Na velikosti záleží

Představují big data slibný technologický směr, nebo se začnou pomalu vytrácet, aby mohla být nahrazena jinými trendy? Na otázky související s touto problematikou odpovídají specialisté na big data společnosti Sophia Solutions, Petr Švarc a Pavel Pilař.

Text / **Richard Jan Voigts**

Stala se analýza a zpracování tzv. big dat reálnou IT disciplínou, nebo jde jen o další buzzword?

Pilař: Zpracování big dat, tj. velkých objemů dat, je v současné době reálnou aktivitou. O big datech se ale hovoří i v situacích, kde to nedává úplný smysl a stejná analýza je řešitelná i prostředky klasických databázových technologií. Respektive přidaná hodnota big data řešení není až tak velká. Z toho tedy občas plyne pocit, že big data jsou buzzword, který se dobře prodává.

Co tedy znamená řešení pro big data a řešení databázové?

Pilař: Základním rozdílem mezi těmito dvěma přístupy je, že v řešení big data můžeme ukládat a následně analyzovat data v té podobě, v jaké jsou na vstupu. Oproti databázovému řešení, kdy je třeba mít předem jasnou představu, jaký typ analýz budeme chtít provádět.

To tedy znamená, že big data mohou být i nestrukturovaná a v databázi strukturovaná?

Pilař: Ano, databázová data jsou vždy strukturovaná, zatímco big data mohou být zcela nestrukturovaná. Hovoříme zde o data lakes (jezera dat), což je další termín používaný v této oblasti. Data ukládáme tak, jak jsou pořizována. Po nějaké době fungování big data řešení se můžeme zpětně vrátit k analýze těchto surových dat, která jsou k dispozici.

Švarc: Ještě bych odpověď doplnil ohledně toho, zda jsou big data buzzword, či nikoliv. S příchodem big dat se začalo hovořit o nových formách datové analýzy. Díky tomuto přístupu je možné ve větší míře provádět prediktivní analýzu a složitější statistické modely nad vstupními daty. Přijde mi, že většina lidí si myslí, že aby tyto typy analýz mohli dělat, potřebují k tomu (technologie pro) big data. Přitom tomu tak není. Prediktivní analýzu a statistické modelování lze provádět i nad těmi daty, která firmy už většinou mají. Tyto dvě věci dnes pouze tak nějak přišly dohromady, proto si je lidé spojují. Jsem však přesvědčen, že spousta zákazníků zjistí, že pro to, co chce analyzovat, nepotřebují technologie pro big data, a navíc zjistí, že big data v podstatě vůbec nemají.

Kdy jsou to ona velká data, a čím se tedy vlastně odlišují od těch ostatních dat?

Švarc: Obecná definice hovoří o tom, že big data jsou taková data, která není možné zpracovat standardními technologiemi, jež byly vyvinuty a k dispozici v posledních desetiletích. To znamená různé relační databáze. Další definice zase říká, že big data mají tři V, a sice velocity, variety a volume (rychlost, rozličnost a objem). Existují však i use cases, kdy data nemusí být nestrukturovaná, ale je jich tolik a „letí" na vás na vstupu tak rychle, že je nedokážete zpracovávat pomocí relačních databází, a v tu chvíli se už začíná o big datech také hovořit.

Setkáváte se i s jinou interpretací termínu big data? V čem se podle vás lidé nejčastěji mýlí?

Švarc: Dost lidí chce pracovat s big daty jen proto, že se o nich dnes mluví.

Může tedy jít o stav, že mají větší objem dat v databázi, a myslí si proto, že mají big data?

Pilař: Myslím, že právě to bývá ne zcela správná interpretace. Máme zákazníky, kteří mají miliardy řádek a terabajtové objemy dat v klasických relačních databázích. Jedná se ale o strukturovaná data a nemá smysl hovořit zde o big datech. Ne každá velká data jsou tedy big data. Petr Švarc zmínil otázku složitějších, prediktivních analýz a vazbu na big data. Pro tyto typy analýz v podstatě vždy potřebujeme strukturovaná data. Ne vždy však dopředu víme, která data budeme chtít analyticky zpracovávat. Zde dávají smysl bigdatová úložiště.

