BIG DATA 1

Technológie spracovania veľkých dát

Peter Bednár, Martin Sarnovský

Úvod – organizácia predmetu

- Stránka predmetu:
  - http://web.tuke.sk/fei-cit/sarnovsky/tsvd/
- Cvičenia
  - 40 bodov = 20 bodov zadanie + 2 * 10 bodov test – znalosti z cvičení aj z prednášok
- Skúška
  - 60 bodov, test

Úvod – prehľad tém na prednáškach

- Paralelné výpočty
- Distribuované súborové systémy a databázy
- Distribuované výpočty
- Architektúry pre spracovanie Veľkých dát
- Technológie pre spracovanie Veľkých dát
- Distribuované spracovanie prúdov dát
- Distribuované metódy strojového učenia

Charakteristika Veľkých dát

- 3V Model

- Objem (Volume) – veľkosť zhromaždených a spracovávaných dát v GB/TB/PB
- Rýchlosť (Velocity) – rýchlosť s akou sú dáta generované a ako rýchlo ich treba spracovať (dáta sú rýchlo aktualizované - samotné aktualizácie však môžu mať malý objem)
- Rôznorodosť (Variety) – je potrebné spracovať dáta rôznych typov (štruktúrované dáta z databáz, texty, multimédia, senzorické dáta, atď.), typ dát sa môže meniť

Charakteristika Veľkých dát

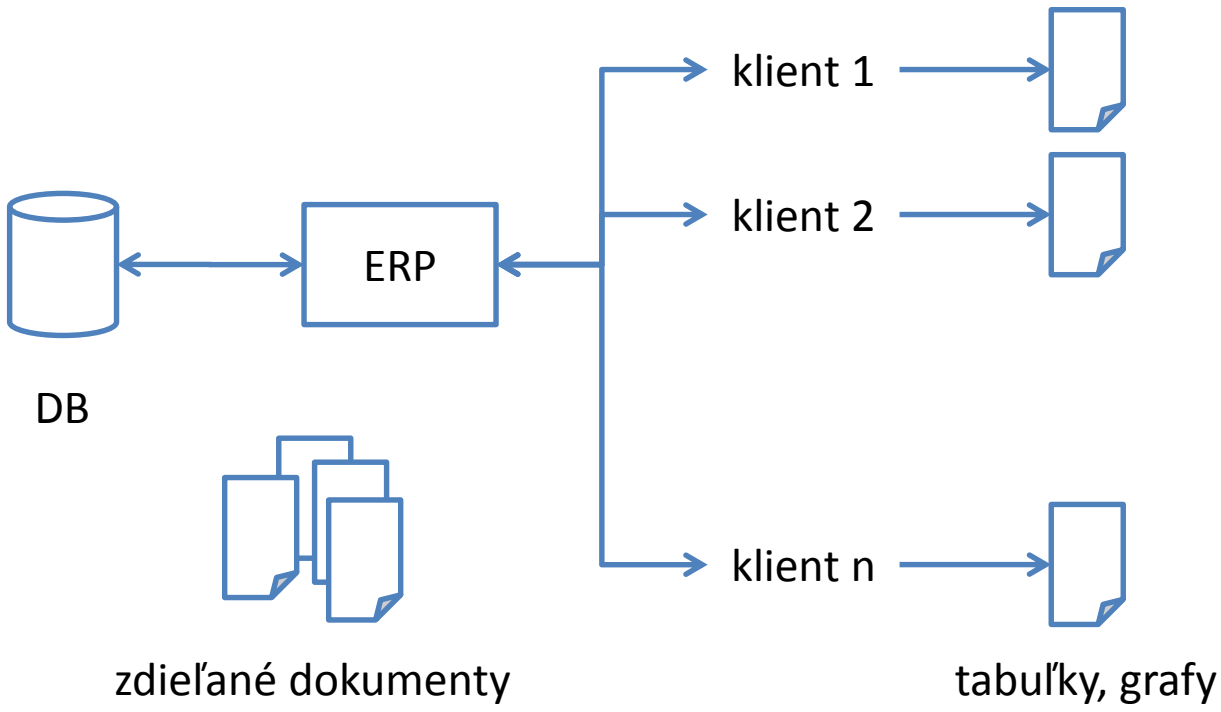
- 5V Model = 3V Model + ďalšie dve charakteristiky

- Pravdivosť (Veracity) – dáta môžu byť nekonzistentné, chybné, zdroj je nedôveryhodný
- Hodnota (Value) – dáta zhromažďujeme a spracúvame aby sme získali nové znalosti, ktoré vieme efektívne aplikovať - zhromažďované dáta musia byť potencionálne užitočné
  - Na druhej strane je niekedy ťažké odhadnúť na koľko môžu byť dáta užitočné v budúcnosti

Podniková analytika 1.0

- Využíva sa architektúra klient-server
- Viacero klientov sa pripája na centralizovaný server na ktorom beží podnikový informačný systém (Enterprise Management System – ERP) pripojený na relačnú databázu
- Neštruktúrované dáta sú uložené v textových dokumentoch v sieťových zdieľaných zložkách/adresároch
- Pre analýzu dát slúžia hlavne agregované reporty a grafické zostavy implementované priamo v klientskych aplikáciách
- Pre zložitejšie analýzy je potrebné dáta exportovať z databázy do externých nástrojov

