



TECHNICAL UNIVERSITY OF LIBEREC
Faculty of Mechatronics, Informatics
and Interdisciplinary Studies



BIG DATA

Lukáš Matějů

30.8.2024 | TPB





ČÁST I.: ORGANIZACE PŘEDMĚTU TECHNOLOGIE PRO BIG DATA



ORGANIZACE

- základní informace
 - přednášející i cvičící
 - Lukáš Matějů
 - lukas.mateju@tul.cz
 - rozsah předmětu 2+2
 - veškeré materiály zveřejňovány na [elearningu FM](#)
- přednášky
 - každé pondělí od 10:40
 - budova A, místnost A-A0110
 - účast nepovinná, ale vítaná



ORGANIZACE

- cvičení
 - každé pondělí od 12:30 a 14:20
 - budova A, místnost A-A0304
 - samostatné práce volně doplňující přednášky
 - 10 povinných a 10 bonusových úloh
 - každá bonusová úloha je za **0,5** bodu
 - na vypracování a odevzdání úloh je **1** týden
 - odevzdává se **výhradně** na cvičeních (s osobním vysvětlením)
 - za každý týden opožděného odevzdání je **-0,5** bodu
 - finální počet úloh může být ovlivněn odpadnutím výuky
 - 2 povolené absence
 - každá další absence je za **-3** body (onemocnění lze snadno omluvit)
 - absence automaticky prodlužuje dobu odevzdání o **1** týden



ORGANIZACE

- zápočet
 - odevzdané a správně vyřešené povinné úlohy ze cvičení
- zkouška
 - prezenční
 - písemná
 - max 20 bodů
 - 10 otázek po 2 bodech
 - body ze cvičení jsou přenášeny ke zkoušce
 - zaměřená na základní koncepty probírané v rámci předmětu



ORGANIZACE

- hodnocení
 - dvě varianty
 - jen za bonusové body ze cvičení...
 - 5,0 bodů -> 1
 - 4,5 bodů -> 2
 - 4,0 body -> 3
 - povinná docházka na přednášky i cvičení (2 povolené absence)
 - absolvování písemné zkoušky...
 - maximum 25 bodů (20 + 5)
 - ≥ 21 bodů -> 1 ≥ 19 bodů -> 1-
 - ≥ 17 bodů -> 2 ≥ 15 bodů -> 2-
 - ≥ 13 bodů -> 3 < 13 bodů -> 4
 - v případě odpadnutí výuky budou potřebné body upraveny





ČÁST II.: BIG DATA



„Až 3,3 miliard uživatelů smartphonů.“
[2019]

„Každou minutu je na YouTube nahráno 300 hodin videí.“ [2017]

„Denně je posláno 140 milionů tweetů.“ [2019]

„Google zaznamenává více jak 63 000
vyhledávání každou vteřinu.“ [2019]

„Uživatelé každý den nahrají na Facebook
více jak 300 milionů fotek.“ [2017]

„Reddit – 25 milionů hlasů každý den.“ [2019]

„PornHub přenese každou vteřinu 147 GB dat.“ [2018]



„Denně je odesláno 294 miliard emailů.“ [2019]

„Instagram zaznamenává denně 4,2 miliardy liků.“ [2018]

„NASA má k dispozici 32 PB dat pozorování a simulací klima.“ [2015]

„Za sekundu je odhadem vytvořeno 1,7 MB dat pro každého člověka.“ [2019]

„Za dva dny vytvoříme tolik informací jako bylo vygenerováno od počátku věků do roku 2003.“ [2014]

„Jen 0,5 % big data je ve skutečnosti analyzováno.“ [2014]

DEFINICE BIG DATA

„Soubory dat, jejichž velikost je mimo schopnosti zachycovat, spravovat a zpracovávat data běžně používanými softwarovými prostředky v rozumném čase.“ [Gartner]

„Big data jsou tam, kde jsou potřeba paralelní výpočty pro zpracování dat.“ [2018]

„Nový nástroj pro vyhledávání relevantních dat a jejich analýzu.“ [Forbes]

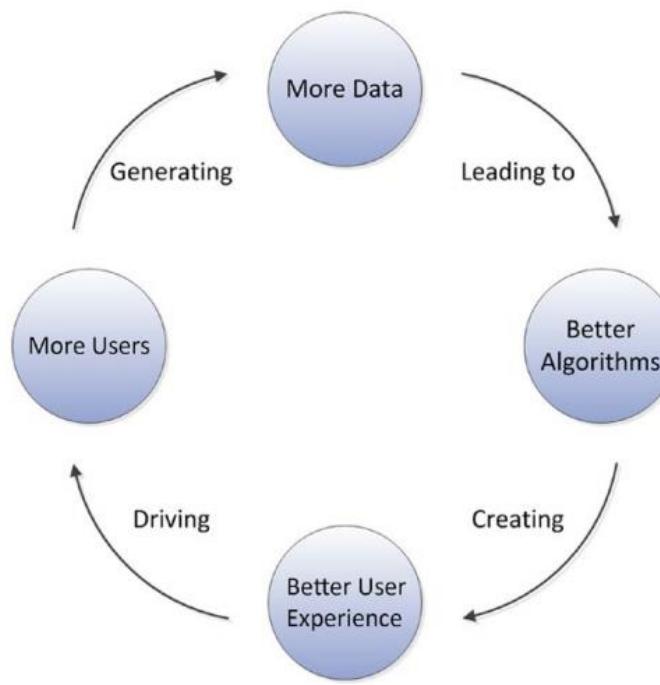


ÉRA BIG DATA

- dnes žijeme v éře big data
 - výrazně ovlivňuje nás každodenní život
- vychází ze dvou základních předpokladů
 - existuje stálý (rostoucí) a rychlý příspun nových dat
 - v roce 2017 proteklo internetem 1.5 ZB (10^{21}) dat
 - pro rok 2022 je odhad 4.8 ZB
 - inovace ve způsobu zpracování
 - cloud computing
 - propůjčení výpočetních serverů
 - umožňuje provádět výpočty kdekoliv a odkudkoliv
- umožňuje dynamickou a škálovatelnou analýzu dat



ÉRA BIG DATA



HARRISON, Guy. *Next generation databases: NoSQL, NewSQL, and Big Data*. ISBN 978-1484213308.



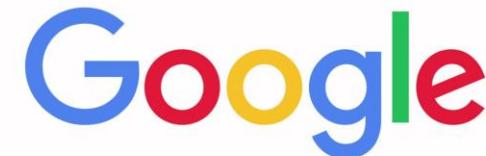
APLIKACE BIG DATA

- obecná aplikace
 - big data oproti klasickým modelům umožňují
 - kvalitnější modely
 - přesnější výsledky
 - aplikace cílené na jednotlivce
 - aplikace cílené na skupinu
- snaha o pochopení a vytěžení informací z velkého množství dat
 - big data analytics



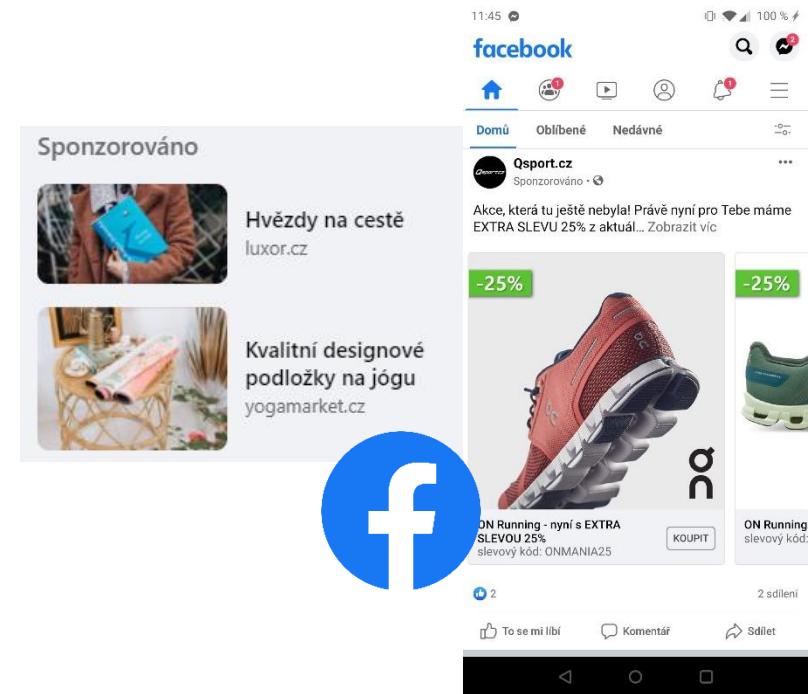
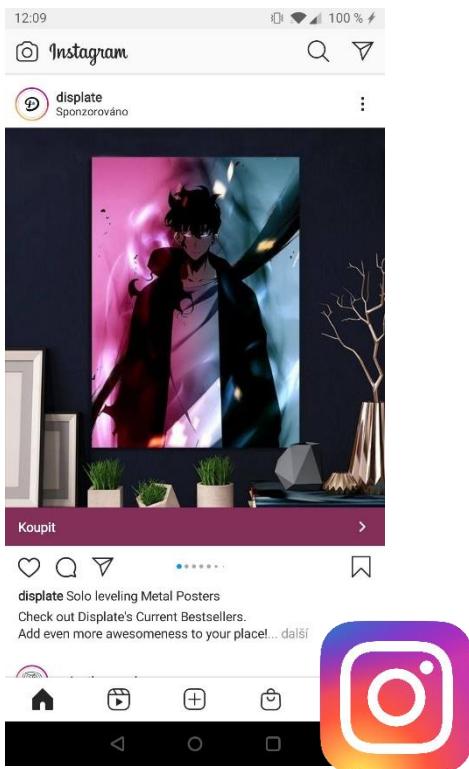
CÍLENÝ MARKETING

- cílené doporučení zboží
- cílená reklama
- určeno na základě aktivity uživatele
 - navštívené stránky
 - historie vyhledávání
 - historie objednávek
 - zhlédnutá videa
 - social media...





