Big data na bežnom notebooku a za sekundy?

Veľké objemy dát sú pre mnohé firmy stále väčší problém. Dynamika biznisu spôsobuje, že ich dátové sklady „praskajú vo švíkoch" a tradičné analytické nástroje BI, ako dátové sklady či kancelárske nástroje typu Excel, nedokážu tento nápor zvládnuť. Pritom „odomknutie" vlastných dát môže byť impulzom na ďalší rast.

Paradoxne sa téma big data stala kľúčovou v obchodnej stratégii mnohých veľkých IT firiem – práve tých, ktorých de facto zlyhané projekty DWH vyvolávali u klientov frustráciu pri analýze vlastných dát. Cieľom týchto „mamutov" nie je nič iné ako dostať sa naspäť do hry a vylákať od klienta viac peňazí na rozvoj dátových skladov, infraštruktúry či dokúpenie ďalších modulov BI.

Pritom efektívne riešenia, ako zvládnuť veľké objemy dát, môžu byť aj pre bežného človeka oveľa dostupnejšie, ako by ste si mysleli. Stále viac organizácií zveruje svoje firemné dáta nástrojom data discovery, čo je dnes technologicky najvyspelejšia oblasť BI. Sú dostupné za zlomok ceny tradičných riešení BI, dokážu si však poradiť aj s veľkými objemami a na počkanie.

V uplynulých rokoch sa tieto nástroje stali pomerne bežnými už aj v slovenskej praxi, etablovali sa napríklad v oblastiach, ako je maloobchod či poisťovníctvo. Práve tie totiž disponujú rozsiahlymi a rýchlo rastúcimi databázami. Analýzy veľkých dát prebiehajú na bežne dostupnom hardvéri, pričom práca s nimi si nevyžaduje hlboké technologické znalosti. Keďže netreba budovať dátové sklady a vytvárať zložité IT prostredie, takéto riešenie je nasaditeľné v priebehu týždňov a jeho návratnosť je rýchla.

Tradičné riešenia nestačia s dychom

V čom je vlastne v tradičnej analytike pri big data problém? Popri spomínanej vysokej cene a dlhej implementácii je to aj chýbajúce zameranie na potreby zákazníka. Často totiž ide o zdĺhavé „technokratické" projekty, ktoré reálne nepomôžu zlepšiť každodennú analytickú prácu ľudí, tí tak zostávajú odkázaní na svoje „excely" a skripty SQL.

Ak si počas procesu uvedomíme chybu a treba report prerobiť, prípadne sa vynorí nejaká ad hoc otázka, s ktorou sa pôvodne nerátalo, znamená to problém. To je už potom parketa pre interné IT oddelenia či externú firmu. Inak povedané, bude to stáť viac času alebo financií.

Pri big data nemôže byť ani reči o použití inak obľúbeného analytického nástroja Excelu. Vyhodnocovať databázu s veľkosťou čo i len 10-tisíc záznamov môže Excelu trvať aj hodiny či dokonca dni. A to ide len o základnú analytiku. Detailné kalkulácie a niektoré štatistické pohľady pri väčšom objeme dát nemožno realizovať vôbec.

Dobrý príklad spoločností, ktoré efektívne spravujú obrovské balíky dát vďaka data discovery, ponúkajú zdravotné poisťovne. Jedna z tých, ktoré pôsobia na našom trhu, potrebuje na dennej báze pracovať s produkčnými systémami obsahujúcimi miliardy záznamov, pričom každý rok pribudne štvrť miliardy nových údajov. Pri takomto množstve záznamov nemožno ani len pomyslieť na to, že by sa to dalo zvládnuť tradičnou cestou. V poisťovni tak boli nútení databázy čiastkovať a zamerať sa len na určitú oblasť či kratšie obdobie.

Ak v poisťovni chcú analyzovať dáta za 4 roky, systém musí zvládnuť až miliardu záznamov. Znie to možno až neuveriteľne, ale technológie data discovery dokážu aj v takomto dlhom rade záznamov nájsť jeden konkrétny aj za menej ako sekundu. Dokonca ho ešte odprezentujú aj v širšom kontexte, takže informácia je komplexnejšia.

Vďaka tomu je poisťovňa schopná pripojiť k existujúcim produkčným systémom aj ďalšie databázy, ktoré sú tiež dostupné ad hoc a na pár kliknutí. Takto možno jednoducho skombinovať demografické či geografické záznamy s produkčnými dátami, číselníkmi a podobne.

