AVex

Akademie věd České republiky

expertní stanovisko AV ČR | 1/2019

Big Data

- Big Data (velká data) jsou rozsáhlé a složité soubory dat, s nimiž nelze pracovat pomocí tradičních metod.
- Představují nové technologické a společenské výzvy týkající se jejich sběru, ukládání, analýzy, vyhledávání, sdílení, přenosu, vizualizace, dotazování, aktualizace, ochrany zdrojů dat a soukromí a zejména jejich interpretace a následného využití.
- Tyto směry výzkumu mají obrovský a zatím zdaleka ne plně využitý potenciál pro dosažení kvalitativních změn v mnoha oblastech, například v průmyslu, finančnictví, zdravotnictví, životním prostředí, sociálních sítích, vzdělávání, národní bezpečnosti atp.
- Interpretace a využití těchto dat souvisí s rozvojem umělé inteligence a informačních věd.
- Změny budou výraznější zejména s nástupem automatizace a robotizace, povedou k vytváření nových pracovních míst a ke ztrátě jiných, tradičních příležitostí v oblastech, kde je lidská práce nahraditelná pomocí umělé inteligence.
- Potřebám nastupujících změn je nutné přizpůsobit výuku a výchovu odborníků a přeškolování zaměstnanců a podpořit příslušné směry výzkumu.
- Pro využití potenciálu těchto změn a zamezení jejich zneužití jsou nutná nastavení vhodných legislativních opatření.

Schopnost ekonomik a států velká data sbírat, zpracovávat a správně je interpretovat poskytne do budoucna významnou informační a případně též technologickou výhodu v rámci strategického rozvoje znalostních ekonomik.

Tento AVex přináší charakteristiky velkých dat z hlediska objemu, rychlosti datových proudů a různorodosti. Zabývá se také sběrem a zpracováním velkých dat a především analýzou, interpretací a zabezpečením, jež jsou klíčovými otázkami a výzvami, které sběr a uchovávání velkých dat přinášejí.

TRILIONY GIGABYTŮ DAT

Označení Big Data je stále relativně nový pojem, nicméně koncept sběru a uchovávání velkého množství dat (informací) za účelem následné možnosti jejich analýzy je starý mnoho staletí.

Na přelomu 20. a 21. století dosáhl tento koncept určité renesance a zvýšeného zájmu, plynoucího mj. z nové filozofie ukládání dat na distribuovaných serverech namísto na lokálních počítačích a z nově se otevírajících technologických možností, jako jsou např. centralizované datové sklady (data warehouses), cloudová úložiště dat (cloud data storage) nabízená jako komerční řešení i řešení pro domácnosti a jednotlivce.

Již v roce 2013 bylo odhadováno, že globálně bylo v oběhu zhruba 4,4 trilionu GB dat, což v přepočtu na každého člověka na planetě představovalo v průměru objem dat 120 DVD filmů.

Fenomén nárůstu objemu globálně sbíraných dat souvisí s mnoha globálními výzvami, jako jsou např. bezpečnostní aspekty (prevence hacking, identifikace fake news apod.), s požadavkem na udržitelný rozvoj znalostních ekonomik, který klade důraz na přidanou hodnotu v průmyslové produkci, čímž se data samotná dostávají do popředí zájmu a jejich analýza se stává strategickým cílem nejen velkých, ale též středních a malých podniků.

Není tedy výjimkou, že ve firmách již běžně nacházíme pracovní pozice, jako je CIO, tj. Chief Information Officer, který je zodpovědný za rozvoj informačních technologií a zpracování dat.

ZRANITELNÝ „INTERNET VĚCÍ“

Velká data přinášejí řadu výzev a otevřených otázek zejména v oblasti bezpečnosti.

Jednak dochází k prudkému rozvoji zdrojů velkých dat, které jsou čím dál propojenější s naším životem. Tento trend souvisí s rozvojem tzv. „internetu věcí“ a znamená, že jednotlivé prvky jsou dynamičtější, chytřejší, navzájem propojené, ale zároveň zranitelnější. Případné narušení integrity – úmyslné či neúmyslné – může mít velmi závažné dopady.