CÍLENÝ MARKETING



DOPORUČOVACÍ SYSTÉMY

- doporučení produktu na základě předchozí aktivity uživatele
 - Netflix navrhuje uživateli filmy / seriály
 - Steam doporučuje hry
 - Amazon zboží
 - cílem je využít znalosti o uživateli k doporučení ideálního produktu
 - udržení pozornosti uživatele
 - prodej produktu



DOPORUČOVACÍ SYSTÉMY

PROTOŽE JSTE HRÁLI ROCKET LEAGUE®



Fotbalové (kopačá) Kompetitivní S online kooperací S nekonečným běháním

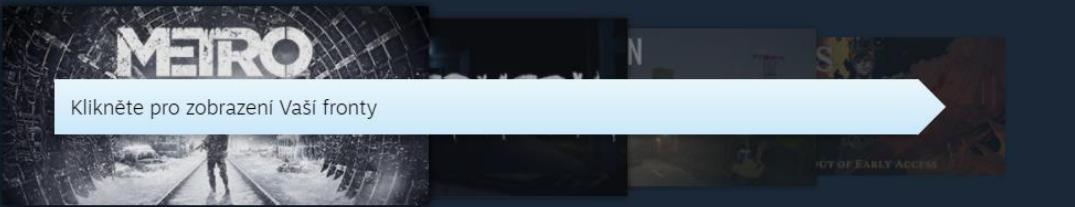
Fotbalové (kopačá) S online kooperací Realistické E-sportovní

Fotbalové (kopačá) Kompetitivní Závodní S důrazem na týmovou hru

Fotbalové (kopačá) S online kooperací Zábavné Kompetitivní

< >

VAŠE FRONTA DOPORUČENÍ



Klikněte pro zobrazení Vaši fronty

ZJISTIT VÍCE



CO BY SE VÁM MOHLO LÍBIT



NA ZÁKLADĚ HER, KTERÉ JSTE HRÁLI PROCHÁZET A PŘIZPŮSOBIT

< >



DOPORUČOVACÍ SYSTÉMY



Top Picks for Lukáš



Doporučené playlisty



Posluněný folk
Playlist • YouTube Music



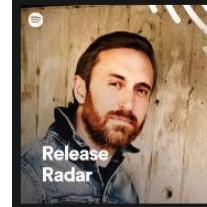
Presenting Ed Sheeran
Playlist • YouTube Music



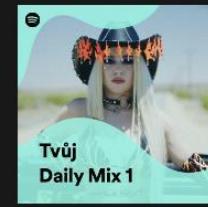
Presenting Škwar
Playlist • YouTube Music

Speciálně pro uživatele Lukáš Matějů

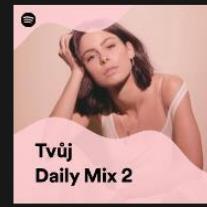
Čím víc budeš poslouchat, tím lepší budou naše doporučení.



Release Radar
Nejnovější hudba umělců, které sleduješ, a nové singly vybrané speciálně...



Daily Mix 1
Ava Max, Sia, Alessia Cara a další žánry

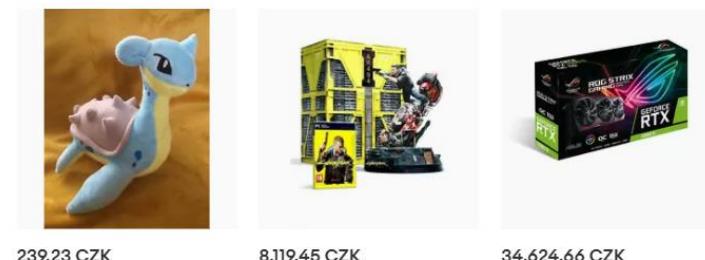


Daily Mix 2
Lena, Ana Kohler, Mike Williams a další žánry



DOPORUČOVACÍ SYSTÉMY

Your Recently Viewed Items | See all →



239.23 CZK 8,119.45 CZK 34,624.66 CZK

Pokémon Go | See all →
Recommended for you



Doporučeno přímo pro Vás



59,- **od 31,-**



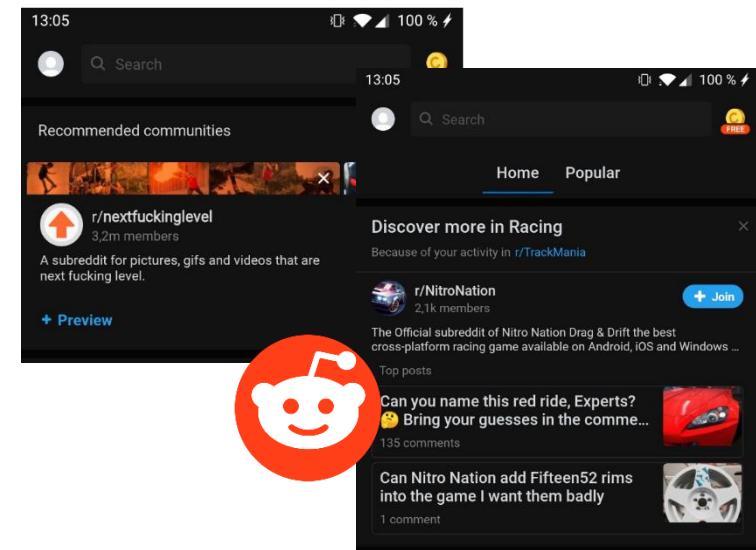
89,- **od 66,-**



499,- **359,-**



199,- **175,-**



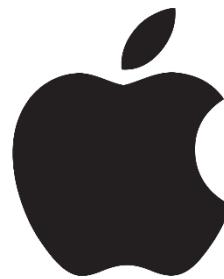
ANALÝZA SENTIMENTU

- založená na hodnocení / komentářích uživatelů
 - pozitivní / negativní (NLP)
 - analýza zboží na základě všech recenzí
 - negativní
 - nedoporučovat
 - analýza uživatele na základě jeho recenzí
 - návrh zboží
 - mínění veřejnosti o firmě
 - je potřeba něco změnit?
 - reakce veřejnosti na nastalou událost
 - je potřeba reakce?
 - reakce veřejnosti na nově uvedený produkt
 - je o něj zájem?



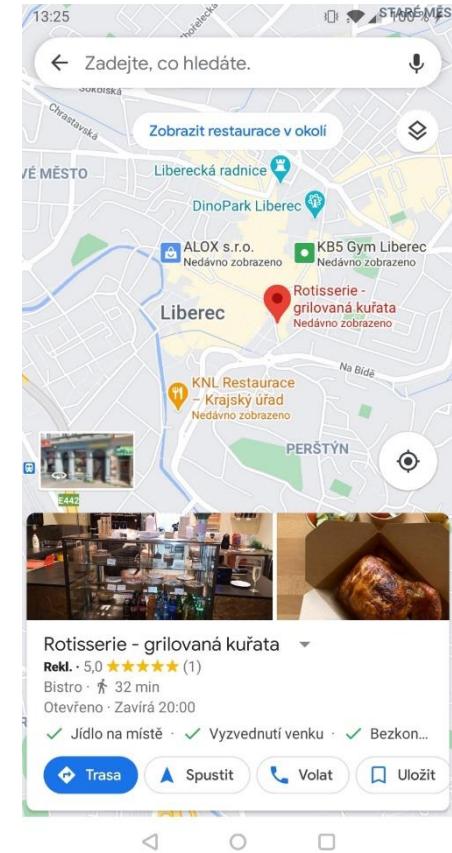
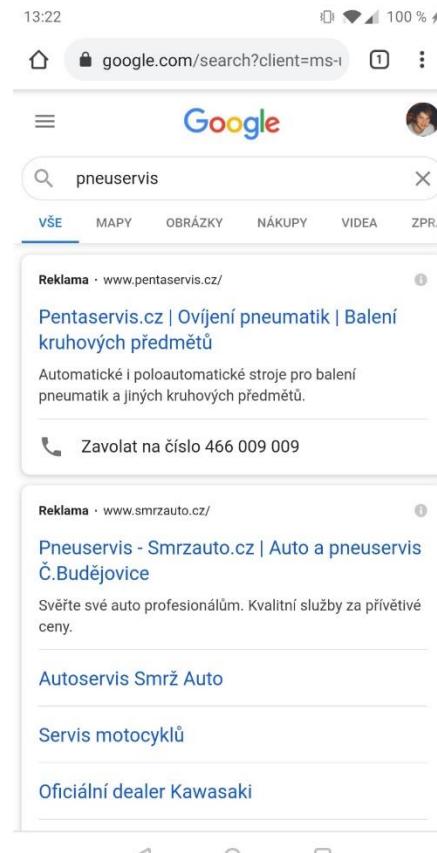
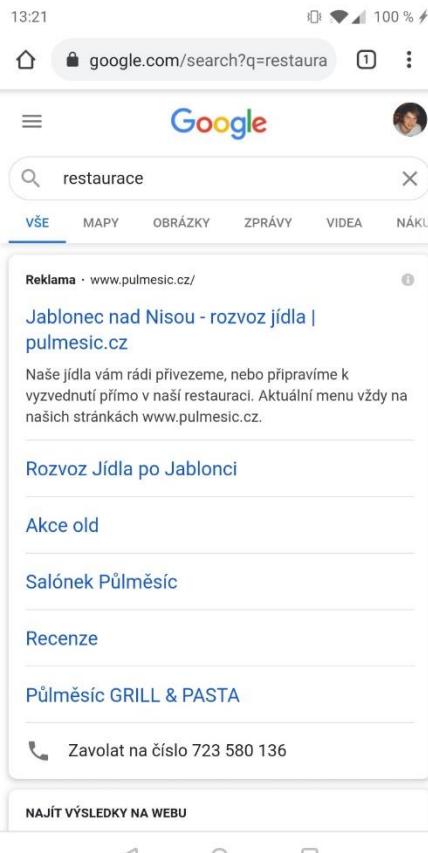
MOBILNÍ REKLAMY

- kombinují informace o uživateli se senzory v telefonu
 - primárně GPS
 - slevové akce určené přímo pro zákazníka
 - doporučení obchodu / restaurace v okolí
 - na základě předchozího vyhledávání zboží
 - na základě stravovacích návyků
 - ...



