Big data

Nové způsoby zpracování a analýzy velkých objemů dat

Ondřej Dolák

„Big data“ je fráze, která se poprvé objevila v oblasti high performance computingu (HPC). Stále častěji se tento pojem začal objevovat v prezentacích HPC dodavatelů v souvislosti s vizualizačními platformami, cloudovými řešeními a úložišti. Co přesně však tato fráze znamená?

Pokud si člověk přečte deset prezentací dodavatelů technologií, přijde na zhruba patnáct různých definic. Každá z nich má podle očekávání tendenci podporovat produkty či služby toho daného dodavatele, nicméně většina vyhovuje definici, se kterou přišla poradenská firma Gartner: big data je termín aplikovaný na soubory dat, jejichž velikost je mimo schopnosti zachycovat, spravovat a zpracovávat data běžně používanými softwarovými nástroji v rozumném čase.

Pojem „velikost“ dat je chápán nejen z hlediska objemu dat měřeného giga-, tera- či petabyty, ale i z hlediska rychlosti jejich tvorby a přenosu a z hlediska různorodosti jejich typů. Jako příklad je často citováno množství údajů o počasí, které získává každý den Národní úřad pro oceán a atmosféru (NOAA) nebo NASA. I komerční sektor má své premianty, jako například energetické, telekomunikační nebo farmaceutické společnosti shromažďující obrovská množství dat. Velké organizace čelí stálé potřebě udržovat rozsáhlé soubory strukturovaných i nestrukturovaných dat. V souladu s vládními nařízeními a s postupnou digitalizací narůstá objem archivovaných elektronických dokumentů, e-mailových zpráv a dalších záznamů o elektronické komunikaci.

Klasický způsob využití dat z datového skladu

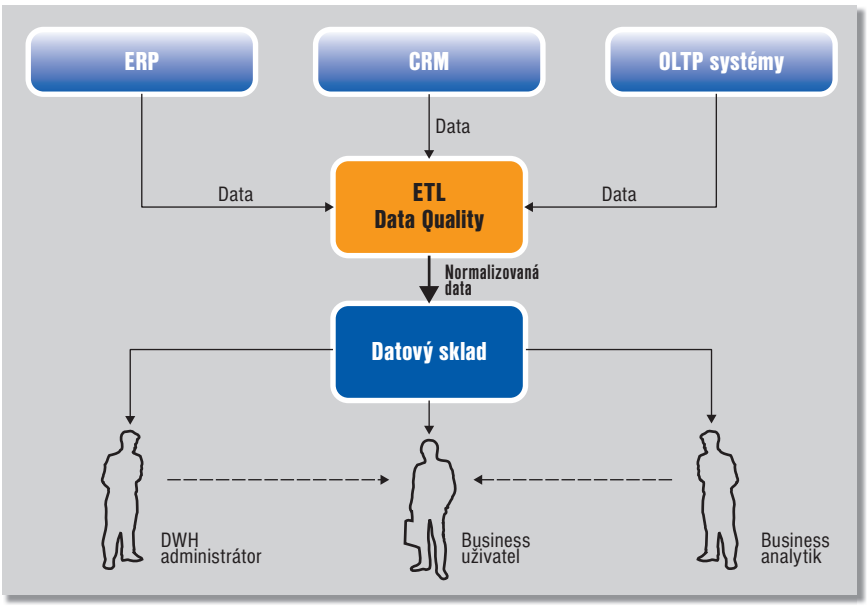
Až do nedávné doby bylo zpracování dat pro analytické účely poměrně statickou úlohou. Konkrétně podniky produkují zejména strukturovaná data ze stabilních datových modelů prostřednictvím podnikových aplikací, jako je CRM, ERP a finanční aplikace. Pomocí ETL nástrojů se data z těchto systémů přenáší do pracovní oblasti, kde se kontroluje kvalita údajů a provádí se jejich normalizace, finálně se pak ukládají do datového skladu. Tento proces obvykle běží v pravidelném cyklu – zpravidla na denní nebo týdenní bázi.

