Big Data

Big Data je termín popisující masivní objemy dat (strukturovaných i nestrukturovaných), která se každodenně valí firmami a internetem. V kontextu Big Data není nejdůležitější množství dat samo o sobě, ale to, co s nimi organizace dělají – jak je analyzují pro lepší rozhodování, predikci trendů a automatizaci.

Charakteristika: Model 5V

Pro definici Big Data se používá model „V“, který se postupně rozšiřoval z původních tří na pět (i více) základních charakteristik:

1. Volume (Objem)

Množství generovaných dat. Mluvíme o terabytech (TB), petabytech (PB) a exabytech (EB). Data pocházejí z logů, sociálních sítí, senzorů a transakcí.

2. Velocity (Rychlost)

Rychlost, jakou jsou data generována a jak rychle musí být zpracována. Příkladem jsou data z burzy nebo senzory v autonomních vozidlech, kde i milisekunda hraje roli.

3. Variety (Rozmanitost)

Data přicházejí v mnoha formátech:

- **Strukturovaná:** Tabulky, databáze.
- **Polostrukturovaná:** XML, JSON soubory.
- **Nestrukturovaná:** Textové dokumenty, e-maily, video, audio, obrázky.

4. Veracity (Věrohodnost)

Kvalita a přesnost dat. U obrovských souborů je problémem „šum“ – neúplná nebo chybná data, která mohou zkreslit výsledky analýzy.

5. Value (Hodnota)

Nejdůležitější bod. Data jsou k ničemu, pokud z nich nedokážeme získat užitečnou informaci, která pomůže byznysu nebo vědě.

Architektura a technologie

Tradiční SQL databáze na Big Data nestačí. Proto vznikly nové přístupy:

Hadoop a MapReduce

Apache Hadoop je open-source framework, který umožňuje distribuované ukládání a zpracování obrovských souborů na klastrech běžných počítačů.

- **HDFS (Hadoop Distributed File System):** Rozdělí data na malé bloky a uloží je na různé uzly v klastru.
- **MapReduce:** Algoritmus, který rozdělí úlohu na menší části, ty zpracuje paralelně a výsledek složí dohromady.

NoSQL Databáze

Databáze, které nevyžadují pevné schéma (tabulky) a skvěle škálují horizontálně.

- **MongoDB:** Dokumentová databáze.
- **Cassandra:** Širokosloupcová databáze (původně z Facebooku).

Real-time Processing (Proudové zpracování)

Nástroje pro analýzu dat v okamžiku, kdy vznikají.

- **Apache Spark:** Mnohem rychlejší než MapReduce, protože zpracovává data v operační paměti (In-Memory).
- **Apache Kafka:** Systém pro distribuované zasílání zpráv a sběr dat z tisíců zdrojů.

Analytické metody

Zpracování Big Data se dělí do čtyř úrovní podle toho, co nám říkají:

```
1. **Deskriptivní (Co se stalo?):** Reporty o prodejích za minulý měsíc.
2. **Diagnostická (Proč se to stalo?):** Hledání příčin poklesu výkonu.
3. **Prediktivní (Co se stane?):** Předpověď odchodu zákazníků ke konkurenci pomocí [[machine_learning|strojového učení]].
4. **Preskriptivní (Co máme dělat?):** Algoritmus sám navrhne nejlepší trasu pro kamion nebo cenu letenky.
```

Praktické využití

- **Personalizovaný marketing:** Netflix nebo YouTube vám doporučují obsah na základě analýzy chování milionů jiných uživatelů.
- **Zdravotnictví:** Analýza genomu, predikce epidemií nebo vývoj léků díky simulacím na obrovských vzorcích.
- **Smart Cities:** Řízení dopravy v reálném čase podle dat ze semaforů, GPS a kamer.
- **Bankovnictví:** Okamžitá detekce podvodných transakcí (fraud detection).

Výzvy a rizika

- **Soukromí a etika:** Sběr tak velkého množství dat o lidech vede k obavám ze sledování (GDPR).
- **Bezpečnost:** Big Data soustřeďují obrovské množství informací na jednom místě, což je lákavý cíl pro útočníky.
- **Nedostatek expertů:** Práce s těmito technologiemi vyžaduje „Data Scientists“ – odborníky na statistiku i programování.

Související pojmy: Hadoop, Spark, NoSQL, Machine Learning, Data Warehouse, Data Lake, GDPR.