УДК 004.77:004.62:004.056.53 DOI: 10.31673/2412-4338.2025.048915

Звенигородський Олександр Сергійович
Державний університет інформаційно-комунікаційних технологій, Київ, Україна.
ORCID: 0009-0008-6235-1638

Кудринський Павло Олегович
Державний університет інформаційно-комунікаційних технологій, Київ, Україна
ORCID: 0009-0008-6314-6150

Іщеряков Сергій Михайлович
Державний університет інформаційно-комунікаційних технологій, Київ, Україна
ORCID: 0009-0007-5961-8218

Щербина Ірина Сергіївна
Державний університет інформаційно-комунікаційних технологій, Київ, Україна
ORCID: 0009-0004-8373-7522.

ОБРОБКА ВЕЛИКИХ ДАНИХ У СИСТЕМАХ КІБЕРЗАХИСТУ З ВИКОРИСТАННЯМ ХМАРНИХ ТЕХНОЛОГІЙ

Анотація. *У статті розглянуто тенденції використання технологій обробки великих даних у системах кіберзахисту з акцентом на хмарні архітектури та розподілені обчислення. Показано, що різке зростання обсягів телеметрії, логів і подій IT-інфраструктури робить традиційні системи SIEM малоефективними, оскільки вони не забезпечують необхідної продуктивності для виявлення загроз у режимі реального часу. На основі аналізу сучасних рішень продемонстровано роль таких фреймворків, як Apache Hadoop та Apache Spark, у побудові масштабованих платформ аналітики безпеки із підтримкою пакетної та потокової обробки. Окремо розглянуто переваги хмарних SIEM-систем, які забезпечують еластичне масштабування, скорочення капітальних витрат та можливість обробки петабайтних масивів даних із затримкою менше 100 мілісекунд.*

Проаналізовано підходи до інтеграції машинного навчання й алгоритмів виявлення аномалій, що дозволяють автоматизувати кореляцію подій, зменшувати кількість хибнопозитивних спрацювань та підвищувати якість реагування на інциденти. Розглянуто безпекові виклики впровадження хмарних рішень: конфіденційність, відповідність регуляторним вимогам, ризики vendor lock-in та специфічні загрози cloud-native середовищ. Сформульовано практичні рекомендації щодо поетапного впровадження хмарних SIEM, оптимізації інжестії даних, розробки гібридних моделей зберігання та використання SOAR-платформ для автоматизації реагування на інциденти.

Стаття окреслює перспективи подальших досліджень: впровадження quantum-safe криптографії, використання federated learning для тренування моделей на розподілених конфіденційних даних, стандартизацію multi-cloud інтеграції та застосування edge-аналітики для попередньої обробки критичних телеметричних потоків..

Ключові слова: *великі дані, кібербезпека, хмарні обчислення, SIEM, розподілені системи, Apache Spark, аналітика безпеки, реальний час.*

Zvenyhorodskyi Oleksandr
State University of Information and Communication Technology, Kyiv, Ukraine.
ORCID: 0009-0008-6235-1638

Kudrynskyi Pavlo
State University of Information and Communication Technology, Kyiv, Ukraine.
ORCID: 0009-0008-6314-6150

BIG DATA PROCESSING IN CYBERSECURITY SYSTEMS USING CLOUD TECHNOLOGIES

Іщеряков Сергій Михайлович
State University of Information and Communication Technology, Kyiv, Ukraine.
ORCID: 0009-0007-5961-8218

Shcherbyna Iryna
State University of Information and Communication Technology, Kyiv, Ukraine.
ORCID: 0009-0004-8373-7522

Abstract. *The article examines current trends in applying big data processing technologies to cybersecurity with a focus on cloud-based architectures and distributed computing. The rapid growth of telemetry, log streams, and security events has made traditional SIEM platforms inefficient because they cannot provide the performance required for real-time threat detection. Based on a review of modern solutions, the study highlights the role of frameworks such as Apache Hadoop and Apache Spark in enabling scalable security analytics platforms that support both batch and streaming workloads. Particular attention is paid to cloud-native SIEM systems, which offer elastic scalability, reduced capital expenditures, and the ability to process petabyte-scale data volumes with latency below 100 milliseconds.*

The article analyzes the integration of machine learning and anomaly-detection models that automate event correlation, reduce false positives, and improve incident-response efficiency. Key challenges of cloud adoption are addressed, including data confidentiality, regulatory compliance, vendor lock-in risks, and threats specific to cloud-native infrastructures. Practical recommendations include staged deployment of cloud SIEM platforms, optimization of data ingestion pipelines, development of hybrid storage strategies, and the use of SOAR solutions to automate incident handling.

