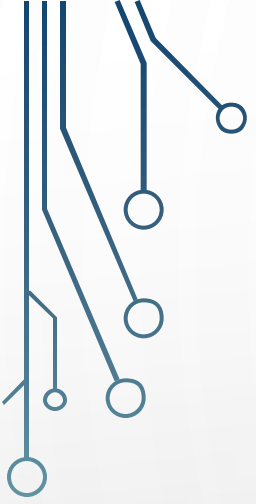


The background of the slide is a dark blue/black field filled with a dense, slightly blurred pattern of white and light blue binary digits (0s and 1s). Overlaid on this are several thin, light blue lines that resemble circuit traces or data paths, with small circles at the junctions and endpoints. These lines are more prominent on the left and right sides of the slide, framing the central text area.

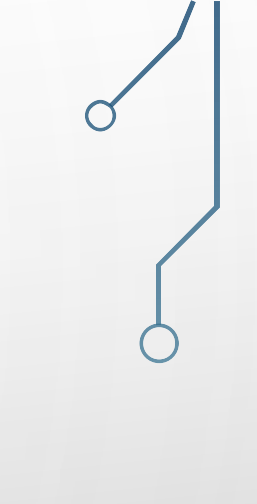
BIG DATA, DATALAKES, DATAWAREHOUSE

DOC. ING. MARCELA HALLOVÁ, PHD.



402,74 miliónov terabytov

**Odhadované množstvo dát,
ktoré sa vyprodukuje
každý deň**



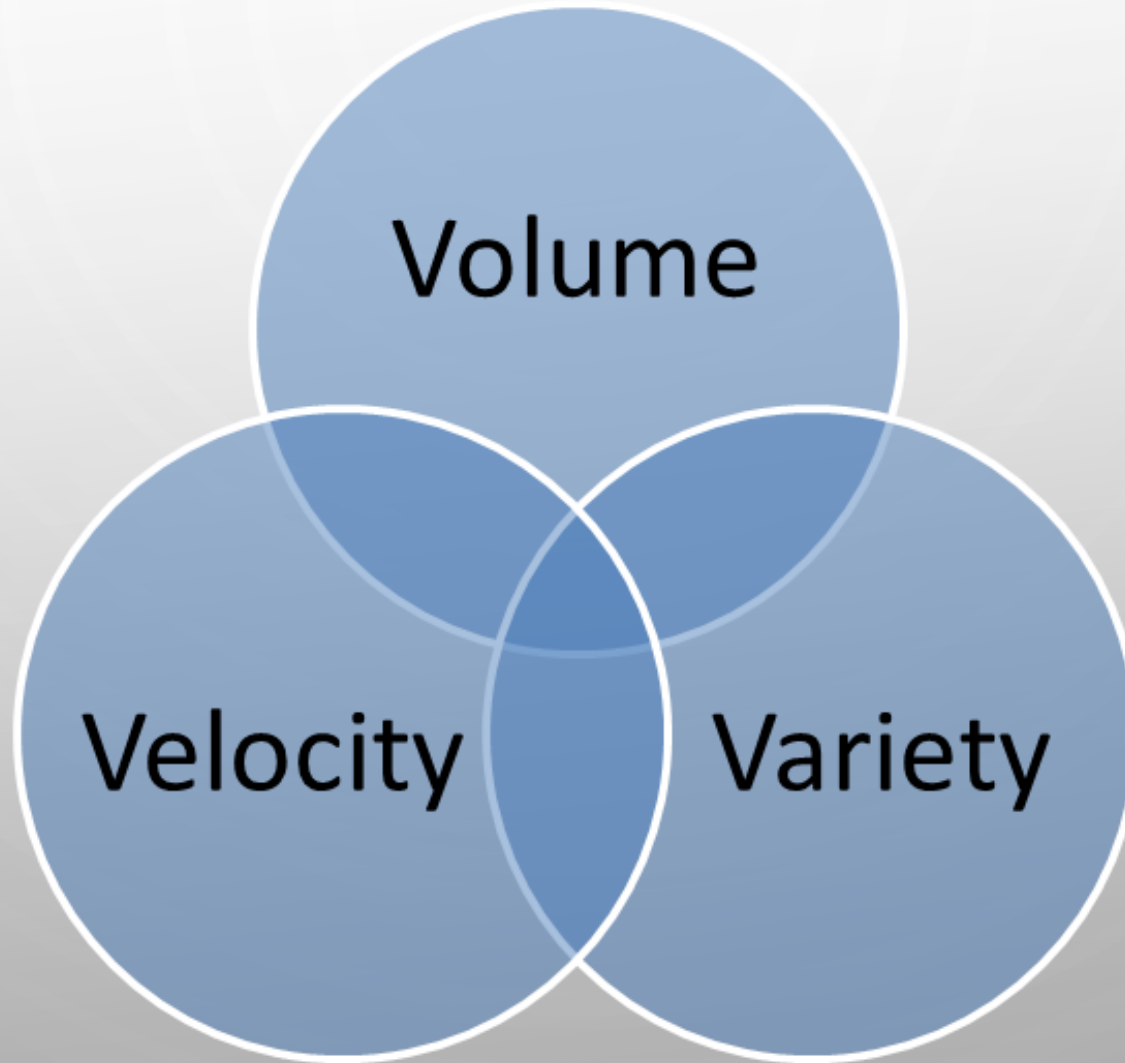
- **Amazon S3:** Hlavná cloudová úložná služba uchováva viac ako 500 biliónov (trillions) objektov a globálne obsluhuje viac ako 200 miliónov požiadaviek za sekundu.
- **Trhový podiel:** Amazon Web Services dominuje globálnemu trhu cloudových úložísk, pričom spotrebováva približne **38% až 41,5% svetovej kapacity** cloudových služieb.
- **Dátové centrá:** Amazon prevádzkuje celosvetovú sieť s viac ako **900 dátovými centrami** v 50 krajinách, pričom len za posledné obdobie zvýšil kapacitu o gigawatty energie kvôli boomu v oblasti umelej inteligencie (AI).