Google

MOBILNÍ REKLAMY



CHOVÁNÍ SKUPINY

- úprava služeb podle chování cílové skupiny
 - letecké společnosti
 - ranní lety kompletně vyprodány
 - večerní lety s poloviční kapacitou
 - přesun některých večerních letů na ráno
 - platí v USA, v Japonsku je ale situace opačná
 - cílení pro konkrétní skupinu / trh



BIOMEDICÍNA

- obrovské množství dat (lidský genom)
 - až 40 EB v roce 2025
- využití
 - výzkum
 - léčebních postupů, léků, atd.
 - cílená léčba
 - určená přímo na míru pacientovi
 - léčba rakoviny





ODKUD BIG DATA POCHÁZÍ?

- většina zdrojů existovala již dříve
- změnil se ale náš přístup a množství analyzovaných dat
- tři hlavní zdroje dat
 - stroje
 - senzory - zdravotnické, průmyslové, environmentální, ...
 - lidé
 - social media, články, blogy, ...
 - organizace
- data mohou být
 - strukturovaná, částečně strukturovaná nebo nestrukturovaná
- hlavní přínos spočívá v jejich kombinaci

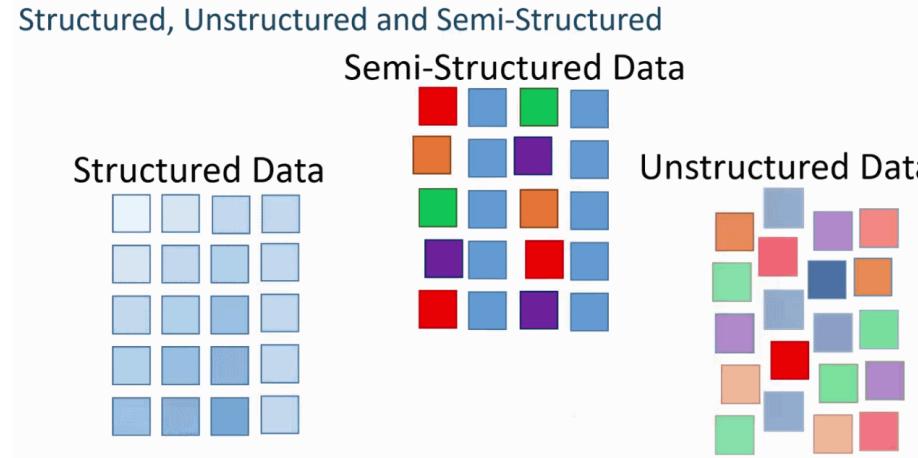


ROZDĚLENÍ DAT

- strukturovaná data
 - informace formátované a převedené do definovaného datového modelu
 - uložené v definovaných polích umožňujících snadný přístup a čtení
 - strojová, lidská i organizační data
 - snadno analyzovatelná
 - např. relační databáze – relace, atributy
- nestrukturovaná data
 - data v nezpracované podobě, nemají žádný specifický formát
 - obtížné na zpracování kvůli komplexnosti a složitosti
 - flexibilní, mnoho podob
 - social media příspěvky, chaty, satelitní snímky, prezentace...
 - ukládána a analyzována v datových skladech

ROZDĚLENÍ DAT

- částečně strukturovaná data
 - na pomezí mezi strukturovanými a nestrukturovanými daty
 - příkladem může být digitální fotografie
 - samotný obraz nemá pevně definovanou strukturu
 - obsahuje ale strukturované atributy typu datum a místo pořízení snímku, ID zařízení, ...



<https://www.astera.com/type/blog/structured-semi-structured-and-unstructured-data/>



STROJOVÁ DATA

- největší zdroj big data
 - velký hadronový urychlovač generuje 40 TB dat každou vteřinu experimentů
 - Boeing 787 produkuje 0,5 TB dat během každého letu
- chytré zařízení
 - zařízení schopná měřit a produkovat big data (pomocí senzorů)
 - proč chytré?
 - schopnost připojení k dalším zařízením / sítím
 - autonomní sběr a analýza dat
 - mají povědomí o prostředí
 - internet věcí (internet of things)
 - např. chytré hodinky



STROJOVÁ DATA

- zpracování v reálném čase
 - vztahy se zákazníky
 - detekce podvodů
 - monitorování systému
 - okamžitá analýza a reakce
- zpracování se přesouvá za daty
- potřeba škálovatelných výpočetních systémů



LIDSKÁ DATA

- lidé vytváří obrovské množství dat na internetu každý den
 - Facebook, Twitter, LinkedIn
 - Instagram, YouTube
 - blogy, komentáře
 - vyhledávání
 - textové zprávy, emaily
 - osobní dokumenty
- většina dat je textových a nestrukturovaných
 - složité zpracování
 - nelze použít předdefinované datové modely (NE relační databáze)
 - komplikace
 - množství formátů
 - množství dat a jejich rychlý růst
 - potvrzení je časově náročné (sběr, uložení, těžba, čištění a zpracování)



LIDSKÁ DATA

- zpracování – několik základních open source frameworků
- nástroje pro zpracování a analýzu jsou vyvíjeny od nuly
 - historicky většina založena na Hadoop
 - zpracování velkého množství dat v distribuovaném výpočetním prostředí
 - často potřeba zpracování dat v reálném čase
 - aktualizace na social media
 - tržní data
 - Apache Storm, Spark, Flink
 - ukládání dat
 - NoSQL databáze
 - ukládání dat typicky na výpočetním cloudu
 - zpracování po vrstvách
 - těžba a uložení, předzpracování, analýza



ORGANIZAČNÍ DATA

- důvěryhodná a užitečná data
- liší se výrazně organizaci od organizace
 - transakce, kreditní karty, bankovnictví, akcie, zdravotní záznamy, senzory, ...
- současné a budoucí použití ale i analýza minulosti
 - predikce prodejů / úspěchů na základě dat a dění ve světě
- vysoce strukturovaná data
 - datový model
 - transakce, referenční tabulky, vazby + metadata doplňující kontext
 - ukládána v relačních databázích (+ SQL)
- riziko
 - datová sila

CHARAKTERISTIKA BIG DATA

„Soubory dat, jejichž velikost je mimo schopnosti zachycovat, spravovat a zpracovávat data běžně používanými softwarovými prostředky v rozumném čase.“ [Gartner]

- 3 základní V's
 - objem (volume)
 - množství dat generovaných každou vteřinu
 - různorodost (variety)
 - neustále rostoucí počet forem dat
 - rychlosť (velocity)
 - rychlosť generování dat
 - rychlosť přesouvání dat z jednoho bodu do druhého



CHARAKTERISTIKA BIG DATA

- další často uváděná V's
 - věrohodnost (veracity)
 - zaujatost, šum, abnormality v datech
 - nejistota v pravdivost a věrohodnost dat
 - valence (valence)
 - propojovatelnost dat formou grafů
- a nezapomenout na
 - hodnotu (value)
 - srdce a lepidlo všech ostatních V's
 - jak z dat získat jejich pravou hodnotu?



<https://suryagutta.medium.com/the-5-vs-of-big-data-2758bfcc51d>

OBJEM (VOLUME)

- množství dat generovaných každou vteřinu
 - PB, EB, ZB (10^{21} bytů)
- zpracování a ukládání velkého objemu dat přináší výzvy
 - škálování
 - horizontální / vertikální / kombinace
 - zajištění kapacity úložiště a výkonu na zpracování dat
 - dostupnost
 - přístup k datům a možnost jejich zpracování
 - bandwidth a výkon
 - přístup k datům v potřebný okamžik
- cílem firem je analýza dat
 - zlepšení poskytovaného produktu / služby
 - náskok před konkurencí

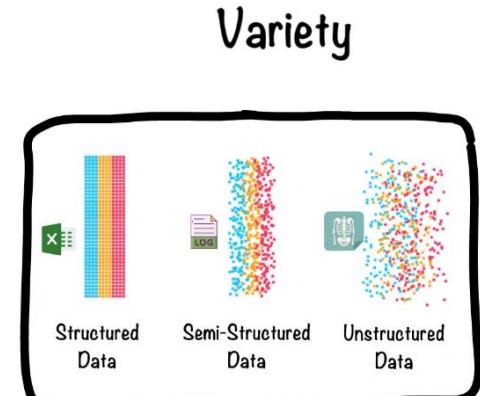
Volume



<https://www.youtube.com/watch?v=bAyrObl7TYE>

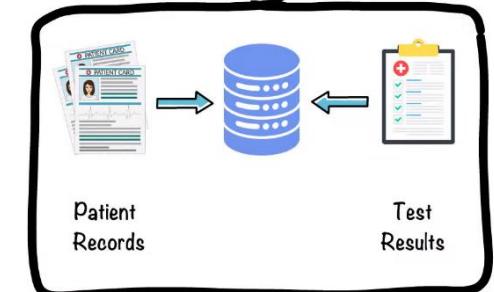
RŮZNORODOST (VARIETY)