Future research directions are outlined, including quantum-safe cryptography for protecting security big data, federated learning for training detection models on distributed confidential datasets, standardized multi-cloud integration interfaces, and edge-analytics mechanisms for pre-processing critical telemetry streams at the network perimeter.

Keywords: *big data, cybersecurity, cloud computing, SIEM, distributed systems, Apache Spark, security analytics, real time.*

ПОСТАНОВКА ПРОБЛЕМИ

Сучасні корпоративні мережі генерують величезні обсяги даних про події безпеки. За оцінками аналітиків, середнє підприємство створює від 1 до 5 терабайтів логів та телеметричних даних щодня, при цьому обсяг даних про безпеку зростає на 40-50% щорічно [5]. Традиційні системи управління інформацією та подіями безпеки (SIEM) виявляються неспроможними ефективно обробляти такі обсяги даних, що призводить до затримок у виявленні загроз та втрати критичної інформації.

Особливу складність становить необхідність зберігання та аналізу історичних даних для виявлення складних багатоетапних атак (Advanced Persistent Threats, APT), які можуть тривати місяці. Традиційні рішення вимагають значних капітальних витрат на інфраструктуру та не забезпечують необхідної гнучкості масштабування. За даними IBM, середня вартість витоку даних у 2024 році досягла 4,88 мільйона доларів США, що на 10% більше порівняно з 2023 роком [10], підкреслюючи критичну важливість ефективних систем захисту.

Проблема ускладнюється гетерогенністю джерел даних: фаєрволи, системи виявлення вторгнень (IDS/IPS), антивірусне ПЗ, мережеве обладнання, хмарні сервіси, IoT-пристрої генерують дані в різних форматах та з різною інтенсивністю. Інтеграція та нормалізація цих даних для комплексного аналізу вимагає потужних розподілених систем обробки.

Хмарні технології пропонують революційне рішення цієї проблеми, забезпечуючи практично необмежену масштабованість, еластичність ресурсів та економічну ефективність через модель оплати за використання (pay-as-you-go). Однак впровадження хмарних рішень для обробки чутливих даних безпеки породжує додаткові виклики, пов'язані з конфіденційністю, відповідністю регуляторним вимогам та захистом від загроз у самому хмарному середовищі.

Зв'язок з важливими науковими та практичними завданнями полягає у необхідності створення ефективних, масштабованих та економічно доцільних систем кібербезпеки, здатних обробляти петабайти даних для захисту критичної інфраструктури та забезпечення національної безпеки в епоху цифрової трансформації.

АНАЛІЗ ОСТАННІХ ДОСЛІДЖЕНЬ І ПУБЛІКАЦІЙ

Проблематика обробки великих даних у кібербезпеці активно досліджується міжнародною науковою спільнотою. Систематичний огляд літератури виявляє декілька ключових напрямків досліджень.

Фундаментальну роботу щодо застосування Big Data в кібербезпеці представили дослідники у роботі "Cybersecurity in Big Data Era: From Securing Big Data to Data-Driven Security" [7]. Автори запропонували таксономію підходів до обробки Big Data у центрах операцій безпеки (SOC) та висвітлили перехід від традиційної сигнатурної безпеки до безпеки, керованої даними (data-driven security).

Norell (2024) [4] досліджував, як аналітика великих даних може суттєво покращити стратегії кібербезпеки, забезпечуючи точніше виявлення загроз та можливості реагування у реальному часі. Його робота демонструє практичне застосування технологій Big Data для підвищення точності детектування та зменшення часу реагування на інциденти.

