






Przetwarzanie danych z wykorzystaniem technologii NoSQL na przykładzie serwisu Serp24

Agenda

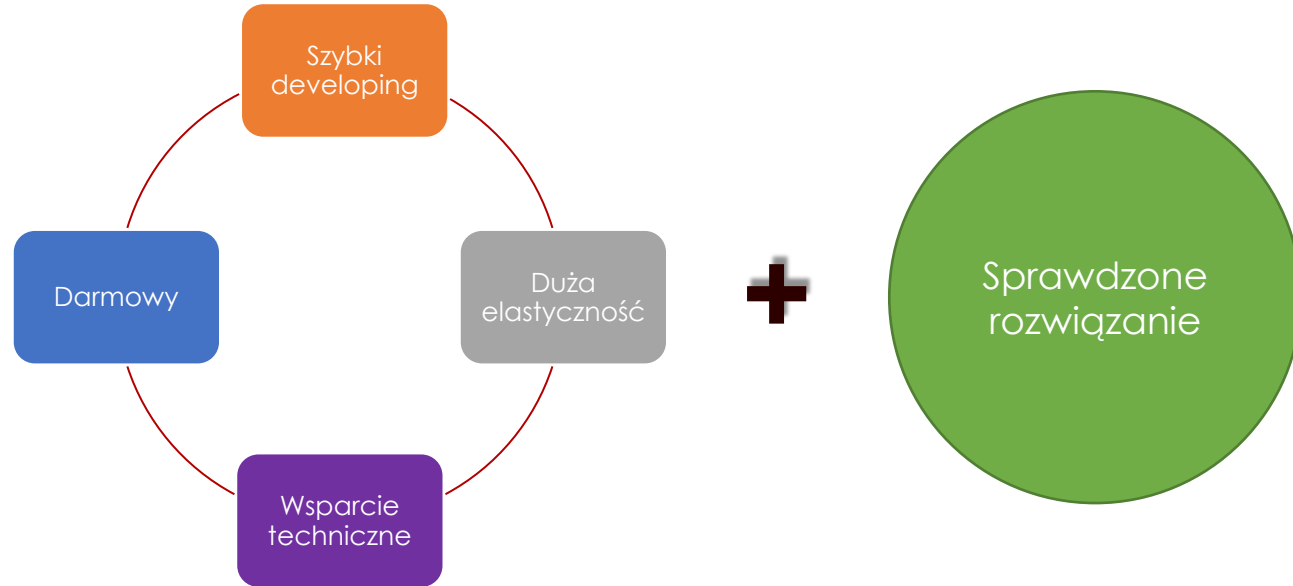
-  Serp24
-  NoSQL
-  Integracja z CMS Drupal
-  Przetwarzanie danych
-  Podsumowanie

Serp24

- Darmowe narzędzie
- Ułatwia planowanie i prowadzenie kampanii SEO
- Monitoruje ponad 8 milionów fraz
- Umożliwia określenie fraz łatwych do pozycjonowania i generujących duży ruch
- Wyszukuje konkurencję na frazy



Dlaczego Drupal?



Dane

- 37 mln - dzienny przyrost rekordów
- 10 GB - dzienny przyrost danych
- Zmienna struktura danych



Big Data = Big problem



SQL problem

- 🔹 Spadek wydajności wraz ze wzrostem rozmiarów bazy
- 🔹 Czasochłonna naprawa uszkodzonych tabel
- 🔹 Konieczność dostosowania struktury tabel do danych
- 🔹 Skalowanie pionowe

