

Apache Cassandra 2.0

Przewodnik instalacji i konfiguracji w systemie Debian Wheezy

Jan Baranowski, Michał Kaik
Politechnika Poznańska

15 czerwca 2014

1 Wstęp

Niniejszy dokument ma za zadanie przedstawić proces instalacji i konfiguracji serwera baz danych NoSQL *Apache Cassandra* w środowisku rozproszonym. Na potrzeby demonstracji zakłada się że środowisko to będzie składać się z kilku węzłów połączonych siecią lokalną.

Apache Cassandra jest serwerem baz danych NoSQL, początkowo rozwijanym przez Facebooka na potrzeby umożliwienia efektywnego przeszukiwania skrzynki odbiorczej. Obecnie Cassandra jest rozwijana przez Apache Foundation (jest jednym z projektów top-level) i stanowi podstawę dla zestawu narzędzi DataStax.

Cassandra powstała jako narzędzie mające w założeniu cechować się:

- * wysoką dostępnością (Cassandra jest określana jako zawsze zapisywalna baza danych)
- * niskim opóźnieniem wykonywanych operacji (ang. latency)
- * odpornością na awarie (możliwością replikacji danych, brakiem komponentów, których awaria może zdestabilizować system (ang. single points of failure))
- * możliwością regulacji kompromisu pomiędzy szybkością działania a odpornością na awarie i spójnością replik
- * relatywnie prostym modelem danych

Zarówno opisanie kolumnowego modelu danych wykorzystywanego w Cas-sandrze, jak i mechanizmów, dzięki którym spełnia ona założenia projektowe nie jest celem tego dokumentu. Autorzy mogą jedynie podać propozycje publikacji, które opisują wspomniane zagadnienia. Zarówno dla modelu danych,

jak i budowy wewnętrznej Cassandra będzie to przede wszystkim dokumentacja udostępniana przez firmę DataStax ([link]). Model danych, przedstawiony z perspektywy osób pracujących z bazami relacyjnymi został doskonale (choć może zbyt obszernie) opisany w Cassandra - The Definitive Guide ([x]). Z kolei architektura Cassandra w dużym skrócie (choć zapewne w sposób wystarczający by służyć jako wstęp do tego dokumentu) została przedstawiona w prezentacji [x].

Procedurę przygotowania klastra przedstawiono na przykładzie systemu operacyjnego Debian GNU/Linux 6, ponieważ uchodzi on za jedno z lepszych rozwiązań dla systemów serwerowych.

2 Wybór dystrybucji

3 Instalacja

3.1 Wybór maszyny wirtualnej Java

3.2 Instalacja i konfiguracja Oracle Java

3.3 Konfiguracja repozytorium

3.4 Instalacja Cassandra

4 Konfiguracja węzła

4.1 Stan systemu po instalacji

4.2 Jak debugować konfigurację?

4.3 Pliki konfiguracyjne

4.4 Weryfikacja stanu węzła

5 Konfiguracja klastra

5.1 Łączenie węzłów w klaster

5.2 Weryfikacja konfiguracji klastra

5.3 Działania dodatkowe

5.3.1 Synchronizacja zegarów

5.3.2 Czyszczenie

5.3.3 Konfiguracja uwierzytelniania

6 Zarządzanie klastrem

6.1 Monitoring węzłów i klastra

6.2 Przyłączanie i odłączanie węzłów

6.3 Backup danych

Źródła