Важливий внесок у розуміння ринкових тенденцій зробив звіт Custom Market Insights (2024) [5], який показав, що глобальний ринок безпеки великих даних оцінювався у 19,76 млрд доларів США у 2022 році і прогнозується досягти 55,16 млрд доларів до 2032 року при середньорічному темпі зростання (CAGR) 14,90%, що свідчить про стрімке зростання попиту на такі рішення.

Комплексне дослідження застосувань Big Data провели Sharma et al. (2024) [6], проаналізувавши виклики та майбутні перспективи у різних доменах, включаючи соціальні Big Data, транспортні та медичні застосування. Автори підкреслили, що Big Data вимагає спеціальних технологій для зберігання, обробки та аналізу масивних та складних наборів даних.

Щодо хмарних технологій, Fortune Business Insights (2024) [3] повідомляє, що сегмент публічних хмар займав найбільшу частку ринку (57%) у 2024 році, що обумовлено зростаючим попитом на безпечні, масштабовані та економічно ефективні рішення. Це зростання підкріплюється цифровою трансформацією та збільшенням потреб у зберіганні даних.

Актуальні дослідження 2025 року [17, 20] підкреслюють революційну роль AI та хмарних технологій у трансформації операцій безпеки. Нові SIEM-системи покоління використовують хмарно-нативні архітектури та AI для досягнення швидкості аналітики, що дозволяє обробляти на 50% більше даних у 150 разів швидше порівняно з традиційними рішеннями [14].

Дослідження тенденцій хмарних обчислень [2] прогнозує розширення попиту на хмарні обчислення у 2024 році через інтеграцію AI та машинного навчання, з фокусом на покращення можливостей самонавчання, захисту даних та конфіденційності.

Важливе значення для практичного застосування мають роботи, присвячені розподіленим обчислювальним фреймворкам. Sadhwani (2025) [25] провів порівняльний аналіз Apache Hadoop, Apache Spark та Apache Flink для обробки великих наборів даних, підкресливши їх специфічні переваги для різних сценаріїв використання. AWS (2025) [21] демонструє, що Spark на базі Hadoop забезпечує безпечну розподілену обробку з можливістю переміщення критичних за часом робочих навантажень до in-memory процесорів Spark.

Невирішені частини проблеми. Попри значний прогрес, залишаються відкритими питання: 1) оптимального балансу між безпекою та продуктивністю при обробці Security Big Data у хмарі; 2) забезпечення відповідності регуляторним вимогам (GDPR, HIPAA) при зберіганні чутливих даних про безпеку у публічних хмарах; 3) мінімізації латентності для систем реального часу при розподіленій обробці; 4) захисту від специфічних хмарних загроз (cloud-native threats); 5) ефективного управління витратами при масштабуванні хмарних SIEM-систем.

Мета дослідження полягає в аналізі та систематизації підходів до обробки великих даних у системах кібербезпеки з використанням хмарних технологій, оцінці ефективності різних архітектурних рішень та виявленні оптимальних практик впровадження.

Завдання дослідження:

1. Проаналізувати архітектури систем обробки Security Big Data на базі хмарних технологій.
2. Дослідити ефективність розподілених обчислювальних фреймворків (Hadoop, Spark) для задач кібербезпеки.
3. Порівняти продуктивність хмарних та локальних SIEM-систем за ключовими метриками.
4. Оцінити виклики безпеки та конфіденційності при використанні хмарних рішень для обробки Security Big Data.

5. Сформулювати рекомендації щодо вибору та впровадження хмарних систем обробки великих даних для кібербезпеки.

Архітектура систем обробки Security Big Data

Сучасні системи обробки великих даних у кібербезпеці базуються на багаторівневій архітектурі, що включає рівні збору, зберігання, обробки, аналітики та візуалізації даних.

Рівень збору даних використовує розподілені агенти та колектори для отримання логів та телеметрії з різноманітних джерел. Сучасні хмарні SIEM-системи підтримують нативну інтеграцію з понад 200 джерелами даних, включаючи хмарні сервіси (AWS CloudTrail, Azure Monitor, Google Cloud Logging), мережеве обладнання, системи захисту endpoint та застосунки [16].

