Kvalita dat

jako krevní obraz společnosti

Problémem nejsou data samotná, ale nedotažené procesy, které vznik „nepořádku“ v datech umožňují.

Jen těžko byste našli firmu, kde si nikdo nikdy nestěžoval na data. Samozřejmě! Ostatně kolem práce s daty se vytvořilo hned celé odvětví. Co se vlastně děje, když data stojí „za starou bačkoru“? S velkou pompou se obyčejně nastartuje projekt, případně se najme konzultantská firma a provede se rozsáhlé čištění dat. K velkému překvapení se ale po pár měsících zjistí, že v datech je už zase nepořádek (!).

A proč se tenhle scénář opakuje znovu a znovu? Protože většina takových projektů vychází z představy, že nekvalitní data napadají systémy tak nějak sama od sebe, jako hejna kobylek. V MAGNETIC IDEAS nejsme zastánci „teorie kobylek“. Kvalita dat je jako lakmusový papírek, kterým lze snadno zjistit **kondici procesů**. Je to stejné jako krevní obraz. Když vám lékař zjistí vysoký krevní tlak, také vám nepustí žilou, ale začne léčit příčinu. S daty je to stejné.

Najít nekvalitní data je snadné, primárním cílem je ale najít a trvale odstranit příčinu vzniku. Proto by za kvalitu dat neměl být odpovědný datový analytik zběhlý v SQL nebo Pythonu, ale spíše vlastník procesu, který rozumí nastavení procesů ve společnosti a dokáže je dostatečně ovlivnit.

Jak tedy řešit problémy s daty?

1. Je důležité se objektivně podívat na data. Nelze se spoléhat na subjektivní postřehy typu „Zákaznické adresy jsou špatně zadané“. Základní (nikoli však jediná) metoda je profilování dat a identifikace krajních hodnot (outliers) nebo vyložených chyb.
2. Dále je třeba zjistit příčiny a definice „správnosti“ dat. Typická chyba v této fázi nastává, když se účastníci projektu mezi sebou shodnou, jak mají správná data vypadat. To jednoduše nelze, protože data používá celá firma a každý trochu jinak. Je nutné odstranit nejen špatná data, ale i zabránit jejich další tvorbě.
3. Ve třetím kroku podnikáme právě tyto prevenční kroky a úpravy postupů, aby se zamezilo vzniku dalších nekvalitních dat.
4. Teprve ve čtvrtém kroku můžeme přistoupit k čištění dat.
5. Nakonec je potřeba stanovit metriky, podle nichž se bude kvalita dat monitorovat a definovat mezní hodnoty, které případně spustí další opravnou aktivitu.

Zní to jednoduše, že? Tak si vzpomeňte, jaké to bylo, když jste naposledy řešili problém s daty.

Autor: Libor Stenzl