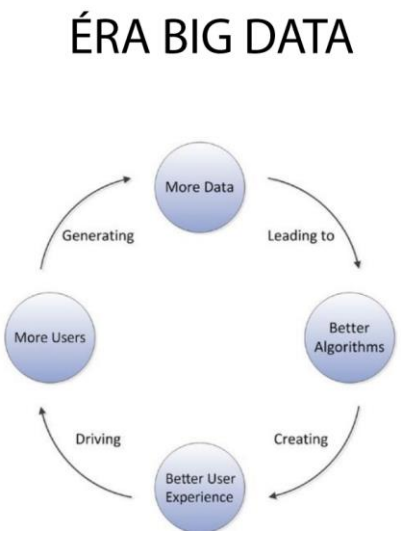
31. BigData – definice, dělení podle struktury a původu, charakteristika, aplikace. Analýza velkých dat – jednotlivé kroky, typy, výhody a výzvy

. BigData – definice.

- Definice:

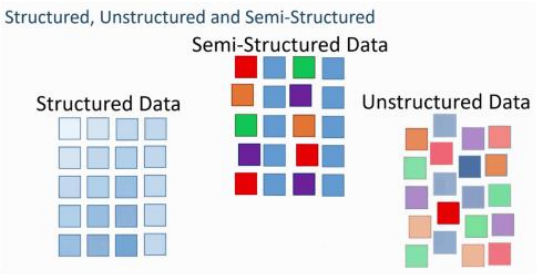
Soubory dat, jejichž velikost je mimo schopnosti zachycovat, spravovat a zpracovávat data běžně používanými softwarovými prostředky v rozumném čase.

- vychází ze dvou základních předpokladů
  - existuje stálý (rostoucí) a rychlý přísun 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



dělení podle struktury

- 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
- 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, …



dělení podle původu

- tři hlavní zdroje dat
  - stroje
    - senzory sbírající data v reálném čase v průmyslu, automobilech, …
    - environmentální senzory
    - zařízení sledující zdravotní stav
  - lidi
    - social media
    - články, blogy, …
  - organizace
    - transakce v databázích

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
- chytrá zařízení
  - zařízení schopná měřit a produkovat big data (pomocí senzorů)
  - proč chytrá?
    - schopnost připojení k dalším zařízením / sítím
    - autonomní sběr a analýza dat
    - mají povědomí o prostředí
    - ➢ internet věcí (internet of things)
    - např. chytré hodinky
      - měření teploty, tepu, kvality spánku, nachozené vzdálenosti, počtu schodů, atd.
      - co když je bude mít každý?
      - co když jich každý bude mít víc?

STROJOVÁ DATA

- zpracování v reálném čase
  - vztahy se zákazníky
  - detekce podvodů
  - monitorování systému
  - ➢ okamžitá analýza a reakce
- zpracování se přesouvá za daty
  - In-Situ
- potřeba škálovatelných výpočetních systémů
- SCADA
  - vzdálený monitoring a ovládání průmyslových procesů

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 jejích rychlý růst
      - potvrzení je časově náročné (sběr, uložení, těžba, čištění a zpracování)

- zpracování – několik základních open source frameworků
- nástroje pro zpracování a analýzu jsou vyvíjeny od nuly
  - 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 organizace od organizace
  - transakce, kreditní karty, bankovnictví, akcie, zdravotní záznamy, senzory, atd.
- 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

ORGANIZAČNÍ DATA

- hlavní užitek spočívá v kombinaci s ostatními big daty
  - efektivita provozu
  - zvýšené prodeje
  - vylepšený marketing
  - zlepšená bezpečnost
  - vyšší zákaznická spokojenost
  - příklad: UPS, Walmart
- firmy dnes mohutně investují do big data
  - výsledky ve všech sektorech
  - prvotní firmy získaly náskok nad svými konkurenty

charakteristika

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

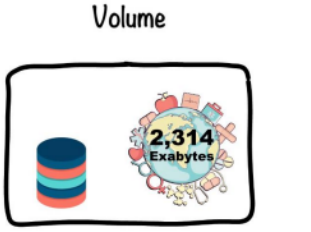
CHARAKTERISTIKA BIG DATA

- 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

Volume (objem)

