SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE
FAKULTA ELEKTROTECHNIKY A INFORMATIKY

DATABÁZY: PREHĽAD A POROVNANIE RÔZNYCH TYPOV NOSQL-DATABÁZ

TÍMOVÝ PROJEKT

2022

Maksim Mištec, Ladislav Rajcsányi,
Alexander Sárközy, Tomáš Kukumberg

SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE
FAKULTA ELEKTROTECHNIKY A INFORMATIKY

DATABÁZY: PREHĽAD A POROVNANIE RÔZNYCH TYPOV NOSQL-DATABÁZ

TÍMOVÝ PROJEKT

Študijný program:	Aplikovaná informatika
Predmet:	I-ASOS – Tímový projekt
Konzultant:	Ing. Stanislav Marochok

2022

Maksim Mištec, Ladislav Rajcsányi,
Alexander Sárközy, Tomáš Kukumberg

Obsah

Zoznam obrázkov a tabuliek

1 Úvod

Cieľom tohto projektu je porovnať viacero druhov NoSQL databáz z hľadiska ich fungovania a rýchlosti, potom implementovať softvér, ktorý bude pracovať so všetkými databázami, ktoré sme si vybrali.

	SQL Databáza	NoSQL Databáza
Typ databázy	Relačné databázy	Nerelačné alebo distribuované databázy
Jazyk dopytov	Structured Query Language (SQL)	Nemá deklaratívny dopytovací jazyk
Schéma	Schéma databázy je pevne stanovená	Schéma databázy nie je pevne stanovená a je dynamická
Škálovateľnosť	Vertikálne škálovateľná	Horizontálne škálovateľná
Model	Používa model ACID	Používa model BASE
Najvhodnejšie pre	Ideálna voľba pre komplexné dotazovacie prostredie	Vhodné pre hierarchické dátové úložisko, pretože podporuje dvojice kľúč-hodnota
Dôležitosť	Mala by sa používať, keď je mimoriadne dôležitá platnosť údajov	Mala by sa používať, keď sú rýchle údaje dôležitejšie ako správne údaje
Najlepšia voľba	Keď potrebujete podporu dynamických dotazov	Keď potrebujete škálovacie schopnosti pre budúce požiadavky
Príklady	Oracle, Postgres, MS-SQL	MongoDB, Redis, Neo4j, Cassandra, Hbase...

Tabuľka 1: Rozdiely medzi SQL a NoSQL sú uvedené v nasledujúcej tabuľke

2 Databázy

2.1 MongoDB

MongoDB je open source program na správu databáz NoSQL. NoSQL sa používa ako alternatíva k tradičným relačným databázam. Databázy NoSQL sú celkom užitočné na prácu s veľkými súbormi distribuovaných údajov. MongoDB je nástroj, ktorý dokáže spravovať informácie orientované na dokumenty, ukladať alebo vyhľadávať informácie.

MongoDB je nerelačná dokumentová databáza, a niektoré z jej vlastností sú:

- Všetky dokumenty sú navzájom nezávislé.
- Dokumenty sú bez schémy, a preto sú flexibilné a ľahko sa upravujú.
- Na ukladanie dokumentov sa používajú kolekcie s cieľom zoskupiť rôzne druhy údajov.
- Dokumenty môžu mať vnorené dokumenty, rôzne dvojice kľúč-hodnota alebo dvojice kľúč-pole.

Obr. 1: MongoDB logo

Prípady použitia MongoDB:

- SEGA ho používa na správu herných účtov.
- Aer Lingus ho používa na správu leteniek a interných aplikácií.
- ly a Sourceforge ho používajú na správu údajov.

2.2 CouchDB

Apache CouchDB je databázový systém s open source kódom orientovaný na dokumenty, napísaný v programovacom jazyku Erlang a navrhnutý na lokálnu replikáciu a jednoduchú horizontálnu škálovateľnosť na rôznych zariadeniach. CouchDB podporuje komerčné subjekty CouchBase a Cloudant. Podobne ako MongoDB, CouchDB je nerelačná dokumentová databáza, a teda jej vlastnosti sú podobné ako MongoDB.