Výstupy „dozajtra" už nie sú problém

Často neriešiteľným problémom v tradičnom svete bývajú nečakané požiadavky. Ak bolo v poisťovni treba urobiť niečo „dozajtra", pri miliónoch záznamov v pôvodom IT prostredí to bolo prakticky nemožné.

Problematická bola aj flexibilita. Ak sa počas spracúvania analýzy objavila nejaká nezrovnalosť, bolo potrebné vytiahnuť z databáz inú vzorku dát – iný typ starostlivosti, iné kódy výkonov a podobne. Celý cyklus analýz sa tak natiahol, reporty sa dokončovali na poslednú chvíľu, v strese a často s chybami. Výstupy tak neboli dostatočne dôveryhodné. Technológie data discovery však problémy s nekvalitnými dátami vyriešili, a to až na úroveň ich vstupu do systémov.

Zo šesť hodín na osem minút

Bežný deň v poisťovni priniesol napríklad aj potrebu zistiť, či predpísaným antibiotikám predchádzali návštevy u lekára. Bolo treba dať do súvisu dva zdroje, z ktorých každý obsahuje dva milióny záznamov. Znamenalo to prácne dohľadávanie v databázach, ktoré sa často opakovalo. Analýzy trvali päť aj šesť hodín, preto sa spravidla spúšťali v noci. Ak nastal nejaký problém alebo zaznamenali nejakú chybu, celý proces sa musel opakovať, čo trvalo naozaj dlho.

Naproti tomu pri riešení data discovery sa len pridali nové dimenzie, analýza sa vykonala nanovo a výsledok bol na svete. Trvalo to osem minút a v prípade, že bolo treba niečo zmeniť, opravná analýza trvala opäť len osem minút, čo je výrazný skok oproti hodinám tradičnou cestou. Takýmto spôsobom bola poisťovňa schopná ušetriť aj dni či dokonca týždne času a desiatky tisíc eur na prípadných „change requestoch" pre externé IT firmy.

Spočítajte včerajšie pivá bez čakania

Podobné problémy ako poisťovne riešia aj v maloobchode. S každým nákupom, každou novou skladovou položkou či novou kampaňou prichádzajú do systémov retailových reťazcov nové dáta. Bez moderných technológií by bolo prakticky nemožné ich sledovať a naplno využívať v prospech podnikania.

Predstavte si, že ste manažér maloobchodného reťazca a potrebujete zistiť, ako sa včera v konkrétnej lokalite predávalo pivo. Aj napriek tomu, že vaša databáza obsahuje 350 miliónov záznamov a máte dvesto pobočiek po celej krajine, na odpoveď nebudete čakať dlhšie ako sekundu. Potrebný pohľad si stačí jednoducho „vyklikať" a želaný výber sa zobrazí v prehľadnej forme a hlavne v kontexte. Takúto analýzu môžete spustiť aj na bežnom kancelárskom notebooku, a to aj pri analýze stoviek miliónov záznamov.

Výsledný report zobrazí nielen počet predaných fliaš piva, ale aj to, ako sa predávali jednotlivé značky, o aké druhy bol najväčší či najmenší záujem, či bol tovar v akcii, či bola daná akcia úspešná a koľko ste vlastne zarobili. Takto možno priebežne vyhodnocovať predaj, plánovať zásoby, prípadne riadiť „za behu" promoakcie na podporu predaja.

Len na porovnanie, pri tradičnom prístupe by vygenerovanie tej istej informácie aj s kontextom trvalo dni. V takom prípade môžete na priebežné riadenie kampaní a vyhodnocovanie výsledkov zabudnúť.

Kúzlo je v komprimácii dát

Za rýchlosťou data discovery sa skrýva technológia, ktorá pri niektorých nástrojoch dokáže komprimovať dáta až na 10 % ich pôvodnej veľkosti, pričom ich kvalita zostáva zachovaná. Vďaka tomu možno aj miliardy dát uložiť do operačnej pamäte servera.

To vyriešilo problém s diskovými operáciami, ktoré sú pomalé a spôsobovali časové straty. Databázy s veľkosťou stoviek miliónov záznamov sa tak dokonca dajú analyzovať aj v bežnom notebooku bez potreby servera. Dá sa teda povedať, že táto zdanlivo komplikovaná téma sa vďaka data discovery dostala do rúk bežného človeka, a to pri minimálnej investícii.

» MARTIN **KOSTIČ**,
Riaditeľ spoločnosti EMARK