Rozwiązanie

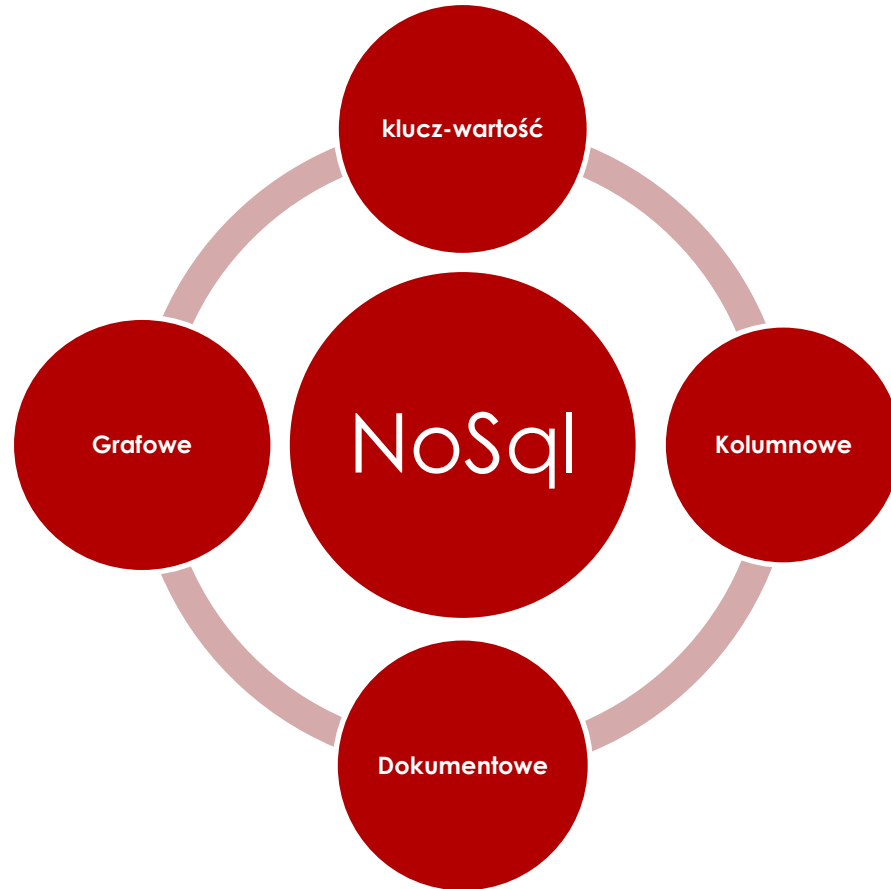


NoSQL

- 🔹 Elastyczna struktura danych
- 🔹 Duża wydajność
- 🔹 Tańsza infrastruktura
- 🔹 Łatwe skalowanie poziome



NoSQL





- Baza klucz-wartość
- Określenie czasu życia danych
- Transakcje
- Dane giełdowe, komunikacja w czasie rzeczywistym, cache



Instagram
Fast beautiful & photo sharing





Cassandra

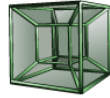
- 🔹 Baza kolumnowa
- 🔹 Integracja rozwiązania Amazon Dynamo i modelu danych Google BigTable
- 🔹 Przechowywanie logów, analiza danych



 **mongoDB**

- Dokumentowa baza danych
- Wsparcie agregacji danych
- Przechowywanie plików
- Polecana gdy potrzeba dużej wydajności przy dużej bazie danych, ograniczeniem jest definiowanie kolumn w strukturze danych





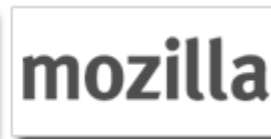
HYPERTABLE

- Baza kolumnowa, wzorowana na Google BigTable
- Struktura danych odwzorowuje wiersze
- Nacisk na łatwe skalowanie poziome i wydajność
- Zalecana do skanowania dużych dwu-wymiarowych tabel, silników wyszukiwania, analizy danych





- 🔹 Grafowa baza danych
- 🔹 Zoptymalizowana pod kątem odczytów
- 🔹 Posiada transakcje
- 🔹 Wyszukiwanie powiązań w sieciach społecznościowych, połączenia w transporcie publicznym, mapy drogowe



Trochę z innej beczki

- Silniki wyszukiwania pełnotekstowego jako baza danych
- Alternatywa dla NoSQL'a
- Duża wydajność









- 🔹 Rozwijany przez Apache Foundation
- 🔹 Licencja open source
- 🔹 Łatwe skalowanie i replikacja
- 🔹 Faceted search