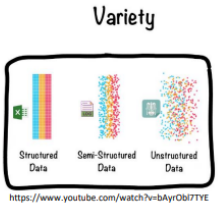
- zpracování a ukládání velkého objemu dat přináší výzvy
  - škálování
    - horizontální / vertikální / kombinace
    - zajištění kapacity úložiště a výkonu na zpracování dat
  - dostupnost
    - přístup k datům a možnost jejich zpracování
  - bandwidth a výkon
    - přístup k datům v potřebný okamžik
- cílem firem je analýza dat
  - ➢ zlepšení poskytovaného produktu / služby
  - ➢ náskok před konkurencí



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

Variety (různorodost)

- 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)



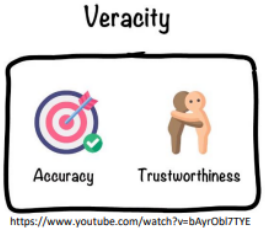
- hybridní data – např. emaily

Velocity (rychlost)

- rostoucí rychlost vytváření big data
  - z různých zdrojů se liší (aktualizace jednou za den vs. každou sekundu)
- rostoucí rychlost ukládání a analýzy big data (zpracování)
- cílem je zpracování v reálném čase
  - vytvoření reklamy na základě uživatelské historie a zobrazení při hledání
  - ➢ pomalá reakce vede ke ztrátě příležitosti
  - např. kempování – zajímá mě dnešní počasí, ne počasí, které bylo před rokem
  - např. neštěstí – okamžitý zásah záchranných složek
- potřeba zvážit rychlost vytváření i rychlost zpracování dat
  - zpracování může čekat na data
  - data mohou čekat na zpracování
  - ➢ rovnováha

Veracity (věrohodnost)

- 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 (× Google Flu Trends)

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

Value (hodnota)

- 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

aplikace

- Cílený marketing
- Doporučovací systémy
- 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 (primárně z GPS)
- Chování skupiny
- Biomedicína
- Big city města

Analýza velkých dat – jednotlivé kroky, typy, výhody a výzvy

- analýza velkých dat je prováděna ve 4 hlavních krocích
  - sběr dat
  - zpracování dat
  - čištění dat
  - analýza dat

Sběr dat

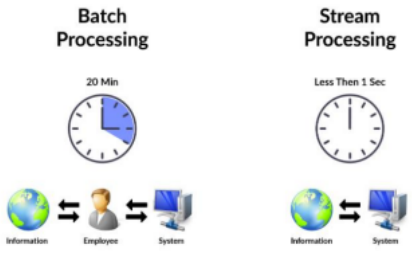
- data collection
- data jsou sbírána z různých zdrojů
  - social media, internetová data (kliky), logy ze serverů, aplikace v cloudu, mobilní aplikace, zákaznické emaily, průzkumy, telefonní hovory, strojová data ze senzorů, …
  - záleží na organizaci

https://www.jotform.com/data-collection-methods/

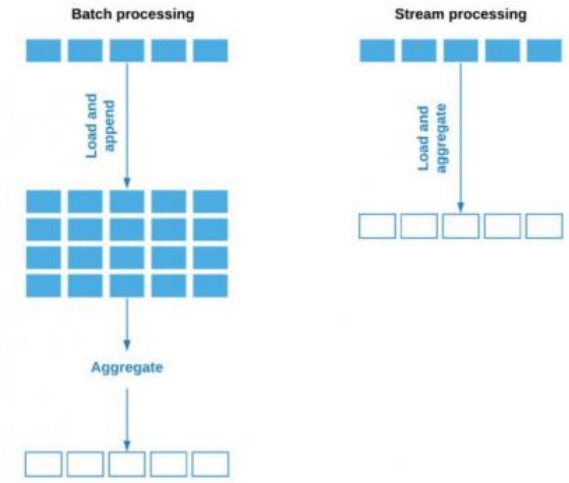
- data jsou v různých formách
  - strukturovaná, nestrukturovaná, částečně strukturovaná
- data mohou být uložena do datových skladů pro snadný přístup
- nezpracovaná data v datových jezerech s přiřazenými metadaty
  - případně i složitá nestrukturovaná data