Obr. 2: CouchDB logo

Prípady použitia CouchDB:

- Spoločnosť United Airlines používa CouchDB pre zábavné systémy počas letu vo viac ako 3 000 lietadlách.
- Red Cross používa aplikáciu iDAT na elektronické vyplňovanie prípadov v oblastiach postihnutých katastrofou. CouchDB sa tu používa ako viacuzlová databáza typu peer-to-peer offline-first.
- BBC pre dynamickú CMS platformu.

2.3 Neo4j

Neo4j je open-source databázový systém, ktorý ukladá informácie vo forme orientovaných grafov. Takéto usporiadanie dát poskytuje veľkú mieru flexibility v organizácii a manipulácii s dátami.

Niektoré z funkcionalít, ktoré poskytuje Neo4j:

- 3 typy údajov: Uzly, vzťahy a atribúty. Uzly môžu mať ľubovoľný počet atribútov a sú prepojené ľubovoľným počtom vzťahov.

- Pre dopyty sa používa jazyk Cypher, ktorý je podobný SQL a je špeciálne prispôsobený na interakciu s grafovými databázami.
- Vysoká miera flexibility a škálovateľnosti.
- ACID compliant režim transakcii.

Obr. 3: Neo4j logo

Prípady použitia Neo4j:

- eBay používa Neo4j na spracovávanie zákazníckych preferencii a vyhodnocovanie odporúčaní.
- Cisco používa Neo4j na analýzu problémov zákazníckej podpory s cieľom predvídať chyby.
- Walmart používa Neo4j na poskytovanie relevantných propagácií a odporúčaní produktov.

2.4 Redis

Redis je open-source NoSQL úložisko dátových štruktúr. Dáta uchováva v systémovej pamäti a preto je určený pre aplikácie, kde je potrebná maximálna rýchlosť spracovania dopytov.

Redis poskytuje niekoľko výnimočných funkcionalít:

- Dáta sú vo forme párov kľúčov a hodnôt.
- Pre hodnoty je dostupných niekoľko abstraktných dátových typov, ako napr. reťazec, množina, zoznam, usporiadaná množina a ďalšie.
- Počas chodu sa dáta, s ktorými sa pracuje, uchovávajú v pamäti. Na disk sa ukladajú periodicky formou memory dumpu alebo append-only záznamu.

Obr. 4: Redis logo

Prípady použitia Redis:

- Pinterest používa Redis na ukladanie zoznamov obrázkov a galérií.
- Coinbase používa Redis na autorizáciu kurzových bodov.
- Twitter používa Redis na správu časovej osi.
- GitHub používa Redis na distribúciu a smerovanie užívateľských dopytov a správu súvisiacich dát.

3 Implementácia

3.1 Docker

Docker je open-source platforma na vývoj, dodávanie a spúšťanie aplikácií. Docker umožňuje oddeliť aplikácie od infraštruktúry, aby bolo možné rýchlo dodávať softvér. Pomocou nástroja Docker je možné spravovať svoju infraštruktúru rovnakými spôsobmi, ako spravujete svoje aplikácie. Využitím metodológie Dockeru na rýchle odosielanie, testovanie a nasadzovanie kódu je možné výrazne skrátiť oneskorenie medzi napísaním kódu a jeho spustením vo výrobe.

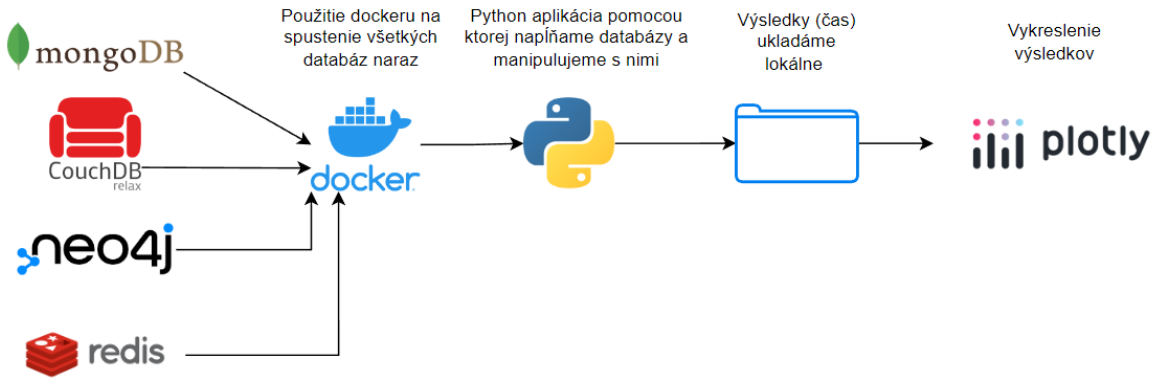
Obr. 5: Docker logo

V našom projekte sme použili Docker na nasadenie viacerých databáz súčasne.

