УДК 004.75

Є.С. Сакало

Харківський національний університет радіоелектроніки, Харків

ЗАСТОСУВАННЯ ХМАРНИХ ОБЧИСЛЕНЬ ДЛЯ ПОШУКОВИХ СИСТЕМ

У статті досліджується проблема обробки великої кількості даних, зберігання даних в розподілених кластерних системах. Аналізуються існуючі способи обробки даних за допомогою хмарних обчислень, принципи їх роботи. Пропонується розробка сервісу, котрий вирішить такі завдання, як зміна та розподіл даних між процесорами, балансування навантаження, обробка відмов, збір і агрегація проміжних результатів.

Ключові слова: *хмарні обчислення, модель програмування MapReduce, пошукові системи, розподілені кластерні системи.*

Вступ

Сучасні завдання та програми, пов'язані з аналізом даних, пред'являють особливі вимоги до обчислювальних ресурсів, значно перевищують можливості окремих комп'ютерів. В даний час Web нараховує десятки мільярдів сторінок. Роботи пошукової системи цілодобово завантажують петабайт даних з вмістом нових і змінених Web-сторінок. Завантажені дані піддаються різним процедурам обробки, пов'язаних з побудовою індексу Web, обчисленням індексів цитування окремих сторінок і т.д.[1] Іншою областю, де поширені програми, пов'язані зі зберіганням і обробкою великих обсягів даних, є наукові обчислення. Найчастіше це виражається в необхідності проведення трудомісткого аналізу зібраних масивів експериментальних даних для отримання нових наукових результатів. [1],Постійне зростання обсягів оброблюваних даних вимагає відповідного нарощування обчислювальних ресурсів, у зв'язку з чим використовується обчислювальна середа, яка повинна мати високу масштабованість. В даний час в області масованої обробки даних спостерігається перехід від спеціалізованих суперкомп'ютерних архітектур до більш економічних і масштабованих, але менш надійних кластерним системам з недорогих серверів масового виробництва. [1,2]

Необхідність безпосередньої реалізації даних механізмів при програмуванні процедури обробки даних є серйозною перешкодою на шляху широкого впровадження подібних систем. Тому необхідно, щоб відповідні технології вже містили в собі реалізації даних механізмів та надавали користувачеві високорівневі моделі програмування[2], що приховують від нього деталі реалізації обчислень в ненадійному розподіленому середовищі. У статті будуть розглянуті моделі програмування та технології розподіленої обробки даних, розподілені системи зберігання даних та інфраструктурні сервіси, буде обрана хмарна платформа та спроектована програмна система. Для цього необхідно розробити сервіс, користувачами якого будуть розробники. Сервіс повинен мати виключно програмну реалізацію. Він також повинен представляти собою набір методів для роботи з Amazon Web Services.

При використанні цього сервісу має з'явитись можливість розробляти програми для обробки даних, таких як: web indexing, data mining, log file analysis, data warehousing, machine learning, financial analysis, scientific simulation, and bioinformatics research на будь-якій зручній мові програмування: Java, Ruby, Perl, Python, PHP, R, or C + +.

1. Приклади сценаріїв використання хмарних обчислень

Сценарій 1: Хмарний споживач може запросити послугу (сервіс) у хмарного брокера замість прямого контактування з хмарним провайдером. Хмарний брокер може створити новий сервіс, комбінуючи набір сервісів або розширюючи існуючий сервіс. У цьому прикладі хмарний провайдер невидимий хмарному споживачеві (рис. 1).

Рис. 1. Участь хмарного брокера
у взаємодії споживача з провайдером

Сценарій 2: Хмарний оператор зв'язку надає послуги підключення і транспорт "доставки" хмарних послуг від хмарного провайдера хмарному споживачеві. Хмарний провайдер встановлює угоду про рівень обслуговування SLA з хмарним оператором і може запитувати виділені і захищені з'єднання (рис. 2).

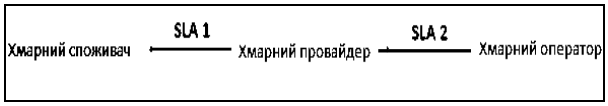

Рис. 2. Участь хмарного оператора
у наданні провайдером послуг

Сценарій 3: Хмарний аудитор проводить незалежну оцінку обслуговування і безпеки реалізації хмарної послуги (рис. 3).

Рис. 3. Участь хмарного аудитора
в оцінці надання провайдером послуг

2. Хмарний сервіс-менеджмент

Хмарний сервіс-менеджмент (Cloud Service Management) включає всі пов'язані з сервісом функції, необхідні для управління і функціонування сервісів, необхідних або пропонованих хмарним споживачам [5,6].