EVOLÚCIA BIG DATA

1. Fáza – štruktúrovaný obsah Obdobie 1970 - 2000	2. Fáza – neštruktúrovaný obsah založený na údajoch z webov Obdobie 2000 - 2010	3. Fáza – Údaje založené na mobilných a senzorických dátach Obdobie 2010 - súčasnosť
✓ Relačné databázové systémy, dátové sklady ✓ Analytické spracovanie online ✓ Dashboardy a výsledkové karty ✓ Data mining a štatistické analýzy	✓ Získavanie a extrakcia údajov ✓ Získavanie názorov – opinion mining ✓ Webové analýzy a web inteligencia ✓ Analýza sociálnych médií ✓ Analýza sociálnych sietí ✓ Priestorová časová analýza	✓ Lokalizačné analýzy ✓ Personálne analýzy ✓ Analýzy relevancie obsahu ✓ Mobilné vizualizácie ✓ Interakcie človek-počítač

CHARAKTERISTICKÉ VLASTNOSTI BIG DATA



CHARAKTERISTIKA BIG DATA

3V

ZÁKLADNÉ
VLASTNOSTI



VOLUME

Obrovské objemy dát
v **petabajtoch** až
exabajtoch.



VELOCITY

Rýchlosť, akou dáta
vznikajú a musia byť
spracované.



VARIETY

Rôznorodosť dátových
typov a formátov –
štruktúrované aj
neštruktúrované.



7V

ROZŠÍRENÉ
VLASTNOSTI



VERACITY

Dôveryhodnosť, kvalita
a presnosť zozbieraných
dát.



VARIABILITY

Premenlivosť dát – menia
sa v čase, štruktúra aj
význam podľa kontextu.



VALUE

Hodnota, ktorú dáta
prinášajú pri správnom
použití a analýze.



VISUALIZATION

Prehľadné zobrazenie
a interpretácia výsledkov
analýz.



VOLUME

Stále rastúce objemy dát
z rôznych zdrojov.



VELOCITY

Spracovanie dát v reálnom
čase alebo takmer v reálnom
čase.