Podniková analytika 1.0 – schéma



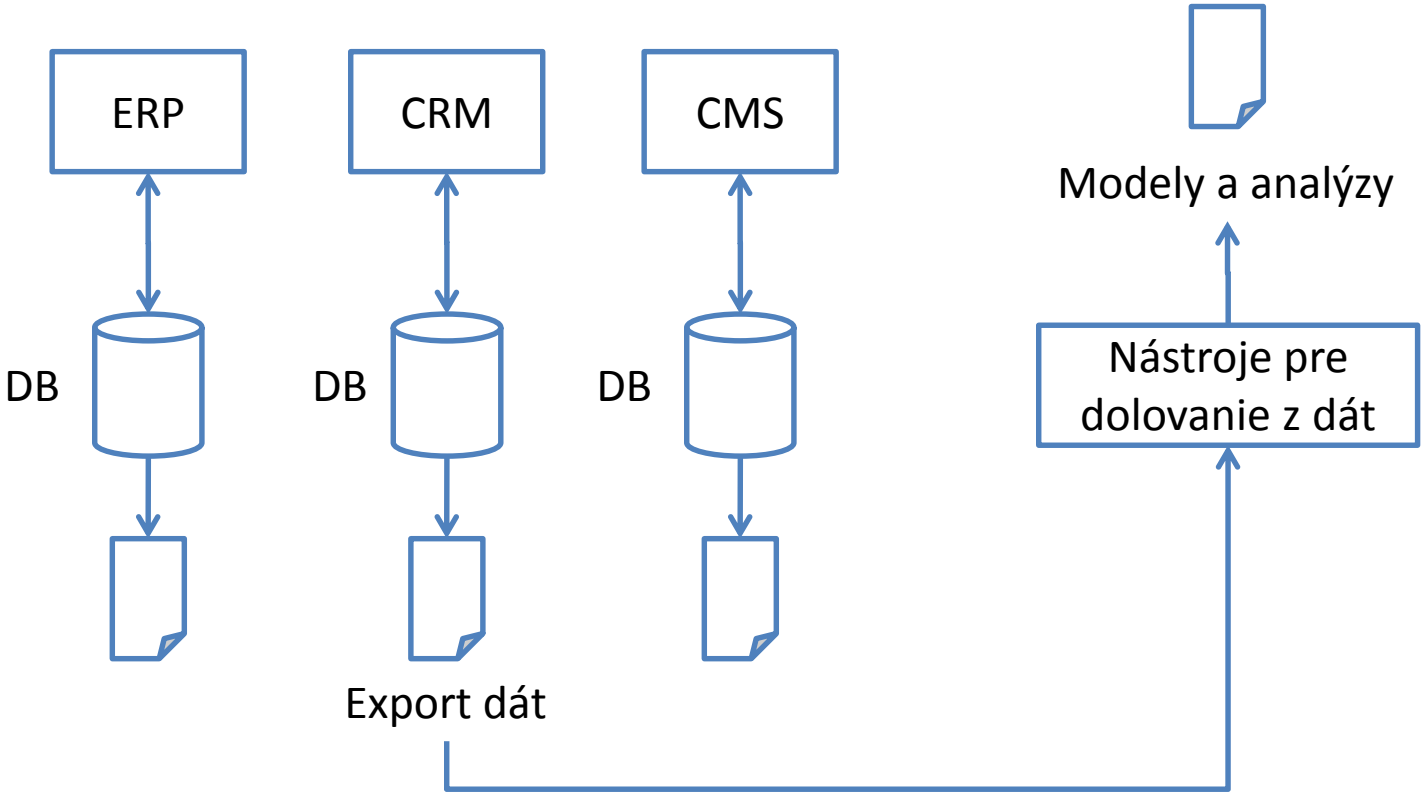
Podniková analytika 1.5 (1)

- ERP systémy sa stávajú zložitejšie a začínajú sa špecializovať, vznikajú napr. systémy pre manažment vzťahov so zákazníkmi (CRM - Customer Relationship Management)
- Neštruktúrované dáta sa uchovávajú v systémoch pre správu obsahu (CMS - Content Management System)
- Zložitejšia infraštruktúra s viacerými servermi – dáta sú fragmentované vo viacerých systémoch a databázach

Podniková analytika 1.5 (2)

- Narastá objem aj rôznorodosť dát, ktoré firmy spravujú, narastajú požiadavky na analytické metódy využívané pri bežnej činnosti firmy
- Začínajú sa využívať metódy dolovania znalostí v databázach
  - Prediktívne modely – snažíme sa predpovedať budúce udalosti
  - Názornejšie popisné modely – snažíme sa lepšie pochopiť historické dáta