- rostoucí různorodost forem dat
 - textová data, obrazová data, síťová data, geografické mapy, simulace, ...
- různá různorodost
 - strukturovaná různorodost
 - rozdíl ve struktuře dat (EKG vs. novinový článek)
 - nosičová různorodost
 - forma, ve které jsou data získána (audio nahrávka vs. text)
 - sémantická různorodost
 - jak data interpretovat
 - různorodost dostupnosti
 - data generovaná v reálném čase (senzory) nebo uložená (záznamy)
 - dostupná neustále (kamery) nebo jen příležitostně (sondy, satelity)
- hybridní data – např. emaily



<https://www.youtube.com/watch?v=bAyrObI7TYE>

RYCHLOST (VELOCITY)



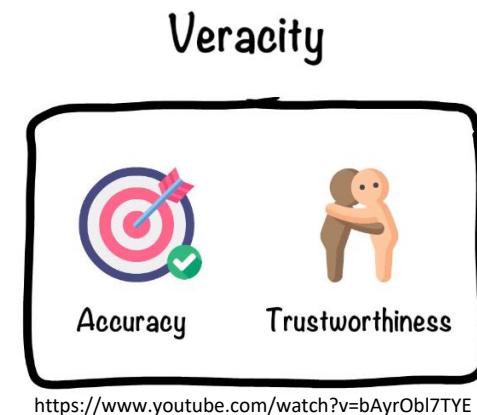
<https://www.youtube.com/watch?v=bAyrObI7TYE>

- rostoucí rychlosť vytvárení big data
 - z rôznych zdrojov sa liší (aktualizácia jednou za deň vs. každou sekundu)
- rostoucí rychlosť ukladania a analýzy big data (zpracovanie)
- cieľom je zpracovanie v reálnom čase
 - vytvorenie reklamy na základe užívateľskej historie a zobrazenie pri hľadaní
 - pomalá reakcia vede k ztrátenej príležitosti
 - napr. kempovanie – zajíma mň dnešné počasie, ne počasie, ktoré bolo pred rokom
 - napr. neštěstí – okamžitý zásah záchranných složiek
- potreba zvážiť rychlosť vytvárenia i rychlosť zpracovania dat
 - zpracovanie môže čakať na data
 - data môžu čakať na zpracovanie
 - rovnováha



VĚROHODNOST (VERACITY)

- odpovídá kvalitě dat
- big data jsou často zašuměná, nejistá nebo nepřesná
 - analýza big data může být jen tak dobrá jako vstupní data
 - hlavní problém u nestrukturovaných dat
- kvalita big data závisí na
 - přesnosti dat
 - spolehlivosti zdroje
 - způsobu vygenerování dat
 - analýze kontextu
- potřeba monitorovat posbíraná data, jejich původ a jak byly dříve analyzovány (x Google Flu Trends)

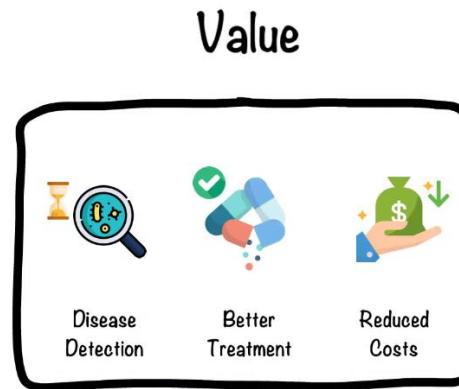


VALENCE (VALENCE)

- propojení dat
 - přímé spojení
 - město – stát
 - zaměstnanec - zaměstnání
 - nepřímé spojení
 - dva vědci jsou propojeni, protože jsou oba fyzici
- poměr množství propojených dat ku možnému počtu propojení
- postupně se časem zvyšuje
 - může vést ke vzniku nových vzorů skupin -> změna
 - potřeba komplexnější analýzy
 - potřeba modelovat a analyzovat valenci
 - detekce skupin, událostí

HODNOTA (VALUE)

- srdce a lepidlo všech ostatních V's
- pravá hodnota dat
 - resp. potenciální hodnota dat z pohledu informací, které obsahují
 - jak ji získat?
- big data ztrácí význam, pokud nenesou hodnotu pro toho, kdo je analyzuje



<https://www.youtube.com/watch?v=bAyrObI7TYE>

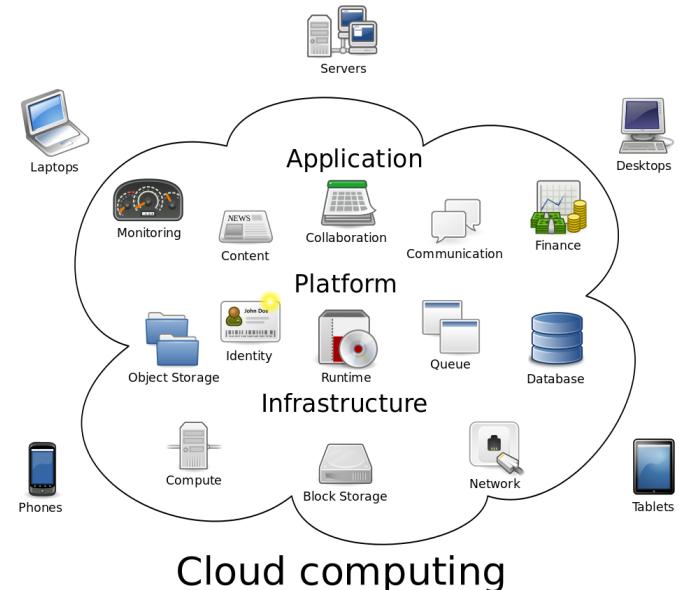


ČÁST III.: CLOUD COMPUTING



CLOUD COMPUTING

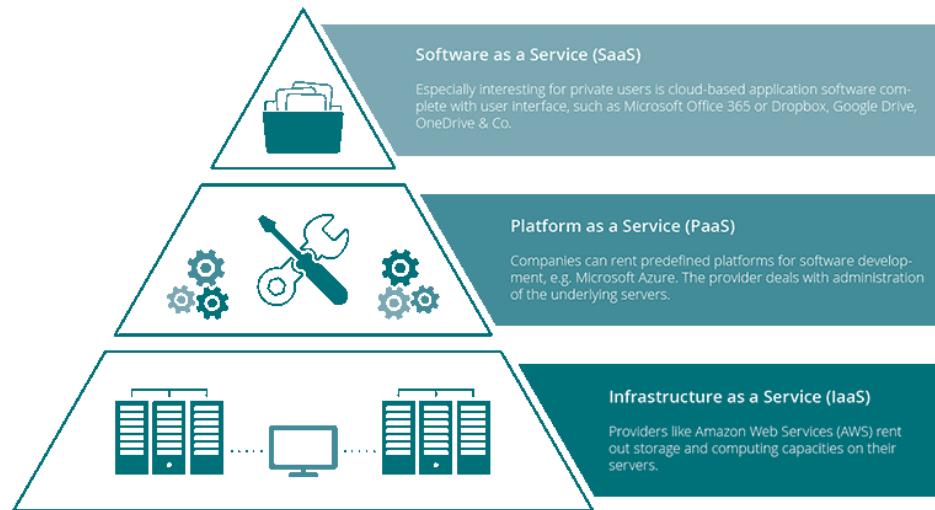
- internetový model vývoje a používání počítačových technologií
 - dostupnost systémových prostředků na vyžádání bez aktivní správy klienta
 - nejčastěji datová (cloudová) úložiště a výpočetní výkon
 - systémové prostředky hostované na vzdálených datových centrech
 - spravované cloudovými poskytovateli služeb
 - uživatelé k nim přistupují přes internet
 - prostředky jsou uživatelům pronajímány
 - charakteristika
 - více nájmů – zdroje sdíleny všemi uživateli
 - škálovatelnost – změna prostředků za běhu
 - pay as you go – platba za využité prostředky
 - aktuálnost – zajištěna poskytovatelem
 - přístup přes internet



https://en.wikipedia.org/wiki/Cloud_computing

CLOUD COMPUTING

- distribuční model
 - vyjadřuje, co je v rámci služeb nabízeno
 - v základu tři modely s rostoucí mírou abstrakce
 - definice od NIST
 - často znázorňovány jako vrstvy

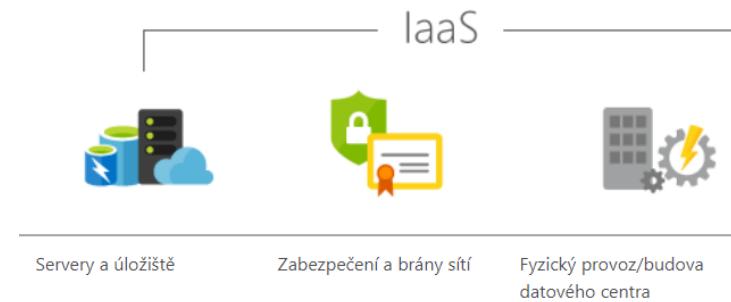


<https://doublehorn.com/cloud-computing-service-models/>



CLOUD COMPUTING

- distribuční model
 - IaaS
 - infrastruktura jako služba (Infrastructure as a Service)
 - předem připravená výpočetní infrastruktura poskytovaná a spravovaná přes internet
 - virtualizované komponenty
 - virtuální stroje, servery, úložiště, síť, ...
 - umožňuje rychle vertikálně navýšovat kapacitu podle aktuální potřeby
 - Amazon Web Services, Rackspace, Windows Azure