Zároveň je potřeba věnovat otázkám bezpečnosti speciální pozornost při skladování, zpracování a přenosu velkých dat, protože zavedené bezpečnostní modely, které se používají při práci se standardními daty (např. zajištění konzistence, dostupnosti a odolnosti proti výpadkům), nejsou automaticky přenositelné do oblasti zpracování velkých dat v distribuovaných systémech. Samozřejmostí je pak využití metod moderní kryptografie pro zajištění dat před neoprávněným přístupem.

Velmi důležitou otázkou je otázka soukromí a právních a etických aspektů zpracování velkých dat. Už nyní dochází ke zpracovávání osobních dat ve veliké míře a často lze propojením zdánlivě neškodných dat dospět k velmi osobním informacím.

Do budoucna lze předpokládat, že i nadále poroste automatické zpracování zdravotních dat, osobních finančních dat, dat o pohybu (např. z mobilních telefonů a kamer s automatickou identifikací obličejů), dat o spotřebitelském chování.

Bude přibývat způsobů, jak tato data využít, a to z různých důvodů – od komerčních až po zajištění bezpečnosti. U mnoha takových způsobů bude docházet buď ke zjevnému narušování soukromí, nebo naopak k nenápadnému posouvání zavedených hranic a etických norem.

CHYBNÁ INTERPRETACE, HROZBA PRO STÁT

Velká data mohou mít svá omezení v tom, jak je možné je použít.

Data mohou vypovídat o korelacích (tj. souvislostech), ovšem nikoliv o příčinném vztahu. Korelace mohou být extrémně užitečné pro predikce, resp. identifikaci dříve nepozorovaného chování, pokud jsou spolehlivé. Mohou však také být zavádějící, pakliže souvislost naznačuje jiná proměnná korelovaná s oběma proměnnými, u kterých korelace studujeme.

Nesprávná interpretace velkých dat může představovat zásadní bezpečnostní hrozby pro stát. Značná ekonomická rizika jsou spojená s nesprávnou analýzou vývoje komoditních a akciových trhů. Velká data však souvisejí také s předpovědí počasí, s environmentálním modelováním a předpovědí živelných či průmyslových katastrof, např. včetně konceptu jaderné energetické bezpečnosti.

METADATA JAKO VÝZVA

Velká data je možné pořídit v obrovském množství různých datových formátů a ze stále rostoucího počtu datových zdrojů.

Zahrnují obrazovou informaci (např. fMRI skeny), záznamy kamer, zvukové záznamy, záznamy o uživatelské aktivitě na internetu, data z počítačových simulací apod. Klíčem ke zpracování podobných dat jsou metadata, tj. data o datech. Tvorba metadat pro velká data může být velkou výzvou, může být obtížné zachytit všechny potřebné detaily týkající se těchto dat.

Obrovskou výzvou je i zpracování velkých dat. Jako u každých dat je nutná jejich extrakce (výběr dat, o která se aktuálně zajímáme, a to ve strukturované podobě), čištění (detekce a oprava či odstranění poškozených nebo nesprávných záznamů), standardizace dat (formátování dat tak, aby bylo možné je přenášet mezi různými systémy) a napojování souvisejících záznamů z různých zdrojů.

CO OVLIVŇUJE NÁRŮST VELKÝCH DAT

Nové zdroje dat:

- digitalizace procesů a služeb (internetové bankovnictví, e-mail, elektronická evidence tržeb, elektronický zdravotní záznam o pacientovi /zatím spíše mimo ČR/),
- automatické generování dat (webové servery zaznamenávající návštěvy na stránkách),
- nízká pořizovací cena senzorů v letadlech, autech, budovách, životním prostředí,
- výroba nových (často miniaturizovaných) zařízení a přístrojů, které zaznamenávají a vysílají data (GPS signál na mobilních telefonech nebo chytré popelnice, které dávají informaci o zaplnění).