Podniková analytika 1.5 – schéma



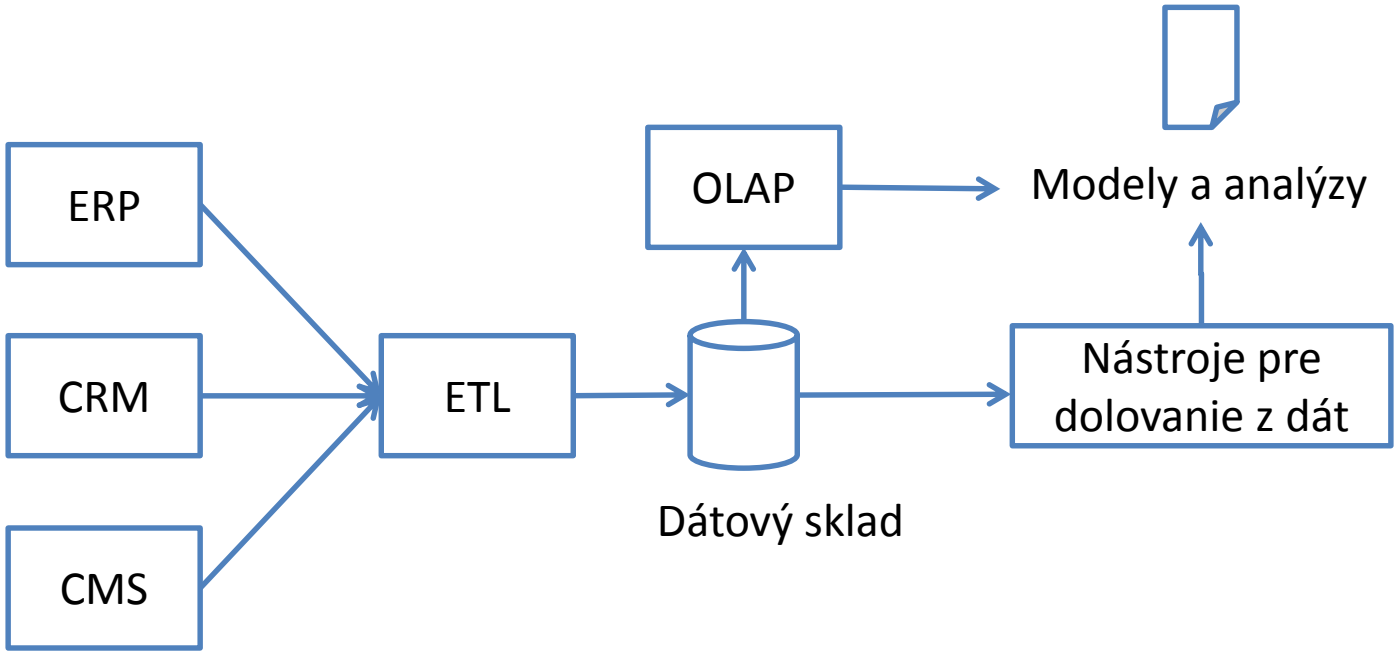
Podniková analytika 2.0 (1)

- Aby bolo možné dáta vyťažiť čo najlepšie, dáta z rôznych zdrojov musia byť integrované
- Zavádzajú sa dátové sklady – dáta z rôznych systémov sa integrujú do centrálneho dátového skladu postaveného na technológii relačných databáz
  - Proces integrácie prebieha v krokoch extrakcie, transformácie a načítania dát – (ETL operácie – Extraction Transformation Loading)

Podniková analytika 2.0 (2)

- Nad dátovými skladmi sa vytvárajú pred-počítané viacrozmerné dátové pohľady s ktorými manažéri môžu interaktívne pracovať – OLAP analýza Online Analytical Processing
- Ak sa dáta zmenia, dáta treba znova agregovať a prepočítať
- S rastúcim objemom dát neúmerne narastá doba spracovania
  - Spracovanie ETL + prepočítanie OLAP sa predlžuje na týždne, ak sa zaradia aj zložitejšie metódy dolovania dát na mesiace – pri niektorých obchodoch však závisí na veľmi krátkych časových intervaloch (napr. na burze sekundy)

