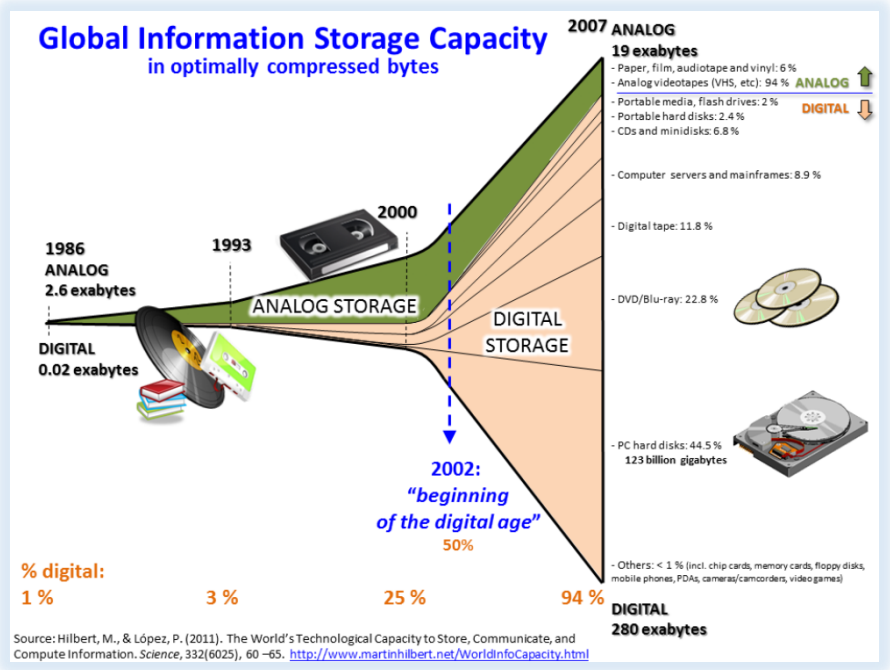
RDB kontra NoSQL



1 Big Data

Historia

- DB
- Data Warehouses
  Experimenty
- BigData

Hlavné zdroje Big Data

- mobilný Internet
- videa cez Internet
- vyhľadávanie
- sociálne siete
- vedecké pozorovania

Vlastnosti 3V

- volume
- velocity - generovanie, spracovanie
- variety

Použitie

- dostupnosť
- analýza
- segmentácia – reprezentatívnosť

Technológie

- Ukladanie - distribuované, škálovanie
- Spracovanie - distribuované a paralelné

Problémy

- **zber** a zhromažďovanie údajov a ich **ukladanie** nie je to isté, ako ich **pochopenie**
- zisťovanie toho, čo znamená Big Data, nie je to isté, ako **interpretácia** málo údajov
- chápanie správania sa **stáda** psov nevysvetľuje štekanie osamelého vlčiaka
- **nehomogenita** údajov - kombinácia informácie z mnohých zdrojov, v rôznych časoch
- vysoká **rozmernosť** môže viesť k nesprávnym a nepravdivým štatistickým a vedeckým záverom

2 RDB kontra NoSQL

- A.A.Coddov článok z 1970 *A relational model ...*
- Carlo Strozzi v 1998, pojem NoSQL
- Dwight Merriman, Eliot Horowitz, 2007, MongoDB

Viacgeneračný úspech **RDB** (relačných databáz) tkvie v

1) *zabezpečení konzistentnosti dát*
2) *použití sekundárnych kľúčov*
3) *dopytovacom jazyku SQL.*

vďaka *matematickému* modelu. Tabuľky RDB obsahujú *štrukturované dáta*, zázmamy, riadky, kde stĺpce majú svoj preddefinovaný typ. Zoznam stĺpcov spolu s ich nazvami a názvom tabuľky určuje schému. Relačné databázy párujú, spájajú dáta z viacerých tabuliek pomocou spoločných atribútov.

NoSQL (Not Only SQL, Non SQL) ukladá **neštruktúrované dáta** bez schémy, zvyčajne **sa vyhýba** operácie **spojenia** a škáluje **horizontálne**.

Ukazuje sa, že RDB systémy sa menej hodia pre

- aplikácie, pracujúce masívnym objemom dát, typy ktorých sa menia rýchlo - semi- a neštruktúrované dáta
- aplikácie, ktoré musia byť neustále zapnuté, prístupnené z mnohých rôznych zariadení a škálované globálne.

Scale up kontra *Scale out*
Namiesto monolitických serverov a ukladacích infraštruktúr sa dnešné aplikácie uprednostňujú **škálovateľné** architektúry podporované **open source** softvérmi.

NoSQL systémy charakterizujú také kľúčové vlastnosti, ako

1) *flexibilný dátový model*
2) *väčšia škálovateľnosť a*
3) *vyšší výkon.*

Všetko má ale svoju cenu. NoSQL systémy zápasia *konzistentnosťou* alebo menej efektívnym *dopytovacím jazykom*.

NoSQL systémy môžeme rozlišovať na základe nasledujúcich charakteristík

- Dátový model
- Dopytovací model
- Model konzistentnosti
- API
- Podpora a komunita

- **Dátový model**

  - Dokument modely – MongoDB, Elasticsearch
  - Gráf modely – Neo4j, Giraph
  - Key-Value modely a modely širokými stĺpcami – Redis, Riak
  - Široko-stĺpcové modely – Cassandra, Hadoop/HBase

- Dopytovací model

- Dokument model - celkom bohaté dopytovacie možnosti
- Key-Value modely - rýchle vyhľadávanie pomocou primárneho kľúča, slabé dopytovanie

- Model konzistentnosti

Nerelačné DB systémy zvyčajne manažujú viacnásobné kópie dát kvôli dostupnosti a škálovateľnosti. Preto konzistentnosť dát zaručujú na rôznych úrovniach, nové dáta sa v dopytoch odzrkadľujú s istým oneskorením. MongoDB poskytuje laditeľnú konzistentnosť.