<http://drupal.org/project/apachesolr>
http://drupal.org/project/search_api_solr



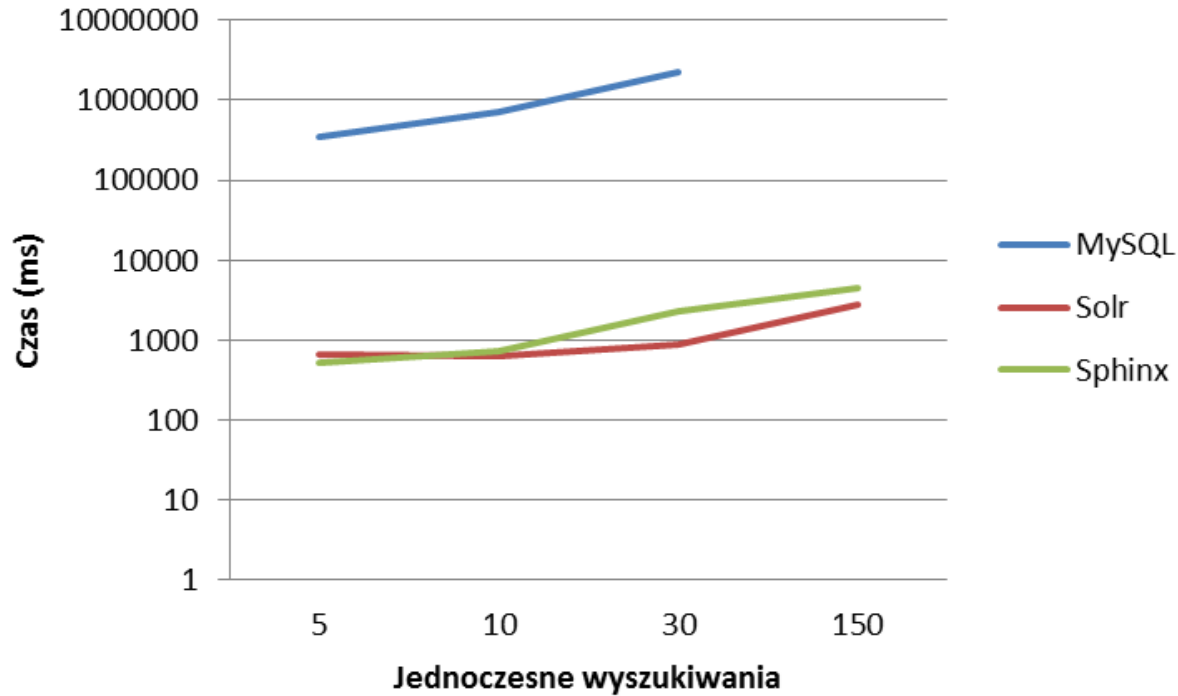
Sphinx

-  Licencja open source
-  Wbudowana obsługa bazy danych MySQL/PostgreSQL
-  Łatwa instalacja
-  Obsługa zapytań SQL