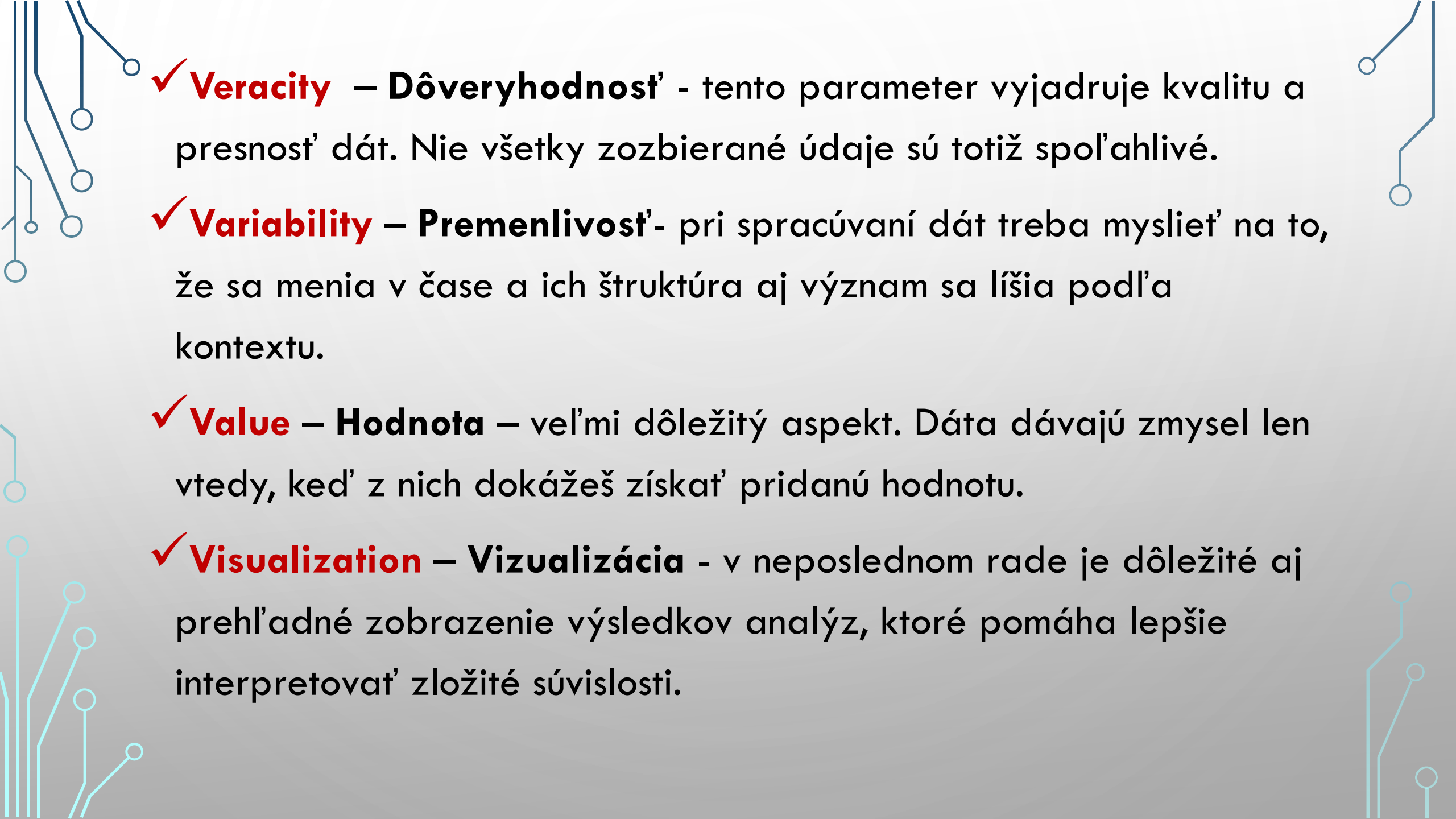


VARIETY

Rôznorodosť formátov,
štruktúr a zdrojov dát.