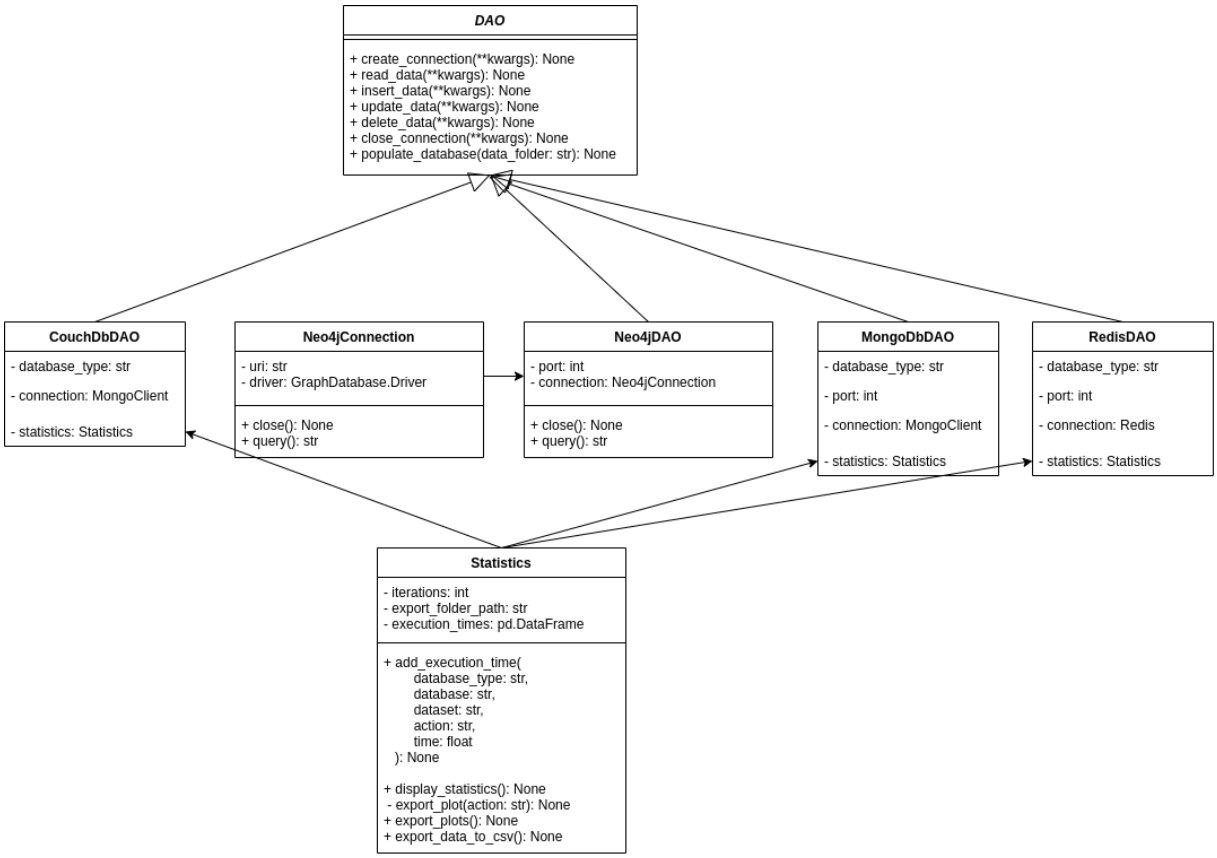
3.2 Postup implementácie

Keďže bolo potrebné, aby všetky vybrané databázy bežali v rovnakom čase, rozhodli sme sa na to použiť Docker. V ňom je veľmi jednoduché nastaviť, na ktorých portoch chceme, aby databázy bežali bez kolízie.

