ДОСЛІДЖЕННЯ ВПЛИВУ АРХІТЕКТУРИ НА МОЖЛИВОСТІ, ВАРТІСТЬ ТА ШВИДКОДІЮ СИСТЕМИ DATA LAKE

Смеляков К. С., Хованський О. О.
Харківський національний університет радіоелектроніки, Харків, Україна

З розвитком нейронних мереж збільшилися вимоги до сховищ даних, у яких зберігають дані для навчання моделей. Сховища повинні не встрачати у швидкості вставки та читання зі збільшенням розміру бази данних. Одним з основних джерел даних э IoT-пристрої, які генерують велику кількість запитів вставки даних. Через ці потреби з'явилося сховище даних Data Lake[1]. Також за останнє десятиліття сильно розвинулися хмарні провайдери, в яких є можливість використовувати різні послуги не витрачаючи час на гроші утримання своїх серверів. Створення Data Lake на потужності хмарного провайдера дозволяє отримати дуже надійне сховище[2]. У всіх провайдерів є безліч сервісів, тому варіантів реалізації архітектури Data Lake безліч.

Метою доповіді є створення сервісу автоматичного розгортання Data Lake на потужностях хмарних провайдерів, створення кількох варіантів архітектури Data Lake та вимірювання швидкості, вартості та доступності цих рішень.

У доповіді наводяться результати вимірів чотирьох архітектур: AWS_1, GCP_1, GCP_2, AZURE_1, на трьох провайдерах: AWS, GCP, Azure. Було виміряно швидкість вставки даних, доступність використаних сервісів у різних регіонах та розрахунки вартості кожного рішення. За результатами вимірів GCP_1 виявився найшвидшим і найдешевшим рішенням, але через обмеження використанного сервісу він вимагає додаткових витрат на утримання MongoDB сервісу.

Без урахування GCP_1 найкращою архітектурою виявилась AWS_1 яка використовує S3 та OpenSearch, трохи повільніше і дорожче виявилася GCP_2 яка використовує CloudStorage та BigTable і абсолютним аутсайдером виявилася AZURE_1 яка використовує BlobStorage, CosmosDB.

Доступність регіонів у всіх провайдерів практично ідентична. У GCP також є програма LOW CO2.

Сервера з такою позначкою мають показник CFE% понад 75%.

Список літератури

1. E. Zagan and M. Danubianu, "Cloud DATA LAKE: The new trend of data storage," 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), 2021, pp. 1-4, doi: 10.1109/HORA52670.2021.9461293.

2. S. Park, B. Cha and J. Kim, "Design and Implementation of Connected DataLake System for Reliable Data Transmission," 2019 23rd International Computer Science and Engineering Conference (ICSEC), 2019, pp. 141-144, doi: 10.1109/ICSEC47112.2019.8974823.