Rozšířené výpočetní schopnosti umožňující rostoucí zájem o velká data:

- zlepšené vlastnosti ukládání dat s větší kompresí, resp. hustotou zápisu, a to za nižší cenu,
- větší výpočetní síla, rychlejší a komplexnější numerické výpočty,
- cloudové výpočty (vzdálený přístup ke sdíleným výpočetním kapacitám pomocí přístroje připojeného na počítačovou síť), levnější přístup k ukládání dat, výpočtům, softwaru a dalším službám,
- pokroky ve statistických a výpočetních numerických technikách, které je možné využít k analýze a extrakci požadované informace z dat,
- vývoj nových prostředků, jako je Apache Hadoop, který umožňuje simultánní distribuované zpracování velkých dat na různých počítačových klastrech a rozšíření schopností existujícího software (např. Excel společnosti Microsoft).

VELKÁ DATA A UMĚLÁ INTELIGENCE

Samotný sběr velkých dat a zabezpečení přístupu k nim není finálním cílem. Tím se stává teprve jejich využití v různých oblastech vědy i společenského života a porozumění a řešení komplexních problémů, zejména v interdisciplinárních oblastech jako např. v inženýrství, aplikované matematice a výpočetní statistice, medicíně, výpočetní biologii, péči o zdraví, v sociálních sítích, finančnictví, obchodě, řízení, výchově, predikci počasí, dopravě a telekomunikacích.

To vše se děje za pomocí umělé inteligence.

Využití umělé inteligence (artificial intelligence, AI) je způsob, jak získat z velkého množství dat znalosti nutné k řešení různých problémů. Umělá inteligence umožňuje delegaci náročných problémů vyhledávání závislostí v datech, opakování jistých vzorů, učení, predikce atp. na počítače. Např. více než polovina burzovních operací se děje za pomoci systémů umělé inteligence.

Podobně jako je využití velkých dat nemožné bez umělé inteligence, tak se ani umělá inteligence nemůže rozvíjet bez velkých dat. To, co byly donedávna dvě disciplíny informatiky, konverguje k jediné disciplíně, které se někdy říká Big AI. Tato konvergence urychluje inovace, jako jsou objev a vývoj nových léčiv, bezpečná samořiditelná auta, alternativní energie atd., a naopak, tlak na řešení těchto problémů urychluje tuto konvergenci.

Je nutné zdůraznit, že investice do informatického výzkumu prováděného na úrovni Akademie věd ČR, vysokých škol a veřejných výzkumných institucí jsou strategickou potřebou dalšího udržitelného rozvoje národní ekonomiky v evropském i celosvětovém kontextu.

ČESKÉ ÚLOŽIŠTĚ CESNET: VÍC NEŽ 21 PETABYTŮ

V České republice se z pohledu infrastruktury nabízí CESNET, z. s. p. o., jehož oddělení datových úložišť zajišťuje provoz a rozvoj národní infrastruktury pro ukládání dat pro vědu a výzkum. Celková hrubá kapacita úložišť přesahuje 21 PB. Tato služba je však nabízena s omezením velikosti úložiště 2 TB na organizaci, po domluvě je možné navýšit.

Prostředí **MetaCentra CESNET**u umožňuje využití zapojených výpočetních a datových zdrojů pro řešení velmi náročných výpočetních úloh, jejichž zvládnutí přesahuje možnosti samostatného pracoviště v ČR. Služba je dostupná i uživatelům bez nutnosti vkladu vlastních zdrojů. Nabízené služby zahrnují gridové výpočty, HPC cloudy, platformy (např. Hadoop), 14 000 výpočetních jader a aplikační licence.

CERIT-SC (CERIT – Scientific Cloud) reprezentuje nejvýznamnější regionální Tier-2 systém v ČR. Jeho výpočtové kapacity jsou částečně včleněny do Metacentra CESNETu. Poskytuje flexibilní úložné a výpočetní kapacity a související služby, včetně podpory jejich experimentálního využití. Současně centrum provádí výzkum a vývoj v oblasti flexibilních e-infrastruktur.