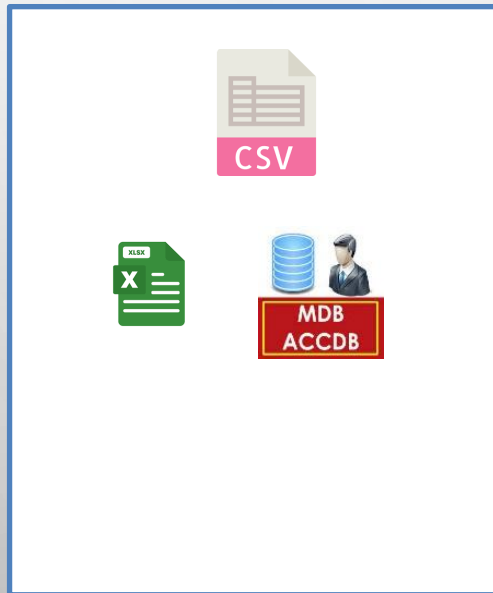
CHARAKTERISTICKÉ VLASTNOSTI BIG DATA

- ✓ **Volume** – **Množstvo** - veľké množstvo dát. Veľkosť údajov hrá rozhodujúcu úlohu pri určovaní ich hodnoty.
- ✓ **Velocity** – **Rýchlosť tvorby** - dáta prichádzajú vysokou rýchlosťou zo strojov, sietí, sociálnych médií, mobilných telefónov a iných zdrojov rýchlosťou veľkých dát.
- ✓ **Variety** – **Rôznorodosť** – sem patrí rozdelenie dát na štruktúrované, polo-štruktúrované a neštruktúrované.

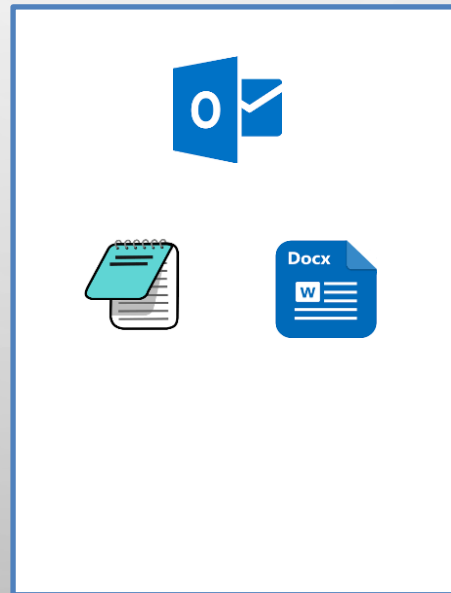
- 
- ✓ **Veracity** – **Dôveryhodnosť** - tento parameter vyjadruje kvalitu a presnosť dát. Nie všetky zozbierané údaje sú totiž spoľahlivé.
 - ✓ **Variability** – **Premenlivosť** - pri spracúvaní dát treba myslieť na to, že sa menia v čase a ich štruktúra aj význam sa líšia podľa kontextu.
 - ✓ **Value** – **Hodnota** – veľmi dôležitý aspekt. Dáta dávajú zmysel len vtedy, keď z nich dokážeš získať pridanú hodnotu.
 - ✓ **Visualization** – **Vizualizácia** - v poslednom rade je dôležité aj prehľadné zobrazenie výsledkov analýz, ktoré pomáha lepšie interpretovať zložité súvislosti.

Typy dát

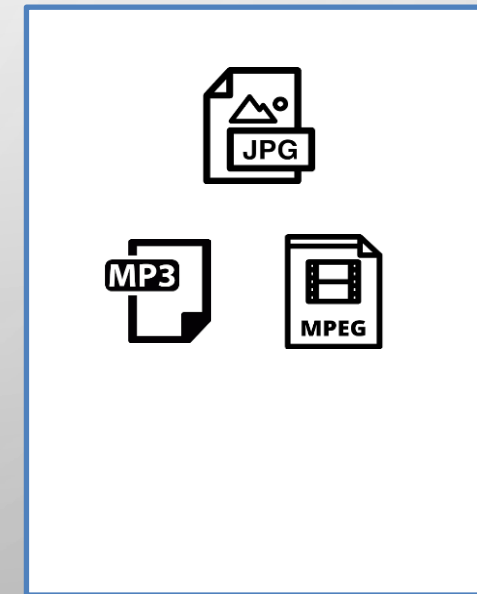
ŠTRUKTÚROVANÉ



POLO - ŠTRUKTÚROVANÉ



NEŠTRUKTÚROVANÉ



FUNGOVANIE BIG DATA

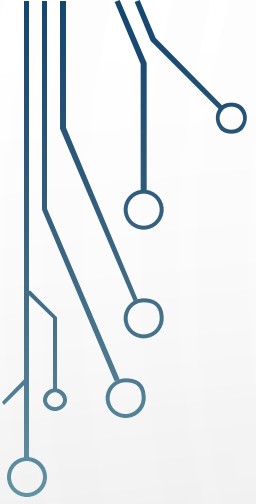
- 1) **Zber dát (data collection)** – zozbieranie dát z rôznych dostupných zdrojov.
- 2) **Ukladanie dát (data storage)** - vo väčšine prípadov putujú do databáz alebo tzv. data lakes, ktoré sú prispôsobené aj na spracovanie obrovských objemov informácií.
- 3) **Spracovanie dát (data processing)** – čistenie, filtrácia, transformácia do vhodného formátu.
- 4) **Analýza dát (data analysis)** - štatistické metódy, dátová analytika, strojové učenie alebo umelá inteligencia – odhaľujú sa vzory, predpovedajú trendy a generujú odporúčania.
- 5) **Vizualizácia a interpretácia** - grafy, prezentácie alebo interaktívne vizualizácie, Google Data Studio, Tableau, Power BI.