Podniková analytika 2.0 – schéma



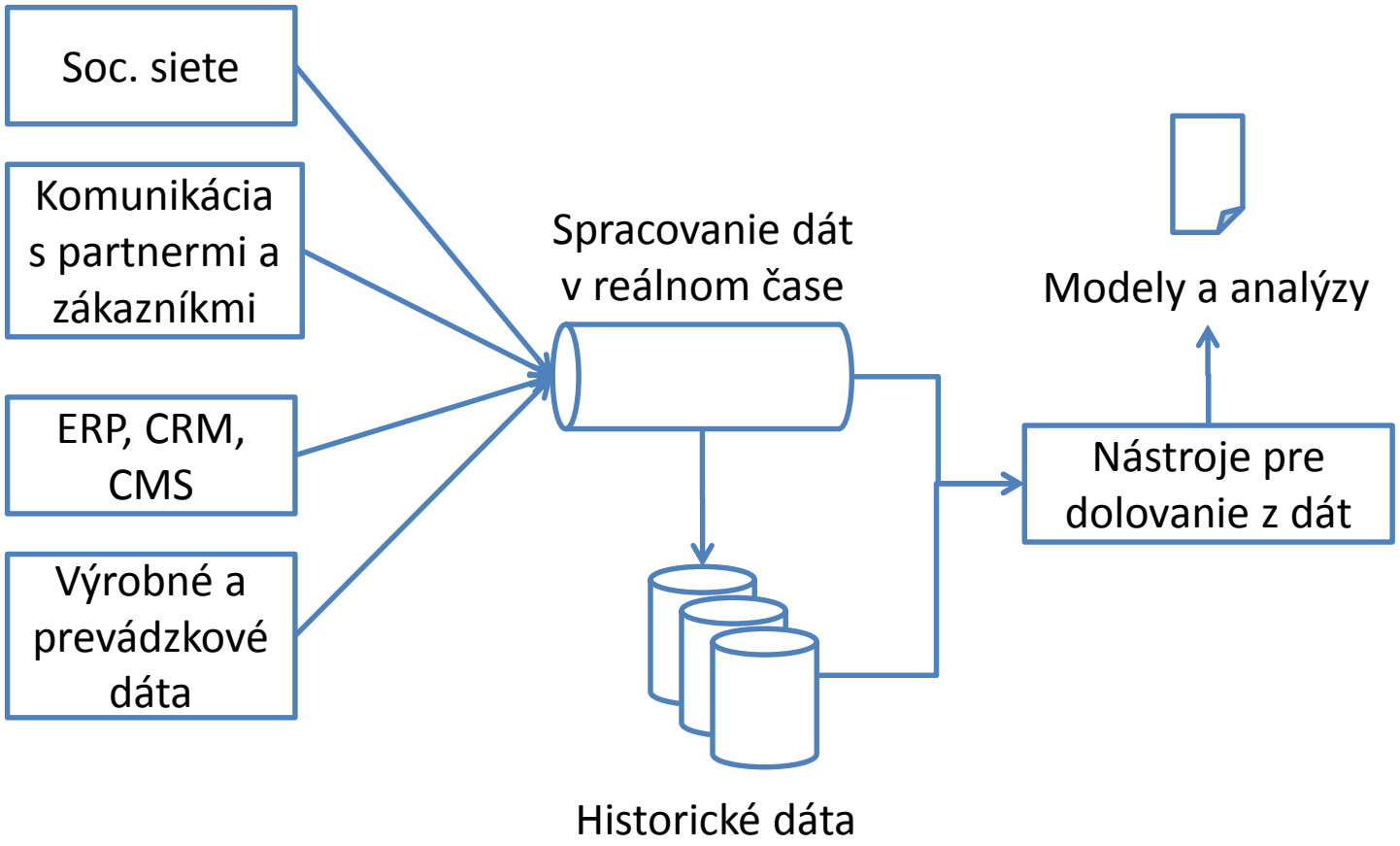
Podniková analytika 3.0 (1)

- Dáta z ERP/CRM/CMS systémov je treba integrovať a spracovať prakticky v reálnom čase
- Je potrebné rozšíriť dáta sledované o zákazníkoch
  - Sociálne siete
  - Komunikácia so zákazníkom cez rôzne kanály
  - Profil zákazníka
- Je potrebné integrovať dáta priamo z výroby, alebo prevádzky
  - Senzorické dáta
  - Diagnostické udalosti

Podniková analytika 3.0 (2)

- Inkrementálne aktualizované predikčné modely
- Interaktívne metódy pre popisnú analýzu dát
  - Vyžaduje si spracovanie veľkého objemu dát v reálnom čase

Podniková analytika 3.0 – schéma



Čo sú to Veľké dáta?

- Definícia 1: Za Veľké dáta sa považujú dáta, ktoré kvôli ich objemu, rýchlosti aktualizovania alebo variabilite nie je možné spracovať bežnými prostriedkami v požadovanom čase

Čo sú to Veľké dáta?

- Aby sme mohli dáta plne využiť, je potrebné zabezpečiť:

- Trvalé uloženie dát
- Spracovanie dát - pri ktorom sa na dáta aplikujú rôzne metódy ktoré ich transformujú na užitočné informácie
- Efektívny prístup k dátam – vyhľadávanie a filtrovanie dát relevantných pre danú úlohu

Výpočtové prostriedky (1)

- Dáta sa tradične spravujú pomocou serverových aplikácií, ktoré bežia na výkonnom počítači – serveri

- Procesor
  - Výkon sa meria v počte operácií za 1s
  - Závisí na zložitosti – počte tranzistorov a tzv. taktovacej frekvencii
  - Výkon jedného procesora je fyzikálne obmedzený (pri súčasne používaných dostupných technológiách)

Výpočtové prostriedky (2)

- Operačná pamäť
  - Procesor dokáže spracovať priamo iba dáta uložené v operačnej pamäti
  - Veľkosť do 100 GB
  - Rýchli prístup – desiatky ns, veľká dátová priepustnosť
  - Dáta nie sú trvalo uložené
- Trvalé pamäťové média
  - V súčasnosti pevné disky (HDD alebo SSD)
  - Slúžia na trvalé uchovanie dát
  - Pomalý prístup – cca 4ms pre HDD, pod 0.1ms pre SSD