Náš postup implementácie je nasledovný:



Obr. 6: Diagram implementačného postupu



Obr. 7: Class diagram testovacej aplikácie

4 Spôsob testovania

Pre testovanie rýchlosti databáz používame Python knižnicu time a túto funkcionalitu sme implementovali v module Statistics. Meranie sme testovali pre CRUD operácie - create (vytvorenie), read (čítanie), update (aktualizácia) a delete (zmazanie) objektov. Namerané dáta najprv ukladáme do csv súboru, z neho následne počítame priemer, medián a smerodajnú odchýlku. Nakoniec dáta zobrazujeme v stĺpcových a koláčových grafoch pomocou knižnice plotly, ktorá podporuje vykresľovanie interaktívnych grafov. V následujúcej sekcii sa nachádzajú ukážky nameraných dát a grafov.

Pre zredukovanie vonkajších vplyvov a náhodnej variancie sme každé meranie vykonali 50-krát a zobrali sme priemer z týchto nameraných hodnôt.

5 Výsledky testovania

Na nasledujúcich obrázkoch sú zobrazené výstupy v rôznych bodoch testovacieho procesu.

	database_type	database	dataset	action	time
2	CouchDB	ASOS_2022	books	insert	7.769653081893921
3	CouchDB	ASOS_2022	countries-small	insert	5.041459083557129
4	CouchDB	ASOS_2022	covers	insert	101.9377670288086
5	CouchDB	ASOS_2022	data	insert	1.5520169734954834
6	CouchDB	ASOS_2022	grades	insert	6.590811014175415
7	CouchDB	ASOS_2022	products	insert	0.2245476245880127
8	CouchDB	ASOS_2022	profiles	insert	32.7408926486969
9	CouchDB	ASOS_2022	restaurant	insert	55.81869029998779
10	CouchDB	ASOS_2022	students	insert	4.460112571716309
11	CouchDB	ASOS_2022	books	read	1.092195749282837
12	CouchDB	ASOS_2022	countries-small	read	0.6903626918792725
13	CouchDB	ASOS_2022	covers	read	8.577019453048706
14	CouchDB	ASOS_2022	data	read	0.09685635566711426
15	CouchDB	ASOS_2022	data	update	1.173919916152954
16	CouchDB	ASOS_2022	grades	read	0.465923547744751
17	CouchDB	ASOS_2022	products	read	0.01963019371032715
18	CouchDB	ASOS_2022	profiles	read	2.4796624183654785
19	CouchDB	ASOS_2022	restaurant	read	4.181886196136475
20	CouchDB	ASOS_2022	students	read	0.32291555404663086
21	CouchDB	ASOS_2022	books	delete	0.06526708602905273
22	CouchDB	ASOS_2022	countries-small	delete	0.06392478942871094
23	CouchDB	ASOS_2022	covers	delete	0.06141042709350586
24	CouchDB	ASOS_2022	data	delete	0.057615041732788086
25	CouchDB	ASOS_2022	grades	delete	0.08223700523376465
26	CouchDB	ASOS_2022	products	delete	0.0616452693939209
27	CouchDB	ASOS_2022	profiles	delete	0.07730913162231445