VÝHODY A PRÍNOSY BIG DATA

- ✓ **Lepšie obchodné rozhodnutia** – veľké dáta umožňujú firmám prijímať rozhodnutia založené na faktoch, nie na dohadoch. Manažéri tak reagujú rýchlejšie a presnejšie na zmeny trhu.
- ✓ **Zvýšená efektivita procesov** – moderná dátová analytika dokáže odhaliť neefektívne procesy, ktoré môžu firmy následne optimalizovať, čím šetria čas, pracovnú silu aj peniaze.
- ✓ **Personalizácia zákazníckej skúsenosti** – vďaka analýze správania klientov a zákazníkov je možné vytvárať ponuky šité na mieru, čo vedie k ich vyššej spokojnosti a lojalite.
- ✓ **Identifikácia nových príležitostí** – big data pomáhajú odhaľovať trendy a potreby, ktoré by inak mohli zostať nepovšimnuté.
- ✓ **Zníženie nákladov** – analýza dát môže viesť aj k zníženiu prevádzkových nákladov, minimalizovaniu chýb a lepšiemu plánovaniu rozpočtov.

VÝZVY PRI SPRACOVANÍ BIG DATA

- ✗ **Ochrana osobných údajov** – spracovanie veľkých objemov dát často zahŕňa citlivé informácie. Je preto nutné zabezpečiť ich anonymizáciu a súlad s legislatívou EÚ.
- ✗ **Rastúci objem dát zvyšuje riziko hackerských útokov a úniku informácií.** Firmy preto musia chrániť svoju infraštruktúru aj prístupové kódy.
- ✗ **Kvalita a presnosť dát** – neúplné, neaktuálne alebo chybné údaje môžu viesť k nesprávnym záverom a následne aj k zlým rozhodnutiam.
- ✗ **Technická náročnosť a náklady** – implementácia big data riešení si vyžaduje špecializované technológie, výkonné servery a často aj vysoké investície.
- ✗ **Nedostatok kvalifikovaných odborníkov** – na trhu chýbajú dátoví analytici, inžinieri a špecialisti schopní efektívne pracovať s veľkými dátami.