Vertikálne škálovanie

- Ak narastajú požiadavky na výpočtové zdroje, pri vertikálnom škálovaní rozšírime konfiguráciu servera
  - Rýchlejší procesor
  - Viac operačnej pamäte
  - Viac diskovej kapacity a rýchlejšie disky

- Dostupné technológie však majú fyzikálne obmedzenia!
  - Nevieme viac integrovať elektronické obvody aby mali väčšiu kapacity a pracovali rýchlejšie

Horizontálne škálovanie

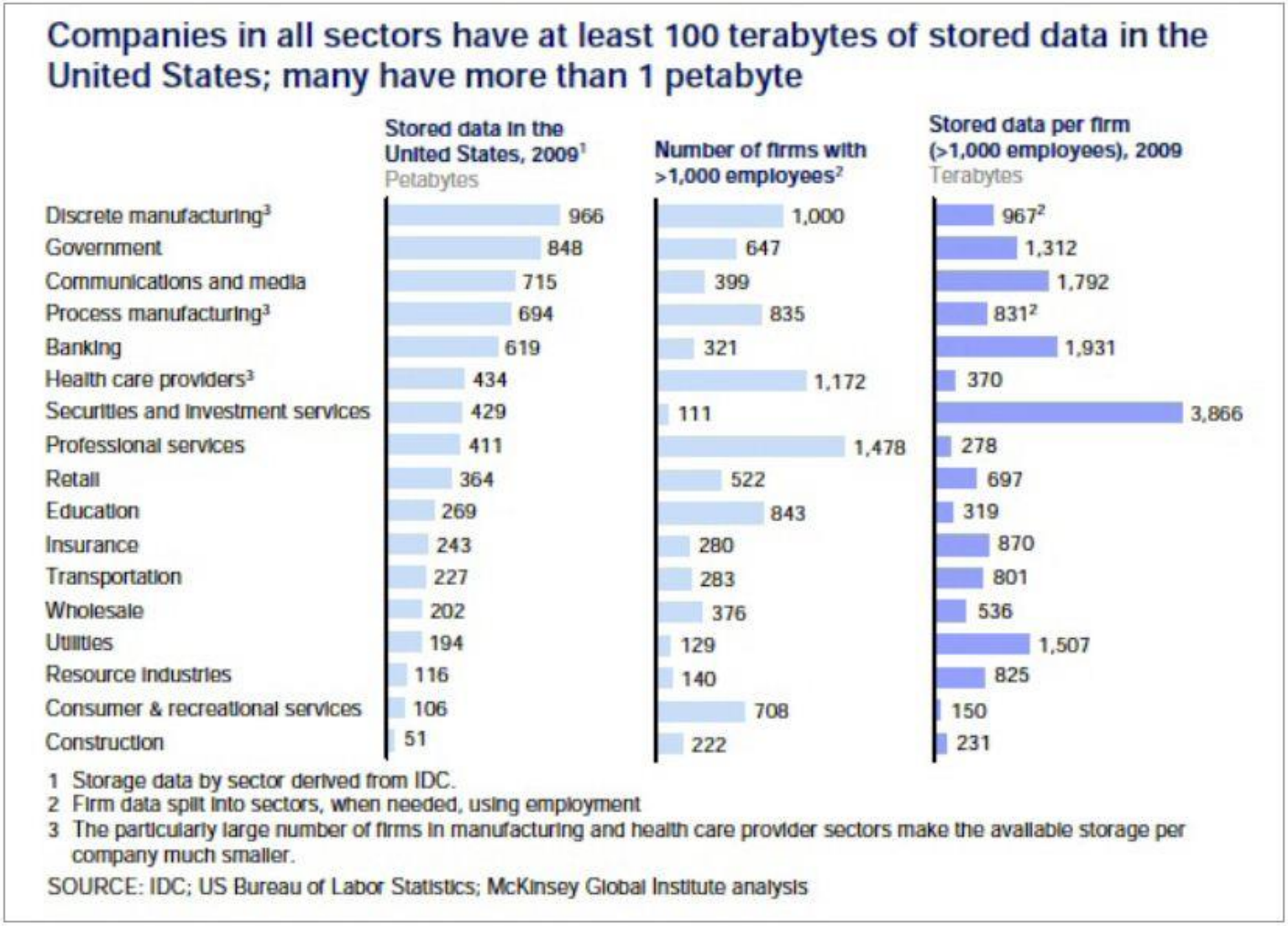
- Pri horizontálnom škálovaní rozdelíme spracovanie dát na viacero fyzických počítačov, ak narastajú požiadavky na výpočtové prostriedky, pridáme ďalší počítač a dáta prerozdelíme – výpočet prebieha distribuovane

- Počítače sú prepojené v sieti
  - Prenosová kapacita v 10-100 GB
  - celkový čas spracovania = čas potrebný na prenos dát a medzivýsledkov cez sieť + čas potrebný na spracovanie podmnožiny dát na najpomalšom uzle