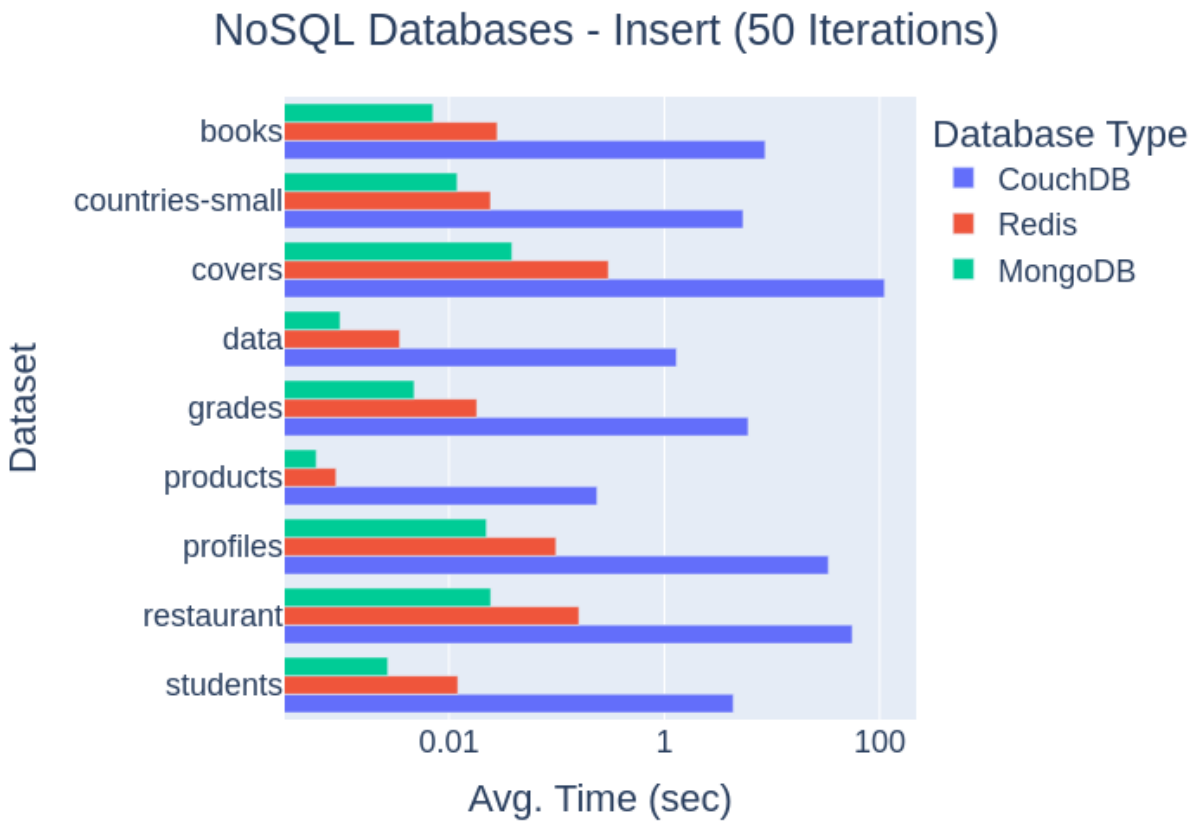
Obr. 8: Ukážka dát z csv súboru

```
CouchDB countries-small delete
      count      mean       std  ...       50%       75%       max
time   50.0  0.055643  0.008413  ...  0.052243  0.062054  0.081397

[1 rows x 8 columns]

CouchDB covers insert
      count        mean       std  ...         50%         75%         max
time   50.0  111.135547  1.513409  ...  111.322697  111.790293  112.918933

[1 rows x 8 columns]

CouchDB covers read
      count      mean       std  ...       50%       75%       max
time   50.0  8.494763  0.094264  ...  8.487459  8.572224  8.666883
```