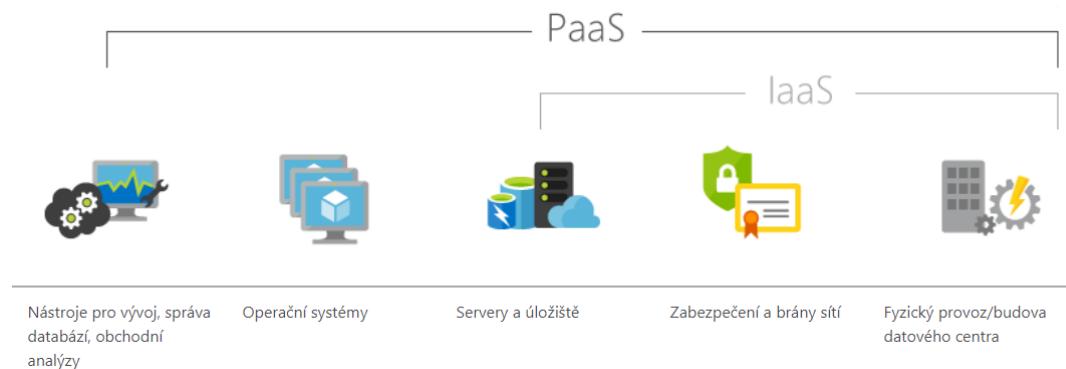


<https://azure.microsoft.com/cs-cz/overview/what-is-iaas/>



CLOUD COMPUTING

- distribuční model
 - PaaS
 - platforma jako služba (Platform as a Service)
 - úplné prostředí pro vývoj a nasazení v cloudu
 - poskytuje prostředky umožňující dodat jednoduché cloudové aplikace ale i propracované aplikace s podporou cloudu
 - zahrnuje nejen IaaS infrastrukturu, ale i další nástroje
 - operační systém, middleware, vývojářské nástroje, systémy správy databází, ...
 - celý životní cyklus webové aplikace – sestavení, testování, nasazení, správa a aktualizace
 - Google App Engine, Heroku, Azure App Services, AWS Elastic Beanstalk

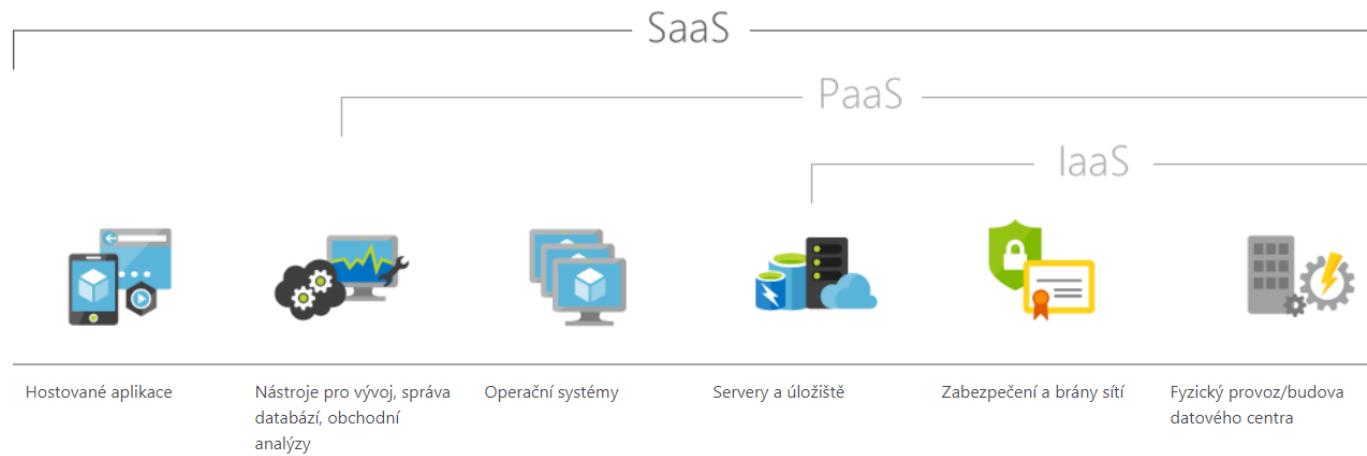


<https://azure.microsoft.com/cs-cz/overview/what-is-paas/>



CLOUD COMPUTING

- distribuční model
 - SaaS
 - software jako služba (Software as a Service)
 - umožňuje uživatelům připojení ke cloudovým aplikacím a jejich používání přes internet
 - úplné softwarové řešení
 - včetně IaaS a PaaS funkcionality
 - Dropbox, Office 365, Google G Suite, Slack



<https://azure.microsoft.com/cs-cz/overview/what-is-saas/>

CLOUD COMPUTING

- distribuční model
 - existují ale i další modely nabízených služeb
 - DBaaS a BDaaS
 - databáze jako služba (Database as a Service)
 - big data jako služba (Big Data as a Service)
 - často SaaS umožňující uživatelům pronajmout databázové služby
 - big data
- cloud virtualizuje zdroje distribuované v clusterech / datacentrech
 - vhodný framework na ukládání a zpracování big data
 - virtualizace velkého množství strojů přináší velká datová úložiště a velkou výpočetní sílu
 - schopnost pojmut, zpracovávat a analyzovat big data



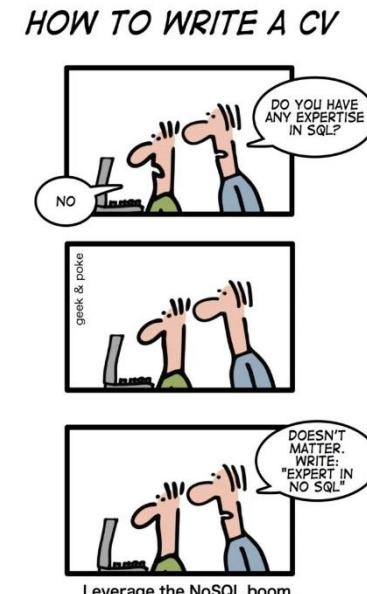


ČÁST IV.: DATABÁZE PRO BIG DATA



NoSQL DATABÁZE

- význam NoSQL
 - často chybně – no SQL (žádné SQL)
 - správně – not only SQL (ne jen SQL)
 - u některých databází je možné SQL používat
- nerelační databáze
 - žádné relace (tabulky)
- technologie známá již od 60. let 20. století
 - snaha vyřešit problémy relačních databází
 - jedno řešení pro všechno
- ale až v současné době nárůst popularity
 - obrovské objemy a široká škála dat
 - generované cloudy, mobilními zařízeními, sociálními médií, ...



<http://geek-and-poke.com/>

NoSQL DATABÁZE

➤ NoSQL databáze se vyvinuly

- dokáží zpracovávat obrovské objemy rychle se měnících nestrukturovaných dat
- pomáhají vývojářům rychle vytvářet databázové systémy pro ukládání nových informací a jejich přípravu pro vyhledávání a analýzu
- umožňují
 - flexibilní vývoj
 - flexibilní zpracování dat
 - provoz v libovolném měřítku

➤ cílem tedy ve výsledku není nahradit relační databáze

- jedná se spíš o alternativu s jinou aplikací (kladivo vs. sekera)
- vhodné např. pro big data a webové aplikace v reálném čase

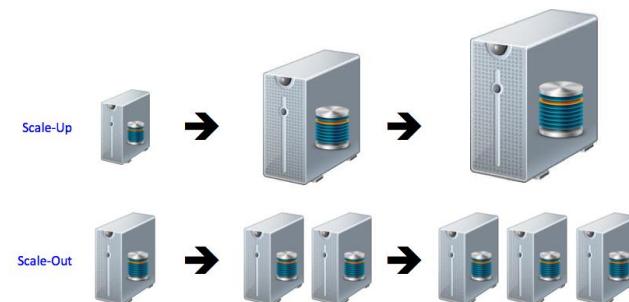


VLASTNOSTI

- celá škála různých technologií
 - vzájemně se výrazně liší
 - neexistuje jednotný interface
 - není standardizovaný jeden jazyk
- volné schéma
 - není nutné předdefinované nastavení
 - flexibilita
- podpora pro všechny data
 - strukturovaná, částečně strukturovaná i nestrukturovaná data
 - vyhýbají se ale (většinou) vazbám mezi daty
 - různé typy big data

VLASTNOSTI

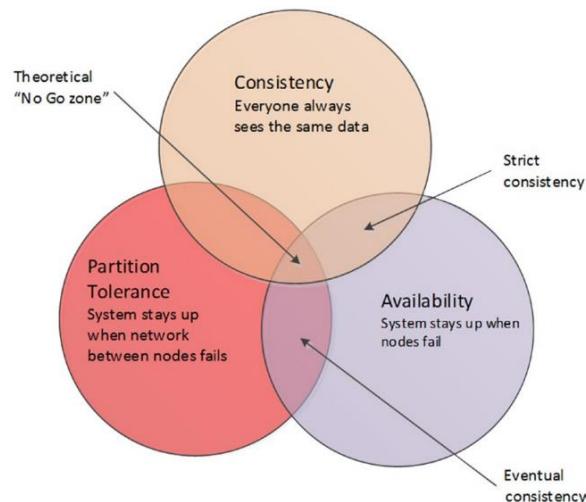
- horizontální škálování (ven)
 - škálovatelnost je vlastnost systému zvládat rostoucí (nebo klesající) množství požadavků přidáváním (nebo odebíráním) dalších prostředků
 - zvyšování kapacity
 - nové komponenty jsou levné
 - přesun k paralelnímu zpracování
 - rozdělení dat mezi uzly – sharding
 - rozdělení záznamů na části umístěné na různých serverech
 - podpora replikace dat
 - automatická distribuce změn v originále do kopíí
 - výkon, flexibilita, redundancy
- vhodné pro analýzu velkého množství dat



