wydział elektrotechniki elektroniki informatyki i automatyki

AMG.net
A Bull Group Company

PROJEKTOWANIE SYSTEMÓW OBIEKTOWYCH I ROZPROSZONYCH

LABORATORIUM 4

Baza danych noSQL - SimpleDB

wersja 1.0

przygotował:
Radosław Adamus

.

HISTORIA WERSJI

DATA	WERSJA	AUTOR	OPIS
29.03.2014	0.1	Radosław Adamus	Pierwsza robocza wersja dokumentu
06.03.2014	1.0	Radosław Adamus	Pierwsza oficjalna wersja dokumentu

Cel:

Celem laboratorium jest:

1. Zapoznanie się z usługą AWS SimpleDB pozwalającej na wykorzystanie prostej nierelacyjnej bazy danych (NoSQL).
2. Praktyczne wykorzystanie API SimpleDB do komunikacji z usługą.

Wymagania wstępne:

1. Posiadanie konta na platformie Github.
2. Skonfigurowane konto AWS
3. Ukończone laboratorium 4.

Narzędzia:

Git, nodeJS, edytor programistyczny (np. Notepad++), konsola SimpleDB(Javascript Scratchpad for Amazon SimpleDB).

Reguły wykonywania ćwiczeń laboratoryjnych:

1. Po ukończeniu laboratorium należy wyłączyć wszystkie działające instancje EC2, i/lub wyzerować ustawienia dotyczące docelowej, minimalnej i maksymalnej liczby instancji w usłudze ASG.
2. Otrzymane dane autoryzacyjne (hasła oraz klucze dostępu) są danymi wrażliwymi i muszą być chronione. W szczególności nie można dodawać do repozytorium kontroli wersji oraz pozwolić na wysłanie na usługi hostujące repozytoria kontroli wersji kodu źródłowego (GitHub) plików zawierających konfigurację autoryzacji dostępu do API AWS.
3. Zmiany zatwierdzane w repozytorium powinny mieć znaczące komentarze.

Opis laboratorium:

1. Informacje podstawowe

Podstawowe informacje na temat nierelacyjnych baz danych znajdują się na końcu instrukcji. Zainteresowanym polecam również obejrzenie prezentacji https://www.youtube.com/watch?v=ASiU89Gl0F0.

2. Usługa SimpleDB

2.1 Cechy

Amazon SimpleDB to wysoko-dostępna usługa, pozwalająca na wykorzystanie mechanizmów nierelacyjnej baza danych w aplikacji i nie wymagająca wykonywania działań administracyjnych. Do jej podstawowych cech należą:

1. Automatyzacja działań administracyjnych (infrastruktura, aktualizacje, replikacje, indeksowanie).
2. Wysoka-dostępność (replikacje)
3. Elastyczność (dodawanie atrybutów bez potrzeby zmiany schematu, wybór pomiędzy pełną a opóźnioną spójnością)
4. Uproszczony, w stosunku do systemów relacyjnych, interfejs umożliwiający zapisywanie oraz wykonywanie zapytań.

2.2 Wykorzystanie

W przeciwieństwie do usługi S3 usługa SimpleDB nie jest dedykowana do przechowywania dużych danych binarnych (plików). W zamian automatycznie dokonuje indeksowania zapisanych danych oraz udostępnia interfejs prostych zapytań. Zaleca się wykorzystywać ją w scenariuszach, w których potencjalnie jest duża ilość danych, które często nie posiadają ustalonej struktury oraz których wykorzystanie i analiza nie wymaga wykonywania skomplikowanych operacji (np. złączeń). Do scenariuszy tych należą: budowanie dzienników systemu, indeksowanie metadanych obiektów przechowywanych w usłudze S3 czy przechowywanie informacji o użytkownikach i stanie rozgrywki w grach online.

2.3 Model danych

Podstawową jednostką strukturalizacji danych w usłudze SimpleDB jest dziedzina (ang. domain), która jest odpowiednikiem tabeli w relacyjnej bazie danych. W ramach dziedziny można dodawać i pobierać dane oraz uruchamiać zapytania. Dane w obrębie dziedziny - obiekty - składają się z jednej lub wielu par klucz-wartość, reprezentujących jego atrybuty. Klucz określa nazwę atrybutu, wartość klucza reprezentuje wartość (lub wartości) atrybutu. Każdy obiekt w dziedzinie może mieć nie tylko różne wartości atrybutów, ale również różne ich zestawy (brak schematu). Dodatkowo wartość atrybutu obiektu może być wielokrotna. Bardziej szczegółowy opis modelu danych można znaleźć na stronie: http://docs.aws.amazon.com/AmazonSimpleDB/latest/DeveloperGuide/DataModel.html

3. Wykonywanie operacji na usłudze SimpleDB

3.1 Interfejs

Operacje na SimpleDB wykonywane są za pośrednictwem usługi internetowej. Interfejs usługi zawiera zestaw operacji przedstawionych na stronie AmazonSimpleDB/latest/DeveloperGuide/SDB_API_Operations.html.

