Data Lake (Datové jezero)

Data Lake je rozsáhlé úložiště, které uchovává data v jejich nativním formátu (raw), dokud nejsou potřeba pro analýzu. Je postaveno na principu **Schema-on-Read**, což znamená, že struktura dat se definuje až ve chvíli, kdy je čteme, nikoliv při jejich ukládání.

Je to klíčová součást moderní MLOps infrastruktury a Big Data architektury, protože umožňuje levně ukládat petabajty informací bez nutnosti okamžitého čištění.

1. Architektura: Od surových dat k informacím

Moderní data lake se obvykle dělí do logických zón (často nazývaných **Medallion Architecture**):

- **Bronze (Raw):** Zóna pro surová data přímo ze zdrojů. Data jsou zde uložena „tak jak jsou“, včetně chyb a duplicit.
- **Silver (Trusted):** Data jsou vyčištěná, normalizovaná a připravená pro datové vědce k experimentování.
- **Gold (Refined):** Data jsou agregovaná a strukturovaná pro konkrétní business potřeby (např. reporting).

2. Data Lake vs. Data Warehouse

Tyto dva koncepty se často pletou, ale slouží k odlišným účelům:

Vlastnost	Data Lake	Data Warehouse
Data	Strukturovaná, polostrukturovaná i nestrukturovaná.	Pouze vysoce strukturovaná (tabulky).
Předpis (Schema)	**Schema-on-Read** (při čtení).	**Schema-on-Write** (při zápisu).
Uživatelé	Datoví vědci, ML inženýři.	Business analytici, manažeři.
Cena	Nízká (levné cloudové úložiště).	Vyšší (optimalizováno pro výkon).
Hlavní účel	Experimentování, trénování AI.	BI, reportování, historické přehledy.

3. Technologie pro Data Lake

Data lake obvykle běží na distribuovaných systémech:

- **Cloudová úložiště:** Amazon S3, Azure Data Lake Storage (ADLS), Google Cloud Storage.
- **On-premise / Open source:** Apache Hadoop (HDFS).
- **Formáty souborů:** Pro efektivní čtení se používají sloupcové formáty jako **Parquet** nebo **Avro**.
- **Správa metadat:** Nástroje jako **Apache Hive** nebo **AWS Glue**, které udržují přehled o tom, co v „jezeře“ vlastně je.

4. Hlavní výhody a rizika

Výhody:

- **Flexibilita:** Můžete uložit data, pro která zatím nemáte využití, ale v budoucnu mohou být cenná.
- **Škálovatelnost:** Snadno roste s objemem dat (petabajty nejsou problém).
- **Demokratizace dat:** Všechny týmy mají přístup k jednomu centrálnímu zdroji pravdy.

Rizika (Data Swamp): Bez správného katalogování, správy metadat a řízení přístupů se data lake může změnit v **Data Swamp** (datovou bažinu) – místo, kde sice data jsou, ale nikdo je neumí najít, pochopit nebo ověřit jejich kvalitu.

5. Budoucnost: Data Lakehouse

Dnes se tyto světy propojují do konceptu **Data Lakehouse** (např. Databricks nebo Snowflake). Ten kombinuje levné úložiště a flexibilitu jezera s výkonem a správou transakcí (ACID), kterou známe z datových skladů.

> **Příklad z praxe:** E-shop ukládá do Data Lake všechna kliknutí uživatelů na webu (miliardy řádků měsíčně). Datoví vědci z těchto surových dat následně trénují doporučovací systém, zatímco vyčištěná data o nákupech posílají do Data Warehouse pro měsíční přehled tržeb.

Zpět na Data a Databáze