B3. Big Data

Termín „data“ je dnes frekventovaným tématem, setkáváme se s ním každý den. A často také s pojmem „big data“. Oba pojmy najdeme i v tomto studijním textu, takže se s nimi seznámíme blíže.

Co jsou data: *Veličiny, znaky* nebo *symboly,* na kterých jsou prováděny operace člověkem nebo počítačem a které jsou uloženy a/nebo přenášeny v písemné formě nebo ve formě elektrických signálů a zaznamenávány na magnetické, optická nebo mechanická záznamová média.

Příklady: čísla 2, -4, 0,257, 3,14, √274 , … , písmena b, xyz, a^2, A+B, …, symboly ♈, ✈, slova „Být či nebýt?“, „Mně už se to nelíbí“, … , .

Co jsou Big Data *(velká data):* Sobory dat, které mají obrovské objemy, jejichž velikost roste exponenciálně s časem. Data o takové velikosti a složitosti nedokáže uložit ani efektivně zpracovat žádný z tradičních nástrojů pro správu dat.

„Freepig.com“

Příklady:

- „Burzy“ jsou příkladem velkých dat, která generují přibližně jeden terabajt nových obchodních dat denně.
- Více než 500+ terabajtů nových dat se každý den dostane do databází stránek sociálních médií, generovaných z hlediska nahrávání fotografií a videí, zpráv atd.

Big Data (velká data) jsou soubory dat, jejichž velikost nebo typ přesahuje schopnost tradičních relačních databází zachytit, spravovat a zpracovávat data s nízkou latencí. Mezi vlastnosti velkých dat patří velký objem, vysoká rychlost a velká rozmanitost. Zdroje Big Dat jsou složitější než zdroje tradičních dat, protože jsou uváděny do činnosti *umělou inteligencí* (AI), *mobilními zařízeními, sociálními médii* a *Internetem věcí* (IoT). Různé typy dat například pocházejí ze senzorů, zařízení, videa/audia provozovatelů, sítí, protokolových souborů, transakčních aplikací, webu a sociálních médií – velká část z nich je generována v reálném čase a ve velmi velkém měřítku.

Velká data jsou soubory dat, jejichž velikost nebo typ přesahuje schopnost tradičních relačních databází zachytit, spravovat a zpracovávat data s nízkou latencí. Mezi vlastnosti velkých dat patří velký objem, vysoká rychlost a velká rozmanitost.

Zdroje dat jsou stále složitější než zdroje tradičních dat, protože jsou poháněny umělou inteligencí (AI), mobilními zařízeními, sociálními médii a internetem věcí (IoT).

Různé typy dat například pocházejí ze senzorů, zařízení, videa/audia, sítí, protokolových souborů, transakčních aplikací, webu a sociálních médií – velká část z nich je generována v reálném čase a ve velmi velkém měřítku.

Díky systematické a podrobné analýze velkých dat můžeme v konečném důsledku podporovat lepší a rychlejší rozhodování, lze modelovat a předpovídat budoucí výsledky a zlepšit marketingové strategie rozhodování. Podniky mohou využívat velké objemy dat a analyzovat širokou škálu zdrojů dat, aby získaly nové poznatky a upravovaly své rozvojové strategie.

Komplexní analýza dat pořízených ze senzorů, videí, testů, protokolů, transakčních aplikací, webu a sociálních médií umožňuje organizacím plně *využívat data pro řízení*.

Kategorie v systému **Big Data:** Existují tři typy Big Data: data strukturovaná, data nestrukturovaná a data semi-strukturovaná.

Strukturovaná data: Jsou to libovolná data, která mohou být uložena, zpřístupněna a zpracována ve formě pevného formátu. V dnešní době velikost takových dat roste do obrovské míry, typické objemy se pohybují v rozmezí několika zettabajtů (jedna miliarda terabajtů tvoří zettabajt).

Příklady: Data uložená v systému správy relačních databází jsou jedním z příkladů „strukturovaných" dat. Údaje v tabulkách, např. jízdní řád, mzdové tabulky, seznam klíčových slov na konci publikace atd.

Tabulka „Zaměstnanci" v databázi je příkladem strukturovaných dat.

Nestrukturovaná data: Jakákoli data s neznámou formou nebo strukturou jsou klasifikována jako nestrukturovaná data. Mají obrovské objemy a představují řadu problémů, pokud jde o jejich zpracování a odvození z nich vyplývajících informací a hodnotových ukazatelů.

Příklad: Heterogenní zdroj dat obsahující kombinaci textových souborů, obrázků, videí atd.

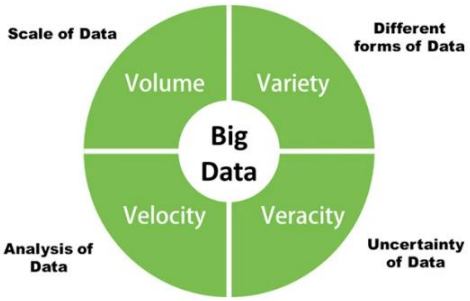
V dnešní době mají organizace k dispozici velké množství nestrukturovaných dat se zajímavými informacemi, ale nevědí, jak z nich odvodit skryté hodnoty, protože tato data jsou v nezpracované formě nebo v nestrukturovaném formátu.

Semi-strukturovaná data: mohou obsahovat obě předchozí formy dat.

Příklady: Osobní webové stránky: obsahují texty, obrázky, videa, odkazy na sociální sítě, odkazy pro komunikaci s přáteli atd. Dalším příkladem jsou webové stránky internetových obchodů s reklamou, platebními terminály.