Další možnosti s ohledem na náročné výpočty poskytuje infrastruktura Národního superpočítačového centra IT4Innovations na VŠB TU Ostrava. Pronájem výpočetního času umožňuje Ředitelství IT4Innovations svým rozhodnutím. Výpočetní čas na superpočítačích Anselm a Salomon je dále rozdělován prostřednictvím veřejných grantových soutěží.

ZÁKLADNÍ CHARAKTERISTIKY VELKÝCH DAT

Koncept velkých dat definoval počátkem roku 2001 průmyslový analytik Doug Laney. Ve svém článku s názvem Application Delivery Strategies artikuloval tři základní charakteristiky velkých dat:

- **Volume (objem dat)**
  Ke sběru dat dochází z mnoha různých zdrojů paralelně a simultánně.

  V mnoha případech se jedná o obrovská množství dat, např. automaticky sbíraná data z průmyslových senzorů, satelitů, kamer apod., která nebylo dříve možno ani uchovávat ani analyzovat.

  Velké objemy dat souvisejí s indexováním souborů pro rychlé vyhledávání na internetu (Google, Yahoo, Bing, Seznam, apod.) a s počítačovou, resp. kybernetickou bezpečností. Velikost datových souborů se může pohybovat řádově v terabytech (1 TB = 1 tis. GB) až petabytech (1 PB = 1 mil. GB), což jsou již velikosti, se kterými se běžný osobní počítač nebo datový server nedokáže vypořádat.

- **Velocity (rychlost datových proudů)**
  Datové informace přicházejí s vysokou intenzitou a rychlostí a je třeba na ně rychle a adekvátně reagovat. Tyto potřeby vynikají zvláště v kontextu bezpečného sdílení dat na internetu (Cyber Security) nebo analýzy dat ze sociálních sítí a potřeby rychlého vyhodnocení případných rizik a reakcí na ně.

  To souvisí též s národní bezpečnostní strategií v oblasti informačních technologií, ochranou citlivých dat, ochranou před útoky na významná datová úložiště a datové servery, např. s daty o transakcích na debetních a kreditních kartách apod. Nově vyvíjené technologie mají za cíl reagovat na citlivá data téměř v reálném čase.

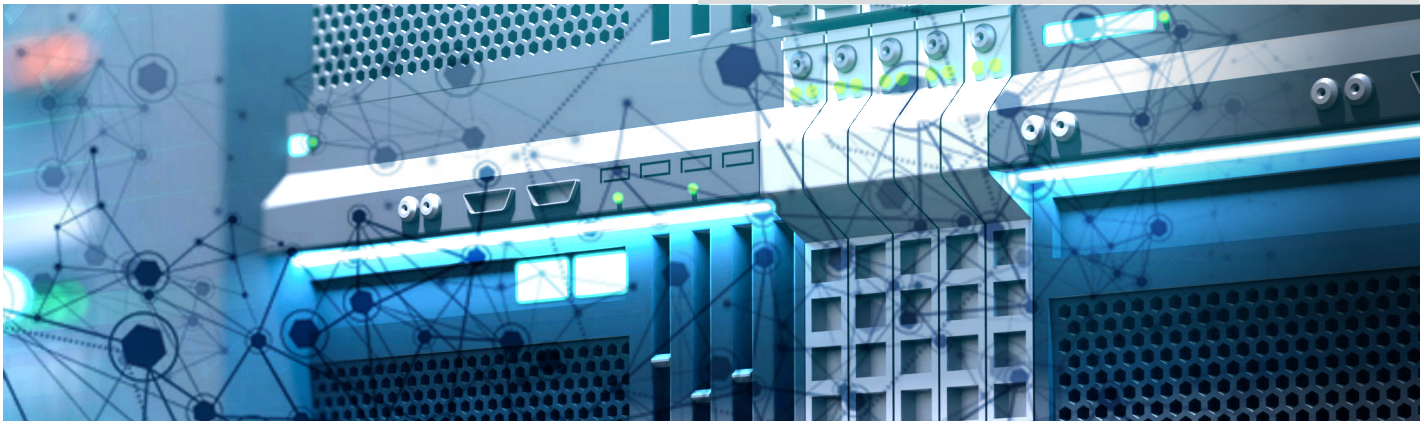
- **Variety (různorodost dat)**
  Dochází k simultánnímu sběru dat v různých formátech, od volných textů a nestrukturovaných dat, e-mailů, videí, audií, dat z akciových trhů a finančních transakcí až k numerickým datům a strukturovaným databázím.

