

Podatkovne baze vrste triplestore

Laboratorij za podatkovne tehnologije

Fakulteta za računalništvo in informatiko

Univerza v Ljubljani

Aljaž Zrnec, Neli Blagus, Marko Janković, Marko Bajec



Cilji predstavitve

- Podatkovna baza triplestore
- Testno okolje
- Analiza
- Ugotovitve

Triplestore...

RDBMS

Hadoop
HBase

Triplestore

- NoSQL – zaradi potreb po zajemanju nestrukturiranih podatkov.
- Razvoj se usmeri v dve smeri:
 - Sistemi za obvladovanje velikih količin podatkov (Hadoop, HBase),
 - Sistemi semantičnega spleta (sistemi za shranjevanje trojčkov - triplestore).
- Primerjava triplestore podatkovnih baz:
 - **D2RQ, Jena SDB, Jena TDB, Neo4j, Owlrim, Sesame in Virtuoso.**

Triplestore

Triplestore

Skalabilnost
Prilagodljivost
Skepanje

- Triplestore: splošno ime za sisteme za shranjevanje podatkov v obliki trojčkov.
- Optimizira za poizvedovanje po trojčkih v RDF obliki.
- RDF shrambe, hibridne RDF shrambe s podporo relacijskih podatkovnih baz in RDF ovojnice.
- Prednosti: prilagodljivost, skalabilnost, shema ni vnaprej opredeljena, uporaba sklepanja, standardi (RDF, SPARQL, OWL,...).

Analiza zmogljivosti...

Scenariji

- Testno okolje: aplikacija za prodajo izdelkov, ki uporablja podatkovno bazo triplestore za hranjenje podatkov o izdelkih.
- Glede na tipično uporabo smo opredelili tri testne scenarije:

Scenarij 1	Scenarij 2	Scenarij 3
Podatkovna baza se uporablja samo za branje:	Podatkovna baza se uporablja za uravnoteženo branje in vnašanje podatkov:	Podatkovna baza se večinoma uporablja za vstavljanje:
• branj 100%, • pisanj 0%, • brisanj 0%.	• branj 50%, • pisanj 50%, • brisanj 0%.	• branj 0%, • pisanj 80%, • brisanj 20%.



Analiza zmogljivosti

Testno okolje

- Navidezni stroj, 3 Gb pomnilnika, Ubuntu Linux 32 bit, gostitelj s Core2Duo 2 GHz procesorjem.
- Podatki generirani z Berlin SPARQL Benchmark Data Generator.
- Testirali smo zmogljivosti z naslednjimi količinami podatkov: 10.000, 100.000, 1.000.000 in 5.000.000 trojčkov.
- Vsak scenarij je obsegal 1000 operacij.
- Merili smo tudi čas vstavljanja podatkov v triplestore-e.

Vstavljanje podatkov

Najboljši:
D2RQ

Najslabši:
Virtuoso

- D2RQ najhitrejši.
- Pri manjših količinah trojčkov so Owlrim, Sesame in Neo4j primerljivi, dočim se pri največji količini trojčkov Neo4j znatno zaostane.
- Virtuoso daleč najpočasnejši, zadnjega testa ni opravil.

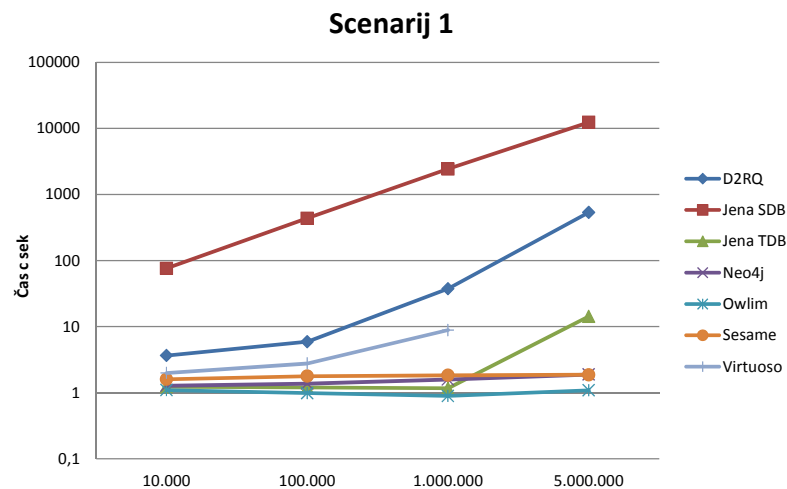
Št. trojčkov	D2RQ	JenaSDB	JenaTDB	Neo4j	Owlrim	Sesame	Virtuoso
10.000	1,100	3,673	1,045	13,239	6,920	8,358	100,054
100.000	2,800	44,597	4,998	58,480	63,302	59,370	963,969
1.000.000	23,500	1359,185	44,148	2222,800	578,833	680,320	9568,958
5.000.000	153,000	17319,072	595,167	34037,544	2969,284	4300,746	/



Scenarij 1

100% branj,
0% pisanj,
0% brisanj.