3.2 Konsola

Usługa nie jest dostępna w z poziomu AWS Management Console. Istnieje natomiast konsola udostępniana na licencji wolnego oprogramowania(open-source) - http://aws.amazon.com/code/1137.

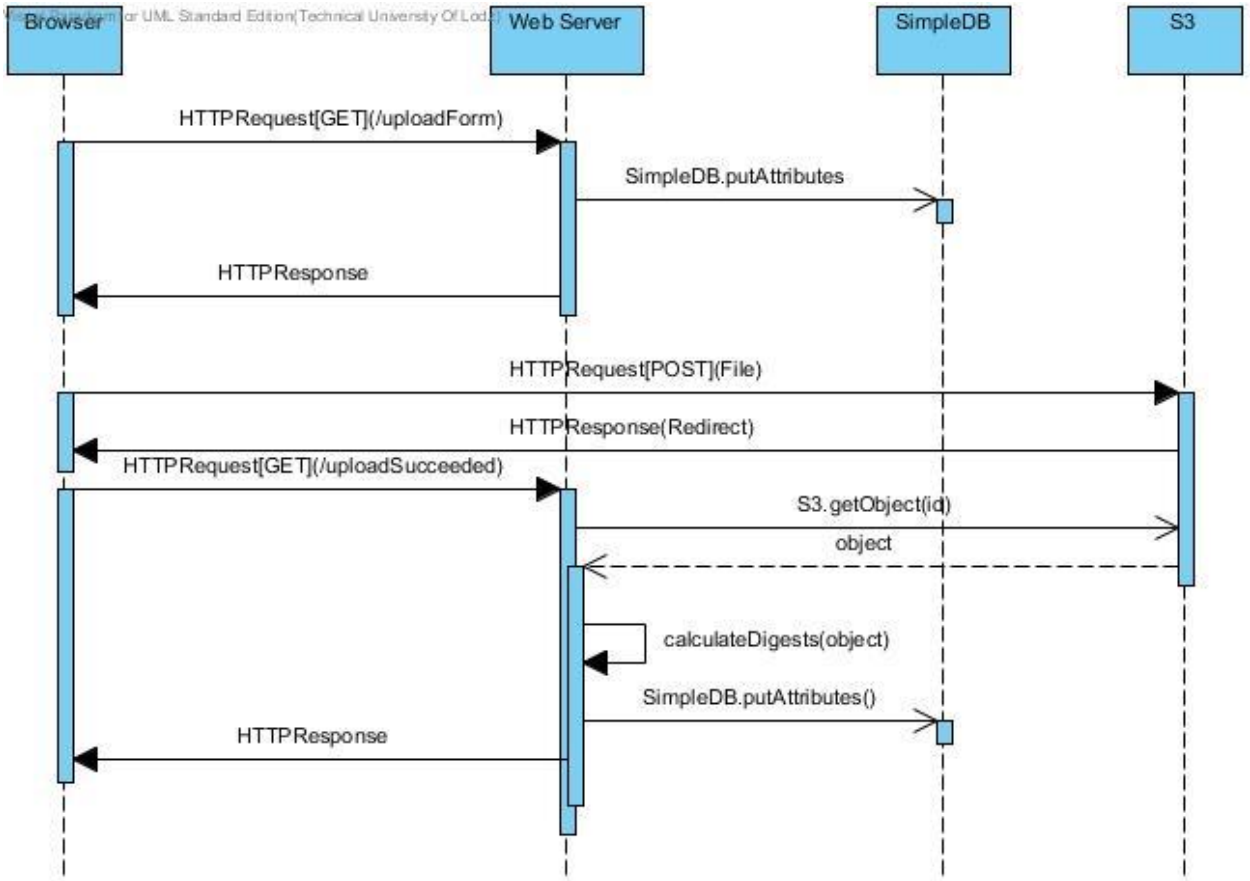
3.3 SDK

Bezpośrednie wywołanie interfejsu (niezależnie od języka programowania) wymaga zbudowania poprawnego żądania zawierającego informacje uwierzytelniające. Tak jak w przypadku innych usług, udogodnienia w zakresie wykorzystania interfejsu usługi dostępne są za pośrednictwem AWS-SDK dla wybranego języka programowania (https://aws.amazon.com/tools/). Opis interfejsu usługi dla platformy NodeJS jest dostępny : http://docs.aws.amazon.com/AWSJavaScriptSDK/latest/frames.html#!AWS/SimpleDB.html

4. Zadania

1. Zapoznaj się z działaniem usługi poprzez utworzenie dziedziny oraz zapis i odczyt obiektu i jego atrybutów. Do tego celu wykorzystaj konsolę SimpleDB:Javascript Scratchpad for Amazon SimpleDB.

2. Rozbuduj aplikację z poprzedniego laboratorium w taki sposób, aby wyliczone skróty przechowywane były w SimpleDB oraz aby każde żądanie wyświetlenia formularza było zapisywane w dzienniku w SimpleDB. Diagram sekwencji na rysunku 1 przedstawia poglądowo zasadę działania procesu wyliczania skrótów dla dokumentu przesyłanego do usługi S3.