http://drupal.org/project/search_api_sphinx



Wydajność silników wyszukiwania



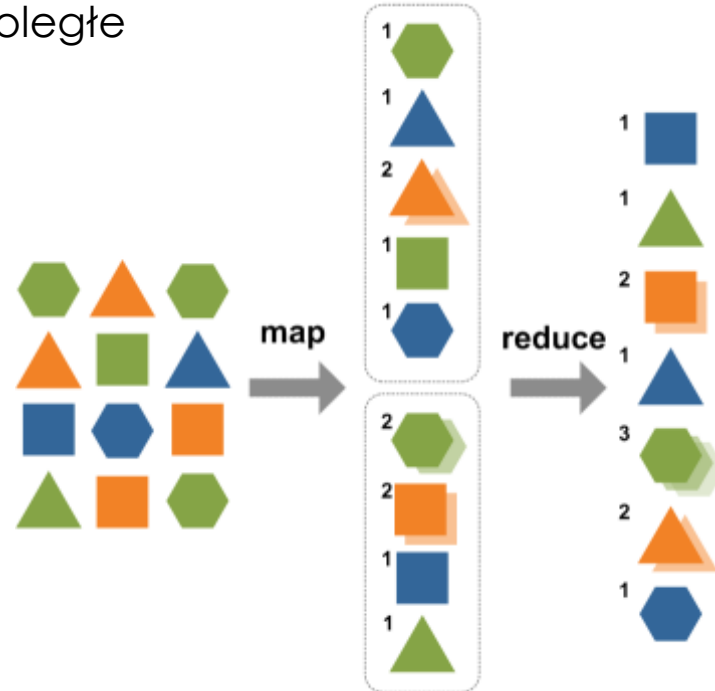
Przetwarzanie danych



Przetwarzanie danych

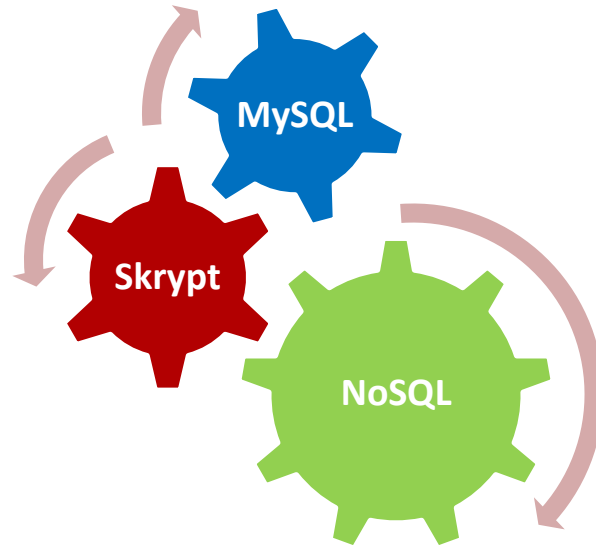
MapReduce czyli przetwarzanie równoległe

- Map() – wybór danych
- Reduce() – agregacja



Przetwarzanie danych c.d.

Buforowanie danych w MySQL'u i archiwizowanie w NoSQL'u już przetworzonych



Przetwarzanie danych c.d.




System hook'ów:

- 🔹 Kompleksowa analiza danych
- 🔹 Przetwarzanie danych przed i po zapisie do NoSQL'a
- 🔹 Modyfikacja przed prezentacją danych



Integracja z CMS Drupal

Moduły implementujące komunikację po API :

-  drupal.org/project/cassandra
-  drupal.org/project/couchdb
-  drupal.org/node/1944834 (Hypertable)



Integracja z MongoDB

Projekt MongoDB (drupal.org/project/mongodb)

MongoDB
cache

MongoDB
field storage

MongoDB
session

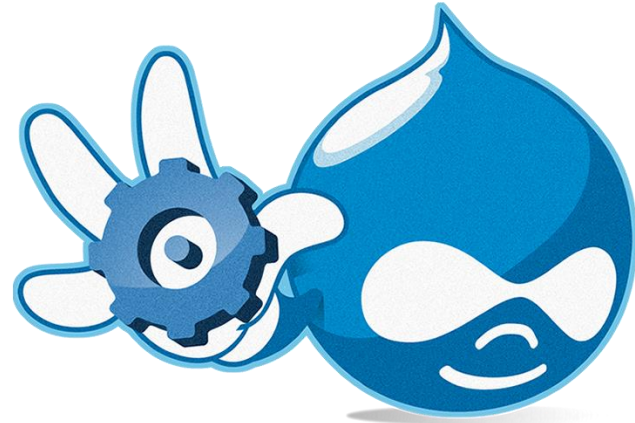
MongoDB
watchdog

MongoDB
block

Search API
MongoDB

Integracja za pomocą views_plugin_query

- Przykład implementacji w module Google Analytics Reports (http://drupal.org/project/google_analytics_reports)
- Doskonały sposób integracji z dowolnym API
- Nie angażuje Entity API



