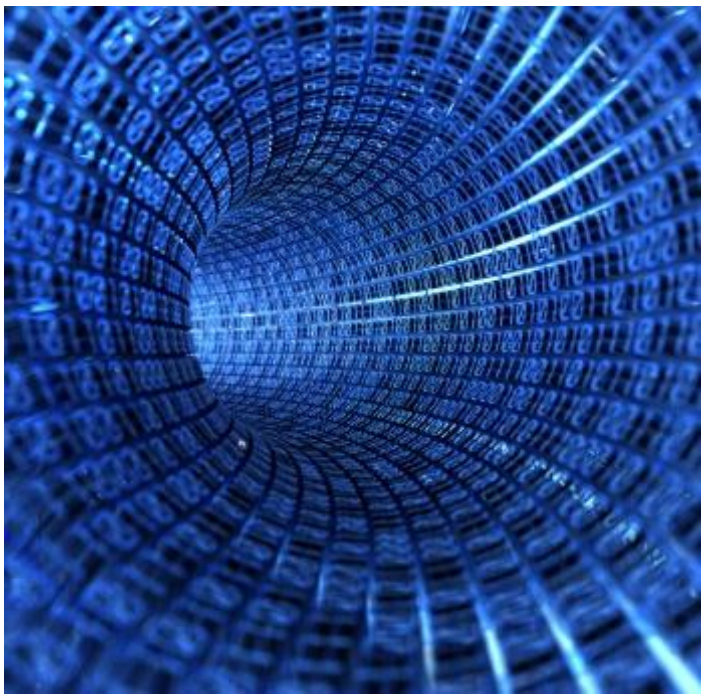
Paradoxne v dobe plnej hesiel ako digitálna technológia, informačná spoločnosť a znalostná ekonomika, je čoraz ťažšie dostať sa k dátam. Teraz nemyslím pod dátami tony papiera, alebo nekonečné tera-bajty diskových polí plných núl a jednotiek, ale dáta s veľkým D na začiatku.

Dáta boli, sú, aj budú základ poznania

Michal Salaj



Experiri est porta ad lumen

LeanPortal

Print 00001

Paradoxne v dobe plnej hesiel ako digitálna technológia, informačná spoločnosť a znalostná ekonomika, je čoraz ťažšie dostať sa k dátam. Teraz nemyslím pod dátami tony papiera, alebo nekonečné tera-bajty diskových polí plných núl a jednotiek, ale dáta s veľkým D na začiatku.

Dáta boli, sú, aj budú základ poznania

Michal Salaj

„Veríme v Boha, všetci ostatní doneste dáta.“
Edward Deming

K tomuto veľmi výstižnému výroku jedného z najväčších mysliteľov minulého storočia a otca manažmentu kvality Edwarda Deminga sa v poslednej dobe čím ďalej, tým viac vraciam a len neveriaco krútim hlavou. Musím sa priznať, že prvý krát som naň narazil až v knihe *Competing on Analytics* (Thomas Davenport, Jeanne Harris, 199X). Ihneď mi však udrel do očí a naštartoval tie správne myšlienkové pochody.

Paradoxne v dobe plnej hesiel ako digitálna technológia, informačná spoločnosť a znalostná ekonomika, je čoraz ťažšie dostať sa k dátam. Teraz nemyslím pod dátami tony papiera, alebo nekonečné tera-bajty diskových polí plných núl a jednotiek, ale dáta s veľkým D na začiatku.

Dáta, tie pravé, sú dnes čoraz viacej vzácnejším materiálom. Z vlastnej praxe a pozorovaní som si všimol, že to nie je ani tak zapríčinené technológiou, možnosťami ich zdieľania a šírenia, alebo v poslednej dobe čím ďalej obľúbenejšej praktike, či už vedeckých, súkromných, ale aj verejných spoločností a organizácií, spoplatňovania zdrojov a prístupov k dátam, než samotnou prehlbujúcou sa negramotnosťou osôb, ktoré s dátami pracujú.

Je to zaujímavé, ale keď sa človek snaží a skutočne hľadá, predsa len definíciu termínu dáta v literatúre nakoniec nájde. Je však až zarážajúce koľko rôznych prístupov, pohľadov a kvalitatívnych rozdielov medzi jednotlivými definíciami existuje. Na konci dňa môžeme konštatovať, že definície sú veľmi rozdielne a zasahujú často krát aj do takých pojmov, ako informácie, znalosti, niekedy dokonca aj do pojmov, ako vedomosti, či inteligencia.

Pritom dáta predstavujú prvý krok na ceste k poznaniu. Keďže v praxi, alebo v teórii denno denne definujeme, získavame, pracujeme, uskladňujeme a

prezentujeme „dáta“, tieto sa stali neoddeliteľnou súčasťou našich životov. Je preto namieste sa nad definíciou tohto pojmu aspoň na chvíľu pozastaviť a dobre ho premyslieť.