Správci datového skladu pak vytvářejí reporty, které běží nad daty uloženými ve skladu. Datoví analytici používají analytické nástroje pro provádění složitých výpočtů nad daty ze skladu, anebo z důvodu omezení velikosti častěji nad vzorky dat z data martu. Běžným uživatelům je většinou povoleno použití BI nástrojů pro základní vizualizaci dat a omezené analytické výpočty. Objemy dat z tradičních datových skladů zřídka překračují několik terabytů.

Změna podstaty big data

Nástup webu, mobilních zařízení a dalších technologií zapříčinil zásadní změnu charakteru dat a způsobu jejich využití. Již nejsou centralizovaná, vysoce strukturovaná a snadno zvládnutelná, ale více než dříve jsou volně strukturovaná (pokud mají vůbec nějakou strukturu), vysoce distribuovaná a mají vzrůstající objem. Často se v této souvislosti hovoří o trojrozměrnosti velikosti a růstu dat (zkráceně také jako 3V):

- **objem** (volume) – množství dat vznikajících v rámci provozu firem roste exponenciálně každý rok,
- **typ** (variety) – různorodost typů dat vzrůstá, například nestrukturované textové soubory, semi-strukturovaná data (XML), data o geografické poloze, data z logů,



- **rychlost** (velocity) – rychlost s jakou data vznikají a potřeba jejich analýzy v reálném čase vzrůstá díky pokračující digitalizaci většiny transakcí, mobilním zařízením a vzrůstajícímu počtu internetových uživatelů.

Big data mají odlišné vlastnosti, které je odlišují od „tradičních" firemních dat. Tradiční datové sklady a nástroje pro správu dat nejsou připraveny na zpracování a analýzy velkých objemů dat ve velmi krátkém čase (někdy real-time) nebo nákladově efektivním způsobem. Proto je třeba hledat nové způsoby zpracování a analýzy velkých objemů dat.

Nové přístupy k analýze a zpracování big data

Jedním z takovýchto nových nástrojů je Hadoop. Hadoop je open source framework pro zpracování, ukládání a analýzu velkého množství distribuovaných, nestrukturovaných dat. Původně byl vytvořen ve společnosti Yahoo!, jako inspirace byla použita MapReduce, uživatelsky definovaná funkce vyvinutá společností Google pro indexování webu. Hadoop je stavěn pro zvládání petabytů a exabytů dat distribuovaných přes více uzlů současně.

MapReduce je výpočetní vrstva v rámci Hadoopu. Úlohy MapReduce přistupují k datům, která jsou distribuována na webu nebo v datových centrech, rozdělují je do více replikovaných dílů a jejich zpracování pošlou na jednotlivé uzly. Dotazy a další zpracování pak probíhá v každém uzlu paralelně. Výsledky jsou agregovány a ukládány do úložné vrstvy, jako například Hadoop Distributed File System (HDFS). Odtud jsou data načtena do jednoho z několika analytických prostředí pro analýzu. Ekosystém Hadoop se dále skládá z dalších vzájemně se doplňujících projektů. Mezi ně, kromě výše uvedených HDFS a MapReduce, patří NoSQL datová úložiště, jako Cassandra nebo HBase.

Hlavní výhodou Hadoopu je, že umožňuje analyzovat úplné datové soubory údajů, včetně nestrukturovaných a částečně strukturovaných dat, a to z hlediska nákladů i času efektivním způsobem. Mezi nevýhody Hadoopu patří částečná nezralost a hektický vývoj. Kromě toho, zavádění a řízení Hadoop clusterů a provádění pokročilé analýzy na velké objemy dat vyžaduje značné odborné znalosti. Pro firmy je takový model vesměs nepřijatelný, a proto v rámci ekosystému vznikla řada firem, které staví komerční řešení na bázi Hadoopu tak, aby se nasazení a správa technologie stala praktickou realitou tradičního enterprise odvětví.

