

A serverless framework with DeepAR-based predictive autoscaling for distributed computing

Oleksandr Kyrychenko

Yuriy Fedkovych Chernivtsi National University, Chernivtsi
<https://orcid.org/0009-0001-6982-3342>

Serhii Ostapov

Yuriy Fedkovych Chernivtsi National University, Chernivtsi
<https://orcid.org/0000-0002-4139-4152>

Oksana L. Kyrychenko

Yuriy Fedkovych Chernivtsi National University, Chernivtsi
<https://orcid.org/0000-0003-0282-9958>

Abstract. *The paper proposes a framework that implements predictive scaling of computational resources for serverless applications in distributed computing using queues. The predictive scaling mechanism is based on the DeepAR algorithm from AWS SageMaker. The proposed solution allows dynamically allocating and releasing resources as needed while remaining within the serverless computing paradigm. Testing results of the serverless application built on this framework demonstrated reduced cold starts and unprocessed requests.*

Keywords: *serverless, predictive modelling, predictive autoscaling, machine learning, cloud.*

Безсерверний фреймворк із прогнозним масштабуванням на основі ДеєрАР для розподілених обчислень

Олександр Кириченко

*Чернівецький національний
 університет імені Юрія Федьковича, м. Чернівці*
<https://orcid.org/0009-0001-6982-3342>

Сергій Остапов

*Чернівецький національний
 університет імені Юрія Федьковича, м. Чернівці*
<https://orcid.org/0000-0002-4139-4152>

Оксана Кириченко

*Чернівецький національний
 університет імені Юрія Федьковича, м. Чернівці*
<https://orcid.org/0000-0003-0282-9958>

Анотація. *У роботі пропонується фреймворк, що реалізує прогнозне масштабування обчислювальних ресурсів для безсерверних додатків при розподілених обчисленнях з використанням черг. Механізм прогнозного масштабування реалізований на основі алгоритму DeepAR з AWS SageMaker. Запропоноване рішення дозволяє динамічно виділяти та вивільняти ресурси в разі потреби залишаючись при цьому в парадигмі безсерверних обчислень. Результати тестування безсерверного додатку, створеного на основі фреймворку, продемонстрували зменшення кількості холодних стартів та необроблених запитів.*

Ключові слова: *безсерверні обчислення, прогнозне моделювання, прогнозне автоматичне масштабування, машинне навчання, хмарні обчислення.*

Швидке поширення використання безсерверних архітектур при побудові рішень для розподілених обчислень викликає потребу в зовсім інших підходах для вирішення проблеми автоматичного масштабування обчислювальних ресурсів. Однією з відмінностей безсерверних платформ є здатність адаптуватися до змін у навантаженні. Таким чином, обчислювальні ресурси у користувацьких додатках можуть повністю вивільнятися вразі відсутності вхідних даних і знову виділятися при сплесках у трафіку. Такий підхід забезпечує необхідну гнучкість при обчисленнях. Однак, проблемою залишається значний час ініціалізації обробників даних, який витрачається на розгортання та налаштування середовища виконання [1].

Проблема тривалої ініціалізації, відома як проблема холодного старту, є надзвичайно критичною для безсерверних додатків, де час виконання, як правило, є незначним у порівнянні з часом ініціалізації середовища виконання. Хмарні провайдери, зазвичай, не видаляють створені контейнери після попередніх запусків протягом певного періоду часу з метою обробки нових запитів, але такий підхід не завжди є економічно ефективним, оскільки кінцевий користувач сплачує за час фактичного використання ресурсів залишаючи витрати часу простою хмарному провайдеру [2, 3].

Для вирішення цієї проблеми необхідно застосовувати підходи, які дозволяють прогнозувати майбутні робочі навантаження та завчасно розгортати необхідні обчислювальні потужності. Це особливо критично для систем, де події надходять асинхронно через черги повідомлень.

