УДК 004.051

ОЦІНКА ЕФЕКТИВНОСТІ ВЕБ-СЕРВІСІВ ЗА ТЕХНОЛОГІЄЮ БЕЗСЕРВЕРНИХ ХМАРНИХ ОБЧИСЛЕНЬ

Черешня В.Р., Курдеча В.В.
Інститут телекомунікаційних систем КПІ ім. Ігоря Сікорського, Україна
E-mail: vitas.cherry@gmail.com

Evaluating the effectiveness of web services based on the serverless cloud computing technology

The basic metrics of the web services effectiveness are summarized. The results of the selection of criteria that are suitable for web services based on the serverless cloud computing technology are presented.

Безсерверні обчислення, або просто Serverless, є гарячою темою у світі архітектури програмного забезпечення. «Велика трійка» вендорів – Amazon, Google і Microsoft – активно інвестують у безсерверні технології та вдосконалюють свої платформи для розробників [1].

Попри безліч плюсів в використанні безсерверних обчислень, як і в традиційних веб-сервісах в них існують слабкі місця. Щоб ідентифікувати ці слабкі місця потрібно вміти правильно оцінити ефективність веб-сервісу.

Коли йдеться про високу ефективність (або продуктивність) веб-сервісу, може бути задіяно один або декілька таких факторів:

- Короткий час відгуку для даної частини роботи.
- Низьке використання обчислювального ресурсу.
- Висока доступність веб-сервісу (англ. High availability).
- Швидке (або дуже компактне) стиснення і декомпресія даних.

Метрики (показники які можна виміряти) ефективності веб-сервісів включають: затримка, доступність, коефіцієнт стиснення, швидкість обробки, масштабованість, час відгуку тощо.

Затримка. Проявляється у двох випадках – при передачі та обробці данних веб-сервісом. Затримка передачі даних (англ. Round-trip delay time, також відомий як час пінгування) залежить тільки від швидкості та надійності інтернет-з’єднання. Під затримкою обробки запиту веб-сервісом мається на увазі як довго запит повинен чекати в черзі перед тим як його передадуть на виконання.

Доступність. Суть властивості полягає в тому, що потрібний інформаційний ресурс знаходиться у вигляді, необхідному користувачеві, в місці, необхідному користувачеві, і в той час, коли він йому необхідний. Простіше кажучи, доступність – це частка часу, коли веб-сервіс перебуває у стані функціонування.

Коефіцієнт стиснення. Коефіцієнт стиснення характеризує ступінь стиснення та в загальному випадку дорівнює відношенню обсягу пам'яті, необхідної для зберігання вихідної (результуючої) послідовності даних, до

обсягу памяті вхідної послідовності даних. Тому чим менше значення коефіцієнта стиснення, тим ефективніший метод стиснення.

Стиснення відбувається на трьох різних рівнях:

- спочатку деякі формати файлів стискаються за допомогою конкретних оптимізованих методів,
- тоді загальне шифрування може відбуватися на рівні HTTP (ресурс передається стиснутим від одного кінця до іншого),
- і, нарешті, стиснення може бути визначене на рівні з'єднання між двома вузлами HTTP-з'єднання.

Швидкість обробки (FLOPS). Дана величина визначається шляхом запуску на обчислювальній машині тестової програми, яка вирішує задачу з відомою кількістю операцій та підраховує час, за який вона була вирішена.

Масштабованість. Коли мова йде про веб-сервіси, ця властивість означає, наскільки швидко та легко веб-сервіс може розширитись для задоволення великого обсягу користувачів. Масштабованість може бути по вертикалі – перейти на більш потужний сервер, або по горизонталі – додати більше серверів. Проектування горизонтальної масштабованості більш бажано, якщо планується підтримка великої кількості користувачів. В такому випадку механізм балансування навантаження (англ. Load balancing) використовується для розподілу запитів над набором серверів.

Час відгуку. В загальному випадку, час відгуку – це сума трьох чисел:

- Часу обслуговування – скільки часу потрібно щоб виконати роботу.
- Часу очікування – як довго запит повинен чекати в черзі.
- Часу передачі – скільки часу потрібно щоб доставити запит до веб-сервісу та відповідь назад користувачу по мережі.

Вимірювання часу обслуговування веб-сервіса зображено на рис. 1.

Рис. 1. Вимірювання часу обслуговування веб-сервіса.

Вибір метрик.

Коли мова йде за сумісність метрик з веб-сервісами, то мається на увазі що, деякі з них можуть бути не корисними через особливості реалізації безсерверної архітектури. Вибір критеріїв, що мають найбільший вплив на задоволеність кінцевого споживача інформаційного ресурсу наведено в табл. 1.

Таблиця 1. Сумісність різних метрик ефективності з веб-сервісами за технологією безсерверних хмарних обчислень.

Назва критерію	Сумісність	Пояснення
Доступність	Ні	Хмарні технології гарантують високу надійність, захищеність та доступність прямо «з коробки» [2].
Масштабованість	Ні	Через особливість реалізації безсерверної архітектури (один запит – один вільний екземпляр веб-сервісу) маштабування по горизонталі виконується автоматично.
Швидкість обробки	Ні	Через особливість реалізації безсерверної архітектури за однакових тестових умов результати замірів швидкості обробки не мають відрізнятися.
Затримка	Ні	У випадку затримки передачі даних проблема вирішується автоматичним додованням екземпляра веб-сервіса у потрібний регіон чи зону доступу – для зменшення затримки через відстань, помилки передачі і обмежень пропускної здатності. А через особливість реалізації безсерверної архітектури (кожен запит обслуговується окремим вільним екземпляром веб-сервісу) затримки обробки взагалі не буде.
Коефіцієнт стиснення	Ні	З експертної точки зору приріст у продуктивності при ефективному стисненні даних буде мізерним у порівнянні з головною проблемою – затримкою при ініціалізації нового екземпляра веб-сервіса для нового запиту, якщо всі інші екземпляри вже зайняті.
Час відгуку	Так	Ігноруючи час передачі (вважаємо з'єднання дуже швидким і надійним), час відгуку – це сума часу обслуговування і часу очікування. Так як кожний запит обслуговується тільки одиним вільним інстансом сервісу то і часом очікування можна знехтувати. В результаті маємо: час відгуку дорівнює часу обслуговування.

У результаті, щоб оцінити ефективність веб-сервісу потрібно виміряти час обслуговування (подібні метрики надаються платформою вендора [2]), який залежить від часу ініціалізації середовища та часу виконання коду. Час виконання можна покращити, замінивши код на більш швидкодіючий, а от час ініціалізації – тільки розробивши метод «розігрівання» [3] веб-сервіса.

Література

1. D. Smith, L. Leong, R. Bala, Magic Quadrant for Cloud Infrastructure as a Service, Worldwide / Gartner, 2018.
2. Serverless Architectures. AWS Lambda and Fn Project / F. Munz // presented at Devoxx, Casablanca, Morocco, 2017.
3. Become a Serverless Black Belt: Optimizing Your Serverless Applications – SRV401 / presented at AWS re:Invent, Las Vegas, USA, 2017.