Kilobajt 1 000 bajtov

Megabajt 1 000 000 bajtov

Gigabajt 1 000 000 000 bajtov

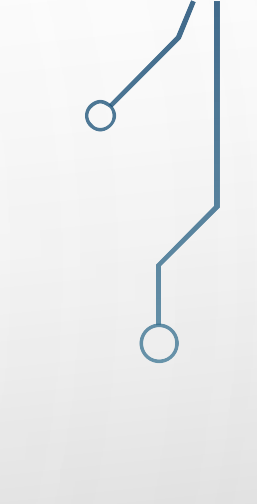
Terabajt 1 000 000 000 000 bajtov

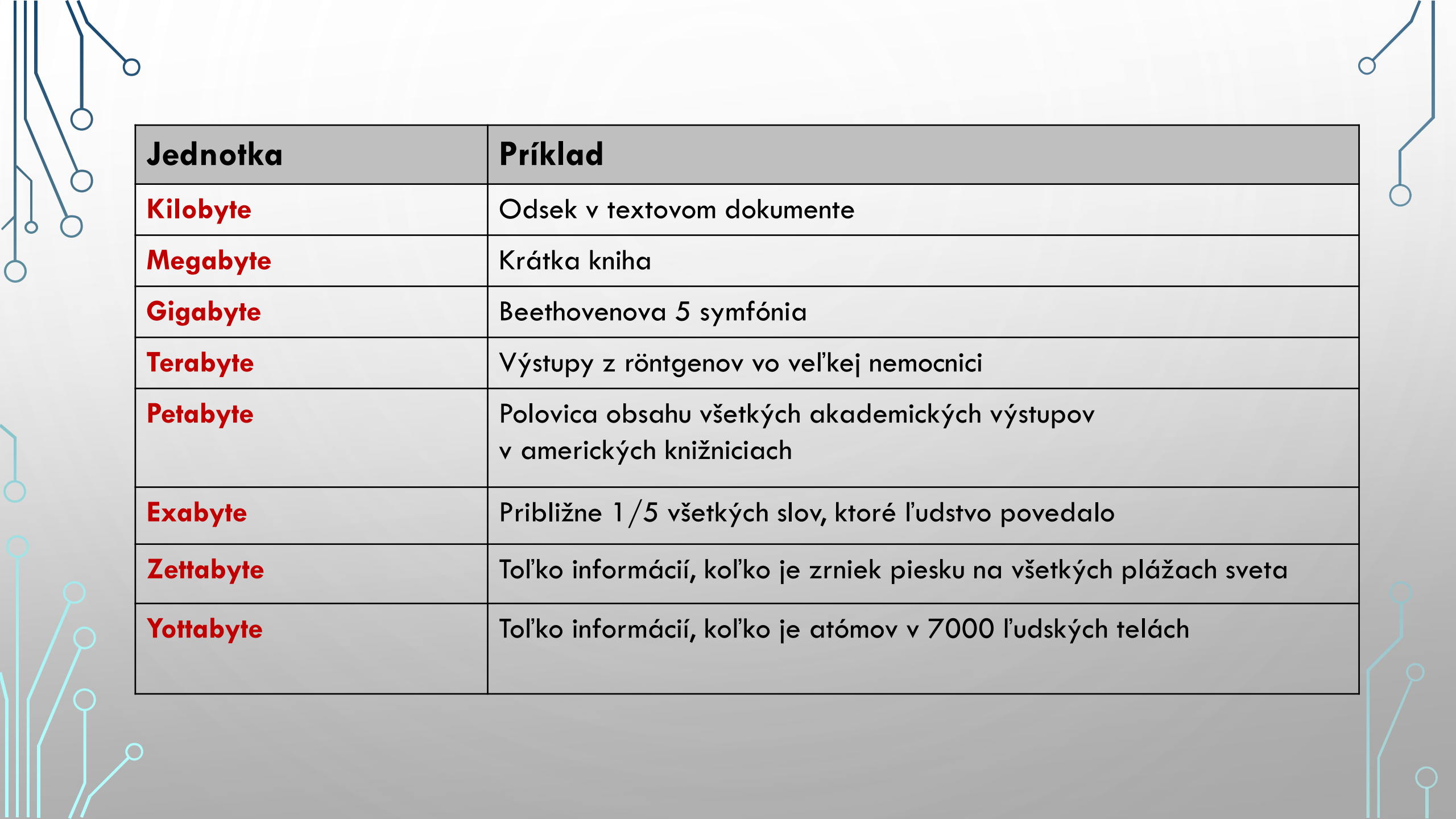
Petabajt 1 000 000 000 000 000 bajtov

Exabajt 1 000 000 000 000 000 000 bajtov

Zetabajt 1 000 000 000 000 000 000 000 bajtov

Yottabajt 1 000 000 000 000 000 000 000 000 bajtov





Jednotka	Príklad
Kilobyte	Odsek v textovom dokumente
Megabyte	Krátka kniha
Gigabyte	Beethovenova 5 symfónia
Terabyte	Výstupy z röntgenov vo veľkej nemocnici
Petabyte	Polovica obsahu všetkých akademických výstupov v amerických knižniciach
Exabyte	Približne 1 / 5 všetkých slov, ktoré ľudstvo povedalo
Zettabyte	Tol'ko informácií, koľko je zrníek piesku na všetkých plážach sveta
Yottabyte	Tol'ko informácií, koľko je atómov v 7000 ľudských telách

OLTP A OLAP

- OLTP - online transaction processing - spracovanie veľkého množstva krátkych operácií nad konkrétnymi záznamami. Dôležitá je konzistentnosť a rýchla odozva.
- OLAP - online analytical processing - slúži na skúmanie trendov, súhrnov a súvislostí. Typický analytický dotaz číta mnoho riadkov a počíta agregácie podľa času, regiónu či kategórie.
- Obe potreby môžu existovať v jednej organizácii, ale ich optimálna architektúra býva odlišná.
- ***Príklad – Vytvorenie objednávky je OLTP. Výpočet tržieb podľa kategórií za posledných 24 mesiacov je OLAP.***