Prístupov, ako Dáta zadefinovať je asi toľko, koľko pokusov bolo uskutočnených a tento je defacto ďalší v poradí. V tomto texte by som rád vychádzal skôr z klasického pohľadu komunity zomknutej okolo databázových systémov, pretože mám k nemu najbližšie a podľa mňa je tento pohľad pre dnešného čitateľa najjednoduchšie predstaviteľný a stráviteľný, keďže väčšina aktivít s dátami sa dnes aj tak deje skrz IT technológie.

Musím sa však priznať, že ďalšia časť textu je skôr súbor definícií viacerých pojmov, ktoré objasňujú celkový obraz okolo tematiky dát, ako len suchá fráza definície. Budem sa snažiť naviesť čitateľa k potrebnému smeru pohľadu a dôležité úvahy rozviesť v jednoduchom jazyku a príkladoch. Nerád by som preto, aby bol tento text braný doslovne, ale skôr, ako vodítko na ceste za poznaním.

Začali by sme reprezentáciou sveta. Svet okolo nás je súbor najrôznejších „vecí“ (hviezdy, planéty, ľudia, môj pracovný stôl a na ňom položený mobilný telefón,…). Z pohľadu dát ich môžeme spoločne pomenovať ako *entity*. Medzi týmito existujú priame a nepriame vzťahy a prebiehajú interakcie. Všeobecne k nim môžeme referovať, ako ku *javom* (rotácia planét spôsobujúca gravitáciu a ňou ovplyvňované entity a javy na ich povrchu, telefonický rozhovor a jeho dôsledky pre zúčastnených, futbalový zápas a interakcia 22 hráčov, divákov, lopty, rozhodcov, výsledkovej tabuľky ligy,…).

Keďže dáta používame predovšetkým na opísanie tohto reálneho sveta, nie je možné a mysliteľné opísať ho absolútne celý. Preto vždy siahame k určitej abstrakcii. Aj mi samotní, v zmysle našich receptorov zbierajúcich dáta a tím aj naše vnímanie sveta okolo nás, funguje vo forme abstrakcie.

Tak napríklad receptory v našich očiach sú schopné zachytiť svetlo s vlnovou dĺžkou v rozmedzí 380-690 nm. Toto nám neumožňuje vidieť napríklad infračervenú, alebo ultrafialovú zložku svetla. Tieto sú pre nás takpovediac „neviditeľné“ a preto aj pre naše vedomie neexistujúce, resp. abstrahované faktory. Preto, keď náš mozog skladá obraz okolia, ako ho vidíme, s týmito faktormi nepočíta. Vlastne tu hovoríme o obmedzení rozlišovacej schopnosti systému. Rovnakým spôsobom fungujú všetky receptory, ako napríklad ušný bubienok, jazykové papily, teplomery, svetlo-citlivé polovodičové prvky vo fotoaparátoch (máte doma 8 megapixelový, alebo len 6 megapixelový?) a ďalšie.

Podobne pristupujeme aj pri výstavbe dátového modelu, ktorým popisujeme súbor entít a javov, ktoré budú pre nás relevantné, teda dôležité. Všetko ostatné bude abstrahované a defacto pre náš model neexistujúce.

Dátový model je teda model, ktorý mapuje a asociuje entity a javy s reálnym svetom a dejmi, ktoré v ňom prebiehajú. Obsah takéhoto modelu môže popisovať ako *fyzické*, tak aj *abstraktné* entity a javy. Fyzickou entitou je napríklad auto, dom,

človek, šváb, zatiaľ čo abstraktnou entitou je napríklad politická strana, názory, atď. Charakteristiky takto modelovo popísaných entít a javov nazývame atribútmi.

Atribút je teda vlastnosť entity, alebo javu, ktorá nadobúda hodnoty z množiny, ktorú informatici radi nazývajú doména, alebo *obor funkčných hodnôt*. Táto množina nie je v žiadnom smere obmedzená čo do obsahu a usporiadania hodnôt. Množina môže obsahovať čísla, písmena, obrazovú reprezentáciu, dátumy, celé pasáže textu, jednoducho ľubovoľné symboly, resp. prvky, podľa ktorých sme schopní rozpoznať odlišné stavy vlastností sledovanej entity, alebo javu. Entity, ktoré majú rovnaké atribúty môžeme zhlukovať do tzv. *typov entít*.

Pre jednoduchú ilustráciu vyššie spomenutých definícií si ich názorne rozoberieme v nasledujúcom krátkom príklade.

Predstavme si, že máme vytvoriť dátový model reprezentujúci našu kanceláriu. Kanceláriou rozumieme prvú a nie nemalú obmedzujúcu podmienku dôležitú pre náš dátový model. Celý známy aj neznámy vesmír zredukovať na jednu kanceláriu je vcelku veľká abstrakcia. Ďalším krokom bude zadefinovanie súboru atribútov, ktoré sú pre nás relevantné. Tieto môžeme chápať ako ďalšie obmedzujúce podmienky.

Takýto prístup redukciou je najčastejší pri výstavbe dátových modelov, kedy ako prvé sa zadefinuje oblasť pôsobnosti modelu a potom v rámci tejto oblasti, sa dodefinujú dôležité vlastnosti. Tu však číha aj najväčšie nebezpečenstvo.

