Крещук М.С. *— рецензент Харченко К.В.*
ННК "Інститут прикладного системного аналізу" НТУУ "КПІ", Київ, Україна

Використання бібліотек MapReduce для декомпозиції задач в Google App Engine

Сучасні хмарні системи отримали широке розповсюдження серед користувачів з різним рівнем підготовки. Найпростішими прикладами використання є поштові сервіси та онлайн офісні пакети. Хмарні обчислення представляють послуги, які поділяються на такі категорії: SaaS (Програмне забезпечення як сервіс), PaaS (Платформа як сервіс) та IaaS (Інфраструктура як сервіс). Для організації декомпозиції обчислювальних задач великої ємності в системах, що масштабуються, можна використовувати категорії PaaS та IaaS. Найбільш зручним з точки зору прикладних додатків є використання PaaS [1, 2, 3]. Основними характеристиками PaaS є:

- Оплата лише за ті послуги, які необхідні і лише за ті, що використовуються;
- Відсутність витрат на придбання, підтримку і модернізацію програмного забезпечення і устаткування;
- Масштабованість - автоматичне виділення та звільнення необхідних ресурсів в залежності від кількості користувачів, що обслуговуються;
- Інтеграція веб-сервісів і баз даних, використання поширених веб-стандартів, можливість інтеграції сервісів, розташованих у приватних мережах;
- Послуга доступна всюди, де є інтернет [4].

Найпопулярнішими PaaS є Amazon S3, Windows Azure та Google App Engine. Для досліджень задач декомпозиції обрану останню, оскільки вона надає можливість використовувати безкоштовний доступ, якщо квоти не перевищуються. На відміну від багатьох IaaS для розміщень додатків на віртуальних машинах, таких як Amazon EC2, платформа Google App Engine тісно інтегрована з додатками і накладає на розробників деякі обмеження. Вона забороняє використовувати системні функції. Запит від користувача в системі повинен обробитися і повернути користувачеві результат менше ніж за 30 секунд, інакше він буде примусово завершений. Ще одне обмеження PaaS Google App Engine – це встановлені щохвилинні квоти, при перевищенні яких запит чи завдання також завершується [5].

Сучасна обчислювальна парадигма MapReduce дозволяє ефективно використовувати метод декомпозиції задачі. Згідно неї, задача поділяється на певну кількість однакових елементарних завдань, які виконуються на вузлах кластера і потім зводяться в кінцевий результат.[6] Найбільш розповсюдженою реалізацією цієї парадигми є Apache Hadoop.[7] Для використання на PaaS Google App Engine існує бібліотека AppEngine-MapReduce.[8] На Рис. 1 показана ілюстрація роботи бібліотеки. Ця бібліотека розроблена з урахуванням особливостей Google App Engine, а саме оптимізована для роботи з квотами системи. Вона представляє такі можливості:

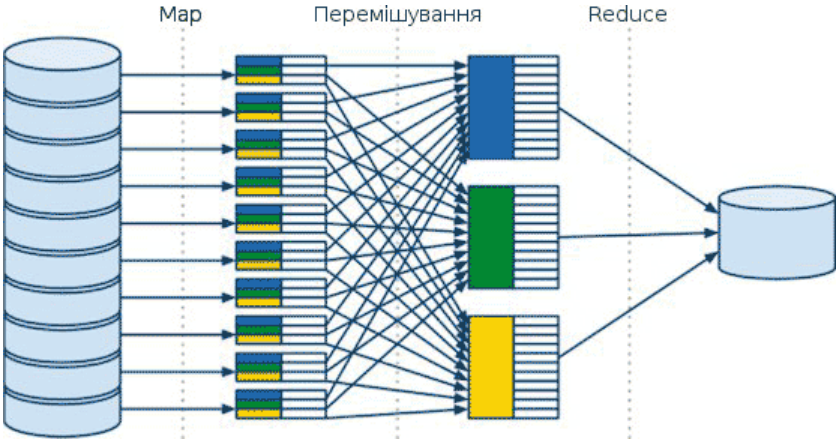

Рис. 1. Ілюстрація роботи AppEngine-MapReduce.

