УДК 044.056

Hadoop-рішення для захисту даних великих обсягів

Наталія Маслова[1], Ольга Половинка[2]

[1]к.т.н., доцент, доцент кафедри прикладної математики та інформатики Донецького національного технічного університету, пл.Шибанкова, 2, м. Покровськ, Україна, e-mail: nataliia.maslova@donntu.edu.ua
[2]аспірант кафедри прикладної математики та інформатики Донецького національного технічного університету, пл.Шибанкова, 2, м. Покровськ, Україна, e-mail: olga.polovinka1@gmail.com

Досліджено одну з проблем даних великого обсягу – забезпечення захисту в процесі накопичення та обробки. Розглядається випадок застосування технології Hadoop в її останній модифікації Apache Hadoop 3.3.0. Пропонується рішення з посиленням захисту оброблюваних даних й підключенням інструментів Apache Knox Gateway, Apache Ranger та Apache Atlas. Передбачена можливість застосування даних, отриманих в результаті роботи локальних баз, електронних архівів, систем управління базами та даних окремих користувачів. Особливостями рішення є також застосування приватної хмари й криптографічних алгоритмів. Наведено приклад реалізації захищеного рішення до розв'язання задачі глибинного аналізу на прикладі паралельної версії задачі пошуку асоціативних правил при роботи з неструктурованими даними великих обсягів.

Ключові слова: *технологія Hadoop, обробка даних, дані великих обсягів, Big Data, безпека даних, захист інформації.*

Вступ: Кожну мить у світі генерується значна кількість даних, які мають найрізноманітніші джерела виникнення, темпи накопичування, структуру, формати представлення. У 2020 компанія Arcserve надрукувала дослідження Cybersecurity Ventures, згідно якого загальний обсяг даних, які зберігаються в хмарах різного типу, у тому числі в приватних хмарах, зростає експоненціально й до 2025 р. досягне 100 цетабайт [1]. Розповсюджені технології обробки, накопичення та зберігання для даних великих обсягів виявляються неефективними. У цій галузі застосовують паралельні архітектури обробки даних, технології побудови спеціалізованих хмарних сховищ (data warehouses), рішення GreenPlum, Apache Spark, Hadoop MapReduce.

Найбільш проблемними задачами при роботі з даними великих обсягів є задачі обробки, аналізу, зберігання та забезпечення безпеки.

1.Проблеми захисту даних великих обсягів

Сучасна реальність створення гібридних сховищ пов'язана з вимогами ретельного ставлення до безпеки, конфіденційності та цілісності даних. Головна задача методів Big Data – обробка величезних обсягів даних й побудова на їх основі

прогнозних моделей, виявлення прихованих зав'язків та взаємодій. Термін Big Date пов'язують з структурованими і неструктурованими даними великих обсягів. Їх формування, накопичення та обробка, як правило, виконуються в on-line режимі, завдяки роботі розподілених систем, застосуванню хмарних сервісів. Паралельно зростанню обсягів даних експоненціально зростає й кіберзлочинність. Серед основних перспективних та проблемних напрямків у забезпеченні захисту даних великих обсягів виділяють [1] обов'язкове застосування заходів з організації захисту даних, у тому числі шифрування; резервне копіювання та відновлення даних; дотримання всіх стандартів відповідності, оцінювання стану безпеки та надання рекомендацій щодо проведення подальших робіт з захисту даних належного, сучасного, рівня. Дані, що використовуються для аналізу часто містять персональну або актуальну комерційну інформацію, вимоги забезпечення конфіденційності можуть стосуватися як даних, які аналізуються, так і отриманих при їх обробці результатів.

Значний внесок у розробку технологій захисту Big Data зроблено міжнародним альянсом Cloud Security Alliance (CSA), Національним інститутом стандартів і технологій США (NIST) та Агентством Європейського Союзу з питань мережевої та інформаційної безпеки (European Union Agency for Network and Information Security, ENISA). Спеціалізовані рішення з захисту великих даних мають практично всі великі компанії. Наприклад, у IBM це рішення Top tips for Big Data Security, у Oracle - Enterprise Security for Big Data Environments, Hewlett Packard Enterprise - HPE Data Protector.

Проблеми захисту Big data виникають на всіх етапах роботи з даними – при формуванні, передачі, накопиченні, зберіганні, аналізі та візуалізації.

Автори вже звертались до проблеми забезпечення безпеки великих даних Так, у [2] показана комплексність проблеми, необхідність багаторівневого захисту, визначено базові ризики. Основною потребою на початкових етапах накопичення даних великих обсягів є шифрування даних. Перспективним методом шифрування даних для даних великих обсягів, які розміщено на хмарі заявлено застосування вітчизняного криптографічного алгоритму, шифру «Калина».