<https://medium.com/faun/scalability-248019b918ed>

VLASTNOSTI

- distribuovaná povaha
 - CAP teorém
 - pro distribuovaný datový sklad nelze zajistit více jak 2 následující vlastnosti
 - konzistence – consistency
 - všechny uzly mají stejná data
 - každé čtení vrací poslední výsledek, nebo chybu
 - dostupnost – availability
 - na každý dotaz je vrácena (nechybová) odpověď
 - odolnost k přerušení – partition tolerance
 - systém funguje dál i při zdržení či ztrátě zprávy
 - nezbytné pro big data
 - eventuálně konzistentní (AP)
 - často volena škálovatelnost a vysoká dostupnost než 100% konzistence
 - striktně konzistentní (AC)

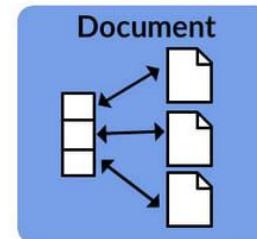
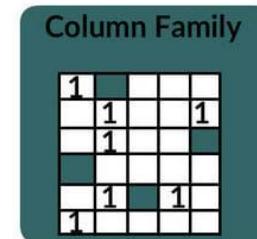
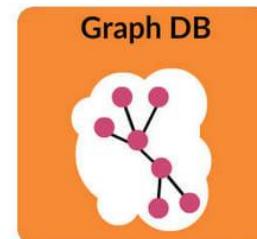
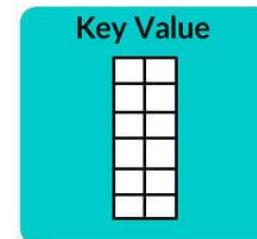


HARRISON, Guy. *Next generation databases: NoSQL, NewSQL, and Big Data*. ISBN 978-1484213308.



TYPY NoSQL DATABÁZE

- key-value úložiště
 - Riak, Redis, Couchbase Server
- dokumentové databáze
 - MongoDB, CouchDB, Elasticsearch
- sloupcová úložiště
 - Apache Cassandra, Druid, HBase
- grafová úložiště
 - Neo4j
- časové řady
 - InfluxDB



<https://dev.to/lmolivera/everything-you-need-to-know-about-nosql-databases-3o3h#key>

KEY-VALUE ÚLOŽIŠTĚ

- nejjednodušší typ NoSQL databází
 - Riak, Redis, Couchbase Server, ...
- data ukládána pomocí klíče (key) a přiřazené hodnoty (value)
 - každá hodnota (blob) přidružena k jedinečnému klíči
 - úložiště použije tento klíč k uložení dat pomocí odpovídající hash funkce
 - funkce je vybírána tak, aby zajistila rovnoměrné rozložení hashovaných klíčů
 - podobné asociativnímu poli / hashi
- optimalizované pro velké objemy dat
 - ale jednoduchých (žádné vazby)
 - velmi rychlé, ale špatně upravitelné
 - dotazy vrací celou hodnotu



REDIS

- Remote Dictionary Server (od 2009)
 - paměťové open-source key-value úložiště
 - v současné době nejpopulárnější zástupce key-value úložišť
 - úložiště je celé udržováno v paměti
 - podpora celé řady abstraktních datových typů
 - replikace, shardování, klastrování



```
127.0.0.1:6379> SET foo 100
OK
127.0.0.1:6379> SET bar "HELLO WORLD"
OK
```

```
127.0.0.1:6379> MSET key1 "Hello" key2 "world"
OK
```

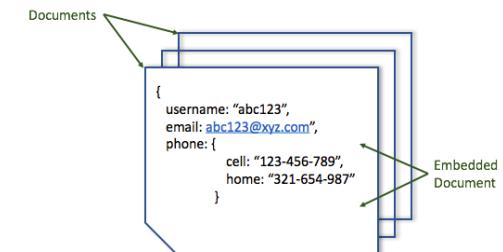
```
127.0.0.1:6379> GET foo
"100"
```

```
127.0.0.1:6379> ZADD users 1981 "Brad"
(integer) 1
127.0.0.1:6379> ZADD users 1990 "Jen"
(integer) 1
127.0.0.1:6379> ZADD users 1991 "Luke"
(integer) 1
```

```
127.0.0.1:6379> SADD auta "Ford" "Skoda" "BMW"
(integer) 3
127.0.0.1:6379> SMEMBERS auta
1) "BMW"
2) "Skoda"
3) "Ford"
127.0.0.1:6379> SCARD auta
(integer) 3
```

DOKUMENTOVÉ DATABÁZE

- pravděpodobně nejpoužívanější NoSQL databáze
 - MongoDB, CouchDB, ...
- v principu podobné key-value úložištím
 - zachován princip key-value (klíč-hodnota)
 - klíč jednoznačným identifikátorem hodnoty
 - ale hodnoty obsahují strukturovaná nebo částečně strukturovaná data
 - tzv. dokumenty (JSON, XML, ...)
 - koncepčně odpovídají objektům v OOP
 - volné schéma, ale vyhýbají se vazbám
 - samotná data mohou být indexována a dotazována
 - indexy nad atributy dat
 - dotazy na strukturu dat i na prvky v této struktuře
 - vylepšuje rychlosť vyhledávání
 - je možné získat jen požadované části dokumentů

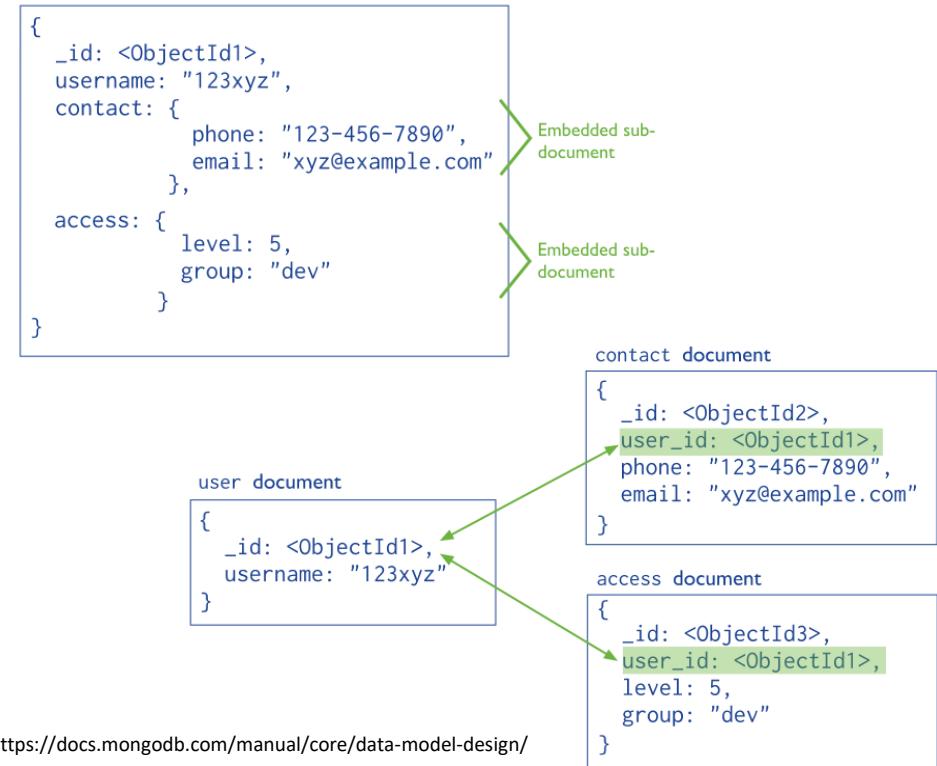


<https://expert.data.blog/category/bigdata/>



DOKUMENTOVÉ DATABÁZE

```
[  
  {  
    "_id" : 1,  
    "artistName" : "Iron Maiden",  
    "albums" : [  
      {  
        "albumname" : "The Book of Souls",  
        "datereleased" : 2015,  
        "songs" : "11"  
      }, {  
        "albumname" : "Killers",  
        "datereleased" : 1981,  
        "genre" : "Hard Rock"  
      }, {  
        "albumname" : "Powerslave"  
      }, {  
        "albumname" : "Somewhere in Time",  
        "datereleased" : [ 1986, 1995 ],  
        "genre" : "Hard Rock"  
      }  
    ]  
  }, {  
    "_id" : 2,  
    "artistName" : "Taylor Swift",  
    "instruments" : [ "vocals", "guitar", "piano" ]  
  }  
]
```



<https://database.guide/what-is-a-document-store-database/>

<https://docs.mongodb.com/manual/core/data-model-design/>



MONGODB

- v současnosti nejpopulárnější dokumentová databáze (od 2009)
 - humongous data
 - ukládání spousty a spousty dat
 - data ukládána v kolekcích dokumentů (BSON)
 - kolekce
 - seznam dokumentů
 - dokument
 - obsahuje data
 - volné schéma
 - reprezentován pomocí vnořených objektů / map
 - ve formátu BSON
 - pole _id je primárním klíčem