- Моніторинг обмежень швидкості роботи, який сповільнює роботу функцій додатку та запобігає перевищенню квот.
- Автоматичне розбиття на задачі для більш швидкого виконання. Це надає можливість запускати одночасно стільки підзадач, скільки потрібно для найшвидшого отримання результатів.
- Стандартні механізми запису/читання даних для роботи додатків з базою даних та BlobStore(механізм зберігання великих об'єктів у двійковому форматі).
- Сторінка статусу для перегляду задач, що виконуються.[8]

Перевагами використання MapReduce в PaaS є можливість обробки значних об'ємів даних та паралельне виконання попередньої обробки даних.[4]

На Рис. 2 показано роботу бібліотеки AppEngine-MapReduce. Тестова програма читає з БД 4000 записів та проводить операції з ними. На графіку видно, що для такої малої кількості записів при збільшенні кількості потоків спостерігається збільшення часу виконання замість зменшення. Це відбувається тому, що 4000 записів - це відносно мала задача і тому бібліотека не розподіляє її на потоки достатньо рівномірно.

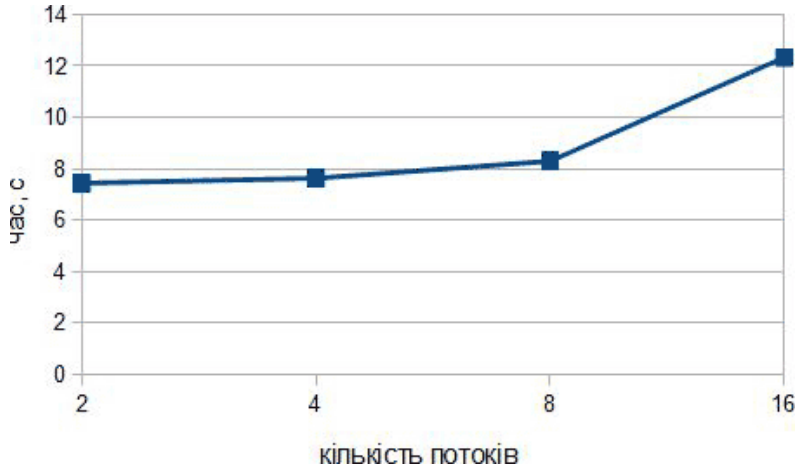

Рис. 2. Залежність часу виконання від кількості потоків

Таким чином, MapReduce можна використовувати для задач з явною паралельною структурою. Найкраще бібліотека працює з порівняно великими об'ємами даних.

Література. **1.** Что такое платформа-как-сервис (Paas), и почему это правильный выбор для новых приложений? [Электронный ресурс]: — Режим доступа: http://www.techdays.ru/videos/4136.html. – Дата доступа 27.02.2012. **2.** Дискуссия на тему полезности PaaS с точки зрения SaaS-предпринимателя. [Электронный ресурс]: — Режим доступа:http://smartsourcing.ru/blogs/itogi_nekonferentsii_aas_predprinimateley_2011/610. – Дата доступа 27.02.2012. **3.** Bernard Golden. Cloud Computing: What You Need to Know About PaaS. / Bernard Golden //CIO Magazine. – 2011. – №07. – p 35. **4.** PaaS — Википедия [Электронный ресурс]: — Режим доступа:http://ru.wikipedia.org/wiki/PaaS. – Дата доступа 27.02.2012. – Название с экрана. **5.** Google App Engine — Википедия [Электронный ресурс]: — Режим доступа:http://ru.wikipedia.org/wiki/Google_App_Engine. – Дата доступа 27.02.2012. – Название с экрана. **6.** MapReduce — Википедия [Электронный ресурс]: — Режим доступа:http://ru.wikipedia.org/wiki/MapReduce. – Дата доступа 27.02.2012. – Название с экрана. **7.** Apache Hadoop[Электронный ресурс]: — Режим доступа:http://hadoop.apache.org. – Дата доступа 27.02.2012. **8.** MapReduce Overview[Электронный ресурс]: — Режим доступа:http://code.google.com/intl/ru/appengine/docs/python/dataprocessing/overview.html. – Дата доступа 27.02.2012.