Big Data na běžném notebooku a za sekundy? Realita už i u nás.

Velké objemy dat jsou pro mnohé firmy stále větším problémem. Dynamika byznysu způsobuje, že jejich datové sklady praskají „ve švech“ a tradiční analytické BI nástroje jako datové sklady či kancelářské nástroje typu Excel nedokážou tento nápor zvládnout. Přitom „odemknutí“ vlastních dat může být impulzem k dalšímu růstu.

Paradoxně, téma Big Data se stalo klíčovým v obchodní strategii mnohých velkých IT firem – právě těch, jejichž de facto ztroskotalé DWH projekty vyvolávaly u klientů frustraci při analýze vlastních dat. Cílem těchto „mamutů“ není nic jiného, než dostat se zpět do hry a vylákat od klienta více peněz na rozvoj datových skladů, infrastruktury či dokoupení dalších modulů BI.

Přitom efektivní řešení, jak zvládnout velké objemy dat, mohou být i pro běžného člověka mnohem dostupnější, než byste si mysleli. Stále více organizací svěřuje svoje firemní data nástrojům Data Discovery, což je dnes technologicky nejvyspělejší oblast BI. Jsou dostupné za zlomek ceny tradičních BI řešení, dokážou si však poradit i s velkými objemy a na počkání.

V uplynulých letech se tyto nástroje staly poměrně běžné už i ve slovenské praxi, etablovaly se například v oblastech jako maloobchod či pojišťovnictví. Právě ty totiž disponují rozsáhlými a rychle rostoucími databázemi. Analýzy velkých dat probíhají na běžně dostupném hardwaru, přičemž práce s nimi si nevyžaduje hluboké technologické znalosti. Jelikož není nutné budovat datové sklady a vytvářet složité IT prostředí, takové řešení je nasaditelné v průběhu týdnů a jeho návratnost je rychlá.

Tradiční řešení nestačí s dechem

V čem je vlastně u tradiční analytiky s Big Daty problém? Vedle zmiňované vysoké ceny a dlouhé implementace je to i chybějící zaměření na potřeby zákazníka. Často se totiž jedná o zdlouhavé „technokratické“ projekty, které reálně nepomůžou zlepšit každodenní analytickou práci lidí, ti tak zůstávají odkázaní na svoje „excely“ a SQL skripty.

Pokud si během procesu uvědomíme chybu a je potřeba report předělat, či vyvstane nějaká ad-hoc otázka, se kterou se původně nepočítalo – znamená to problém. To je už potom parketa pro interní IT oddělení, či externí firmu. Jinak řečeno, bude to stát víc času anebo financí.

U Big Dat nemůže být ani řeč o použití jinak oblíbeného analytického nástroje – Excelu. Vyhodnocovat databázi o velikosti i jen 10 tisíc záznamů může Excelu trvat i hodiny, či dokonce dny. A to je řeč jen o základní analytice. Detailní kalkulace a některé statistické pohledy při větším objemu dat není možné realizovat vůbec.

Most z Moskvy do Londýna

Dobrým příkladem společností, které efektivně spravují obrovské balíky dat pomocí Data Discovery, jsou zdravotní pojišťovny. Jedna z těch, co působí na našem trhu, potřebuje na denní bázi pracovat s produkčními systémy, které obsahují miliardy záznamů, přičemž každý rok přibyde čtvrt miliardy nových údajů. Takovéto množství záznamů neumožňovalo ani jen pomyslet na to, aby bylo zvládnutelné tradiční cestou. V pojišťovně tak byli nuceni databáze dělit a zaměřit se jen na určitou oblast či kratší období.

Takovýto objem si už vyžaduje opravdu účinný nástroj. Vždyť posuďte sami – jestliže vycházíme z toho, že jeden řádek v Excelu má 1 centimetr a všech 250 miliónů záznamů bychom seřadili na jeden papír, výsledný seznam by byl dlouhý 2500 kilometrů. Jinak řečeno, dokázal by přemostit celý evropský kontinent a spojil by Moskvu s Londýnem.

Když chtějí v pojišťovně analyzovat data za 4 roky, systém musí zvládnout až miliardu záznamů. Zní to možná až neuvěřitelně, ale Data Discovery technologie dokážou i z takto dlouhé řady záznamů najít jeden konkrétní i za méně než sekundu. Dokonce ho ještě odprezentují i v širším kontextu, takže informace je komplexnější.