Zpracování dat

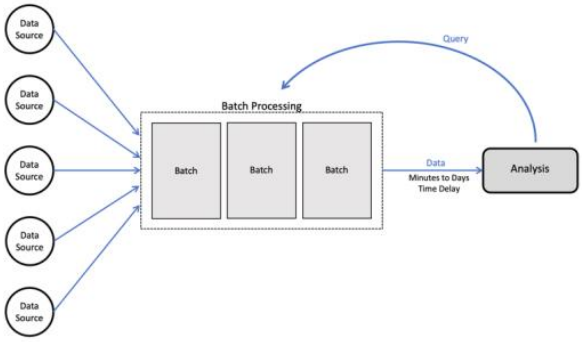
- data processing
- data jsou organizována, tříděna a rozdělena pro analytické dotazy
- vhodné zpracování výrazně zlepšuje výkon dotazů
- obtížné se stále rostoucím množstvím dat
- různé způsoby zpracování dat
  - batch processing
  - stream processing
  - distributed processing
  - real-time processing
  - a další
    - online processing
    - commercial data processing
    - multiprocessing



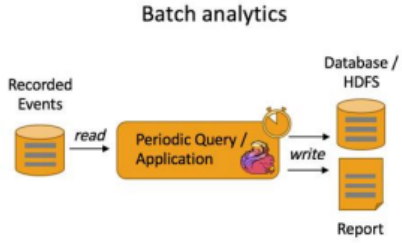
- batch processing vs. stream processing



  - batch processing

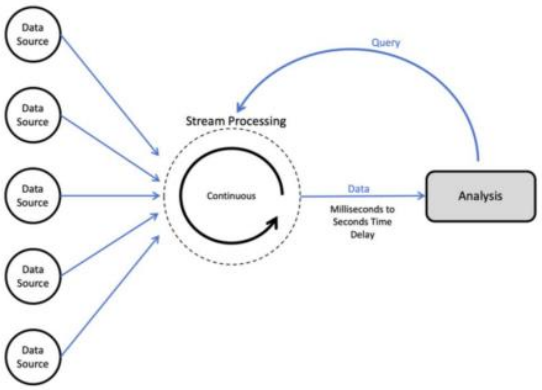


- batch processing
- zpracování velkého objemu dat najednou nebo v dávkách
  - data sbírána určitý časový interval
- často plánované
  - časový interval nebo určité množství dat
  - není potřeba manuální zásah
- často off-line zpracování
  - zpracování není okamžité
- často velmi časově náročné
  - obrovské množství dat
  - ale i tak zpracování v dávkách šetří čas
- např. výplatní systémy, faktury



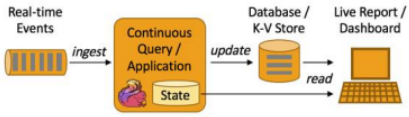
https://dzone.com/articles/towards-a-unified-data-processing-framework-batch

- stream processing



- stream processing
- okamžité zpracování dat v momentě jejich vytvoření / sběru
  - téměř žádné časové zpoždění mezi přijetím a zpracováním dat
  - zpracování menších množství dat, ale okamžité
- důležité pro real-time aplikace
  - data generovaná v reálném čase
- umožňuje vylepšení v reálném čase
- komplexnější a dražší
- např. detekce podvodů
- např. Uber – nejbližší auta