Zákazník často začne s tématem big data. Když lépe poznáme jeho současnou situaci a potřebu, často se ukáže, že klasické technologie splní požadavky lépe. S nadsázkou lze tedy říci, že ne každý to má tak velké, jak si zprvu myslel.

Jaké technologie se ke zpracování big dat využívají?

Švarc: Základem je cluster Hadoop, kde je úplně základní komponentou HDFS (Hadoop Distributed File System). Jde o distribuované úložiště dat, které je možné lineárně rozšiřovat, škálovat. Tím je položen základ k tomu, abychom big data vůbec mohli začít sbírat. Máme tedy distribuované úložiště, a pokud přestává stačit, jednoduše ho rozšíříme tak, že do clusteru přidáme další servery. Jde také o jeden z příkladů, kde klasická relační databáze takto nemůže fungovat, protože když se u ní dostanete na nějaký velikostní limit, už ji dále při zachování efektivity nerozšíříte. Další věc, kterou HDFS nabízí, je, že zároveň představuje i výpočetní framework. Začalo se s joby (aplikacemi) Map Reduce, teď se pro spouštění aplikací používá Yarn. Tyto výpočetní frameworky využívají distribuovaného úložiště (clusteru) k tomu, že zde spouštějí výpočty na všech serverech najednou. Výpočet je proto teoreticky lineárně škálovatelný, když přestane hardware stačit, Hadoop se rozšíří o další servery.

Tímto distribuovaným úložištěm může být třeba internet?

Švarc: Obecně může být distribuovaným úložištěm cokoliv, avšak v enterprise řešení například internet nedává smysl, především

Petr Švarc, specialista na big data, senior consultant, Sophia Solutions

Pavel Pilař, team leader a managing consultant pro reporting a analýzy, Sophia Solutions

Foto: Radek Vebe

z důvodu ochrany dat. Enterprise řešení potřebuje, aby byla nad daty governance (správa, kontrola), aby se vědělo, kdo k nim má přístup apod.

Z hlediska enterprise řešení to tedy může být něco od relační databáze přes archiv dokumentů v textové podobě až po zvukové záznamy a videozáznamy?

Pilař: To, co jste vyjmenoval, jsou řekněme různé typy dat, které můžeme do datového úložiště uložit.

A je to za zdí, mřížemi a zámkem firmy…

Švarc: Přesně tak, je to v chráněné interní podnikové síti, která je onou zdí.

Co z toho umí dodat Sophia Solutions? Můžete jmenovat řešení, které používáte pro zpracování a analýzu velkých dat?

Švarc: Jsme partnery společnosti MapR, což je americká firma, která nabízí hadoopovskou distribuci stejného jména. V podstatě je to sada nástrojů, které se používají pro zpracování big dat, počínaje HDFS, Yarn přes HBase, Hive až po analytické nástroje, jako je Spark, Drill apod. Když jsme se před pár lety začali o big data zajímat, vybrali jsme si ji z velké trojice MapR, Hortonworks a Cloudera, což jsou další distribuce. MapR distribuce obsahuje rozšířené a vylepšené softwarové komponenty Hadoopu (MAPR-FS místo HDFS, MAPR-DB místo HBase apod.), přidala vlastní know-how, a díky tomu jejich cluster dokáže být rychlejší a hardwarově méně náročný než ty dvě další zmíněné, se kterými ovšem v případě požadavku klienta umíme pracovat též.

Máte zde nějaké vlastní unikátní kompetence?

Pilař: Za naši unikátní kompetenci považuji to, že máme nejen technologické odborníky, ale i specialisty data science, kteří s big daty dokážou pracovat analyticky. To znamená, že nad úrovní úložiště dat, o kterém hovořil Petr Švarc, pracují s programovacími jazyky R a Python. Díky tomu lze přímo nad big data úložištěm vytvářet pokročilé analýzy, jako jsou segmentace či prediktivní modely.