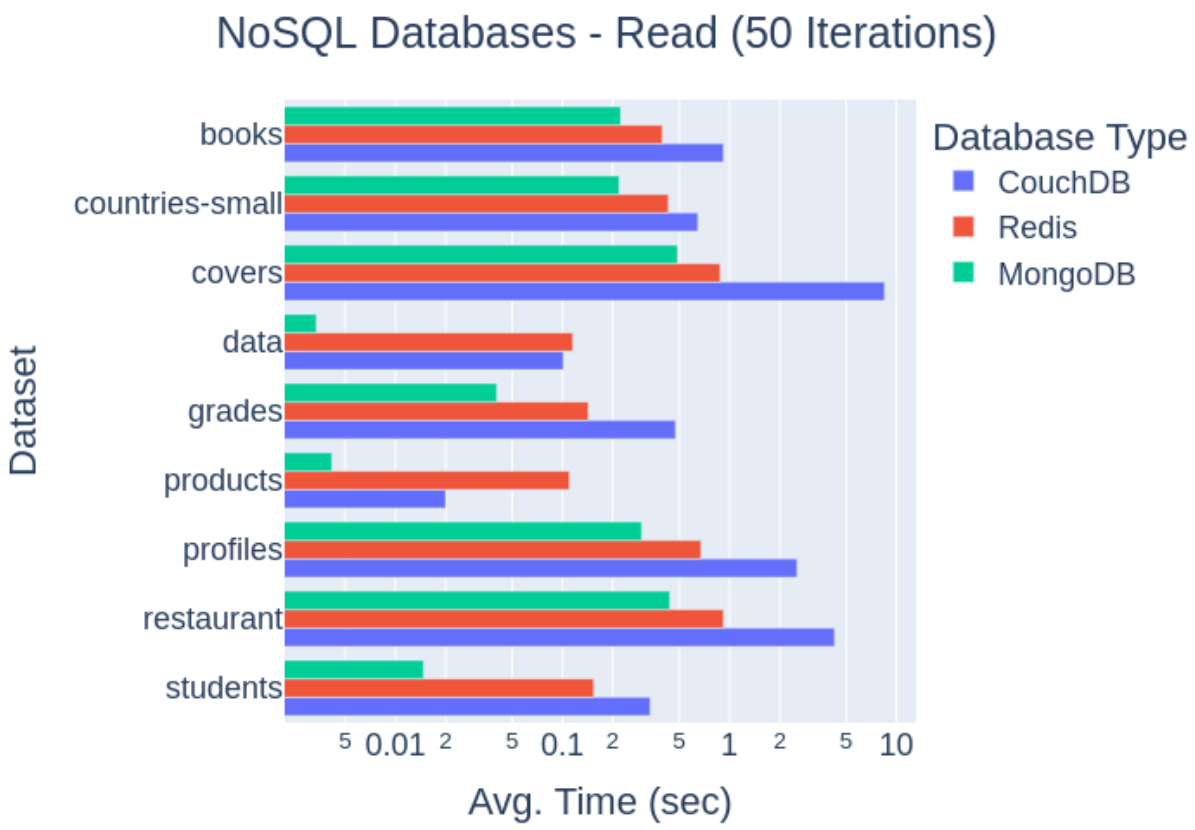
Obr. 9: Vypočítané dáta

5.1 MongoDB, Redis, CouchDB

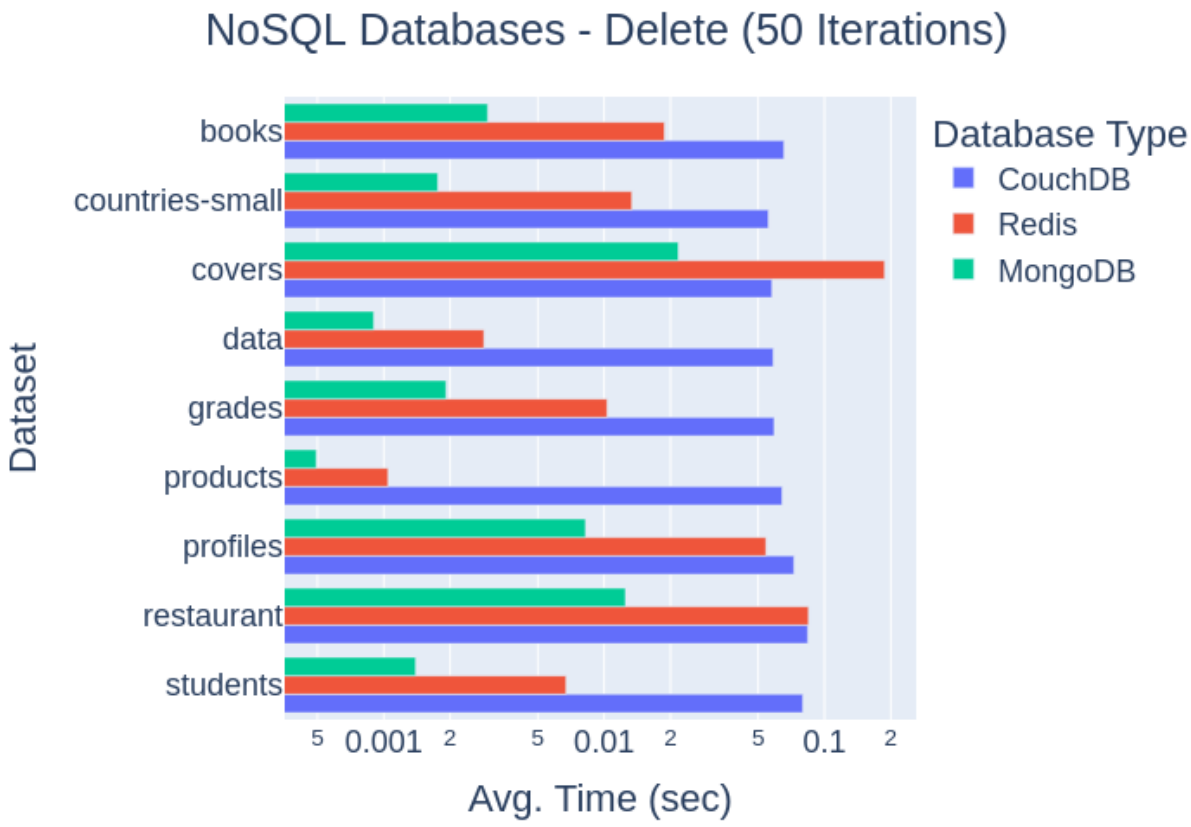
Nasledujúce grafy zobrazujú výsledky pre jednotlivé operácie, databázy, a testovacie datasety.