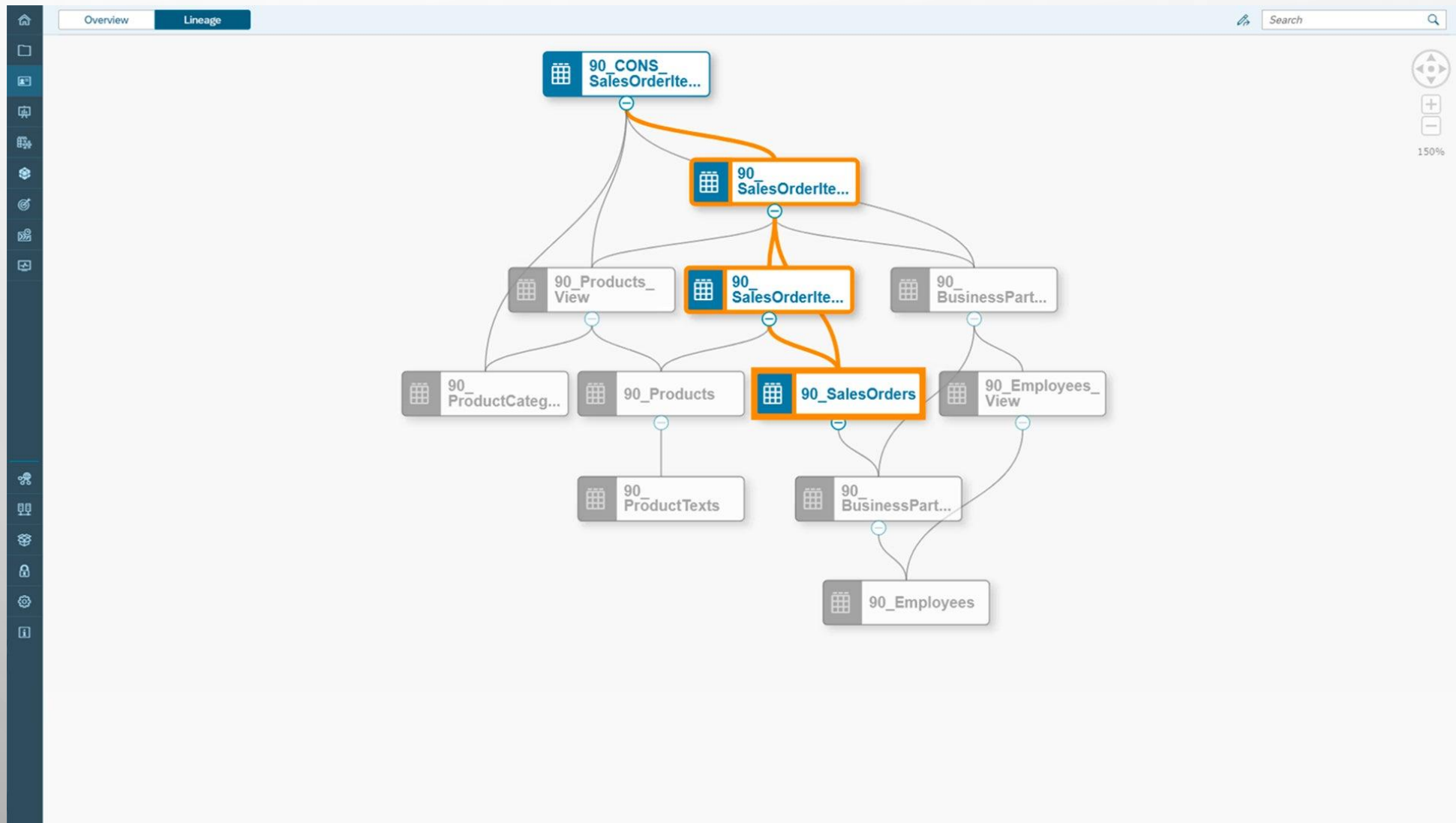
DATA WAREHOUSE – DÁTOVÝ SKLAD

- Data warehouse je analytické úložisko, do ktorého sa integrujú údaje z viacerých zdrojov.
- Jeho význam spočíva najmä v tom, že poskytuje zjednotené, dôveryhodné a historicky porovnateľné údaje.
- Pred uložením alebo pri príprave analytických tabuliek sa riešia rozdielne formáty, duplicity, identifikátory a obchodné pravidlá.
- Dátový sklad sa často používa pre business intelligence, pravidelné reporty a firemné KPI.
- Dôležitá je jednotná definícia metrík: napríklad tržba môže byť pred zľavou, po zľave, s DPH alebo bez DPH.