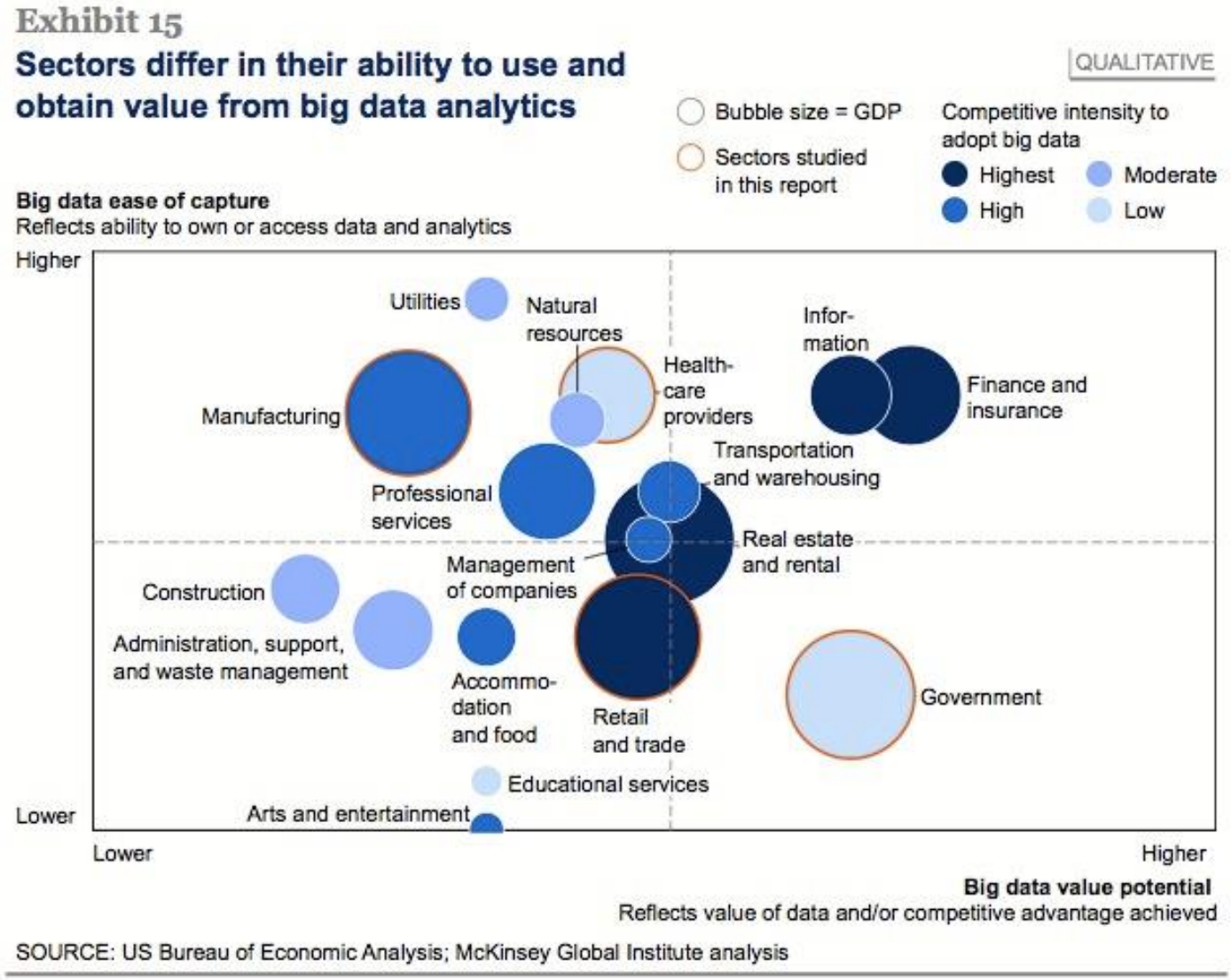
Čo sú to Veľké dáta?

- Definícia 2: Za Veľké dáta sa považujú dáta, ktoré kvôli ich objemu, rýchlosti aktualizovania alebo variabilite nie je možné spracovať bežnými prostriedkami v požadovanom čase a vyžadujú spracovanie v distribuovanom prostredí