Rysunek 1 Diagram sekwencji - architektura przetwarzania żądania wyliczania skrótu dla dokumentu w docelowej aplikacji

Materiały:

1. Wprowadzenie do SimpleDB: http://www.slideshare.net/hungryblank/simpledb-an-introduction
2. Jak działa usługa SimpleDB http://www.slideshare.net/robtweed/developing-nodemdb (slajd 10).
3. Konsola SimpleDB Javascript Scratchpad for Amazon SimpleDB
4. Krytyka SimpleDB: http://cloudcomments.net/2011/06/22/7-reasons-why-people-dont-use-simpledb/
5. SimpleDB a DynamoDB: http://www.allthingsdistributed.com/2012/01/amazon-dynamodb.html

Suplement (TLDR):

NoSQL - nierelacyjne bazy danych.

W ostatnich latach jesteśmy świadkami eksplozji ilości danych, dla których istnieje potrzeba przetwarzania i przechowywania. Decyzje biznesowe wielu firm podejmowane są na podstawie rezultatów tej analizy. Dotyczy to szczególnie firm internetowych, które muszą przechowywać i analizować dane takie jak dzienniki (logi) aktywności, strumienie kliknięć pochodzących z wielu udostępnianych usług, itp.. Jedną z cech tych danych jest częsty brak (lub ograniczony zakres) stabilnej strukturalizacji zawartej informacji. Oznacza to, że zdefiniowany schemat danych może podlegać częstym i znacznym zmianom.

Istniejące rozwiązania w zakresie baz danych nie były przystosowane do tego rodzaju wymagań. Relacyjne systemy baz danych wymagają ustalonego schematu oraz cechują się wysokimi kosztami skalowania. Powstała sytuacja, która wymusiła proces poszukiwania nowych rozwiązań. Jego rezultatem jest nurt noSQL[1] - nierelacyjnych baz danych. W jego ramach powstały systemy baz danych, których podstawowym założeniem jest udostępnianie mechanizmów efektywnego przechowywania dużej liczby danych (nie posiadających schematu) oraz minimalizacja kosztów skalowania. Odbywa się to zazwyczaj kosztem braku zaawansowanych mechanizmów wyszukiwania oraz brakiem, lub znacznym ograniczeniem, narzędzi zapewniania spójności współbieżnych operacji wykonywanych na danych.

Typowa baza NoSQL jest bazą klucz-wartość (ang. key-value) lub zorientowaną na dokumenty (ang. document oriented)[2]. Różnica polega zazwyczaj na stopniu “rozumienia” przechowywanych danych co przekłada się na możliwości wyszukiwania i indeksacji. W pierwszym przypadku mamy do czynienia z twałym odpowiednikiem tablicy mieszającej (asocjacyjnej) - aby pobrać dane z bazy danych należy znać wartość klucza (unikatowy indeks), który im odpowiada. W drugim przypadku oprócz wyszukiwania po kluczu możliwe jest wykorzystanie dodatkowych indeksów. Nierelacyjne bazy danych stanowią obecnie silny ekosystem rozwiązań o coraz większym znaczeniu biznesowym. Do przykładów należą bazy typu klucz-wartość: Redis , Riak czy bazy zorientowane na dokumenty: MongoDB, CouchDB.

[1] Zainteresowanych historią powstania nazwy oraz głębszym wprowadzeniem w temat odsyłam do prezentacji https://www.youtube.com/watch?v=qI_g07C_Q5I

[2] Podział ten nie jest kompletny, do pejzażu nierelacyjnych baz danych można również zaliczyć systemy baz danych oparte na składach kolumnowych (ang. column-oriented) czy też grafowe bazy danych.

Z punktu widzenia chmur obliczeniowych, nierelacyjna baza danych jest po prostu kolejną usługą, którą można udostępniać, tak jak każdą inną. Należy zwrócić jednak uwagę, że w tym wypadku mamy do czynienia z usługą platformą (system bazy danych) działającą na dostępnej infrastrukturze chmury. Użytkownik usługi nie definiuje swoich potrzeb w kategoriach infrastruktury (np. liczby jednostek obliczeniowych czy dostępnej pamięci operacyjnej). Platforma udostępnia zestaw cech, takich jak wysoka-dostępność, przezroczysta skalowalność, czy odporność na awarie, jako integralny element usługi. Zarządzanie infrastrukturą, w celu osiągnięcia tych cech, leży po stronie usługi. Do minimum ograniczone są również działania administracyjne - programista korzystający z usługi nie musi np. samodzielnie aktualizować oprogramowania platformy[3].

Amazon, posiada dwie usługi udostępniające funkcjonalność nierelacyjnej bazy danych w chmurze obliczeniowej - SimpleDB oraz DynamoDB. DynamoDB jest usługą, która dedykowana jest rozwiązaniom o wysokich wymaganiach w zakresie wydajności (wykorzystuje dyski SSD) oraz skalowalności. SimpleDB ma ograniczony rozmiar dziedziny (max. 10 GB) oraz ograniczoną przepustowość.

[3] Jest to istotna cecha komponentów w modelu chmury PaaS. Oczywiście nie ogranicza się to tylko do systemów baz danych.