VÝHODY SKLADOVANIA DÁT

- ✓ **Lepšia podniková analytika** - dátový sklad zjednocuje údaje z viacerých systémov do jedného konzistentného zobrazenia podniku.
- ✓ **Rýchlejšie dotazy a prehľady** - používatelia môžu spúšťať komplexné dotazy na veľké množiny údajov oveľa rýchlejšie, čo urýchľuje vykazovacie cykly a znižuje závislosť od IT.
- ✓ **Zlepšená kvalita a konzistentnosť dát** - dáta sa pred vstupom do skladu vyčistia, overia a štandardizujú, čím sa zabezpečí, že analytika bude založená na vysoko kvalitných a spoľahlivých informáciách.
- ✓ **Hlbší historický prehľad** - bohaté historické údaje, čo uľahčuje rozpoznávanie dlhodobých vzorov, hodnotenie výkonu a vytváranie presnejších prognóz, ktoré posilňujú strategické plánovanie.

SNÍMKA OBRAZOVKY SKLADU ÚDAJOV ZOBRAZUJÚCA PÔVOD ÚDAJOV



DÁTOVÝ SKLAD VERZUS DATABÁZA

- Databázy aj dátové sklady ukladajú údaje, ale slúžia rôznym rolám.
- Databáza spravuje informácie pre konkrétny pracovný úsek v reálnom čase, zatiaľ čo dátový sklad kombinuje aktuálne a historické údaje z celej organizácie na podporu výkazníctva a analýzy.
- Hoci funguje na technológii databázy, dátový sklad pridáva nástroje na integráciu, modelovanie a správu dát v priebehu času.
- Databázy udržiavajú každodenné operácie spustené spracovaním transakcií a rýchlou aktualizáciou záznamov.
- Dátové sklady podporujú analytiku, pomáhajú tímom rozpoznávať trendy, porovnávať výkon a prijímať strategické rozhodnutia.

DATA LAKE – DÁTOVÉ JAZERO

- ✓ Centralizované úložisko, ktoré ukladá štruktúrované, pološtruktúrované a neštruktúrované údaje vo svojich natívnych formátoch.
- ✓ Dátové jazero akceptuje nespracované dáta tak, ako sú, a zachováva svoju pôvodnú štruktúru a formát, kým nie sú potrebné pre prípady použitia pokročilej analýzy, umelej inteligencie (AI) a strojového učenia (ML).
- ✓ Zahŕňa konsolidáciu údajov z viacerých zdrojov na jedno prístupné miesto – dátové jazero, čo znamená, že analytici, analytici a inžinieri strojového učenia môžu všetci skúmať, experimentovať s informáciami, ktoré by inak mohli ostať uväznené v nesúrodých systémoch.

ZDROJE ÚDAJŮV, KTORÉ BY MOHLI BYŤ ULOŽENÉ V DÁTOVOM JAZERE

- Databázy
- Súbory
- Aplikčné protokoly
- Informačné kanály sociálnych médií
- Protokoly snímačov IoT

TYPY DÁTOVÝCH JAZIER

- ✓ **Cloudové dátové jazerá** - hostené výlučne na cloudových platformách, neobmedzená škálovateľnosť, priebežné stanovenie cien a jednoduchú integráciu s cloudovou natívnou analytikou a službami umelej inteligencie.
- ✓ **Lokálne dátové jazerá** - sú nasadené v rámci vlastných dátových centier organizácie a poskytujú úplnú kontrolu nad infraštruktúrou, bezpečnosťou a suverenitou dát.
- ✓ **Hybridné dátové jazerá** - kombinujú cloudové a lokálne úložisko, čo organizáciám umožňuje uchovávať niektoré údaje lokálne a zároveň využívať cloudové zdroje na škálovateľnosť a pokročilú analýzu.
- ✓ **Viaccloudové dátové jazerá** - zahŕňajú viacerých poskytovateľov cloudu, čo pomáha organizáciám vyhnúť sa odhláseniu dodávateľov, optimalizovať náklady pomocou najlepších služieb od každého poskytovateľa a zabezpečiť kontinuitu podnikania prostredníctvom redundancie.



**ĎAKUJEM ZA
POZORNOST!**