  Jak vyplývá z publikace Application Delivery Strategies, velká data jsou symptomem informační exploze ve třech výše uvedených dimenzích.

Ke třem výše zmíněným charakteristikám velkých dat, které se označují také jako 3V, můžeme ještě přidat další dvě:

- **Variabilita**
  Lze zmínit vysokou variabilitu dat, např. v rámci dlouhodobého sledování a vyhodnocování. Variabilita může souviset s globálními či lokálními událostmi, kdy dochází k extrémnímu nárůstu datových vstupů, které není snadné dobře zvládat.

- **Komplexita**
  Problém komplexity odráží potřebu spojit data z heterogenních zdrojů. Výzvy spočívají ve správném napojení datových struktur, identifikaci korespondujících jednotek, čištění dat zahrnujících různou chybovost v závislosti na zdroji a transformaci dat při jejich přenosu mezi nekompatibilními datovými systémy. Tato vlastnost velkých dat je v angličtině též nazývána Veracity a týká se především možné chybovosti, resp. nespolehlivosti dílčích datových vstupů.



11 000 SERVERŮ V CERNU

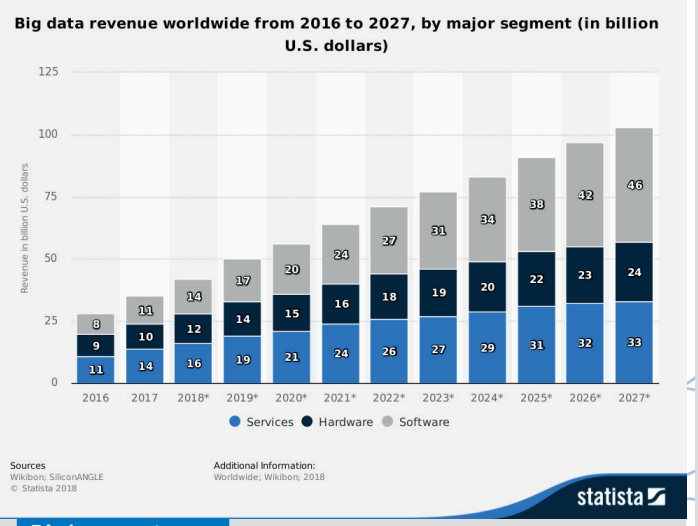
Jako vhodný příklad ze zahraničí uveďme ženevský CERN, ve kterém se provádějí fyzikální pokusy na lineárních urychlovačích (Large Hadron Collider, LHC).

Zhruba 600milionkrát za sekundu dochází ke kolizi částic, které se poměrně komplexním způsobem rozpadají do ještě většího počtu částic. Elektronické obvody zaznamenávají průchod každé částice detektorem jako časovou řadu elektronických signálů. Ročně zpracovávají zdejší fyzikové zhruba 30 PT dat.

Ani CERN nemá dostatek výpočetních kapacit ke zpracování těchto dat. Od roku 2002 tedy přešel na gridové výpočty, aby sdílel další světově dostupné výpočetní kapacity. Distribuovaná víceúrovňová výpočetní infrastruktura umožňuje komunitě 8000 fyziků přístup k datům LHC téměř v reálném čase. Gridová technologie je postavena na technologii World Wide Web (WWW), kterou vynalezl CERN v roce 1989. Tyto technologie přestavují obrovské energetické požadavky. V roce 2013 navýšil CERN energetickou kapacitu z 2,9 MW na 3,5 MW, aby bylo možné nainstalovat více počítačů. Datové centrum CERNu zpracovává 1 PT dat každý den, ekvivalent cca 210 000 DVD.

V centru je využíváno 11 000 serverů se 100 000 výpočetními jádry.