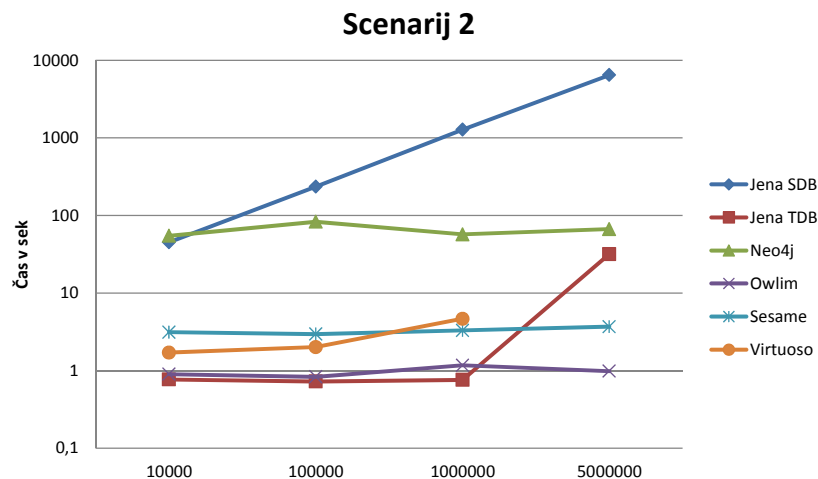
- Prvi scenarij obsega le branje podatkov in najbolj primerni za take operacije so Neo4j, Owlim in Sesame.
- Jena SDB in D2RQ sta z dolgimi povprečnimi časi branja neprimerna za uporabo v aplikacijah, kjer je odzivnost zelo pomembna.



Scenarij 2

50% branj,
50% pisanj,
0% brisanj.

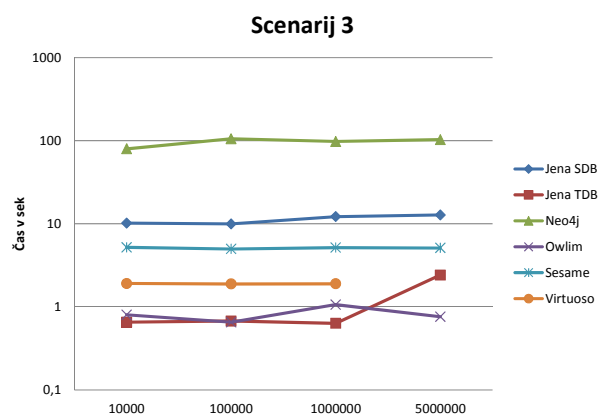
- V drugem scenariju, z enako količino branja in vstavljanja, sta se za najprimernejšo izbiro izkazala Owlrim in Sesame.
- Jena SDB je z nekaj sekund dolgim časom branja neprimerna za uporabo v aplikacijah, kjer je odzivnost pomembna.



Scenarij 3

0% branj,
80% pisanj,
20% brisanj.

- V aplikacijah, kjer je veliko spreminjanja podatkov, je glede na rezultate tretjega scenarija za uporabo najprimernejši Owlrim.
- Za uporabo sta primerna tudi malenkost počasnejša Jena TDB in Sesame.
- Neo4j zaradi počasnega vstavljanja podatkov ne more zagotoviti zadovoljivih odzivnih časov.



Sklep

Ugotovitve

- Rešitve ki uporabljajo za shranjevanje trojčkov relacijske podatkovne baze, po zmogljivostih ne morejo konkurirati namenskim triplestore bazam.
- Ker so triplestore podatkovne baze z vidika implementacije tako zelo različne, je težko izbrati univerzalno rešitev, saj ima vsaka od njih svoje prednosti in slabosti.
- Pred izbiro je smiselno preizkusiti nekaj rešitev, ki po funkcionalnostih najbolj ustrezajo našim potrebam.