Метою дослідження є розробка безсерверного рішення для розподіленої обробки даних з використанням черг повідомлень, яке на основі метрик та алгоритмів машинного навчання здійснює автоматичне масштабування обчислювальних ресурсів.

Розроблений фреймворк складається з наступних компонент:

AWS SQS для управління обробкою подій [4];

AWS Lambda для обробки даних [5];

AWS CloudWatch для моніторингу метрик черги [6];

AWS EventBridge для запуску обробників даних згідно розкладу [7];

AWS SageMaker для побудови та розгортання моделей машинного навчання [8].

Для якості моделі використовується DeepAR [9], алгоритм навчання для прогнозування часових рядів на основі рекурентних нейронних мереж. Перевагою зазначеного алгоритму є можливість вивчати типову поведінку на основі кількох пов'язаних часових рядів, що сприяє покращенню точності прогнозу [10].

Прогнозування здійснюється на основі двох обраних метрик:

- ApproximateNumberOfMessagesVisible – довжина черги.
- NumberOfMessagesSent – кількість отриманих повідомлень [4].

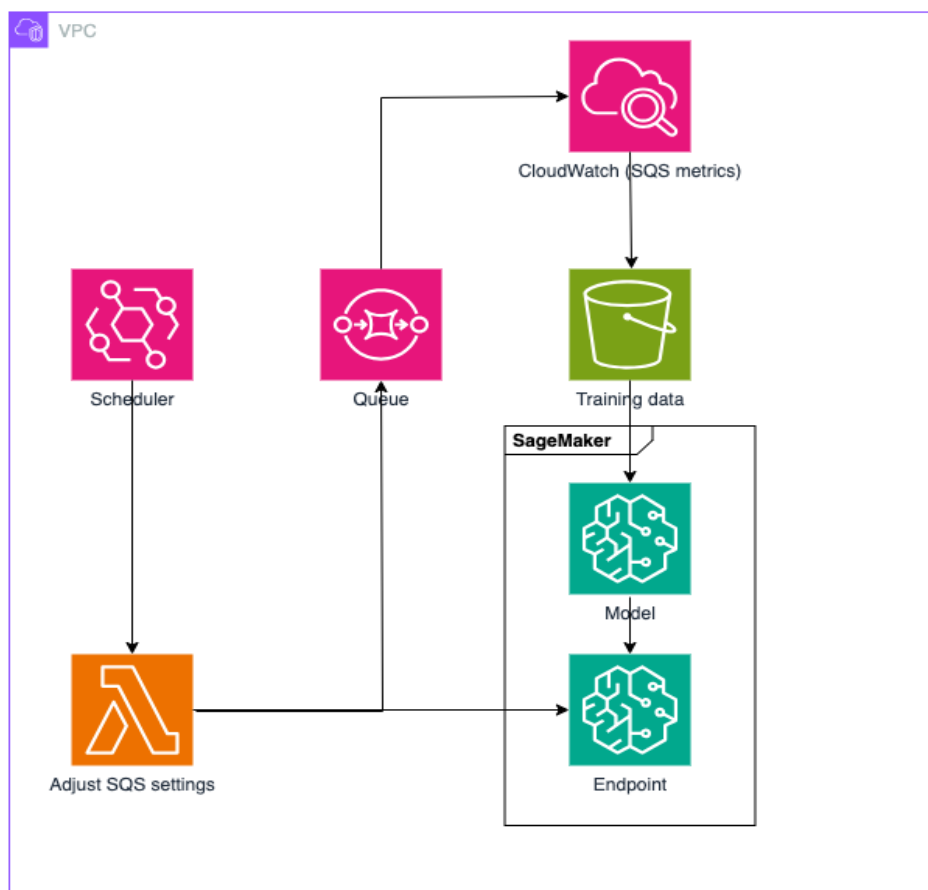


Рис. 1. Архітектура фреймворку для автоматичного масштабування обробників даних

Результати прогнозування використовуються для зміни параметрів Provisioned Concurrency функцій AWS Lambda залежно від очікуваного навантаження.