Moja strona

Pokaż wspólne frazy



pl.wikipedia.org

I Konkurent

en.wikipedia.org

II Konkurent

es.wikipedia.org

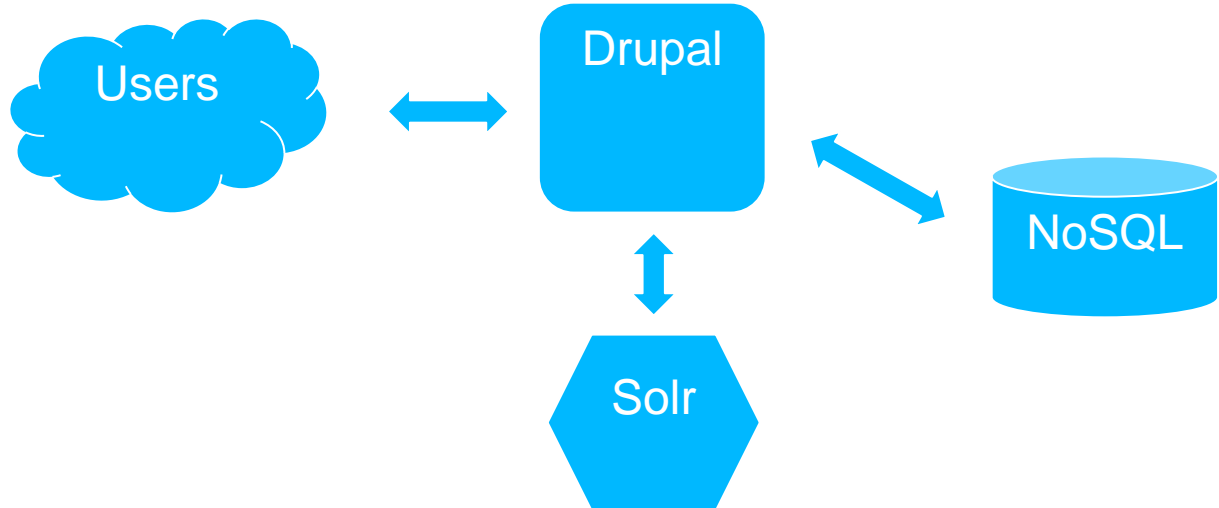
keyword

ZASTOSUJ

SŁOWO KLUCZOWE	KONKURENCYJNOŚĆ	CPC	WYSZUKIWANIA GLOBALNE	WYSZUKIWANIA LOKALNE	WYNIKI	PL.WIKIPEDIA.ORG	EN.WIKIPEDIA.ORG
nk	0	0.31	30400000	24900000	314000000	13	22
allegro	1	0.31	20400000	20400000	227000000	18	27
youtube	2	0.31	755000000	13600000	6760000000	8	9
o2	0	0.31	7480000	5000000	327000000	12	23
alegro	8	0.31	1830000	1830000	226000000	18	28
pko	0	1.28	1830000	1830000	16500000	13	22
orange	8	0.34	37200000	1500000	1570000000	12	35
tvn24	0	0.31	1830000	1500000	6410000	8	41
t-mobile	7	0.31	5000000	823000	5200000000	9	23
face	0	0.32	151000000	823000	2590000130	11	33
ikea	1	0.51	20400000	823000	173000000	8	19

Integracja za pomocą Search API

- 🔹 Indeksowanie części danych w Solr'ze
- 🔹 Pobieranie ciężkich danych z NoSQL
- 🔹 Rozszerzenie funkcjonalności wyszukiwania w NoSQL





MOJE ULUBIONE

→ [dodaj tą stronę](#)

Znaleziono 12393 elementów

▶ [\[wszystkie wyniki\]](#)

FILTROWANIE: SUBDOMENA

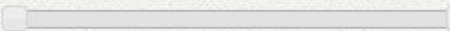
- ▶ [drupal.org](#) (10503)
- ▶ [groups.drupal.org](#) (1121)
- ▶ [api.drupal.org](#) (346)
- ▶ [localize.drupal.org](#) (226)
- ▶ [munich2012.drupal.org](#) (53)
- ▶ [denver2012.drupal.org](#) (29)
- ▶ [association.drupal.org](#) (21)
- ▶ [portland2013.drupal.org](#) (20)
- ▶ [qa.drupal.org](#) (17)

FILTROWANIE: POZYCJA



zakres od 1 do 50

Od Do



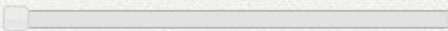
IDŹ

FILTROWANIE: LICZBA WYNIKÓW



zakres od 1 do 25.269.999.616

Od Do



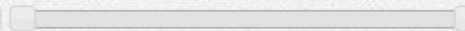
IDŹ

FILTROWANIE: WYSZUKIWAŃ LOKALNIE



zakres od 0 do 74.000

Od Do



IDŹ

Problemy związane z przetwarzaniem

- Niska wydajność w przypadku bardzo dużych pakietów danych
- Denormalizacja = bardzo szybki przyrost liczby rekordów
- Wykorzystywanie MapReduce wymaga znacznej wiedzy i doświadczenia
- Brak join'ów wymaga przestawienia z MySQL'lowej logiki projektowania aplikacji



Aplikacje hybrydowe

- Elastyczność
- Wydajność
- Skalowalność
- Świadomy wybór

Przydatne strony

- 🔗 <http://kkovacs.eu/cassandra-vs-mongodb-vs-couchdb-vs-redis>
- 🔗 <http://www.techrepublic.com/blog/10things/10-things-you-should-know-about-nosql-databases/1772>
- 🔗 <http://nosql.mypopescu.com/>
- 🔗 <http://nosql.findthebest.com/>



Pytania?

Dziękuję za uwagę



Sławomir Sokół

e: slawomir.sokol@semtec.pl