Rozlišujeme čtyři charakteristiky Big Data.

Objem dat (volume) se vztahuje k velikosti datových souborů, které je třeba analyzovat a zpracovat a které jsou nyní často větší než terabajty[1] a petabajty[2]. Obrovský objem dat vyžaduje odlišné a netradiční technologie zpracování v porovnání s tradičním ukládáním a zpracováním dat v laptopech a standardních databázích.



Rychlost (velocity) se týká rychlosti, s jakou jsou data generována. Vysokorychlostní data jsou generována takovým rychlým tempem, že jejich zpracování vyžaduje odlišné (distribuované) techniky zpracování. Příkladem dat, která jsou generována vysokou rychlostí, jsou zprávy na Twitteru nebo komunikace na Facebooku.

Rozmanitost (variety) dělá Big Data opravdu velká. Velká data pocházejí z mnoha různých zdrojů a obecně jsou jedním ze tří typů: strukturovaná, semi-strukturovaná a nestrukturovaná data. Rozmanitost typů dat často vyžaduje odlišné přístupy k jejich zpracování a tvorbu speciálních algoritmů. Příkladem velmi rozmanitých datových souborů mohou být audio a video soubory, které jsou generovány na různých místech ve městě.

[1] Terabyte: Jeden bit (binary digit) má jednoduchou binární hodnotu buď 0 nebo 1; jedná se o nejmenší jednotku popisující data v počítači. Jeden terabyte je nejmenší objem paměti, který poskytují dnešní media na trhu. Existují jednotky větší než terabyte: petabyte, exabyte, zettabyte, yottabyte and brontobyte. A geopbyte je 10^{30} bytes. Prakticky jeden terabyte dat je roven: 728 177 CD, nebo 40 single-layer Blu-ray discs, 85 899 345 stran textu ve Wordu, 500 hodin videa, 310 000 fotografií, nebo 17 000 hodin hudby.

[2] Petabyte: 1 PB = 1.000 TB, tj. 1 PB je o něco větší než 1 kvadrilion bytů.

Pravdivost (veracity) se týká kvality dat, která jsou analyzována. Vysoce pravdivá data obsahují mnoho záznamů, které jsou cenné pro analýzu a které významným způsobem přispívají k celkovým výsledkům. Data s nízkou pravdivostí naopak obsahují vysoké procento nesmyslných dat. Nehodnotná data v těchto souborech se označují jako šum. Příkladem vysoce pravdivého souboru dat mohou být data z lékařského experimentu nebo studie.

Data, která jsou charakteristická velkým objemem, vysokou rychlostí a velkou rozmanitostí, musí být zpracována pomocí pokročilých nástrojů (analytik a algoritmů), aby z nich bylo možné odhalit smysluplné informace. Kvůli těmto vlastnostem dat byla znalostní doména, která se zabývá ukládáním, zpracováním a analýzou těchto datových sad, označena jako velká data (Big Data)**.**

Výhody z analýzy a zpracování Big Data:

- Podniky mohou při rozhodování využívat vnější inteligentní informace.
- Přístup k sociálním datům z vyhledávačů a webů jako Facebook, Twitter umožňuje organizacím vyladit jejich obchodní strategie.
- Lepší zákaznický servis.
- Včasná identifikace rizika pro produkty a služby.
- Lepší efektivita provozů.
- Big Data technologie mohou být použity pro vytvoření datového skladu.

Souhrn:

Big Data je termín používaný k popisu souboru dat, který je obrovský, a ještě s časem exponenciálně roste. Existují tři typy Big Data: data strukturovaná, data nestrukturovaná a data semi-strukturovaná. Strukturovaná data jsou libovolná data, která mohou být uložena, zpřístupněna a zpracována ve formě pevného formátu. V dnešní době se dat pohybují v rozmezí několika zettabajtů. Nestrukturovaná data mají neznámou formu nebo strukturu. Mají obrovské objemy a představují řadu problémů, pokud jde o jejich zpracování a odvození z nich vyplývajících informací a hodnotových ukazatelů. Semi-strukturovaná data mohou obsahovat obě předchozí formy dat. Rozlišujeme čtyři charakteristiky Big Data. Objem dat (volume) se vztahuje k velikosti datových souborů, které je třeba analyzovat a zpracovat a které jsou nyní často větší než terabajty a petabajty. Rychlost (velocity) se týká rychlosti, s jakou jsou data generována. Vysokorychlostní data jsou generována takovým rychlým tempem, že jejich zpracování vyžaduje odlišné (distribuované) techniky zpracování. Příkladem dat, která jsou generována vysokou rychlostí, jsou zprávy na Twitteru nebo komunikace na Facebooku. Rozmanitost (variety) dělá Big Data opravdu velká. Pravdivost (veracity) se týká kvality dat, která jsou analyzována. Vysoce pravdivá data obsahují mnoho záznamů, které jsou cenné pro analýzu a které významným způsobem přispívají k celkovým výsledkům.

***

Odkazy na relevantní témata:

https://cs.wikipedia.org/wiki/Big_data

https://www.guru99.com/what-is-big-data.html

https://www.gartner.com/en/information-technology/glossary/big-data

***

Klíčová slova:
množství
znaky
symboly
Velká data – Big Data
umělá inteligence
mobilní zařízení
sociální média
Internet věcí
strukturovaná data
nestrukturovaná data
semi–strukturovaná data
čtyři typy Big Data