Jeden z možných scénářů budoucího vývoje: příjmy z velkých dat do roku 2027



Zdroj: *www.statista.com*

Graf ukazuje výhled celosvětových příjmů z velkých dat podle hlavních sektorů – údaje v miliardách dolarů. Příjmy v roce 2027 by mohly přesáhnout 100 miliard dolarů. Největší podíl (téměř polovina budoucích možných příjmů) by přitom měl připadnout na sektor software.

PROJEKT HADOOP UMOŽŇUJE DISTRIBUOVANÉ UKLÁDÁNÍ A ANALÝZU DAT

Na konci 20. a začátkem 21. století došlo k výraznému rozšíření webu ve smyslu globálního nárůstu počtu webových stránek. Z tohoto důvodu byly navrženy vyhledávače a vyhledávací indexy, které umožňují vyhledat relevantní informace pomocí textového dotazu. Zpočátku na vyhledávací dotazy odpovídali často lidé, ale postupem času se stala zásadní potřebou automatizace, neboť web se již rozrostl do obrovského množství webových stránek.

První vyhledávače vznikly jako univerzitní projekty, ale později již vznikaly nové společnosti jako Yahoo, AltaVista apod. Jedním z takových projektů byl vyhledávač Nutch, který měl dosáhnout rychlejšího webového vyhledávání založeného na distribuovaném ukládání dat a distribuovaných výpočtech prováděných simultánně na několika různých počítačích. Tehdy také započal nový projekt webového vyhledávače Google, vycházející ze stejného konceptu distribuovaného ukládání a automatické distribuované analýzy dat, aby výsledků vyhledávání na internetu mohlo být dosaženo v kratším čase. Část projektu Nutch spojená se zpracováním dat dala vznik projektu Hadoop.

V roce 2008 uvolnilo Yahoo Hadoop jako „open-source“, tedy volně dostupný projekt, který dnes zastřešuje a řídí nezisková nadace Apache Software Foundation (ASF), což je globální komunita softwarových vývojářů.

Čím je projekt Hadoop významný v kontextu Big Data?

- Objem dat. Hadoop umožňuje rychle ukládat a zpracovávat obrovské objemy dat. To je důležité zejména v kontextu rychle se rozvíjejících sociálních sítí.
- Výpočetní síla. Distribuovaný výpočetní model Hadoop zpracovává velká data rychleji. Čím více výpočetních uzlů je k dispozici, tím větší výpočetní sílu můžeme použít.
- Chybová tolerance. Data i výpočty a zpracování dat jsou chráněny proti selhání hardwaru. Jestliže selže nějaký uzel (počítač, server), úkoly jsou automaticky přesměrovány na jiné uzly tak, aby distribuované výpočty neselhaly. Automaticky jsou ukládány mnohonásobné kopie zdrojových dat.
- Flexibilita. Na rozdíl od tradičních distribuovaných relačních databází již není třeba předzpracovat data před jejich uložením. Je možné ukládat tolik dat, kolik si přejeme, a až později se rozhodnout, jak je použít, resp. analyzovat. Data mohou mít různé formáty, od textů přes obrázky až po videa.
- Nízká cena. Otevřený software je možné používat bezplatně. Ten je využíván na zcela obyčejných počítačích s nízkou pořizovací cenou (tzv. commodity hardware) k uložení velkého objemu dat.
- Rozšiřitelnost. Je snadné rozšířit stávající systém přidáním dalších uzlů, aby bylo možné zpracovat více dat.

Přehled použité literatury: www.avcr.cz/avex.

AVex je nezávislé a nestranné expertní stanovisko, které Akademie věd České republiky připravuje pro legislativní potřeby zákonodárců Poslanecké sněmovny a Senátu Parlamentu České republiky.

Připravila Akademie věd ČR, odborným garantem je Ústav informatiky AV ČR.

Odpovědná redaktorka: Markéta Růžičková, **e-mail:** avex@kav.cas.cz, **www.avcr.cz/avex.**

Kontaktní osoba: prof. Ing. Emil Pelikán, CSc., e-mail: pelikan@cs.cas.cz.