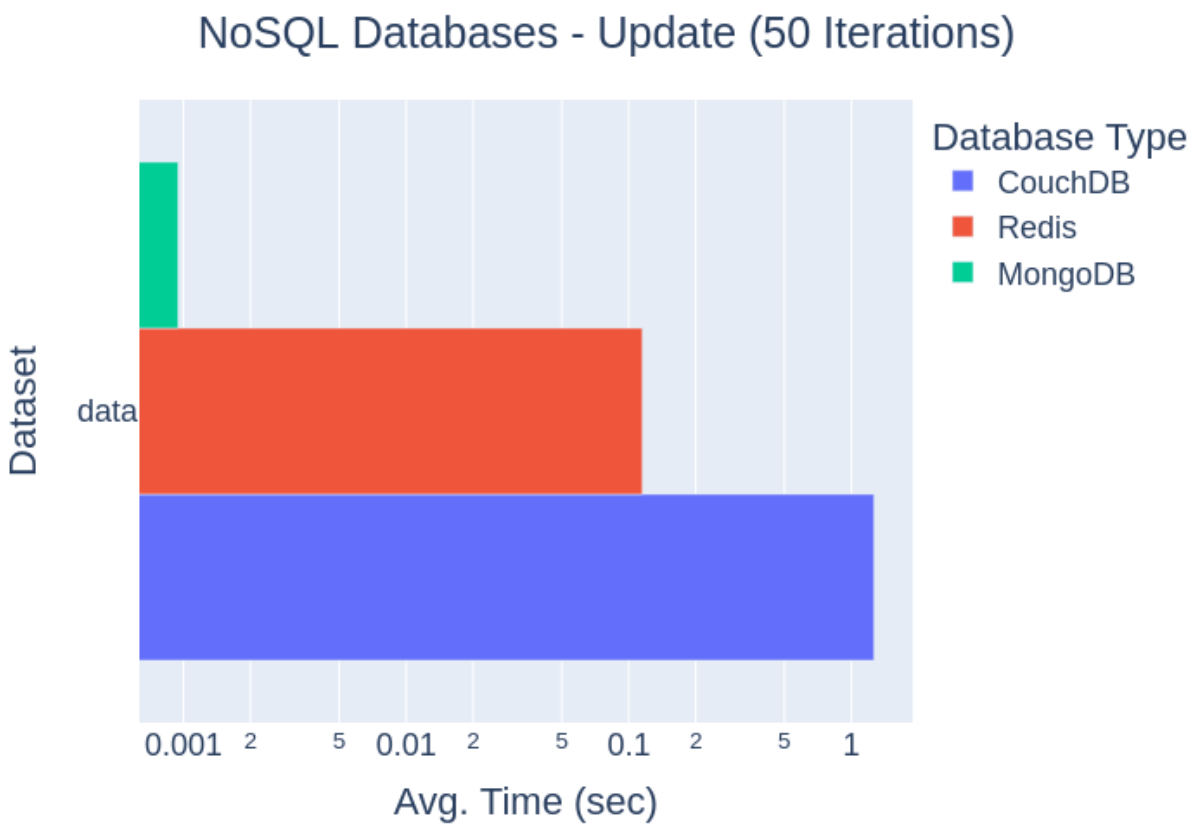
Obr. 10: Porovnanie rýchlosti vykonania operácie Insert



