# BIG DATA

www.trigama.eu

trigama

## CO JSOU TO BIG DATA?

**Jsou to soubory dat, které je vzhledem k jejich velikosti obtížné zpracovávat obvyklými IT prostředky. Stejný pojem ale označuje i postupy a technologie, které zpracování takových souborů dat umožňují.**

## BIG DATA
## KDY JSOU MÁ DATA OPRAVDU VELKÁ?

**Velká data splňují kritérium některého ze „tří V“**

### VOLUME

**Data potřebná ke zpracování úlohy jsou tak objemná, že jejich uchovávání v tradičních databázích by bylo příliš drahé.**

#### PŘÍKLAD

Detailní zdrojová data o vztazích a interakcích zákazníků s jiným uživateli, aplikacemi nebo systémy (demografické údaje, komunikační data, nákupní chování)

#### MOŽNOSTI VYUŽITÍ

- Vyhodnocení geografických a časových závislostí.
- Rozpoznání vztahů, identifikace vlivných jedinců.
- Analýzy spotřebitelských preferencí.

### VELOCITY

**Úloha vyžadující zpracování velkého a neustálého přírůstku dat v reálném čase.**

#### PŘÍKLAD

Strojově generovaná data sloužící pro sledování stavu systémů a zařízení (systémové logy, data ze senzorů, údaje o provozu komunikačních sítí)

#### MOŽNOSTI VYUŽITÍ

- Monitoring, prevence chyb a výpadků, předcházení škodám.
- Předzpracování objemných dat před plněním do relačních databází.

### VARIETY

**Vysoká různorodost, variabilita dat, nutnost kombinovat data z více zdrojů v různých strukturách a formátech.**

#### PŘÍKLAD

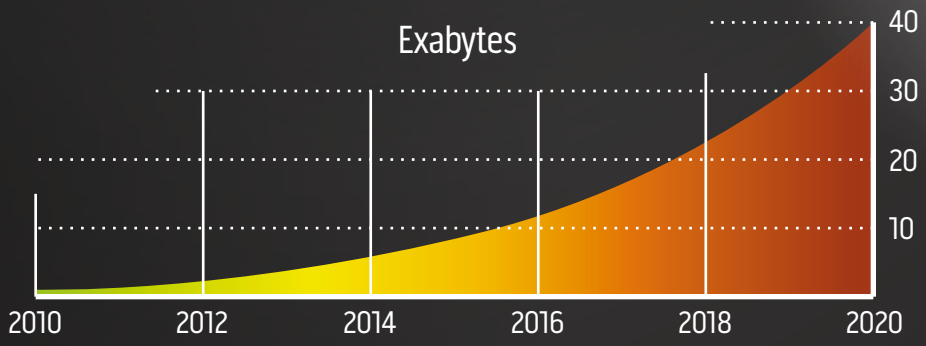
Nestrukturovaná data vytvářená velkou skupinou přispěvatelů (veřejně dostupné internetové zdroje, zpravodajství, sociální sítě)

#### MOŽNOSTI VYUŽITÍ

- Analýzy velkých objemů nestrukturovaných dat, např. textové prohledávání.
- Archivace dat z různých zdrojů v rozmanitých formátech.

## VÝVOJ OBJEMU DAT OD ROKU 2010

Objem generovaných dat v posledních letech exponenciálně narůstá. Spolu s tím roste i potřeba analýzy velkých dat a nároky na jejich zpracování.
Zdroj dat: http://www.unece.org



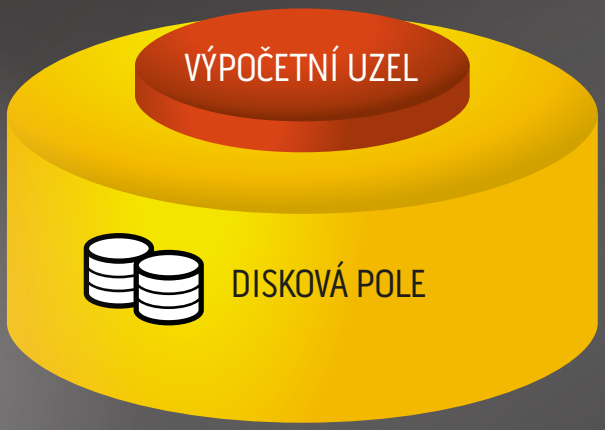
trigama

# Zpracování velkých dat

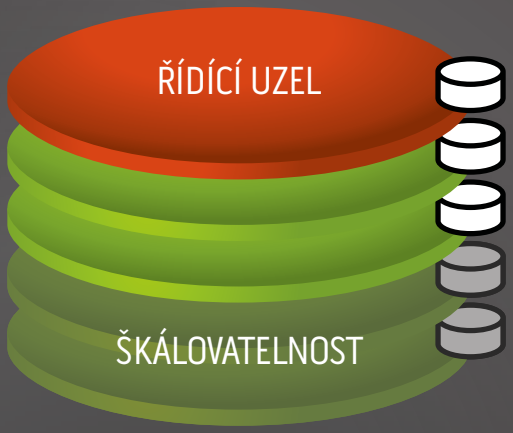
www.trigama.eu

**Tradiční databázové technologie jsou optimalizované pro rychlý přístup k údajům uloženým v předem dané struktuře. Pro ostatní typy úloh, kde je třeba zpracovávat velké objemy dat, je vhodnější používat specializované aplikace běžící v rámci výpočetního clusteru. Ten umožní rozdělit úlohu na menší části, každou část zpracuje paralelně a výsledek složí z dílčích odpovědí ve výrazně kratším čase.**

## TRADIČNÍ RELAČNÍ DATABÁZE




## BIG DATA CLUSTER




| Datový sklad | Big data cluster |
|---|---|
| ? Oddělené výpočetní jednotky a disky | ? Disková kapacita přímo na výpočetním uzlu |
| − Vyžaduje specializovaný HW | + Vyhovuje běžně dostupný HW |
| − Obtížná škálovatelnost | + Jednoduché připojení dalších uzlů zvýší výpočetní výkon i kapacitu |
| − Zpracování velkého objemu dat neefektivní | + Velká data zpracovávají všechny uzly najednou |
| − Uložení strukturovaných dat | + Uložení strukturovaných i nestrukturovaných dat |
| + Jednoduché dotazy zpracuje ihned | − I dotaz na malá data má velkou režii |

## BIG DATA A HADOOP

hadoop

**Apache Hadoop je programové prostředí, které umožňuje paralelní běh big data aplikací v rámci výpočetního clusteru. Zahrnuje sadu nástrojů pro distribuované pořizování, ukládání a zpracování velkých dat.**

Je to open-source systém volně dostupný i pro komerční použití (podobně jako třeba webový server Apache). Existují ale také komerční distribuce, součástí jejichž licence je i provozní podpora (např. Hortonworks, Cloudera, MapR).