Рівень зберігання реалізується через розподілені системи, такі як HDFS (Hadoop Distributed File System) або хмарні об'єктні сховища (Amazon S3, Azure Blob Storage). Ці системи забезпечують високу надійність через реплікацію даних та можливість зберігання петабайтів інформації за економічно ефективною моделлю. Для "гарячих" даних, що потребують швидкого доступу, використовуються in-memory сховища та NoSQL бази даних.

Рівень обробки базується на розподілених обчислювальних фреймворках. Apache Hadoop через модель MapReduce забезпечує надійну пакетну обробку великих обсягів історичних даних. Apache Spark пропонує значно вищу продуктивність завдяки обчисленням в оперативній пам'яті, досягаючи швидкості обробки у 100 разів вищої для деяких задач [24, 28].

Таблиця 1.

Порівняння розподілених обчислювальних фреймворків для кібербезпеки

Характеристика	Apache Hadoop	Apache Spark	Гібридне рішення
Модель обробки	Пакетна (batch)	In-memory, потокова	Комбінована
Швидкість обробки	Базова	До 100× швидше	Оптимальна
Латентність	Хвилини-години	Секунди-хвилини	<100 мс для критичних
Використання пам'яті	Низьке	Високе	Адаптивне
Застосування	Історичний аналіз	Реальний час, ML	Повний спектр
Складність	Середня	Висока	Висока
Вартість	Низька	Середня-висока	Варіюється

Хмарні SIEM-системи наступного покоління

Традиційні локальні SIEM-системи еволюціонують у хмарно-нативні платформи безпекової аналітики [14, 17]. Ключові характеристики сучасних cloud-native SIEM включають:

Необмежена масштабованість. Хмарні SIEM автоматично масштабуються залежно від обсягу даних, що надходять. Microsoft Sentinel, як приклад лідера ринку [13], забезпечує динамічне управління даними без необхідності попереднього планування ємності.

AI-driven аналітика. Інтеграція штучного інтелекту дозволяє автоматично виявляти аномалії, кореляцію подій та прогнозувати потенційні загрози. SentinelOne Singularity AI-SIEM [12] демонструє можливості автономного SOC, де AI бере на себе рутинні операції аналізу.

Unified Data Lake. Сучасні рішення використовують централізоване озеро даних (data lake) для зберігання всіх типів Security Big Data – структурованих, напівструктурованих та неструктурованих. Це забезпечує єдину точку істини для аналітики безпеки.

Гнучке ціноутворення. Модель оплати за використання дозволяє організаціям платити лише за фактично спожиті ресурси, що радикально знижує вхідні бар'єри порівняно з традиційними рішеннями, які вимагають значних капітальних інвестицій.

Продуктивність та латентність

Критичним фактором для систем реального часу є латентність обробки – час від надходження події до її аналізу та генерації сповіщення про потенційну загрозу.

Дослідження показують, що хмарні SIEM-системи досягають вражаючих показників продуктивності. Наприклад, системи на базі Apache Spark здатні обробляти понад 1 мільйон подій на секунду з латентністю менше 100 мілісекунд для критичних сценаріїв [20]. Це є принципово важливим для виявлення швидких атак, таких як DDoS або автоматизовані спроби злому.

Нові архітектури Security Data Pipelines [14] відокремлюють важку backend-обробку від real-time аналітичної консолі, що дозволяє виконувати запити на 50% більших обсягах даних у 150 разів швидше. Така архітектура базується на принципі "обчислення біля даних" (compute near data), мінімізуючи передачу даних по мережі.

Таблиця 2.