Для тестування фреймворку реалізовано безсерверний додаток для розподіленої обробки даних з розрахунковим навантаженням 1 млн. записів. Навантаження емулювалося за допомогою Locust – ефективного інструмента для навантажувального тестування розподілених систем [11]. Модель DeepAR навчалася на зібраних метриках AWS SQS. Мінімальний обсяг історичних даних, який модель повинна використовувати для прогнозування становив 30 кроків, а довжина вихідного прогнозу – 50 хвилин [12]. Точність моделі оцінювалася за значеннями середньої абсолютної похибки (MAE) та кореня середньоквадратичної помилки (RMSE).

Варто зазначити, що застосування прогнозного автомасштабування дозволило зменшити кількість холодних стартів до 27% та кількість необроблених запитів до 14%. При цьому побудована модель продемонструвала прийнятну якість прогнозування зі значенням середньої абсолютної похибки (MAE) 10019,21 та кореня середньоквадратичної похибки (RMSE) 13140,91.

Таким чином, використання запропонованого фреймворку при розробці безсерверних додатків сприяє оптимальному використанню обчислювальних ресурсів не виходячи за рамки концепції безсерверних обчислень.

Список використаних джерел

1. Thumala, S. (2020). Building Highly Resilient Architectures in the Cloud. *Nanotechnology Perceptions*, 16(2), 251-271.
2. Mampage, A. R. (2023). *Autonomous Resource Management for Serverless Computing* [PhD thesis, The University of Melbourne], 248. <http://clouds.cis.unimelb.edu.au/students/AnupamaPhDThesis2023.pdf>.
3. Alharthi, S., Alshamsi, A., Alseiari, A., & Alwarafy, A. (2024). Auto-Scaling Techniques in Cloud Computing: Issues and Research Directions. *Sensors*, 24(17), 5551. <https://doi.org/10.3390/s24175551>.
4. Amazon Web Services. (2025). *Amazon Simple Queue Service*. <https://aws.amazon.com/sqs>.
5. Amazon Web Services. (2025). *What is AWS Lambda?* <https://docs.aws.amazon.com/lambda/latest/dg/welcome.html>.
6. Amazon Web Services. (2025). *What is Amazon CloudWatch?* <https://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/WhatIsCloudWath.html>.
7. Amazon Web Services. (2024). *Amazon EventBridge*. <https://aws.amazon.com/eventbridge>.
8. Amazon Web Services. (2024). *Amazon SageMaker*. <https://aws.amazon.com/sagemaker>.
9. Amazon Web Services. (2025). *Use the SageMaker AI DeepAR forecasting algorithm*. <https://docs.aws.amazon.com/sagemaker/latest/dg/deepar.html>.
10. Januschowski, T., Arpin, D., Salinas, D., Flunkert, V., Gasthaus, J., Stella, L., & Vazquez, P. (2018, April 23). *Now available in Amazon SageMaker: DeepAR algorithm for more accurate time series forecasting*. AWS Machine Learning Blog. <https://aws.amazon.com/blogs/machine-learning/now-available-in-amazon-sagemaker-deepar-algorithm-for-more-accurate-time-series-forecasting>.
11. Heyman, J., Holmberg, L., Baldwin, A., Byström, C., Hamrén, J., & Heyman, H. (2025). *What is Locust?* <https://docs.locust.io/en/stable/what-is-locust.html>.
12. Kyrychenko, O., Ostapov, S., & Kyrychenko, O. L. (2025, April 4). Predictive autoscaling in AWS serverless by means of machine learning and SQS metrics. In *The 6th International Workshop on Intelligent Information Technologies & Systems of Information Security (IntelliTISIS 2025)* (Vol. 3963, pp. 77–87). CEUR-WS. <https://ceur-ws.org/Vol-3963>.