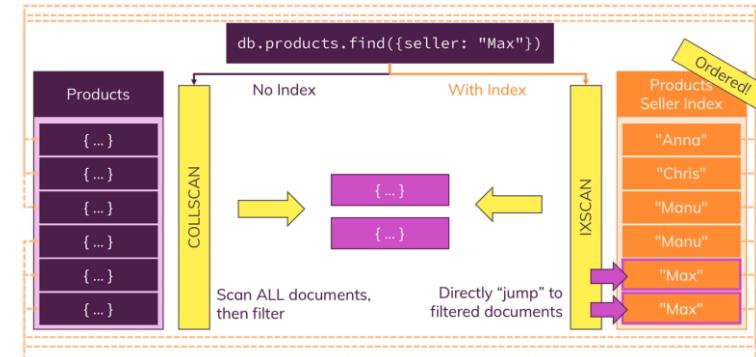
Looks like JSON. Example:

```
{  
    "_id" : ObjectId("7b33e366ae32223aee34fd3"),  
    "title" : "A blog post about MongoDB",  
    "content" : "This is a blog post about MongoDB",  
    "comments": [  
        {  
            "name" : "Frank",  
            "email" : fkane@sundog-soft.com,  
            "content" : "This is the best article ever written!"  
            "rating" : 1  
        }  
    ]  
}
```

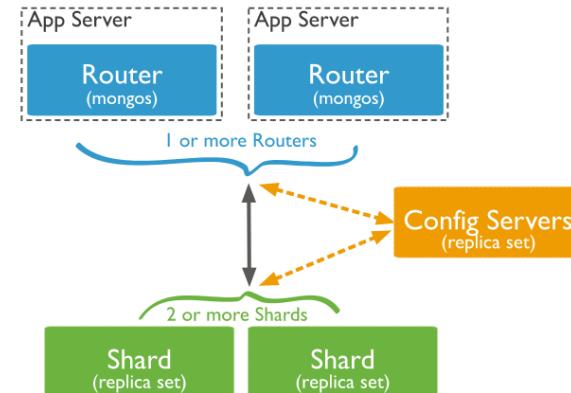
<https://www.udemy.com/course/the-ultimate-hands-on-hadoop-tame-your-big-data>

MONGODB

- hlavní funkce a vlastnosti
 - volné schéma
 - ad-hoc dotazy
 - indexování
 - single field, compound, multikey, ...
 - transakce
 - omezeně od verze 4.0
 - replikace
 - master-slave
 - sharding
 - agregace
 - agregační roura, map-reduce, jednoúčelové
 - souborový systém
 - zabezpečení



<https://www.udemy.com/course/mongodb-the-complete-developers-guide>



<https://docs.mongodb.com/manual/sharding/>



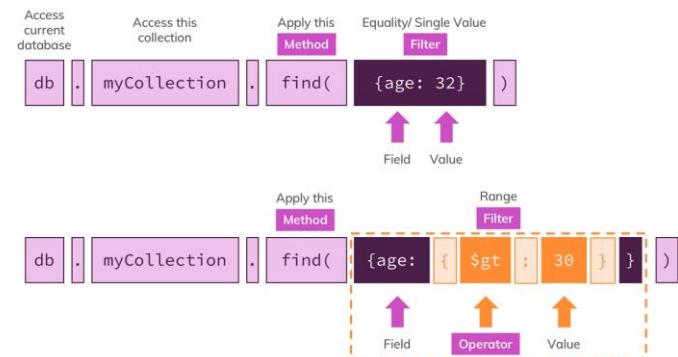
MONGODB

Create
insertOne(data, options)
insertMany(data, options)

Read
find(filter, options)
findOne(filter, options)

Update
updateOne(filter, data, options)
updateMany(filter, data, options)
replaceOne(filter, data, options)

Delete
deleteOne(filter, options)
deleteMany(filter, options)



```
> db.posts.find().pretty()
{
  "_id" : ObjectId("5d2cdcf6f222de7f3b9d281a"),
  "title" : "Post One",
  "body" : "Body of post one",
  "category" : "News",
  "likes" : 4,
  "tags" : [
    "news",
    "events"
  ],
  "user" : {
    "name" : "John Doe",
    "status" : "author"
  },
  "date" : "Mon Jul 15 2019 16:07:18 GMT-0400 (EDT)"
}
{
  "_id" : ObjectId("5d2cded2af222de7f3b9d281b"),
  "title" : "Post Two",
  "body" : "Body of post two",
  "category" : "Technology",
  "date" : "Mon Jul 15 2019 16:08:10 GMT-0400 (EDT)"
}
```

<https://www.udemy.com/course/mongodb-the-complete-developers-guide>

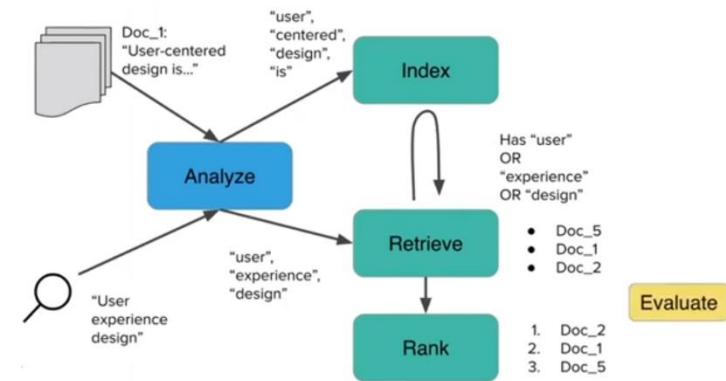
```
Atlas atlas-12nfzk-shard-0 [primary] dpb4> db.prispevky.updateOne({ nazev: 'Zpravy' }, {
$set: { nazev: 'Moje Zpravy' } }, { upsert: true })
{
  acknowledged: true,
  insertedId: null,
  matchedCount: 1,
  modifiedCount: 1,
  upsertedCount: 0
}
```

```
db.users.aggregate(
[
  { $project : { month_joined : { $month : "$joined" } } },
  { $group : { _id : {month_joined:"$month_joined"} , number : { $sum : 1 } } },
  { $sort : { "_id.month_joined" : 1 } }
]
```



SEARCH-ENGINE DATABÁZE

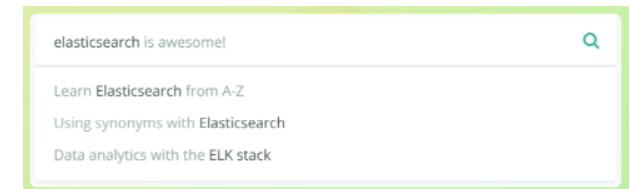
- NoSQL dokumentové databáze zaměřené na vyhledání obsahu
 - Elasticsearch, Solr, ...
- využívají indexování
 - kategorizace podobných vlastností mezi daty
 - urychlení (distribuovaného) vyhledávání
- optimalizované pro práci s daty
 - velké množství dat
 - volné schéma
- poskytují speciální funkce
 - full-textové vyhledávání
 - složité vyhledávací výrazy
 - řazení výsledků



<https://www.youtube.com/watch?v=dqRDyeFJuvk>

ELASTICSEARCH

- engine pro full-textové prohledávání a analýzu (od 2010)
 - open source
 - založený na Apache Lucene
 - dokumentová databáze
 - data ukládána do dokumentů s poli (JSON)
 - dotazování pomocí REST API
 - distribuovaný
 - sharding, routing, replikace
 - součástí Elastic Stack (Kibana, Logstash, X-Pack, Beats)
 - architektura
 - základem je **uzel** (node)
 - každý uzel je součástí **clusteru**
 - základní jednotkou pro ukládání dat je **dokument**
 - data jsou organizována v **indexech**



<https://www.udemy.com/course/elasticsearch-complete-guide/>





ELASTICSEARCH

- (textová) analýza

- vstupní data dokumentu v poli `_source`
 - ta ale nejsou přímo používána pro prohledávání

➤ při indexaci jsou textová pole analyzována

- o analýze se stará analyzátor skládající se ze tří komponent
- znakový filtr (character filter)
- tokenizér (tokenizer)
- token filtr (token filter)

➤ výsledky analýzy uloženy v polích efektivních pro prohledávání (invertovaný index, ...)

➤ reálná aplikace jen na textová pole / hodnoty



<https://www.udemy.com/course/elasticsearch-complete-guide/>



ELASTICSEARCH

- prohledávání
 - dotazy na úrovni termínů (term-level queries)
 - vyhledávají přesnou hodnotu vůči invertovanému indexu
 - vyhledávaná hodnota není analyzována
 - fulltextové dotazy (full-text queries)
 - vyhledávaná hodnota je analyzována stejným analyzárem jako invertovaný index
 - je možné najít jen hodnoty v invertovaném indexu
 - dotazy s Booleovskou logikou
 - leaf a compound queries
 - podpora agregací

```
18 GET /order/_search    ▶ 🔍  
19 {  
20   "size": 0,  
21   "aggs": {  
22     "amount_stats": {  
23       "stats": {  
24         "field": "total_amount"  
25       }  
26     }  
27   }  
28 }
```

```
18 "aggregations" : {  
19   "amount_stats" : {  
20     "count" : 1000,  
21     "min" : 10.27,  
22     "max" : 281.77,  
23     "avg" : 109.20961,  
24     "sum" : 109209.61  
25   }  
26 }  
27 }  
28 }
```

```
1 GET /recipes/_search    ▶ 🔍  
2 {  
3   "query": {  
4     "multi_match": {  
5       "query": "pasta",  
6       "fields": ["title", "description"]  
7     }  
8   }  
9 }
```

```
1 GET /products/_search    ▶ 🔍  
2 {  
3   "query": {  
4     "term": {  
5       "is_active": true  
6     }  
7   }  
8 }  
9  
10 GET /products/_search  
11 {  
12   "query": {  
13     "term": {  
14       "is_active": {  
15         "value": true  
16       }  
17     }  
18 }  
19 }
```

```
1 {  
2   "took" : 29,  
3   "timed_out" : false,  
4   "_shards" : {  
5     "total" : 1,  
6     "successful" : 1,  
7     "skipped" : 0,  
8     "failed" : 0  
9   },  
10   "hits" : {  
11     "total" : {  
12       "value" : 487,  
13       "relation" : "eq"  
14     },  
15     "max_score" : 0.7194644,  
16     "hits" : [  
17       {  
18         "_index" : "products",  
19         "_type" : "_doc",  
20         "_id" : "1",  
21         "_score" : 0.7194644,  
22         "_source" : {  
23           "name": "Pasta alla Norma",  
24           "category": "Italian",  
25           "difficulty": "Medium",  
26           "prep_time": 30,  
27           "cooking_time": 60,  
28           "servings": 4,  
29           "calories": 500,  
30           "protein": 25,  
31           "carbohydrates": 75,  
32           "fat": 15,  
33           "fiber": 5,  
34           "sugar": 5,  
35           "salt": 2,  
36           "spices": "Garlic, basil, tomato sauce",  
37           "instructions": "Boil pasta until al dente. In a separate pan, saut\u00e9 onions and garlic. Add tomato sauce and spices. Toss pasta with the sauce.",  
38           "image": "https://example.com/images/pasta-norma.jpg"  
39         }  
40       }  
41     ]  
42   }  
43 }
```

```
1 GET /recipes/_search    ▶ 🔍  
2 {  
3   "_source": "created",  
4   "query": {  
5     "match_all": {}  
6   },  
7   "sort": [  
8     { "created": "desc" }  
9   ]  
10 }
```



