

ДОСЛІДЖЕННЯ ШВИДКОДІ СИСТЕМИ ОЗЕРО ДАНИХ В ЗАЛЕЖНОСТІ ВІД РІЗНИХ ПОСТАЧАЛЬНИКІВ ХМАРНИХ СЕРВІСІВ

Смеляков К. С., Кузьменко О. В.

Харківський національний університет радіоелектроніки, Харків, Україна

Навчання нейронної мережі та аналіз великих даних вимагають великих наборів даних. Звичайні СКБД не підходять для таких завдань, оскільки для навчання часто необхідно зберігати сирі дані, щоб згодом над ними вже робити перетворення, досліджувати залежності та експериментувати. Стандартні СКБД сильно втрачають у швидкості пошуку та вставки даних при зростанні розміру БД. Тому була створена архітектура зберігання даних Озеро Даних [1], що не має таких недоліків. Не кожен має можливість створити сховище Озеро Даних на власних серверах, тому останнім часом стало популярним використовувати для цієї мети потужність хмарних провайдерів. На даний момент немає швидких способів розгорнути сховища Озера Даних у більшості основних постачальників хмарних послуг.

Метою доповіді є створення інформаційної системи з автоматичного розгортання Озера Даних на 3-х хмарних провайдерах: GCP, AWS, Azure, та порівняння варіантів реалізації Озера Даних на цих провайдерах за швидкістю додавання даних.

Розглядаються результати реалізації сховища Озеро Даних на трьох хмарних провайдерах і результати порівняння їх один з одним з точки зору швидкості додавання даних. Результати показують, що найшвидшою реалізацією Озера Даних є реалізація з використанням хмарного провайдера GCP та таких сервісів як BigQuery та Cloud Storage.

Другою за швидкістю була реалізація з використанням технологій OpenSearch та S3 від хмарного провайдера AWS [2]. Третьою та останньою за швидкістю стала реалізація за допомогою технологій Blob Storage та CosmosDB від Azure.

В загальному випадку рекомендовано для найкращою швидкості використовувати реалізацію від GCP, та ця реалізація має обмеження, що дані, які додаються у BigQuery не можуть змінюватися перші кілька годин після додавання. Через це в деяких випадках, де можливі зміни даних в перший час після їх додавання, рекомендується використовувати реалізацію озера даних з використанням технологій AWS.

Список літератури

1. The Enterprise Big Data Lake: Delivering the Promise of Big Data and Data Science / A. Gorelic. – London, 2019. – 224 с.
2. Aws: Amazon Web Services: The simple complete guide to mastering Amazon web services and become a cloud guru / J. Hack. – Birmingham, 2020. – 229 с.