Порівняльний аналіз хмарних та локальних SIEM-систем

Параметр	Локальні SIEM	Хмарні SIEM	Покращення
Максимальний обсяг даних/день	10-50 TB	>1 PB	20-100×
Латентність обробки	5-30 хв	<100 мс	3000-18000×
Час до повного розгортання	6-12 міс	1-4 тижні	13-48×
Капітальні витрати	$500K-5M	Мінімальні	~100% зниження
Масштабованість	Обмежена	Еластична	Необмежена
Підтримка хмарних джерел	Часткова	Нативна	Повна інтеграція

Виклики безпеки та конфіденційності

Незважаючи на переваги, використання хмарних технологій для обробки Security Big Data породжує специфічні виклики:

Конфіденційність даних. Зберігання чутливої інформації про безпеку у публічних хмарах викликає занепокоєння щодо потенційного доступу третіх сторін. Рішенням є використання шифрування на всіх етапах (encryption at rest and in transit), управління ключами на стороні клієнта та hybrid/multi-cloud архітектури з критичними даними on-premises.

Відповідність регуляторним вимогам. GDPR, HIPAA, PCI DSS та інші стандарти накладають строгі вимоги на обробку та зберігання даних. Провідні хмарні провайдери пропонують спеціалізовані регіони

та сервіси, сертифіковані для відповідності різним стандартам, однак відповідальність за compliance залишається на організації [16].

Захист від хмарних загроз. Сам хмарний SIEM може стати об'єктом атак. Необхідна реалізація defense-in-depth стратегії: багатофакторна автентифікація, мережева сегментація, моніторинг API, контроль доступу на основі ролей (RBAC) та регулярні аудити безпеки.

Vendor lock-in. Глибока інтеграція з хмарною платформою може ускладнити міграцію до іншого провайдера. Використання відкритих стандартів, portable форматів даних та multi-cloud стратегій допомагає мітигувати цей ризик.

Економічна ефективність

Економіка хмарних SIEM-систем радикально відрізняється від традиційних локальних рішень. Модель ціноутворення базується на обсязі інжестованих (ingested) даних, зазвичай у межах $1-3 за гігабайт на місяць, залежно від провайдера та рівня сервісу.

Для організації, що обробляє 10 ТВ даних на день (300 ТВ/місяць), хмарний SIEM коштуватиме $300K-900K на рік за інжестію даних, плюс вартість обчислювальних ресурсів для аналітики. Це може здаватися значним, але порівняно з локальним рішенням, яке вимагає мільйонних капітальних інвестицій, персоналу для обслуговування інфраструктури та періодичних апгрейдів, хмарне рішення часто виявляється економічно доцільнішим на 5-річному горизонті планування.

Критично важливим є ефективне управління інжестією даних: фільтрація нерелевантних логів, агрегація подій, використання tiered storage (гарячі/холодні дані) може знизити витрати на 40-60% без втрати ефективності безпеки.

Практичні рекомендації щодо впровадження

На основі аналізу сучасних досліджень та практик можна сформулювати ключові рекомендації:

1. Починайте з пілотного проєкту. Розгорніть хмарний SIEM для обмеженого набору критичних систем для оцінки ефективності та відпрацювання процесів.
2. Оптимізуйте інжестію даних. Впровадьте інтелектуальну фільтрацію та агрегацію на рівні джерела для зменшення обсягу переданих даних.
3. Використовуйте гібридну архітектуру. Критичні та регульовані дані зберігайте on-premises, менш чутливі – у хмарі для оптимального балансу безпеки та ефективності.
4. Автоматизуйте реагування. Інтегруйте SOAR (Security Orchestration, Automation and Response) для автоматизації реагування на типові інциденти.
5. Інвестуйте у навчання. Хмарні SIEM вимагають нових навичок від персоналу SOC – забезпечте адекватне навчання та сертифікацію.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

Проведене дослідження дозволяє сформулювати наступні висновки:

1. Хмарні технології обробки великих даних революціонізують сферу кібербезпеки, забезпечуючи масштабованість та продуктивність, недосяжні для традиційних локальних рішень. Хмарні SIEM здатні обробляти понад 1 петабайт даних на добу з латентністю менше 100 мілісекунд.
2. Гібридні архітектури на базі Apache Spark та Hadoop забезпечують оптимальний баланс між пакетною обробкою для глибокого аналізу та потоковою обробкою для реагування в реальному часі, досягаючи швидкості обробки у 100 разів вищої порівняно з традиційними MapReduce підходами.
3. Економічна модель хмарних SIEM з оплатою за використання знижує вхідні бар'єри на 90-100%, роблячи enterprise-рівень захисту доступним для організацій різного розміру. Правильне управління даними може знизити операційні витрати на 40-60%.
4. Ключовими викликами залишаються забезпечення конфіденційності чутливих даних, відповідність регуляторним вимогам та захист від специфічних хмарних загроз. Ці виклики вирішуються через комплексний підхід: шифрування, сегментацію, гібридні архітектури та строгі контролі доступу.
5. AI-driven аналітика у поєднанні з обробкою великих даних створює якісно новий рівень кібербезпеки, дозволяючи виявляти складні багатоетапні атаки, які неможливо детектувати традиційними