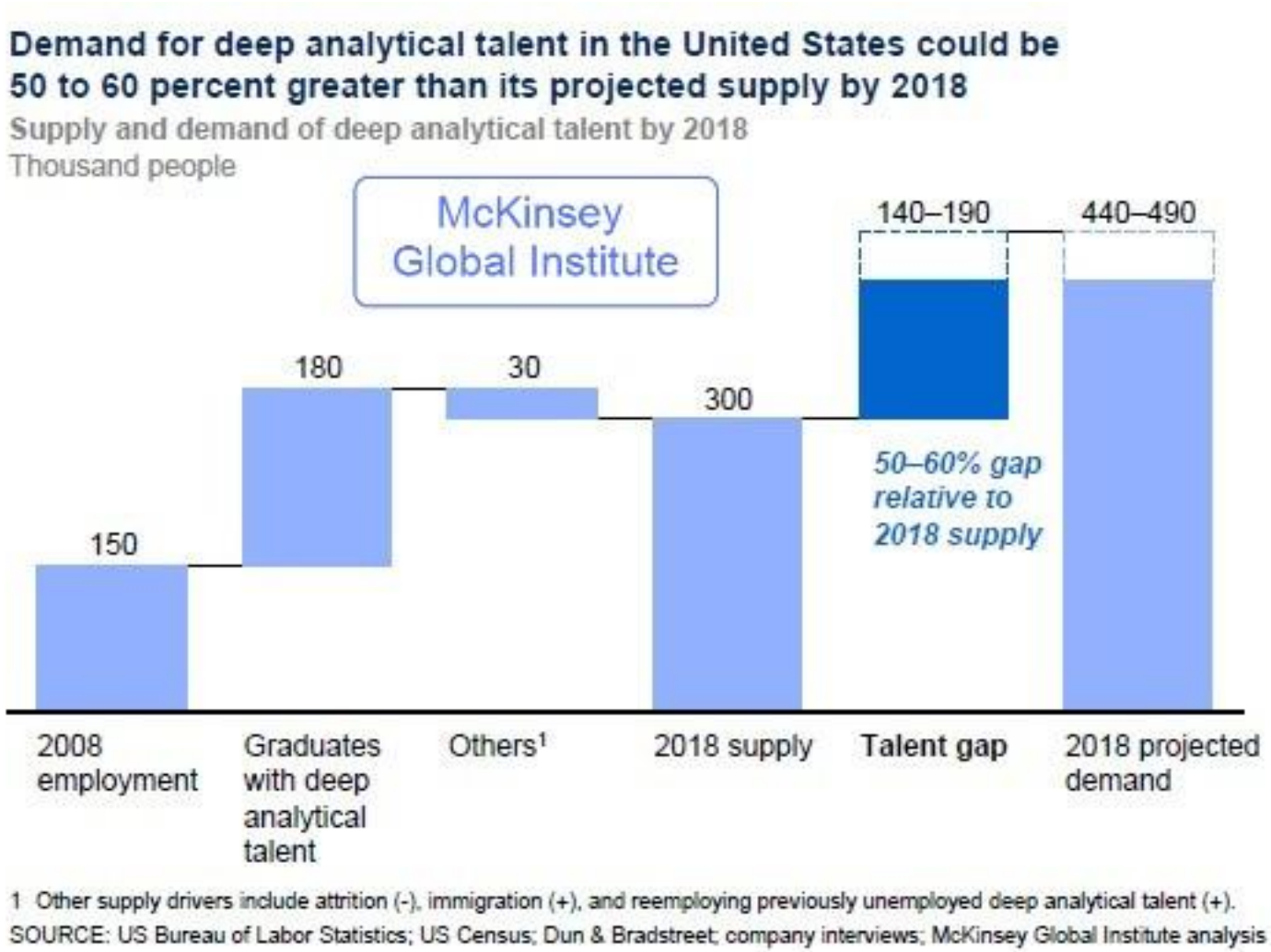
Objem dát podľa sektorov



Zhodnotenie dát podľa sektorov



Odhadovaný nedostatok odborníkov



Komunikácia a sociálne siete (1)

- Odhadom sa odošle každú minútu:
  - 204 mil. emailov
  - 1,8 mil. likov na Facebooku
  - 278 000 Tweetov
  - 200 000 fotografií (iba na Facebook)
  - 100 h videí (trvalo by asi 15 rokov ak by sme si chceli pozrieť všetky videá odoslané za 1 deň)
- 3 miliardy hovorov a 6 miliárd SMS denne len v USA

Komunikácia a sociálne siete (2)

- Využitie v marketingu, cielenej reklame, starostlivosti o zákazníka
  - Odhaduje sa, že firmy ktoré analyzujú sentiment správ na Tweetri musia spracovať 12 TB dát denne
- Ale aj na "crowdsourcing" informácií o tom ako a v akom prostredí žijeme
  - Ekológia, sociálne vedy, kultúrne dedičstvo, ...
  - Wikipédia
- A bezpečnosť
  - Odhaduje sa že americká bezpečnostná agentúra NSA sleduje asi 1,6% všetkých komunikácií na internete (30 PB denne!)