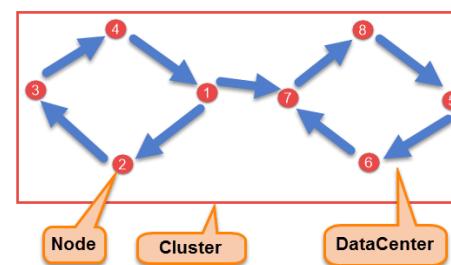
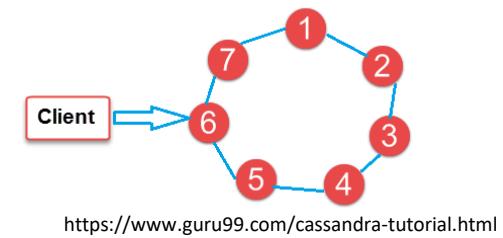
SLOUPCOVÁ ÚLOŽIŠTĚ

- NoSQL databáze optimalizované pro čtení a zápis ve sloupcích
 - Apache Cassandra, Druid, HBase, ...
- sloupcově orientovaný model pro ukládání dat
 - data ukládána do záznamů schopných pojmut velké množství dynamických sloupců
 - klíče záznamů ani jména sloupců nejsou fixní
- optimalizované pro rychlý přístup k velkým objemům dat
- výhody sloupcových databází
 - horizontální škálování a replikace
 - volné schéma
 - rychlé načítání dat, dotazování a agregace

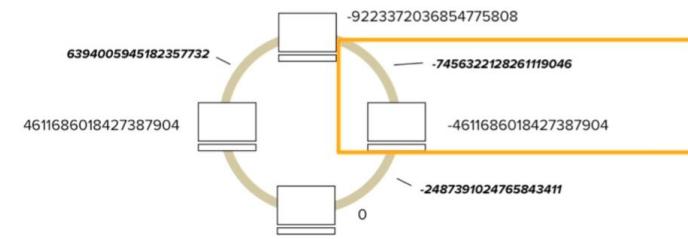


CASSANDRA

- v současnosti nejpopulárnější sloupcová databáze (od 2008)
 - open source
 - hlavní vlastnosti
 - distribuovaná databáze - horizontální škálování, replikace
 - odolnost vůči chybám
 - vysoká dostupnost
 - vysoký výkon
 - dotazovací jazyk CQL
 - sekundární indexy
 - technologie pro big data
 - architektura
 - uzly, datová centra, cluster
 - každý uzel plní stejnou roli
 - možnost virtuálních uzlů
 - snitch, gossip, partitioner, replikační faktor a strategie



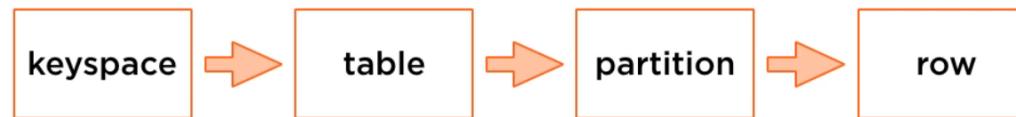
<https://www.guru99.com/cassandra-architecture.html>



<https://www.udemy.com/course/apache-cassandra>

CASSANDRA

- sloupcově orientovaný model
 - v Cassandře je databáze definována jako **keyspace**
 - keyspace obsahuje **tabulky** (tables)
 - všechna data mají přiřazený **partition** klíč
 - určuje jejich umístění v clusteru
 - může být i složený z více sloupců (composite)
 - všechna data v partition uložena spolu
 - data v partition reprezentována jako jeden nebo více **řádků**
 - každý řádek jednoznačně identifikovaný primárním klíčem



<https://app.pluralsight.com/library/courses/cassandra-developers>

CASSANDRA

```
D:\Prezentace>docker exec -ti cas1 nodetool status mykeyspace
Datacenter: datacenter1
=====
Status=Up/Down
|/ State=Normal/Leaving/Joining/Moving
--  Address      Load    Tokens  Owns (effective)  Host ID            Rack
UN  172.17.0.3   253.25 KiB  256        46.1%           6a05ed73-667b-4842-ae9d-e446cbbabba  rack1
UN  172.17.0.2   297.64 KiB  256        53.9%           82811efa-4821-49c5-ad66-3c7fe2154df9  rack1
Datacenter: datacenter2
=====
Status=Up/Down
|/ State=Normal/Leaving/Joining/Moving
--  Address      Load    Tokens  Owns (effective)  Host ID            Rack
UN  172.17.0.4   247.19 KiB  256        100.0%          c784b328-6fff-42fc-a56a-001baec8e420  rack1
```

```
cqlsh> DESCRIBE KEYSPACES;
system_traces  system_schema  system_auth  system  system_distributed
```

```
CREATE TABLE activity(
  home_id text,
  datetime timestamp,
  event text,
  code_used text,
  PRIMARY KEY (home_id, datetime)
) WITH CLUSTERING ORDER BY (datetime DESC);
```

```
INSERT INTO activity (home_id, datetime, event,
code_used) VALUES ('H01474777', '2014-05-21 07:32:16',
'alarm set', '5599');
```

```
cqlsh> SELECT * FROM cycling.cyclist_category;
```

```
cqlsh> SELECT category, points, lastname FROM cycling.cyclist_category;
```

```
UPDATE home_security.home SET phone = '310-883-7197'
WHERE home_id = 'H01474777';
```

```
UPDATE home_security.home SET phone = '310-883-7197',
contact_name = 'Mr. Drysdale' WHERE home_id = 'H01474777';
```

```
cqlsh> DELETE FROM cycling.calendar WHERE race_id = 200;
```

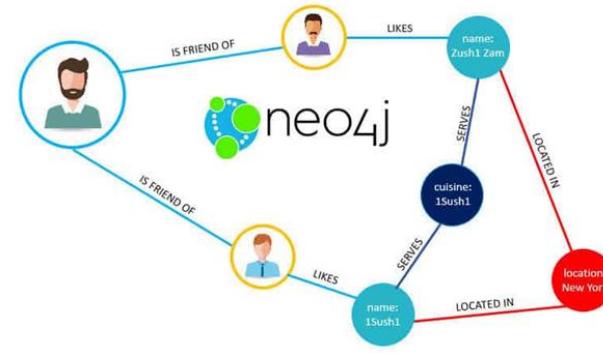
```
TRUNCATE cycling.user_activity;
```

```
DROP TABLE cycling.cyclist_name;
```



GRAFOVÁ ÚLOŽIŠTĚ

- NoSQL databáze považující vztahy mezi daty za minimálně stejně důležité jako samotná data
 - Neo4j, ...
- grafový datový model
 - základem jsou vrcholy a hrany doplněné o vlastnosti
 - hrany mohou být orientované
 - modeluje snadno vazby
 - a umožňuje jejich snadné procházení
- použití pro velké objemy dat
 - s velkým množstvím propojení (vazeb)
 - sociální sítě, dopravní sítě, ...

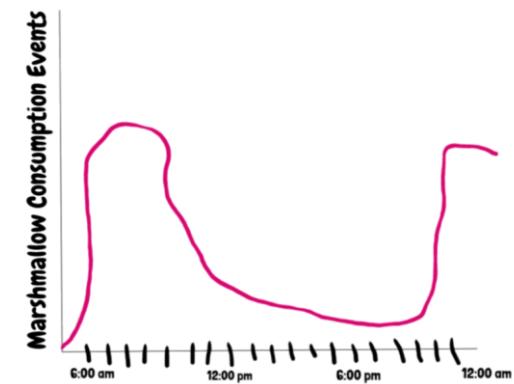


<https://www.bmc.com/blogs/neo4j-graph-database/>

neo4j

DATABÁZE ČASOVÝCH ŘAD

- NoSQL databáze optimalizované pro časové řady
 - InfluxDB, Prometheus, ...
- časové řady
 - soubor pozorování získaný opakovaným měřením v čase
 - při nanesení na graf je jednou z os vždy čas
 - samotné body nenesou důležité informace, ty pochází z celé časové řady
- přináší speciální funkcionality
 - ukládání a komprese dat s časovou značkou
 - správa životního cyklu dat
 - summarizace
 - skenování velkého množství záznamů
 - dotazování na časové řady



<https://www.influxdata.com/blog/what-is-time-series-data-and-why-should-you-care/>

A PŘÍŠTĚ?

- analýza velkých dat
- technologie pro big data





Děkuji za pozornost.
Otázky?