У цій роботі зупинимось на побудові захищеного Hadoop-рішення.

2.Технології обробки даних значних обсягів

Дослідники Big Data застосовують різні технології та програмні засоби бізнесової аналітики. Основними інструментами та технологіями роботи з великими даними при наявності вимоги застосування паралельних методів обробки є бібліотеки масово-паралельної обробки, системи управління базами даних категорії NoSQL, алгоритми MapReduce, проекти Hadoop та MPI.

Hadoop – це платформа з відкритим кодом , в основу якої покладено розподілену файлову систему. На базі Hadoop функціонують близько тисячі різноманітних сервісів для обробки інформації. Microsoft Azure, Amazon і Google та інші Cloud-гіганти застосовують платформу для зберігання й обробки великих

даних. На обсягах інформації більш терабайту платформа працює ефективніше класичних реляційних БД.

Безпосередньо Hadoop з 2013 року став платформою для збору, зберігання і аналізу великих масивів даних. Hadoop – достатньо складна технологія. Створені на її базі продукти вимагають наявності спеціальних навичок та знань. На допомогу Hadoop та MapReduce намагаються прийти такі продукти, як Hive, Pig, Kubernetes, Impala, Tez, Spark й бази типу NoSQL. Вони доповнюють Hadoop та долають його недоліки. Але такі недоліки Hadoop, як недосконалі засоби забезпечення захисту даних залишаються актуальними й вимагають подальшого рішення.

3. Структура захищеного Hadoop-рішення

Сучасна версія Apache Hadoop 3.3.0 підтримує ARM-архітектуру, Java 11, систему каталогу YARN-додатків, файлову систему Tencent Cloud COS для доступу до об'єктного сховища COS і планування запуску контейнерів за розкладом. Анонсовано полегшення роботи з DNS і IP, а також стабілізація HDFS RBF (Router-based Federation). У версії 3.3.0 вдосконалено засоби керування безпекою, а саме – підтримку безпеки маршрутизатором HDFS [3].

Слід виділити необхідні напрямки захисту кластера Apache Hadoop:

- запобігання атак і несанкціонованого доступу до Big Data ззовні;

- забезпечення безпеки використання великих даних внутрішніми клієнтами (у тому числі й автоматизованими системами);

- жорстке адміністрування та моніторинг всіх завдань, пов’язаних з безпекою Big Data.

Прояві означених проблем у базовій версії Hadoop сприяють можливість входу в систему в якості системного адміністратора без автентифікації, відсутність за замовченням заборони роботи сервісів Hadoop без SSL, відсутність захищеного контуру кластера Hadoop.

Перша проблема вирішується налаштуванням автентифікації для кожного сервісу окремо. Для кластеру того ж ефекту можна досягнути включенням мережевого протоколу автентифікації Kerberos. Процес вимагає уважного настроювання й контроля, трудомісткий, хоча є звичайним при організації автентифікації.

Рішенням другої проблеми є вимога включення SSL для кожного сервісу.

Класично третя проблема вирішується обмеженням користувачів засобами мережевого екранування, або організацією єдиного входу до системи спеціалізованими засобами Hadoop.

Існують технічні рішення і додаткові інструменти реалізації корпоративних моделей безпеки. Це системи моніторингу, засоби шифрування, інструменти підтримки політик доступу, захищені протоколи передачі даних. Але це здебільше комерційні наробки великих постачальників дистрибутивів і хмарних рішень Big Data (Cloudera, HortonWorks, ArenaData, Amazon EMR, MCS і т.д.).

В умовах невеликих та середніх підприємств актуальним є і використання загальнодоступних інструментів. Прикладами таких рішень є використання та настроювання рішень Apache Knox Gateway, Apache Ranger та Atlas. Перше рішення забезпечує єдину точку доступу для всіх HTTP з'єднань з кластерами Apache Hadoop і систему єдиної автентифікації (Single Sign On) для сервісів і користувацького інтерфейсу компонент Apache Hadoop.

Apache Ranger забезпечує моніторинг та управління комплексною безпекою даних на платформі Hadoop.

Apache Atlas доповнює набір базових сервісів управління та допомагає розв’язати питання організації комплексного та захищеного захисту Hadoop, інтеграції корпоративних систем і користувацьких даних.

Загальний вигляд схеми захисту рішення Hadoop з урахуванням вищеозначеного наведено на рис. 1.