Pri zlom výbere obmedzujúcich podmienok, kedy na prvý pohľad nedôležité vlastnosti entít a javov majú veľmi kritickú váhu, no je od nich abstrahované, klesá vypovedacia hodnota celého modelu.

Vypovedacia hodnota dát, resp. dátového modelu, je hodnota, resp. váha pri tvorbe výpovedí, extrapolácií a informácií, ktorú môžeme takýmto dátam a modelu prikladať. *Hodnotu* môžeme definovať aj ako mieru spôsobilosti atribútu uspokojiť potrebu, kvôli ktorej bol tento zadefinovaný, svojou vnútornou kvalitou, ktorá ho robí žiadúcim.

Skutočný problém nastáva práve v tomto momente, pretože často krát nemáme ako túto hodnotu posúdiť, pretože nekvalitu atribútu nemáme ako a s čím porovnať. Z vlastnej skúsenosti a skúseností mnohých ďalších expertov z celého sveta s ktorými som mal možnosť spolupracovať na rôznych projektoch môžem povedať, že prvú definíciu dátového modelu môžeme chápať iba ako referenčným nástrelom a je potrebných x iterácií a zmien, aby dosiahol potrebnú „kvalitu“. *Preto jedinou dobrou radou je neustále sa vracajte späť k svojmu dátovému modelu a dopytujte sa „Je to to pravé orechové?“.*

Ale vráťme sa späť k nášmu príkladu. Rozhodli sme sa v rámci kancelárie sledovať len zamestnancov a to pomocou atribútov meno, priezvisko, identifikačné číslo, dátum narodenia a fotografia.

Každý z atribútov má presne definované rozsahy a typy symbolov, prvkov, ktoré môžu nadobudnúť. Tak napríklad atribúty Meno a Priezvisko môžu byť definované, ako textové reťazce s prvým písmenom vždy veľkým a maximálne 100 znakov dlhými. Identifikačné číslo je

definované napríklad, ako pole 5 po sebe idúcich číslic. Identifikačné číslo v tomto prípade môže plniť úlohu *jednoznačného odlíšenia* entít, resp. zamestnancov, keďže podmienkou nadobúdania hodnôt je, že každý zamestnanec má *jedinečnú kombináciu*. Dátum je definovaný v tvare DD.MM.RRRR. Fotografia je v našom modeli definovaná, ako súbor .jpg tvaru, ktorý je výstupom fotografovania digitálnym fotoaparátom o rozlišovacej schopnosti 5 megapixelov, pri nástupe zamestnanca do práce v rozlíšení 800x600 pixelov, pričom záber je urobený s bielym pozadím za snímanou osobou. Na chvíľu sa pozastavme nad definíciou fotografie.

„Jednoduché, výstižné a všetci vedia o čom je reč.“ Takto, alebo podobne by určite vyššie zmienenú definíciu okomentoval môj nebohý praded, ktorý nemal ani potuchu o počítačoch, informatike či dátach, no pritom nikdy by nestratil prehľad vo „veciach“. Divili by ste sa však, ako málo sa dozviete o dátach v dnešnej praxi. Možno, že sa ani nedivíte a bojujete tak, ako ja s týmto, nebojím sa povedať, faktom každý deň. Trvá celé dni, niekedy týždne kým sa k podobným informáciám dopátrate. A veľmi často sa k nim nedopátrate vôbec a celé dátové súbory, možno kritických pozorovaní putujú do koša.

Kvalitu dát neovplyvňujú len faktory, ako rozlišovacia schopnosť a rýchlosť receptorov, ktorým bola donedávna pripisovaná hlavná rola pri jej posudzovaní, ale mnohé ďalšie, často zdanlivo nesúvisiace faktory.

Práve preto, aby sme dokončili charakterizáciu pojmov okolo tematiky dát, musíme sa zmieniť ešte o pojmoch ako dátová reprezentácia, dátový záznam a formát.

Dátovou reprezentáciou rozumieme práve súbor pravidiel podľa ktorých sa dáta zaznamenávajú na médium, resp. dátový nosič.

Dátový záznam potom predstavuje fyzický objekt, alebo ich súbor, ktorý reprezentuje dáta na médiu, či už sa jedná o perom napísané písmená a číslice na papieri, alebo zmagnetizované prvky reprezentujúce bity na diskových poliach.

Reprezentácia dát je dosiahnutá skrz *formát*, teda pravidlá, akým spôsobom sa má vykonať dátový záznam. Ide napríklad o pravidlá typu, jeden záznam = jeden riadok, radenie atribútov do poradí, ako napríklad meno pred priezvisko a mnoho ďalších.

Po tom, ako sme si zadefinovali predchádzajúce pojmy začíname asi tušiť o čom pojem Dáta hovorí. Keby sme ho chceli v súhrne zadefinovať, *predstavujú Dáta súbor fyzických dátových záznamov, reprezentovaných na médiu, ktoré majú svoj zadefinovaný formát, pričom ich obsahom sú jednoznačne odlíšiteľné pozorovania vybraných atribútov, resp. vlastností entít a javov za účelom uspokojiť potrebu poznania, kvôli ktorej boli tieto zadefinované.*