Управління замовниками (Customer Management) – це управління обліковими записами користувачів, відкриття або закриття або припинення дії облікових записів, управління профілями користувачів, управління взаємодією з замовниками (customer relationship) на основі надання контактів і вирішення питань та проблем замовників, і т.п.

Управління контрактами (Contract Management) - це управління сервісними контрактами, висновок або закриття або припинення дії контрактів, і т.п.

Управління постачанням (Inventory Management) - ценастройка і управління каталогом послуг, і т.п.

Бухгалтерія і нарахування (Accounting and Billing) – це управління платіжної інформацією за замовниками, відправка рахунків на оплату, обробка отримання платежів, відстеження рахунків, і т.п.

Звітність і аудит (Reporting and Audit) – це моніторинг дій користувачів, генерація звітів, тощо.

Ціноутворення і тарифікація (Pricing and Rating) - це оцінка хмарних послуг і визначення цін, обробка спеціальних пропозицій і правил ціноутворення, заснованих на профілі користувача, і т.п.

Конфігурування полягає в резервуванні або конфігуруванні (Provisioning чи Configuration) - включає всі пов'язані з сервісом функції, необхідні для управління і функціонування сервісів, необхідних або пропонованих хмарним споживачам.

Швидке резервування (Rapid Provisioning) – це автоматичне розгортання хмарних систем на основі запитів сервісів або ресурсів або можливостей.

Модифікація ресурсів (Resource Change) - налаштування конфігурацій або призначення ресурсів для відновлення, оновлення і підключення нових вузлів в хмару.

Моніторинг та звітність (Monitoring and Reporting) - цевиявлення та моніторинг віртуальних ресурсів, моніторинг функціонування (дій і подій) хмари та генерація звітів про продуктивність. Вимірювання показників (Metering) – це надання можливостей кількісних вимірів на рівні абстракції, відповідному типу сервісу (наприклад, засобів зберігання, обробки, пропускної здатності та активних облікових записів користувачів).

Управління рівнем обслуговування (SLA Management) – це визначення параметрів SLA контракту (схема з параметрами якості сервісу - QoS), моніторинг SLA, застосування SLA відповідно до заданих політиками.

3. Модель програмування MapReduce

MapReduce - модель програмування і платформа для пакетної обробки великих обсягів даних, розроблена і використовувана всередині компанії Google для широкого кола додатків. Модель MapeReduce відрізняється простотою і зручністю використання, приховуючи від користувача деталі організації обчислень в ненадійно розподіленому середовищі. Користувачеві досить описати процедуру обробки даних у вигляді двох функцій - map і reduce, після чого система автоматично розподіляє обчислення за кластером з великої кількості машин, обробляє відмови машин, балансує навантаження і координує взаємодії між машинами для ефективного використання мережевих і дискових ресурсів.[3] Реалізація MapReduce в Google орієнтована на обчислювальну інфраструктуру, що складається з великого числа недорогих серверів з масових комплектуючих. Як правило, це двопроцесорні x86-машини з 4-8 Гб оперативної пам'яті, що працюють під управлінням Linux. Для з'єднання машин в кластері використовується комутований Gigabit Ethernet. Кластери складаються з тисяч машин, тому постійно виникають відмови окремих вузлів. Для зберігання даних використовуються недорогі IDE-диски, підключені до кожної з машин. Дані зберігаються під управлінням розподіленої файлової системи GFS. Слід відзначити, що GFS використовує реплікацію для надійного зберігання даних. Запуском MapReduce-завдань [4] на кластері управляє планувальник, який відстежує стан машин і підбирає групу машин для виконання завдання.

Модель MapReduce накладає ряд обмежень на програму для того, щоб автоматизувати розпаралелювання, запуск і керування обчисленнями на кластері. З одного боку, це значно спрощує задачу програміста і практично не вимагає від нього спеціальної кваліфікації. З іншого боку обмеження, що накладаються системою не дозволяють реалізувати в ній рішення довільних завдань. Наприклад, в рамках описаної моделі не можна простим чином реалізувати операції типу JOIN і SPLIT або організувати взаємодію між паралельними процесами так, як це робиться в технології MPI.

4. Опис програмної реалізації

Для сборки проекту(сервісу) використовується Maven. Apache Maven - фреймворк для автоматизації складання проектів, специфікованих на XML-мові POM. Maven забезпечує декларативну, а не імперативну збірку проекту. Тобто, в файлах проекту pom.xml міститься його декларативне опис, а не окремі команди. Усі завдання з обробки файлів Maven виконує через плагіни. Основним фреймворком сервісу є Spring. Spring Framework забезпечує вирішення багатьох завдань, з якими стикаються Java розробники та організації, які хочуть створити інформаційну систему, засновану на платформі Java. Через широку функціональність важко визначити найбільш значущі структурні елементи, з яких він складається [7]. Spring Framework не цілком пов'язаний з платформою Java Enterprise, незважаючи на його масштабну інтеграцію з нею, що є важливою причиною його популярності.

