Spása jménem Big Data?

BI projekty honosící se názvem Big Data se těší u zákazníků v poslední době velké popularitě. Nezřídka je nemožné prosadit projekt, který toto slovní spojení nemá v názvu bez ohledu na to, že obsah projektu nemá dle definice s Big Daty nic společného. Podstatně horší na tom ale je nesmyslné očekávání, které management od těchto projektů mívá, plynoucí většinou z nepochopení přínosu Big Data technologií.

Ilustrovat to lze mj. i na následujících otázkách, které pochází od skutečných zákazníků:

1. Mám nahradit svůj stávající datový sklad technologií Big Data?
2. Vyřeším BI problém v oblasti XY použitím Big Data nástroje?
3. Dokážou se Big Data nástroje přizpůsobit procesům v naší firmě?

Co jsou to Big Data

Společnost Gartner před pár lety přišla na základě nových požadavků na správu dat s definicí, která klasifikovala data do tří dimenzí – Objem (Volume), Typ (Variety) a Rychlost (Velocity) - 3V. Bohužel tyto veličiny jsou v zásadě relativní a nepostihují všechny charakteristiky dat. Postupem času jednotliví výrobci i sama společnost Gartner definici rozšířila o další charakteristiky - Různorodost (Variability), Důvěryhodnost (Veracity), Komplexnost (Complexity) apod. Bohužel i tyto charakteristiky zcela nepopisují to, co Big Data jsou, vždy je totiž nutné vzít v úvahu souvislost daných dat a úloh, které pomocí nich chceme řešit. Teprve v okamžiku, kdy máme data a úlohu, kterou nelze efektivně vyřešit pomocí „standardních“ technologií a je zde nutnost použít „nové“ technologie, je možné hovořit o Big Data (a to ještě v kontextu aktuálního okamžiku).

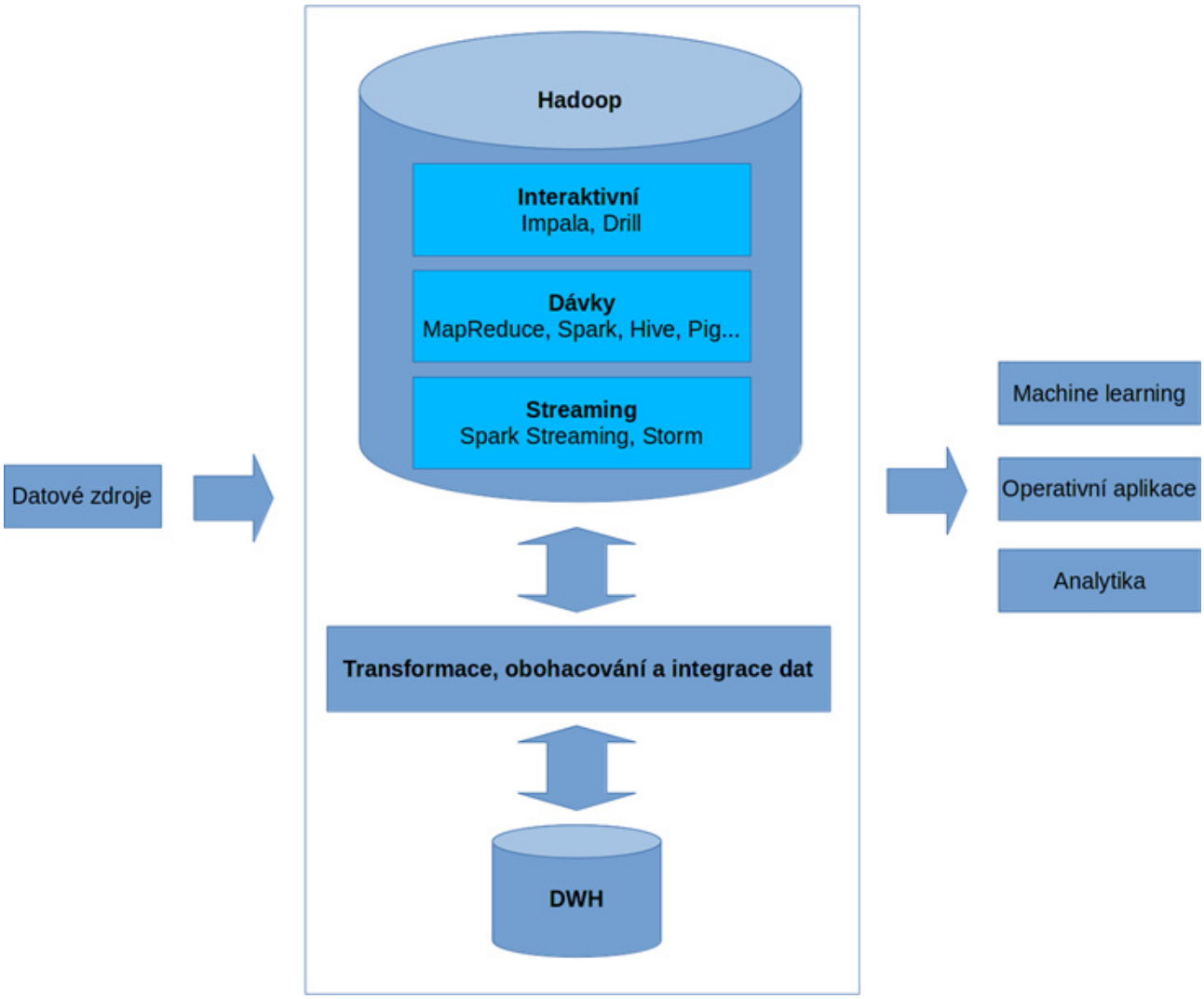
Mám nahradit svůj stávající datový sklad technologií Big Data?