- API

Po dokončení aplikácie, postavenej na konkrétnom type DBS, preniesť ju na iný DB základ je nákladné, náročné a riskantné.

Zrelosť API môže mať značný dopad na čas a náklady potrebné na vývoj a udržiavanie aplikácie a databáz.

- Podpora a komunita, UI

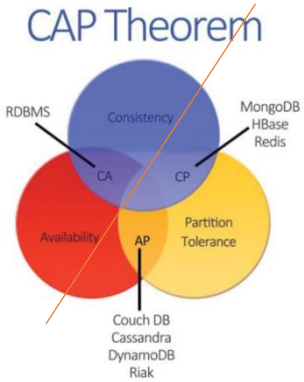
Databáza so silnou komunitou uľahčuje

- nájsť osvedčené postupy, kódové vzory
- najať vývojára, ktorý ovláda daný produkt.

Umelá inteligencia

ACID

- Atomicity - každá transakcia je v celku úspešná alebo vôbec nie
- Consistency - DB z validného stavu sa dostane v dôsledku transakcie do validného stavu (logické požiadavky - obmedzenia, kaskádovitosť, triggery)
- Isolation - konkurenčné transakcie sa chovajú akoby sekvenčne
- Durability - garancia pre prípad havárie, výpadku prúdu



CAP Theorem (Brewer's Theorem - 2000)

- Consistency – (atomicita a lineárna konzistentnosť)
- Availability
- Partition Tolerance – výpadok časti

Problém súbežnej garancie všetkých troch požiadaviek.

Ako zabezpečiť spoľahlivosť systému v prípade straty konzistentnosti?

ACID kontra **BASE** požiadavky

ACID	BASE
Atomicity Consistency Isolation Durable	Basically Available – v zmysle CAP Soft State – zmeny bez vstupu – doľadenie konz. Eventually Consistency – za dĺhší čas

MongoDB poskytuje vysokú dostupnosť, škálovateľnosť a partitioning/rozdeľovanie vďaka dokumentovému modelu s dynamickou schémou na úkor konzistentnosti a podpory transakcie.

Dynamická schéma znamená, že dokumenty v kolekcii môžu mať rôzne štruktúry resp. rovnaké [kľúče] atribúty môžu obsahovať dáta rôzneho typu.

MongoDB je navrhnutý pre prácu s dokumentami, ktoré nepotrebujú vopred definované stĺpce alebo typy, vďaka čomu dátový model je veľmi flexibilný.

Podobne ako primárny kľúč RDBMS (ktorý jednoznačne identifikuje každý riadok), MongoDB musí mať kľúč, ktorý jednoznačne identifikuje každý dokument v kolekcii a sa nazýva _id . Jeho hodnota sa buď zadáva manuálne, alebo sa automaticky vygeneruje. Táto hodnota kľúča je nemenná a môže byť akéhokoľvek typu dát s výnimkou polí.

Rank			DBMS	Database Model	Score		
Apr 2018	Mar 2018	Apr 2017			Apr 2018	Mar 2018	Apr 2017
1.	1.	1.	Oracle	Relational DBMS	1289.79	+0.18	-112.21
2.	2.	2.	MySQL	Relational DBMS	1226.40	-2.46	-138.22
3.	3.	3.	Microsoft SQL Server	Relational DBMS	1095.51	-9.28	-109.26
4.	4.	4.	PostgreSQL	Relational DBMS	395.47	-3.88	+33.69
5.	5.	5.	MongoDB	Document store	341.41	+0.89	+15.98
6.	6.	6.	DB2	Relational DBMS	188.95	+2.28	+2.29
7.	7.	7.	Microsoft Access	Relational DBMS	132.22	+0.27	+4.04
8.	↑ 9.	↑ 11.	Elasticsearch	Search engine	131.36	+2.81	+25.69
9.	↓ 8.	9.	Redis	Key-value store	130.11	-1.12	+15.75
10.	10.	↓ 8.	Cassandra	Wide column store	119.09	-4.40	-7.10
11.	11.	↓ 10.	SQLite	Relational DBMS	115.99	+1.17	+2.19

https://db-engines.com/en/

http://db-engines.com/en/ranking

Rank			DBMS	Database Model	Score	
Apr 2016	Mar 2016	Apr 2015			Apr 2016	Mar 2016
1.	1.	1.	Oracle	Relational DBMS	1467.53	-4.48
2.	2.	2.	MySQL	Relational DBMS	1370.11	+22.39
3.	3.	3.	Microsoft SQL Server	Relational DBMS	1135.05	-1.45
4.	4.	4.	MongoDB	Document store	312.44	+7.11
5.	5.	5.	PostgreSQL	Relational DBMS	303.73	+4.10
6.	6.	6.	DB2	Relational DBMS	184.08	-3.85
7.	7.	7.	Microsoft Access	Relational DBMS	131.97	-3.06
8.	8.	8.	Cassandra	Wide column store	129.67	-0.66
9.	9.	↑ 10.	Redis	Key-value store	111.24	+5.02
10.	10.	↓ 9.	SQLite	Relational DBMS	107.96	+2.19
11.	11.	↑ 14.	Elasticsearch	Search engine	82.58	+2.41

https://db-engines.com/en/