Streaming analytics



https://dzone.com/articles/towards-a-unified-data-processing-framework-batch

Čištění dat

- data cleansing
  - data scrubbing
- kvalita vstupních dat se může značně lišit
  - značně ovlivňuje výsledky analýzy
- kvalita dat je zaručena čištěním
  - zajišťuje i relevanci dat
- odstranění chybných, narušených, duplicitních i chybně formátovaných dat
- obzvláště důležité u dat z více zdrojů
  - více duplikátů, různých formátů značek, …

https://www.iteratorshq.com/blog/data-cleaning-in-5-easy-steps/

- čištění dat
  - data cleansing
  - 5 základních kroků
    - odstranění duplikátů a irelevantních pozorování
    - oprava strukturních chyb
    - odstranění outlierů
    - vložení, označení nebo úplné odstranění chybějících hodnot
    - analýza kvality dat

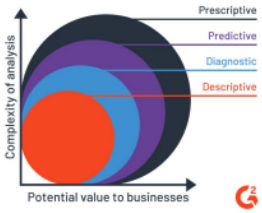
https://www.iteratorshq.com/blog/data-cleaning-in-5-easy-steps/

  - často využívána umělá inteligence a strojové učení
    - čištění nestrukturovaných dat (obrázky, nahrávky, videa)
  - zpracování přirozeného jazyka (NLP) na očištění lidmi generovaných dat

Analýza dat

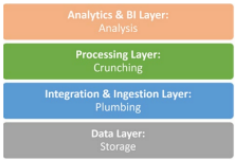
- data analysis
- posbíraná, zpracovaná a očištěná data jsou konečně analyzována
- využívá se různých technik
  - data mining
  - umělá inteligence
  - strojové učení a hluboké učení
  - text mining a statistické analýzy
  - vizualizace
- ➢ cílem je získat důležité informace

- big data analytics
  - 4 základní typy analýzy velkých dat
  - deskriptivní analýza
    - odpovídá na otázku: „Co se stalo?“
    - umožňuje pochopení toho, co se v minulosti událo
  - diagnostická analýza
    - odpovídá na otázku: „Proč se to stalo?“
    - snaží se retrospektivně najít příčiny problému
  - prediktivní analýza
    - odpovídá na otázku: „Co se pravděpodobně stane?“
    - na základě dat minulých a současných předpovídá budoucnost
  - preskriptivní analýza
    - odpovídá na otázku: „Jak to uskutečnit?“
    - na základě predikovaných informací navrhuje další postupy

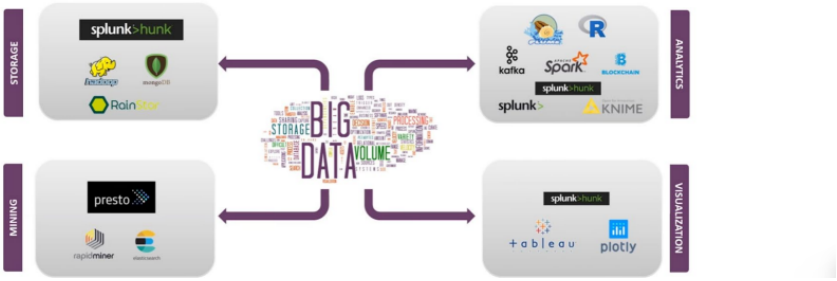


https://www.g2.com/articles/types-of-data-analyti

- technologie pro big data
  - 4 základní směry použití
    - ukládání a správa dat, těžba dat, analýza dat, vizualizace dat
  - big data technology stack
    - sada technologií na vývoj robustního analytického engine
    - dříve jen Hadoop stack



https://blog.panoply.io/



- výhody technologií pro velká data
  - analýza dat pokročilými algoritmy a modely
  - použití na platformách pro velká data nebo v analytických systémech
  - umožňují práci se strukturovanými i nestrukturovanými daty
    - z různých zdrojů
  - vizualizace analyzovaných dat
  - snadná integrace s dalšími technologiemi

➢ umožňují dělat důležitá rozhodnutí na základě dat a predikcí