Obr. 11: Porovnanie rýchlosti vykonania operácie Read



Obr. 12: Porovnanie rýchlosti vykonania operácie Delete

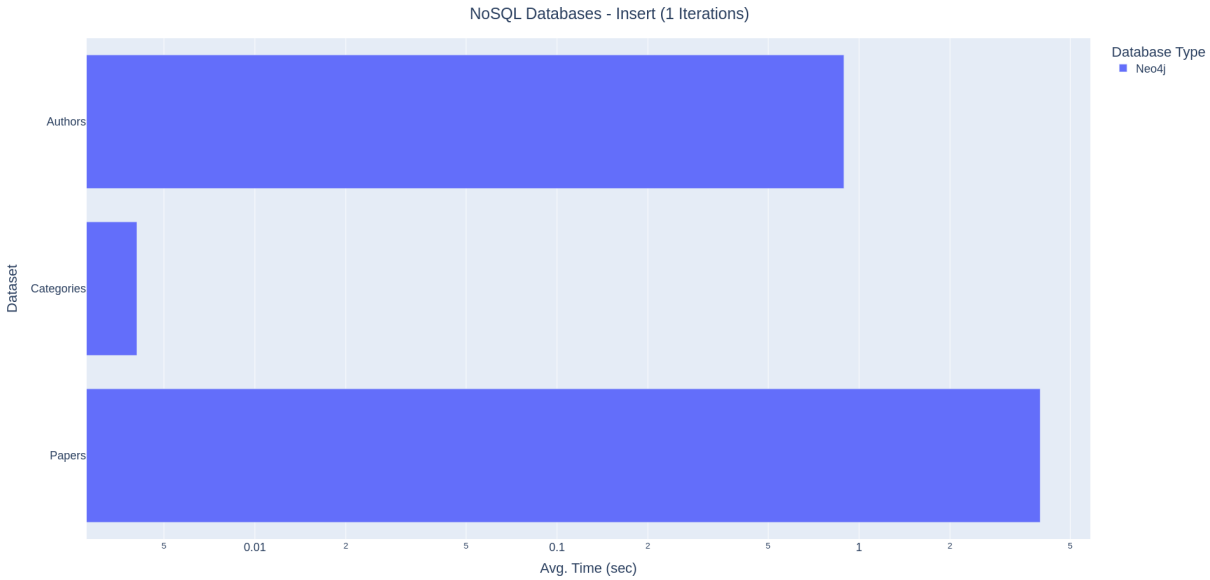


Obr. 13: Porovnanie rýchlosti vykonania operácie Update

dataset_type	time	time
CouchDB	students	4.373265
Redis	students	0.012092
MongoDB	students	0.002707
CouchDB	restaurant	55.680459
Redis	restaurant	0.161091
MongoDB	restaurant	0.024435
CouchDB	profiles	33.417803
Redis	profiles	0.098647
MongoDB	profiles	0.022305
CouchDB	products	0.237487
Redis	products	0.000894
MongoDB	products	0.000586
CouchDB	grades	6.009800
Redis	grades	0.018142
MongoDB	grades	0.004757
CouchDB	data	1.299615
Redis	data	0.003486
MongoDB	data	0.000974
CouchDB	covers	111.135547
Redis	covers	0.302159
MongoDB	covers	0.038421
CouchDB	countries-small	5.393462
Redis	countries-small	0.024307
MongoDB	countries-small	0.011936
CouchDB	books	8.659671
Redis	books	0.028077
MongoDB	books	0.007117

Tabuľka 2: Tabuľka, v ktorej sú uvedené presné časy jednotlivých príkazov pre každý dataset

5.2 Neo4j



Obr. 14: Čas vykonania operácie Insert v Neo4j

dataset	time
Authors	0.890965
Categories	0.004067
Papers	3.977546

Tabuľka 3: Tabuľka času vykonania operácie insert v Neo4j databáze pre každý dataset

6 Záver

Z testov najlepšie dopadla MongoDB databáza. Je to hlavne kvôli tomu, že pôvodné dáta boli v ideálnom formáte pre MongoDB a pre ostatné databázy sme museli dáta upravovať. Výsledky teda ilustrujú, že reálna rýchlosť databázy záleží predovšetkým aj od toho, či je správne zvolená databáza pre dátové štruktúry, s ktorými chceme pracovať.