentials, problems, prospects]. Statystyka Ukrainy – Statistics of Ukraine, 4, 12–19 [in Ukrainian].
8. Jain V. K. Big Data and Hadoop. Khanna Publishing, 2017. Retrieved from https://books.google.com.ua/books?id=i6NODQAAQBAJ&printsec=frontcover&hl=ru&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false
9. Kaufmann M. (2016). Towards a reference model for big data management. Research report, Faculty of Mathematics and Computer Science, University of Hagen, retrieved July 15, 2016. Retrieved from https://ub-deposit.fernuni-hagen.de/receive/mir_mods_00000583
10. Wikipedia. Apache Hadoop. Retrieved from https://uk.wikipedia.org/wiki/Apache_Hadoop
11. Bernard Marr & Co. Amazon: Using Big Data to understand customers. Retrieved from https://www.bernardmarr.com/default.asp?contentID=712
12. Hill K. How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did. Forbes. February 16, 2012. Retrieved from https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/
13. Jellybooks. Retrieved from https://www.jellybooks.com/

Посилання на статтю:

Горобець О. О. Великі дані – джерело статистичної інформації: на прикладі книговидавничої галузі // Науковий вісник Національної академії статистики, обліку та аудиту: зб. наук. пр.. 2019. №1-2. С. 7-13.