V okamžiku, kdy má společnost fungující datový sklad, který je ekonomicky rentabilní a splňuje všechny požadavky i s výhledem do budoucnosti, nemá cenu uvažovat o pořizování Big Data technologie. Je důležité si uvědomit, že Big Data technologie z pohledu informací ve svém důsledku nepřinášejí nic zásadně nového než stávající klasické BI technologie. Rozdíl je pouze ve výkonnosti a způsobu zpracování, nicméně samotný výsledek - informace - má stejnou podobu.

Opačná situace nastane v okamžiku, kdy stávající datový sklad má nějaký problém:

- kapacitní problémy – neumožňuje udržování a analýzu hluboké historie; nesplňuje požadavky na nové typy dat
- výkonnostní problémy - způsobené objemem dat nebo požadavky na rychlost zpracování
- nevhodnou strukturu - datový sklad je obtížné přizpůsobit novým volnějším datovým strukturám

Pokud tyto problém nelze řešit stávající technologií datového skladu, ať už z technických důvodů (méně pravděpodobné), ale především z ekonomických důvodů (nejčastější), má smysl se ptát, zda neexistuje jiné řešení. Pak je vhodně použita technologie Big Data na „odlehčení" datového skladu - přesun historie a analytických pískovišť na Big Data platformu.



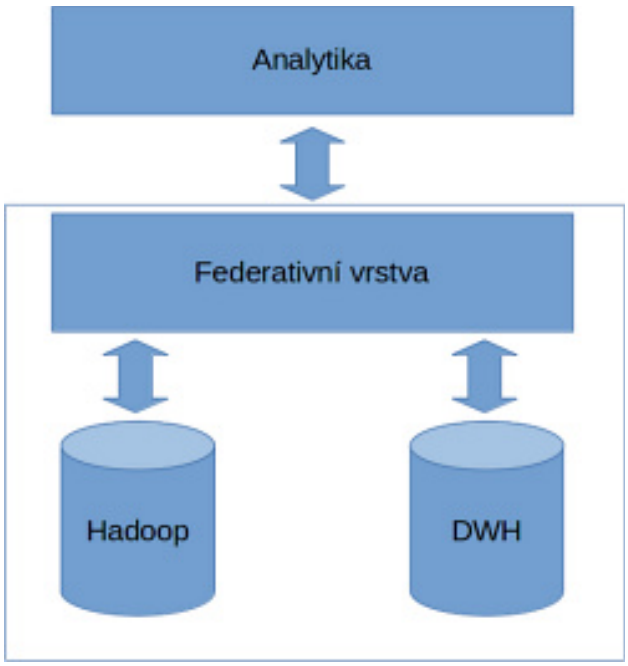
Obr. 1: Použití Hadoopu pro odlehčení datového skladu

Vyřeším BI problém v oblasti XY použitím Big Data nástroje?

Pokud bychom byly v pozici výrobce Big Data softwaru, tak bychom zajisté odpověděli kladně a ani bychom nemuseli příliš lhát. Jak již bylo uvedeno výše, Big Data nástroje umožňují provádět datovou analýzu stejně dobře nebo za určitých podmínek i lépe než tradiční technologie (např. SQL). Pokud jsme v pozici zákazníka, tak bychom na tuto otázku měli odpovědět kladně pouze v okamžiku, kdy známe charakteristiku a hodnotu dat, ze kterých chceme těžit informace a víme, že stávající technologie na zpracování technologicky nestačí nebo je její použití ekonomicky nerentabilní. Z toho vyplývá, že úvahy o použití nových technologií (a to nejen v oblasti Big Data) by při řešení problémů měly přicházet na řadu až jako poslední, protože pořízení technologie samo o sobě žádný přínos nemá.

Dokážou se Big Data nástroje přizpůsobit procesům v naší firmě?

U BI manažerů často oprávněně panuje obava z alternativních systémů vůči hlavnímu datovému skladu, především z důvodů vzniku alternativních zdrojů „pravdy”. Nutno podotknout, že toto může být problém každého BI systému, který se buduje vedle datového skladu, nikoli jen Big Data. Naopak technologie Big Data (především ty založené na technologii Hadoopu) představují ideální doplněk pro stávající datové sklady. Klasický datový sklad bude mít stále své pevné místo v BI řešeních, ale už nebude nutné, aby obsahoval všechna data, nebo bude mít data rozprostřena v různých systémech. Právě řešení rozprostřenosti dat bude čím dál více nabývat na důležitosti, jelikož bude čím dál obtížnější data slévat na jedno místo. Federativní vrstva pak může zajišťovat kombinaci více datových zdrojů například pomocí standardního jazyka SQL, aniž by všechna data musela nutně existovat v jedné relační databázi nebo aby byla dokonce v relační databázi uložena.



Obr. 2: Federativní vrstva

Nasazujte Big Data tam, kde to má smysl

Technologie Big Data je žhavým trendem oblasti zpracování dat a má svá uplatnění i vedle tradičních a ověřených metod, jakou jsou datové sklady. Nicméně by tato technologie měla být nasazována tam, kde to má skutečný význam a nejen proto, že v konkurenčním podniku již Big Data mají.

Dovolím si jedno doporučení na závěr - než začnete vypisovat výběrové řízení na technologii Big Data ve vaší společnosti, nechte si (s mnohem nižšími náklady) udělat studii, zda by měla tato technologie u vás ten patřičný a očekávaný efekt.