Čím se řídí výběr nástrojů pro big data?

Švarc: V první řadě tím, co je dostupné pro Hadoop, protože nad tímto základním distribuovaným systémem vznikají softwarové projekty, které umožňují data distribuovaně zpracovávat. Jde o novou technologii a klasické nástroje pro analýzu a reporting neumějí využít potenciálu tohoto clusteru. Proto vznikají nové nástroje a typickým z nich je právě Spark, což je nástroj pro datové vědce, který se ovládá přes Python, Javu nebo Scalu. Pomocí něj můžete analyzovat nestrukturovaná, surová data v hadoopovském clusteru a využít jeho potenciálu.

Jaký je rozdíl ve využití analytických nástrojů nad dříve zmiňovaným datovým skladem a v případě big dat?

Pilař: Rozdíl mezi analytickými nástroji na těchto dvou platformách, tj. datovým skladem a big daty, je propastný. V datových skladech už vycházíme z toho, že v nich jsou data uspořádaná, konsolidovaná a vyčištěná pro následnou analýzu. Dnes se hovoří o Self Service BI (business intelligence), případně Self Analytics (samoobslužná BI, samoobslužná analytika), kde si i koncový uživatel může připravit report, vizualizaci a v některých případech například i trendový nebo prediktivní model. Aby toto mohlo fungovat, musí být v datovém skladu jeho data konsolidovaná a vyčištěná. ›

S big daty jsme v úplně jiné situaci. Jejich výhodou a zároveň nevýhodou je, že do takového datového úložiště lze dávat v zásadě libovolná data a jejich nesprávnou interpretací se můžeme velmi snadno dostat k zavádějícím výsledkům. Proto jejich analýza vyžaduje trochu jiné technické dovednosti, jak se k nim lze vůbec dostat, ale i analytické a byznysové. Teprve v okamžiku, kdy se tyto tři dovednosti potkají u člověka s know-how data science, pak teprve může dávat analýza nad úložištěm big dat smysl. Často se postupuje tak, že výstupem analýzy úložiště big dat je přidání oblasti data mart (datového tržiště). To je pak k dispozici i analytikům v byznysu, kteří nemají tak hlubokou znalost na technické a analytické úrovni pro big data.

Petr Švarc (40),

v oblasti datových skladů a reportingu se pohybuje 7 let. Do společnosti Sophia Solutions nastoupil v roce 2011, od té doby se zaměřuje na masivně paralelní zpracování dat a následně big data technologie. V současnosti zastává pozici senior consultant.

Technické dovednosti získal dříve jako Java vývojář a ABAP vývojář, byznys zkušenosti následně jako SAP konzultant.

Vystudoval Západočeskou univerzitu v Plzni, Fakultu aplikovaných věd, obor kybernetika a řídicí technika.

Pavel Pilař (44),

se více než 18 let věnuje oblasti business intelligence. Jeho specializací jsou reportingové a analytické systémy. Dlouhodobě se specializuje na oblast data miningu a statistické analýzy dat. V současnosti ve společnosti Sophia Solutions vede z pozice managing consultant a team leader R&A tým, který se touto problematikou zabývá. Od roku 2015 je jedním z partnerů společnosti.

Je absolventem Elektrotechnické fakulty Českého vysokého učení technického v Praze.

S nástroji pro analýzu dat pracovali vždy specialisté. Jak je tomu dnes? Poradí si například s interpretací výsledků i netechnicky založený pracovník?

Pilař: Jde o současný trend v oblasti Self-Service BI. Pokud mám jasně uspořádaná a vyčištěná data v datovém skladu, pak si jednoduché vizualizace a jednoduché reporty může připravit i koncový uživatel sám. V současnosti tak není třeba, aby pro každý nový report, analýzu či dashboard potřeboval byznys zadavatel vývojáře z IT.

Kdybychom to pro přirovnání velmi zjednodušili, mohlo by jít o analogii „vzít v Excelu sloupec hodnot a kliknout na sloupcový nebo koláčový graf“?

