

УДК 004.67

Красько Б.В., аспірант,  
b.v.krasko@nuwm.edu.ua

Грицюк П.М., докт. екон. наук, професор

Національний університет водного господарства та природокористування,  
p.m.hrytsiuk@nuwm.edu.ua

## ЕФЕКТИВНА АНАЛІТИКА BIG DATA ІЗ ЗАСТОСУВАННЯМ AWS EMR

Обробка великих даних (Big Data) є однією з ключових технологій сучасного світу, яка дозволяє компаніям та організаціям отримувати цінну інформацію з величезних масивів даних. AWS EMR (Elastic MapReduce) є одним із найпопулярніших рішень для аналізу Big Data у хмарному середовищі, оскільки надає можливості автоматизації, масштабування та інтеграції з іншими сервісами AWS. Його використання дозволяє обробляти великі обсяги інформації швидко, економічно та ефективно, що є важливим фактором для бізнесу, науки та державних установ.

AWS EMR забезпечує масштабованість, автоматизацію керування кластерами та підтримку різних фреймворків, таких як Apache Spark, Hadoop, Presto та HBase. Завдяки інтеграції з іншими сервісами AWS, такими як S3, Glue, Athena та Lambda, EMR дозволяє створювати високопродуктивні системи для аналізу Big Data. Ці можливості роблять його потужним інструментом для обробки великих обсягів даних, що часто вимагають складних та ресурсомістких обчислень. Інтеграція з такими сервісами, як S3 для зберігання даних або Athena для SQL-запитів без необхідності налаштування інфраструктури, значно спрощує процес обробки та аналізу даних. Окрім того, EMR може бути інтегровано з Amazon EKS (Elastic Kubernetes Service), що дає можливість запускати аналітичні навантаження в контейнеризованому середовищі, забезпечуючи ще більшу гнучкість у використанні ресурсів та можливість більш ефективно управляти інфраструктурою з застосуванням Kubernetes. Це дозволяє зберігати високий рівень продуктивності, навіть коли аналізуємо великі дані, що розподіляються через різні середовища.

Однією з ключових переваг AWS EMR є його здатність масштабувати обчислювальні ресурси залежно від навантаження. Завдяки використанню автоматичного масштабування, системи можуть динамічно додавати або видаляти вузли кластера в залежності від поточних потреб, що дозволяє забезпечити ефективне використання ресурсів і оптимізувати витрати на обчислювальні потужності. Крім того, підтримка Spot Instances значно знижує витрати на обчислення, оскільки ці інстанси можна використовувати за значно нижчою ціною порівняно з On-Demand інстансами. Це робить AWS EMR економічно вигідним рішенням для великих аналітичних завдань, де ефективність витрат має ключове значення [1].

Архітектура AWS EMR (Elastic MapReduce) складається з трьох основних типів вузлів: Master, Core та Task Nodes. Master Node виконує роль головного керівника кластера, керуючи розподілом завдань між іншими вузлами та забезпечуючи управління кластером. Core Nodes виконують основну роботу з обробки даних, а також відповідають за зберігання даних у HDFS (Hadoop Distributed File System) або S3 (Simple Storage Service). Вони здійснюють основну обробку та зберігання даних, що дозволяє зберігати дані на надійному сховищі з високою доступністю. Task Nodes виконують додаткові обчислення, які не потребують зберігання даних, але забезпечують паралельне виконання додаткових завдань для оптимізації продуктивності. Така модульна структура забезпечує гнучкість і дозволяє ефективно управляти ресурсами кластера, зокрема у випадку специфічних завдань або змін у навантаженні. Завдяки цьому можна налаштовувати кластер під конкретні потреби і зменшувати витрати на непотрібні ресурси.

Використання EKS (Elastic Kubernetes Service) у поєднанні з AWS EMR надає ще більше можливостей для масштабування та оптимізації обчислювальних завдань. Завдяки запуску

контейнеризованих обчислювальних завдань у Kubernetes, ми отримуємо можливість створювати більш портативні та ізольовані середовища для обробки даних. Такі контейнери можуть масштабуватися незалежно від інших компонентів системи, що дозволяє зручніше керувати навантаженням, виділяти ресурси та забезпечувати більш ефективний розподіл обчислювальних задач між вузлами.

AWS EMR дозволяє ефективно працювати з Data Lake, використовуючи Amazon S3 для зберігання даних. Зберігання даних у S3 надає значні переваги, оскільки воно розділяє обчислювальні ресурси та ресурси зберігання, що дозволяє більш гнучко управляти даними. Це підвищує загальну ефективність системи та дозволяє безпечно зберігати великі обсяги даних, доступних для обробки. Використання AWS Glue для створення каталогу метаданих спрощує управління великими наборами даних та забезпечує зручний доступ до метаданих, що дає змогу автоматично обробляти дані та використовувати їх для подальших операцій. Інтеграція з Amazon Athena дозволяє виконувати SQL-запити без необхідності розгортання окремих кластерів, що значно спрощує обробку даних. Крім того, AWS Lake Formation дозволяє централізовано керувати доступом до даних, що значно підвищує рівень безпеки, контролю над інформацією та дозволяє відповідати вимогам з конфіденційності [2].

Apache Spark є одним з найбільш потужних інструментів для аналізу даних у AWS EMR. Завдяки високій швидкості обробки даних та підтримці різних мов програмування, таких як Python, Scala та Java, Spark дає можливість ефективно виконувати як пакетну, так і потокову обробку даних. Це дозволяє швидко обробляти великі обсяги інформації в реальному часі, що особливо корисно для застосувань, що потребують швидкого отримання результатів. Крім того, використання таких форматів зберігання, як Parquet та ORC, дозволяє значно зменшити обсяг даних.

Оптимізація продуктивності AWS EMR включає в себе різноманітні методи та стратегії, такі як використання Data Skipping, Partitioning, кешування та налаштування параметрів кластеру. Використання Data Skipping дозволяє зменшити час обробки запитів, оскільки пропускаються частини даних, що не відповідають критеріям запиту. Partitioning дозволяє ефективно організувати дані для швидшого доступу до них, розбиваючи великі обсяги даних на більш дрібні, зручні для обробки частини. Кешування допомагає зменшити час на повторне виконання операцій, що вже були виконані раніше. Налаштування параметрів JVM та розподілу пам'яті в Spark та Hadoop може значно покращити продуктивність, даючи змогу краще управляти обчислювальними ресурсами кластера [3].

AWS EMR є потужним інструментом для аналізу Big Data, який поєднує в собі масштабованість, гнучкість та економічну ефективність. Використання автоматичного масштабування, підтримка різноманітних фреймворків та інтеграція з AWS-сервісами робить його оптимальним вибором для компаній, що працюють з великими обсягами даних. Інтеграція EMR з Kubernetes через EKS надає додаткові переваги у вигляді контейнеризації аналітичних завдань, що покращує їхню ізоляцію, безпеку та продуктивність. У майбутньому очікується подальша автоматизація процесів обробки даних, зростання ролі машинного навчання та розширення можливостей AWS EMR для аналізу та обробки складних аналітичних запитів.

#### Список посилань

1. Amazon Web Services, Inc. AWS Elastic MapReduce (EMR). Документація. URL: <https://aws.amazon.com/emr/> (дата звернення: 04.05.2025).
2. Amazon Web Services, Inc. AWS Glue. Документація. URL: <https://aws.amazon.com/glue/> (дата звернення: 29.03.2025).
3. Amazon Web Services, Inc. Amazon Athena. Документація. URL: <https://aws.amazon.com/athena/> (дата звернення: 04.05.2025).