

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

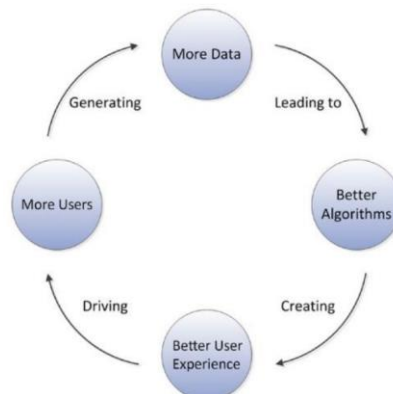
. BigData – definice.

- Definice:

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

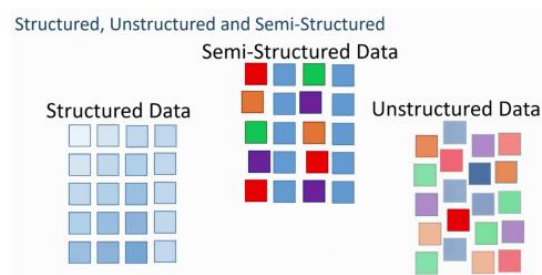
- vychází ze dvou základních předpokladů
 - existuje stálý (rostoucí) a rychlý přísun nových dat
 - v roce 2017 proteklo internetem 1.5 ZB (10^{21}) dat
 - pro rok 2022 je odhad 4.8 ZB
 - inovace ve způsobu zpracování
 - cloud computing
 - propůjčení výpočetních serverů
 - umožňuje provádět výpočty kdekoli a odkudkoli
- Umožňuje **dynamickou a škálovatelnou** analýzu dat

ÉRA BIG DATA



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 nestructurovaných
 - 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)



- 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, bankovníctví, 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 síla

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

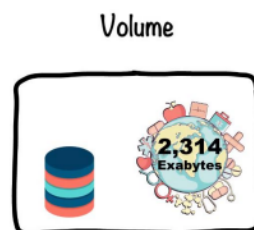
- **další často uváděná V's**
 - **věrohodnost (veracity)**
 - zaujatost, šum, abnormality v datech
 - nejistota v pravdivost a věrohodnost dat
 - **valence (valence)**
 - propojovatelnost dat formou grafů
- **a nezapomenout na**
 - **hodnotu (value)**
 - srdce a lepidlo všech ostatních V's
 - jak z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



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

Volume (objem)

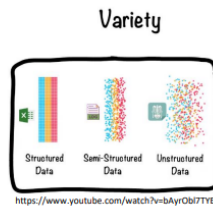
- **zpracování a ukládání velkého objemu dat přináší výzvy**
 - **škálování**
 - horizontální / vertikální / kombinace
 - zajištění kapacity úložiště a výkonu na zpracování dat
 - **dostupnost**
 - přístup k datům a možnost jejich zpracování
 - **bandwidth a výkon**
 - přístup k datům v potřebný okamžik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



<https://www.youtube.com/watch?v=bAyrObi7TYE>

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 uložena (záznamy)
 - dostupná neustále (kamery) nebo jen příležitostně (sondy, satelity)
- hybridní data – např. emaily



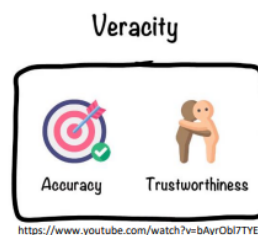
Velocity (rychlost)

- 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

<https://www.youtube.com/watch?v=bAyrObi7TYE>

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)

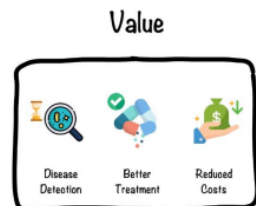


Valence

- **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

Analytická řešení získávají přehledy a predikují výsledky analýzou datových sad. Aby však bylo možné data úspěšně analyzovat, musí být nejprve uložena, uspořádána a vyčištěna řadou aplikací v integrovaném procesu postupné přípravy:

- **Shromažďování.** Data, která mají strukturovanou, částečně strukturovanou a nestrukturovanou podobu, jsou shromažďována z různých zdrojů z webu, mobilních zařízení a cloudu. Pak jsou uložena do úložiště – datového jezera nebo datového skladu – v rámci přípravy na zpracování.
- **Zpracování.** Ve fázi zpracování se uložená data ověřují, řadí a filtrují, což je připravuje pro další použití a zlepšuje výkon dotazů.
- **Čištění.** Po zpracování se data vyčistí. Konflikty, redundance, neplatná nebo neúplná pole a chyby formátování v datové sadě se opraví a vyčistí.
- **Analýza.** Data jsou teď připravená k analýze. Analýza velkých objemů dat se provádí pomocí nástrojů a technologií, jako je dolování dat, umělá inteligence, prediktivní analýza, strojové učení a statistická analýza, které pomáhají definovat a predikovat vzorce a chování v datech.