Основним критерієм для реалізації даного сервісу є Amazon Elastic MapReduce API, який дозволяє програмно реалізувати можливості Amazon, а саме можливість користуватись фреймворком Hadoop.

Для роботи зі своїми сервісами Amazon потребує авторизації. Amazon API для цього надає інтерфейс AWSCredentials. Сервіс буде його використовувати створюючи свій клас для авторизації Amazon Credentials. Його задача користуватися специфічними ключами для авторизації: accessKey та secretAccessKey, які видаються користувачам Amazon.

Висновок

В результаті розробленого проекту були отримані наступні результати :

1. Розроблений сервіс обміну даними за допомогою хмарних обчислень.
2. Досліджена та використана технологія користування Hadoop.
3. Застосована модель програмування MapReduce і платформа для пакетної обробки великих обсягів даних

Після проведених досліджень було виявлено, що на даний момент Amazon Web Services є найкращим IaaS сервісом. У зв'язку з цим для розробки власного сервісу для обробки великих обсягів даних використовувалася саме ця пропозиція моделі IaaS. Основним плюсом є наявність Amazon Elastic MapReduce API, що надає можливість запрограмувати необхідні можливості хмарного сервісу [8].

Внаслідок цього був розроблений сервіс, користувачами якого будуть розробники. Сервіс має виключно програмну реалізацію. Він представляє собою набір методів для роботи з Amazon Web Services.

Список літератури

1. Barroso, L. A. Web search for a planet: The Google cluster architecture / L. .Barroso, J.Dean-IEEE Micro, 2003 – 56 с.

2. GSA, Cloud Computing Initiative Vision and Strategy Document (DRAFT) [Електронний ресурс] - Режим доступу: www/ URL: http://info.apps.gov/sites/default/files/Cloud_Computing_Strategy_0.ppt - 17.03.2012 г. - Загол. з екрану.

3.Dean, J. MapReduce: Simplified data processing on large clusters /J. Dean-Commun ACM, 2008- С. 107-113.

4.Dean, J. MapReduce: Simplified data processing on large clusters / J. Dean, S.Ghemawat - San Francisco, CA, 2004 - С. 137-150.

5.Ghemawat, S.The Google file system. In 19th Symposium on Operating Systems Principles / S.Ghemawat, H.Gobioff - Lake George, NY, 2003 - С. 29-43.

6.Michael, I. Dryad: Distributed Data-Parallel Programs from Sequential Building Blocks/ I. Michael, B. Mihai -Lisbon, Portugal, 2007 -С. 21-23.

7. IBM, "Cloud Computing Reference Architecture v2.0" [Електронний ресурс] - Режим доступу: www/ URL:http://www.opengroup.org/cloudcomputing/doc.tpl?CALLER=documents .tpl&dcat=15&gdid=23840 - 11.02.2012 г.

8. The Hadoop Distributed File System: Architecture and Design [Електронний ресурс] - Режим доступу: www/URL:http://hadoop.apache.org/core/docs/current/hdfs_design.html - 6.01.2012 г. - Загол. з екрану.

Надійшла до редколегії 2.03.2013

Рецензент: д-р техн. наук, проф. Є.В. Путятін, Харківський національний університет радіоелектроніки, Харків.

ПРИМЕНЕНИЕ ОБЛАЧНЫХ ВЫЧИСЛЕНИЙ ДЛЯ ПОИСКОВЫХ СИСТЕМ

Е.С. Сакало

В статье исследуется проблема обработки большого количества данных, хранения данных в распределенных кластерных системах. Анализируются существующие способы обработки данных с помощью облачных вычислений, принципы их работы. Предлагается разработка сервиса, который может решить такие задачи, как изменение и распределение данных между процессорами, балансировка загрузки, обработка отказов, сбор и агрегация промежуточных результатов.

Ключевые слова: *облачные вычисления, модель программирования MapReduce, поисковые системы, распределенные кластерные системы.*

CLOUD COMPUTING APPLICATIONS FOR SEARCH ENGINES

I.S., Sakalo

In this article examines the problem of processing large amount of data, the storage of data in a distributive cluster systems. Analyzes existing methods of processing data using cloud computing, principles of their work. Propose to develop a service that can solve problems such as modification and distribution of data between processors, load balancing, failover, data collection and aggregation of intermediate results.

Keywords: *cloud computing, model of programming MapReduce, search engines, distributive cluster systems.*