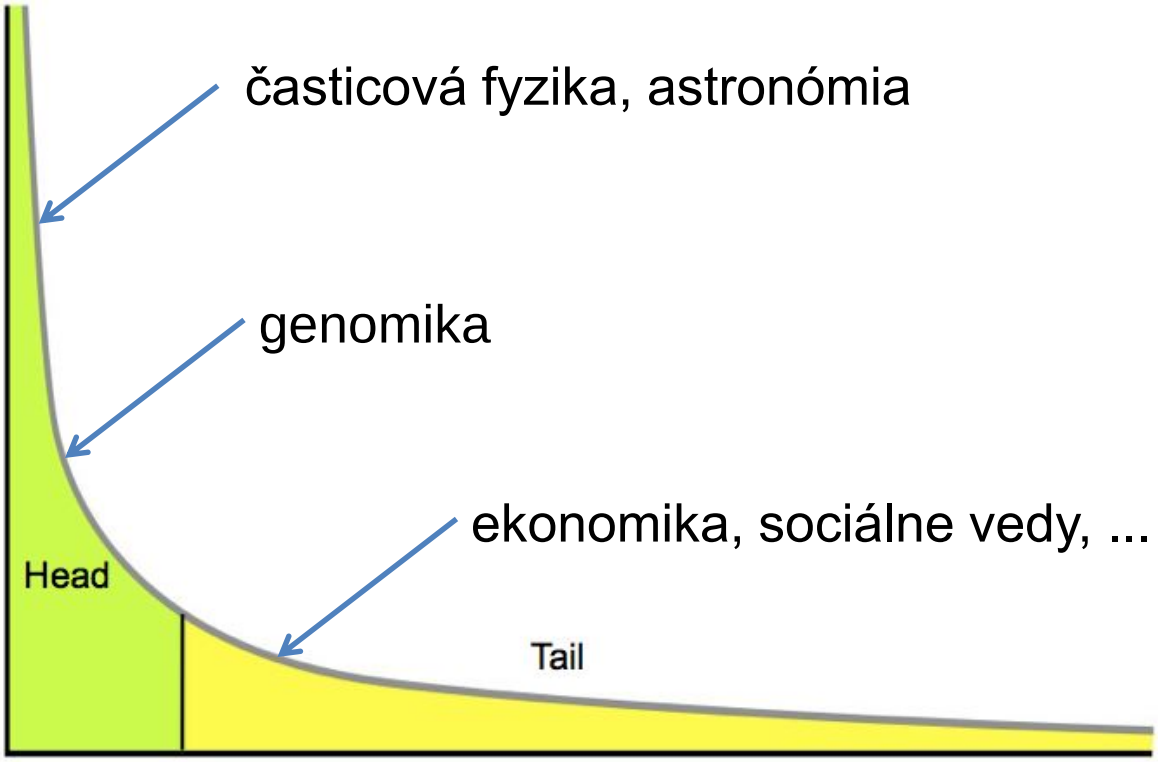
Veda a výskum (2)

- Výskum v oblasti časticovej fyziky
  - LHC urýchľovač vyprodukuje 20 mil. zrážok za 1s, iba cca 1 zrážka z 100 000 je vhodná na analýzu, v reálnom čase je potrebné vyselektovať zaujímavé udalosti
  - Aj tak LHC produkuje ročne 26 000 000 GB (25 PT) dát ročne
  - Dáta sú distribuované do 150 členských organizácií
  - Celkovo je využívaných asi 200 000 procesorových jadier

Veda a výskum (2)

- Výskum v oblasti medicíny a biológie
  - Narastá objem dát v databáza sekvencií ktoré popisujú proteíny, nukleové kyseliny (DNA / RNA) a ďalšie zložité organické zlúčeniny
  - Cieľom je určiť funkciu génov a pochopiť interakcie pri biologických procesoch ktoré ovplyvňujú naše zdravie
  - Veľká variabilita dát – sekvencie, textové dáta, klinické namerané údaje

Veda a výskum (3)



- Pravidlo 80-20: 20% výskumu generuje 80% dát (nemusí to však znamenať 80% znalostí)

Medicína a starostlivosť o zdravie (1)

- Elektronické záznamy pacientov
  - Cieľom je poskytnúť lekárom informácie potrebné pre správnu liečbu pacientov
  - V US zomrie na následky predvídateľných chýb 400 000 pacientov ročne!
  - Heterogénne dáta – texty, laboratórne vyšetrenia, obrazové dáta 2D/3D

Medicína a starostlivosť o zdravie (2)

- 80% znalostí nahromadených v medicíne a príbuzných oblastiach bolo zaznamenaných iba v textovej podobe
  - MEDLINE - digitálna knižnica pre medicínu, 18 mil. záznamov, 2000-4000 denne
- V súčasnosti narastá hlavne objem obrazových dát
  - nárast 20-40% každý rok, v roku 2012 bolo potrebné uchovať asi 1 miliardu obrázkov iba v USA

Doprava a logistika

- Len DHL prepraví 160 mil. zásielok ročne
  - 17 000 vozidiel, 250 lietadiel (714 + 2335 letov denne)
- Optimalizácia času doručenia, pokrytia, nákladov
  - crowdsourcing prepravy
- Spracovanie v reálnom čase + prediktívne metódy
- Nové typy služieb
  - zber dát o dopravnej situácii, komunikáciách a infraštruktúre

Internet vecí (1)

- Internet of Things (IoT)
- Cieľom je poprepájať rôzne zariadenia tak aby umožnili lepšie pochopiť a riadiť
  - Osobný život
  - Domácnosti
  - Mestá
  - Dopravu
  - Výrobu
  - Poľnohospodárstvo

Internet vecí (2)

- Inteligentné senzory a riadiace systémy
  - Mobilné zariadenia – telefóny, tablety, "wearables" (nositeľná elektronika)
  - Inteligentné zariadenia v domácnosti
  - RFID značkovanie
  - Rozsiahle senzorické siete
- V súčasnosti je pripojených na Internet 13 miliárd zariadení (z toho 1,6 miliárd telefónov)
- Odhaduje sa, že do roku 2020 narastie počet pripojených zariadení na 50 miliárd

Internet vecí (3)

- Pre veľa úloh je potrebné vyhodnotiť a spracovať udalosti v reálnom čase
- Zložitá správa a zabezpečenie samotnej infraštruktúry
  - Sledovanie transakcií a diagnostika sietí