Pilař: Dá se to tak říci. Self-Service BI nástroje jsou dnes natolik uživatelsky přívětivé, že je stejně jednoduché vytvořit graf v těchto nástrojích jako ve zmiňovaném Excelu.

Švarc: To se ale týká analýzy strukturovaných a konsolidovaných dat nad datovým skladem. V oblasti big dat právě chybí ona strukturovanost a konsolidovanost dat. Nástroje pro Self-Service BI, které se v současné době používají nad datovými sklady, nad big daty logicky nedávají smysl, protože v nich může být v podstatě cokoliv.

V případě big dat tedy musí report pro manažera nějaký datový odborník dopředu připravit?

Pilař: Přesně tak. Musí znát jeho byznysový problém, připravit mu technickou část a manažer si pak může report zobrazit.

Ve kterých odvětvích průmyslu je největší zájem o technologie pro zpracování big dat?

Švarc: Big data dnes nejvíce „frčí“ asi u telekomunikačních operátorů. Mají spoustu dat a hledají řešení pro big data.

Pilař: Dále jsou to banky a utilities, tj. distributoři energií.

Co je ale na telekomunikačních operátorech nestrukturovaného?

Švarc: Nestrukturovanost u nich sice zase až taková není, ale v konečném důsledku se u nich vyskytuje semistrukturovanost dat. Jedním z jejich use cases je, že sbírají diagnostická data z hardwaru, který řídí jednotlivé buňky (BTS), protože jejich verze se v poslední době střídají nějaké dva až tři roky, dnes už jsme ve čtvrté generaci a je v přípravě pátá. Něco je tedy zčásti na novější verzi, něco na starší, a přitom stále potřebují řídit síť jako celek, mít o ní informace jako o celku.

Pilař: Aplikací big dat je mnoho a přibývají. Když se přidržím telekomunikací, pak Petr Švarc zmiňoval aplikaci, kde BTS hlásí spojení a rozpojení spojení. S tím se ale také otevírá obrovské pole působnosti pro další analýzu. Kdykoliv jedeme autem, pohybujeme se s mobilním telefonem po městě, po světě, mobilní operátoři informace o našem pohybu sbírají a ty se pak promítají například do mapových systémů jako Google Maps. To pak už jsou samozřejmě big data a související technologie pro jejich zpracování. Strukturovanou částí informace je, ke které stanici je který telefon přihlášen, na to už ale navazuje další objem nestrukturovaných dat, jak se kde který telefon pohybuje. Podobnou problematiku jsme řešili u větších komplexů, jako jsou nákupní střediska nebo třeba letištní budovy. I zde je podstatné, kde se lidé zdržují, jakým směrem a jakou rychlostí se uvnitř těchto komplexů pohybují. To opět vede na big data analýzu. Je to podobné jako u telekomunikační sítě, pouze ve vnitřním prostoru.

Máte ještě jiný příklad? Z jiné oblasti?

Pilař: Dalším příkladem může být již zmiňovaný banking, kde jsme nad big data úložištěm analyzovali transakce a vytvářeli prediktivní model. Pomocí big data technologií se řeší i otázka zabezpečení, kdy se analyzují logy síťových komponent. Každá komunikace, přihlášení k počítači, každý odeslaný paket dat za sebou zanechávají nějakou datovou stopu na firewallech a routerech, které je předávají. Množství těchto dat je obrovské, a přestože jde o strukturovaná data, nelze je reálně ukládat do databází a zpětně je v nich analyzovat. Dojde-li k indikaci nestandardní aktivity, je samozřejmě velmi zajímavá možnost zpětné analýzy. Tedy zda jde o nahodilý incident, nebo o výsledek nějaké dlouhodobější soustředěné snahy. Dostáváme se do situace, že v nějakém časovém bodě je třeba provést analýzu, co se dělo několik dnů nebo týdnů zpátky. V té době přitom ještě nebyla známa informace, co je třeba hledat. Takovéto analýzy lze efektivně udělat nad big daty, zatímco klasická databáze pro to není příliš vhodná. ■

Plnou verzi rozhovoru najdete na ictrevue.ihned.cz.