Díky moderním technologiím jsou schopni připojit k existujícím produkčním systémům i další databáze, které jsou též dostupné ad-hoc a na pár kliků. Takto dokážou jednoduše zkombinovat demografické či geografické záznamy s produkčními daty, číselníky a podobně.

Výstupy „do zítřka“ už nejsou problém

Často neřešitelným problémem v tradičním světě bývají nečekané požadavky. Když bylo v pojišťovně nutné udělat něco „do zítřka“, při milionech záznamů původem z IT prostředí to bylo prakticky nemožné.

Problematická byla i flexibilita. Když se během zpracovávání analýzy objevila nějaká nesrovnalost, bylo nutné vytáhnout z databází jiný vzorek dat – ať už jiný typ péče, jiné kódy výkonů a podobně. Celý cyklus analýz se tak natáhl, reporty se dokončovaly na poslední chvíli, ve stresu a častokrát s chybami. Výstupy tak nebyly dostatečně důvěryhodné. Data Discovery technologie však problémy s nekvalitními daty vyřešily a to až na úroveň jejich vstupu do systémů.

Ze šesti hodin na osm minut

Běžný den v pojišťovně přinesl například i potřebu zjistit, jestli předepsaným antibiotikům předcházely návštěvy u lékaře. Bylo třeba dát do souvislosti dva zdroje, ze kterých každý obsahuje dva milióny záznamů. Znamenalo to pracné dohledávání v databázích, které se často opakovalo. Analýzy trvaly pět i šest hodin, proto se zpravidla spouštěly v noci. Pokud nastal nějaký problém anebo zaznamenali nějakou chybu, celý proces se musel opakovat, což trvalo opravdu dlouho.

Naproti tomu u Data Discovery řešení se jen přidaly nové dimenze, analýza proběhla znovu a výsledek byl na světě. Trvalo to osm minut a v případě, že bylo třeba něco změnit, opravná analýza trvala opět jen osm minut, což je výrazný skok oproti hodinám tradiční cestou. Takovýmto způsobem byla pojišťovna schopná ušetřit i dny či dokonce týdny času a desítky tisíc eur na případných „change requestech“ pro externí IT firmy.

Spočítejte včerejší piva bez čekání

Podobné problémy jako pojišťovny se řeší i v maloobchodě. S každým nákupem, s každou novou skladovou položkou či novou kampaní přicházejí do systémů retailových řetězců nová data. Bez moderních technologií by bylo prakticky nemožné je sledovat a naplno využívat ve prospěch podnikání.

Představte si, že jste manažer retailového řetězce a potřebujete zjistit, jak se včera v konkrétní lokalitě prodávalo pivo. A navzdory tomu, že vaše databáze obsahuje 350 milionů záznamů a máte dvě stě poboček po celé zemi, na odpověď nebudete čekat déle než sekundu. Potřebný pohled si stačí jednoduše „vyklikat“ a žádoucí výběr se zobrazí v přehledné formě a hlavně v kontextu.

Takovouto analýzu můžete spustit i na běžném kancelářském notebooku, a to i při analýze stovek milionů záznamů.

Výsledný report zobrazí nejen počet prodaných lahví piva, ale i to, jak se prodávaly jednotlivé značky, o jaké druhy byl největší či nejmenší zájem, jestli bylo zboží v akci, jestli byla daná akce úspěšná a kolik jste vlastně vydělali. Takto je možné průběžně vyhodnocovat prodej, plánovat zásoby, případně řídit „za běhu“ promo akce na podporu prodeje.

Jen pro porovnání, u tradičního přístupu by vygenerování téže informace včetně kontextu trvalo dny. V takovém případě můžete na průběžné řízení kampaní a vyhodnocování výsledků zapomenout.

Kouzlo je v komprimaci dat

Za rychlostí Data Discovery se skrývá technologie, která u některých nástrojů dokáže komprimovat data až na 10 % jejich původní velikosti, přičemž jejich kvalita zůstává zachovaná. Díky tomu je možné i miliardy dat uložit do operační paměti serveru.

To vyřešilo problém s diskovými operacemi, které jsou pomalé a způsobovaly časové ztráty. Databáze o velikosti stovek milionů záznamů je tak dokonce možné analyzovat i v běžném notebooku, bez potřeby serveru. Dá se tedy říci, že toto zdánlivě komplikované téma se díky Data Discovery dostalo do rukou běžného člověka, a to při minimální investici.

Martin Kostič
Ředitel společnosti EMARK