Trh řešení pro big data

Řešit problémy s big data znamená zasahovat do mnoha komponent IT architektury od hardwaru po optimalizaci vzorkování dat. Proto i trh s řešeními orientovanými na big data je velmi pestrý. Jednoduchý přehled obsahuje tabulka.

Hardware

V rámci big data je velké zaměření na hardwarové konsolidace. Většina velkých dodavatelů nabízí integrovaná řešení včetně specializovaného hardwaru s důrazem na výkonnost při snížení celkových nákladů na správu a provoz. Na druhou stranu je zde i druhá skupina dodavatelů, kteří upřednostňují hardwarovou nezávislost, nebo přímo podporují běh svých systémů na komoditním hardwaru.

Big data distribuce

Množství dodavatelů řešení specializovaných na big data se rychle zvětšuje. Řada z dodavatelů vyvinula své vlastní Hadoop distribuce s různým stupněm úprav. Patří mezi ně jak řada firem vzniklých právě na základě potřeby řešit „velká data", tak dnes i většina velkých hráčů na IT trhu.

Data management

V rámci data managementu hrají prim především noSQL databáze jako prostředek pro obsluhu požadavků na čtení a zápis velkých objemů dat. Z pohledu integrace je zde patrná snaha o napojení big data technologií do stávajících nástrojů a zároveň jejich těsnou integraci se stávajícími technologiemi, především RDBMS.

Analýza a vizualizace

Obecně platí, že čím větší vzorek dat, tím přesnější výsledek analýzy. To samozřejmě zvyšuje tlak na zvětšování objemu analyzovaných dat. Dodavatelé analytických nástrojů se snaží vylepšit své produkty, aby velké objemy dat zvládly a ulehčily uživatelům od vymýšlení a tvorby různých „náhradních řešení". Trendem je využívat vlastní, vestavěné databáze jako součást analytických nástrojů, maximum dat nahrávat do operační paměti a pracovat s nimi tzv. in-memory, využívat nové principy uložení dat v databázi (sloupcově orientované databáze) či využívat masivně paralelní systémy. Velký rozvoj se očekává v následujících letech v oblasti zobrazování výsledků analýz (vizualizace dat). Bude se zvyšovat množství zobrazitelných bodů (hodnot), očekávat lze vylepšování animačních schopností analytických nástrojů.

Big data rovná se budoucí příležitost

Více než osmdesát procent všech dat v podniku má nestrukturovanou formu. Nejtěžší je nalézt v nich informace podstatné pro daný byznys. Informační pracovníci tráví dnes téměř čtvrtinu svého času právě vyhledáváním informací, přičemž doba získání těchto informací může mít vliv nejen na vnitřní produktivitu firmy, ale také například na spokojenost a loajalitu zákazníků. Big data přinášejí nový pohled i na samotné projekty datových skladů. Tradiční projekty budování datových skladů trvají i roky, od formulace zadání až po provedení samotných změn na základě výsledků analýz mnohdy uplyne dlouhá doba, což může způsobit značné finanční ztráty. Při použití big data prostředků lze tyto projekty zásadně urychlit a současně dospět k přesnějším výsledkům. Rychlost dosažení přínosů (time-to-value) bude klíčovým ukazatelem úspěšnosti těchto projektů. To bude vyžadovat změnu i na straně dodavatelů, kteří budou akceptovat krátké, intenzivní projekty. ■

	Hardware	Big data distribuce	Data management	Analýza a vizualizace
Požadavky	úložiště servery sítě	komunitní Hadoop enterprise Hadoop non-Hadoop frameworky	noSQL databáze datové integrace datová kvalita a řízení	analytické platformy vizualizace dat BI nástroje
Dodavatelé	DELL HP Oracle Cisco IBM	Cloudera IBM EMC Greenplum Teradata Oracle Kognitio ParAccel SGI	IBM EMC Greenplum Oracle SAP ParAccel Informatica	SAS Oracle Tableau IBM Datameer

Autor působí jako senior consultant společnosti Sophia Solutions.