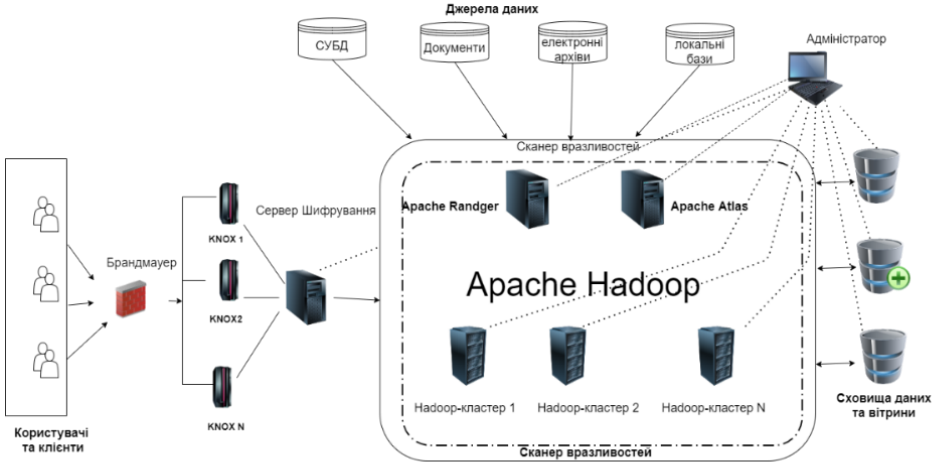

Рис. 1. Схема захисту рішення Hadoop

Apache Knox Gateway э шлюзом, який забезпечує рішення проблеми захисту периметра. Розміщення шлюзу Knox Gateway між клієнтами та кластерами Hadoop розширює доступ нових користувачів до Hadoop, дозволяє використовувати ідентифікаційні дані з корпоративних систем для безпечного доступу до кластерів.

Доповнення структури приватної хмари описаними інструментами та методами шифрування, описаними в [2] забезпечує захист, придатне для практичного застосування при розв’язуванні наукових та практичних задач опрацюванні великих обсягів даних. Так, наприклад, у [4] наведено результати аналізу алгоритмів пошуку асоціацій для роботи з неструктурованими даними великих обсягів з використанням алгоритму апріорної групи та наведена його паралельна реалізація на фреймворку Hadoop MapReduce. Промислове застосування розробки пла-

нується на базі захищеної версії Hadoop, основні моменти котрої описано у цій статті.

Висновки. В роботі досліджено стан питання забезпечення безпеки великих даних. Зроблено огляд відомих у цих сферах рішень та рекомендації щодо їх застосування. Розглянуто застосування технології Hadoop (версія Apache Hadoop 3.3.0). Пропонується рішення з посилення захисту оброблюваних даних. Особливостями рішення є також застосування приватної хмари й криптографічних алгоритмів, як описано в попередній роботі авторів. Наведено приклад реалізації захищеного рішення до розв'язання задачі глибинного аналізу на прикладі паралельної версії задачі пошуку асоціативних правил при роботи з неструктурованими даними великих обсягів. Отже, зроблено ще один крок у напрямку застосування можливостей сучасних технологій у забезпеченні захисту даних великих обсягів й побудови практичних рішень для проведення досліджень та рішення практичних задач.

Література

[1] The 2020 Data Attack Surface Report, Arcserve, 2 Dec 2020, URL https://info.arcserve.com/en/the-2020-data-attack-surface-report#:~:text=Cybersecurity%20Ventures%20research%20shows%20the,world's%20data%20at%20that%20time

[2] *N.Maslova, M.Fedorko* Features of Big Data Protection, January 2018, URL https://www.researchgate.net/publication/328821496 FEATURES_OF_BIG_DATA_PROTECTION DOI: 10.31474/1996-1588-2018-1-26-41-47

[3] The Apache Software Foundation. Apache Hadoop, 28 July, 2020, URL https://blogs.apache.org/hadoop/entry/announce-apache-hadoop-3-3

[4] *Polovynka Olha, Dmitrieva Olga* Research of the efficiency of the apriori group algorithms on different database sizes. - ScientificWorldJournal Issue No6 Part 1 December 2020, p.71-78.

Hadoop solution for large data protection

Nataliya Maslova, Olha Polovynka

Investigated one of large data problems of - providing protection in the process of accumulation and processing. The case of application of Hadoop technology and its latest modification Apache Hadoop 3.3.0 is considered. A solution is proposed with strengthening the protection of processed data, connecting the Apache Knox Gateway, Apache Ranger and Apache Atlas tools. The possibility of using data obtained as a result of the work of local databases, electronic archives, database management systems and individual users is provided. The solution also features the use of a private cloud and cryptographic algorithms. An example of the implementation of a secure solution to the problem of Intelligent Data Analysis is given on the example of a parallel version of the problem of finding association rules when working with unstructured data of large volumes.

Отримано 12.03.21