

Big Data a jejich zpracování

Big Data nejsou definována jen svou velikostí, ale především výzvami, které přinášejí při jejich ukládání, vyhledávání a analýze. Pro popis Big Data se používá model **5V**, který definuje jejich základní charakteristiky.

1. Charakteristika 5V

- **Volume (Objem):** Množství dat generovaných každou sekundu (terabyty až zettabyty).
 - **Velocity (Rychlost):** Rychlost, jakou data vznikají a jak rychle musí být zpracována (např. data ze senzorů v reálném čase).
 - **Variety (Různorodost):** Data mají různé formáty – strukturovaná (tabulky), polostrukturovaná (JSON, XML) a nestrukturovaná (video, audio, texty).
 - **Veracity (Věrohodnost):** Kvalita a přesnost dat. Velká data jsou často „špinavá“ a vyžadují čištění.
 - **Value (Hodnota):** Schopnost proměnit surová data v užitečné informace pro byznys nebo vědu.
-

2. Architektura zpracování

Zpracování velkých dat probíhá ve dvou hlavních režimech:

Dávkové zpracování (Batch Processing)

Data se sbírají po určitou dobu a pak se zpracují najednou jako velký balík. Je to efektivní pro historické analýzy.

- **Hlavní nástroj: Apache Hadoop (MapReduce).**

Proudové zpracování (Stream Processing)

Data se zpracovávají okamžitě, jakmile dorazí. Klíčové pro systémy, kde záleží na milisekundách (např. detekce podvodů u platebních karet).

- **Hlavní nástroje: Apache Spark Streaming, Apache Flink, Apache Kafka.**
-

3. Klíčové technologie a nástroje

Technologie	Popis
Apache Hadoop	Ekosystém pro distribuované ukládání (HDFS) a zpracování dat na stovkách serverů.
Apache Spark	Moderní a velmi rychlý engine, který provádí výpočty v operační paměti (In-memory).
Apache Kafka	Distribuovaná platforma pro přenos zpráv, která slouží jako „centrální nervová soustava“ pro proudy dat.
NoSQL databáze	Databáze jako MongoDB nebo Cassandra navržené pro obrovské objemy nestrukturovaných dat.
Data Lake	Úložiště, které uchovává obrovské množství dat v jejich přirozeném, surovém formátu.

4. Využití v praxi

- **Personalizace:** Doporučovací systémy (Netflix, Amazon, YouTube).
- **Zdravotnictví:** Analýza genomu a predikce šíření epidemií.
- **Smart Cities:** Optimalizace dopravy na základě dat z GPS a senzorů v reálném čase.
- **Finance:** Analýza trhů a automatizované obchodování (algorithmic trading).

Související články:

- [Strojové učení: Jak využít data](#)
- [Python pro datovou vědu](#)
- [Umělá inteligence: Budoucnost analýzy](#)

Tagy: *it big-data analytics hadoop spark kafka data-science infrastructure*

From:

<https://www.serviceit.cz/> - IT ENCYKLOPEDIE

Permanent link:

https://www.serviceit.cz/doku.php?id=it:sw:big_data

Last update: **2026/01/02 19:59**