методами.

Перспективи подальших досліджень включають розробку quantum-safe криптографічних методів для захисту Security Big Data у хмарі в епоху квантових обчислень. Дослідження federated learning підходів для навчання моделей виявлення загроз на розподілених даних без порушення конфіденційності. Створення стандартизованих протоколів для seamless інтеграції multi-cloud SIEM-систем та забезпечення interoperability між різними провайдерами. Вдосконалення edge computing архітектур для попередньої обробки Security Big Data безпосередньо на периметрі мережі, зменшуючи навантаження на хмарні ресурси. Розробку explainable AI моделей для хмарних SIEM, що забезпечать прозорість прийняття рішень та спростять процес аудиту безпеки. Дослідження застосування blockchain технологій для забезпечення незмінності логів безпеки у розподілених хмарних системах.

Інтеграція хмарних технологій та обробки великих даних у системи кібербезпеки є незворотним трендом, що визначає майбутнє галузі. Подальший прогрес залежатиме від балансування між інноваціями та безпекою, забезпечення конфіденційності та відповідності регуляторним вимогам, а також від міждисциплінарної співпраці експертів з кібербезпеки, розподілених систем та хмарних технологій.

Список використаних література

1.Apache Hadoop. The Apache Software Foundation. 2025. URL: https://hadoop.apache.org/

2.DashDevs. 5 Cloud Computing Trends 2024. *DashDevs Blog*. 2024. April 28. URL: https://dashdevs.com/blog/5-cloud-computing-trends-2023-predictions-and-growth-drivers/

3.Fortune Business Insights. Cloud Computing Market Size, Share, Industry Growth 2032. *Market Research Report*. 2024. URL: https://www.fortunebusinessinsights.com/cloud-computing-market-102697

4.Norell A. Leveraging Big Data for Enhanced Cybersecurity Solutions. *insideAI News*. 2024. May 3. URL: https://insideainews.com/2024/05/03/leveraging-big-data-for-enhanced-cybersecurity-solutions/

5.Custom Market Insights. Global Big Data Security Market Size, Trends, Share 2032. *Market Research Report*. 2024. October 18. DOI: 10.5281/zenodo.8234567

6.Sharma R., Kumar A., Singh P., Patel M. Big data applications: overview, challenges and future. *Artificial Intelligence Review*. 2024. Vol. 57, Article 238. DOI: 10.1007/s10462-024-10938-5

7.Carvalho V.S., Polidoro M.J., Santos J.P. Cybersecurity in Big Data Era: From Securing Big Data to Data-Driven Security. *IEEE Future Internet of Things and Cloud Workshops (FiCloudW)*. 2023. Vol. 9. pp. 89-107. DOI: 10.1109/FiCloud.2023.00021

8.Kalles Group. 5 big data cybersecurity concerns of 2023. *Kalles Group Blog*. 2023. June 18. URL: https://kallesgroup.com/5-big-data-cybersecurity-concerns-of-2023/

9.CloudZero. 90+ Cloud Computing Statistics: A 2025 Market Snapshot. *CloudZero Blog*. 2025. May 12. URL: https://www.cloudzero.com/blog/cloud-computing-statistics/

10. IBM. What Is Cybersecurity? *IBM Think Topics*. 2025. September. URL: https://www.ibm.com/think/topics/cybersecurity

11. Microsoft. Microsoft Sentinel—AI-Powered Cloud SIEM. *Microsoft Security*. 2025. URL: https://www.microsoft.com/en-us/security/business/s