EKONOMICKÁ UNIVERZITA V BRATISLAVE
FAKULTA HOSPODÁRSKEJ INFORMATIKY

Evidenčné číslo: 103004/I/2017/1220259998

BIG DATA a možnosti ich spracovania

Diplomová práca

2017 **Milan Ragula**

EKONOMICKÁ UNIVERZITA V BRATISLAVE
FAKULTA HOSPODÁRSKEJ INFORMATIKY

BIG DATA a možnosti ich spracovania

Diplomová práca

Študijný program: Informačný manažment
Študijný odbor: 3.3.24 Kvantitatívne metódy v ekonómii
Školiace pracovisko: Katedra aplikovanej informatiky
Vedúci záverečnej práce: doc. Ing. Gabriela Kristová, CSc.

Bratislava 2017 **Milan Ragula**

Ekonomická univerzita v Bratislave
Fakulta hospodárskej informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta:	Bc. Milan Ragula
Študijný program:	Informačný manažment (Jednoodborové štúdium, inžiniersky II. st., denná forma)
Študijný odbor:	3.3.24 Kvantitatívne metódy v ekonómii
Typ záverečnej práce:	Inžinierska záverečná práca
Jazyk záverečnej práce:	slovenský
Sekundárny jazyk:	anglický

Názov: Big data a možnosti ich spracovania

Anotácia: Cieľom práce je spracovať problematiku spracovania veľkých objemov dát a následne získané poznatky demonštrovať na úlohách z praxe.

Vedúci:	doc. Ing. Gabriela Kristová, CSc.
Katedra:	KAI FHI - Katedra aplikovanej informatiky FHI
Vedúci katedry:	Ing. Mgr. Peter Schmidt, PhD.
Dátum zadania:	18.10.2015

Dátum schválenia: 06.11.2015

doc. Ing. Gabriela Kristová, CSc.
vedúci katedry

Čestné vyhlásenie

Čestne vyhlasujem, že záverečnú prácu som vypracoval samostatne a, že som uviedol všetku použitú literatúru.

Dátum: 05. 05. 2017

.....

(podpis študenta)

Poďakovanie

Týmto spôsobom by som sa chcel poďakovať vedúcej záverečnej práce doc. Ing. Gabriele Kristovej, CSc. za odbornú pomoc, usmernenie a cenné pripomienky, ktoré mi poskytla pri vypracovaní diplomovej práce.

ABSTRAKT

Ragula, Milan: *BIG DATA a možnosti ich spracovania.* – Ekonomická univerzita v Bratislave. Fakulta hospodárskej informatiky; Katedra aplikovanej informatiky. – Vedúci záverečnej práce: doc. Ing. Gabriela Kristová, CSc.

Práca sa zaoberá možnosťou spracovania veľkého množstva neštruktúrovaných dát – Big Data, s cieľom vyťažiť z nich informácie, ktoré by organizáciám lepšie pomohli poznať zákazníka, jeho správanie a záujmy voči ich službám a produktom, ktoré ponúkajú. Táto problematika sa označuje anglickým pojmom „Customer Intelligence“ alebo zákaznícka inteligencia.

V teoretickej časti analyzujeme súčasný stav problematiky Big Data, ako aj právny a morálny aspekt analýzy neštruktúrovaných dát. Taktiež v teoretickej časti práce sme detailne rozobrali problematiku platformy Big Data, jednotlivé charakteristiky, ktorými sa vyznačuje, ako aj možnosti ich využitia. Popísali sme tu konkrétne technológie, ktoré slúžia na analýzu veľkého množstva dát.

V praktickej časti sme popísali metodiku, metódy skúmania a základné ciele tejto práce. V tejto časti sme sa venovali samotnej analýze získaných dát, kde sme popísali návrh technického riešenia a spôsob spracovania, či analyzovania problematiky Big Data pomocou nami zvoleného nástroja. Navrhované riešenie je následne implementované a zároveň sme demonštrovali prínosy Big Data analytics na príkladoch počtu hovorov v oblasti telekomunikácií v rámci Slovenskej republiky za mesiac január, február a marec r. 2017. Analyzujeme kategórie ako sú vek zákazníka, kraj, kde sa daný používateľ nachádza v dobe hovoru, mobilný telefón, z ktorého je hovor uskutočnený, zákaznícky segment a typ siete. Následne sme ponúkli interpretáciu daných výsledkov.

Kľúčové slová:

veľké dáta, python, databázové systémy, neštruktúrované dáta, vizualizácia

ABSTRACT

Ragula, Milan: *BIG DATA and the possibility of processing.* – University of Economics in Bratislava. Faculty of Business Informatics; Department of Applied Informatics – Tutor of thesis: doc. Ing. Gabriela Kristová, CSc.

This paper work deals with the possibility of processing large amounts of unstructured data – so called Big Data - to get information from them that would help organizations better understand their customers, their behavior and interests, especially towards the organizations, their services, and the products they offer. This system is called Customer Intelligence. In the theoretical part of the paper we analyze the current state of the Big Data as well as the legal and moral aspects of the analysis of unstructured data. Also in the theoretical part of the thesis we provide detailed overview of the problems of the Big Data platform, their individual characteristics, as well as the possibilities of their use. We have also described specific technologies that are used to analyze a large amount of data.

In the practical part we have described methodology, methods of investigation and basic objectives of this work. We devoted this section to our analysis of the data obtained, describing the technical solution and how to process or analyze the Big Data using the tool we have selected. The proposed solution is subsequently implemented and demonstrates the benefits of Big Data analytics on examples of telecommunication calls within the Slovak Republic for the period of January, February and March 2017. We provide an analysis of categories such as customer age, region where the call is realized, the mobile phone from which the call was done, the customer segment, and the type of network. Then we interpret and describe the outcomes of our analysis.

Keywords:

Big Data, python, database system, unstructured data, visualisation

Obsah

Zoznam obrázkov

Zoznam grafov

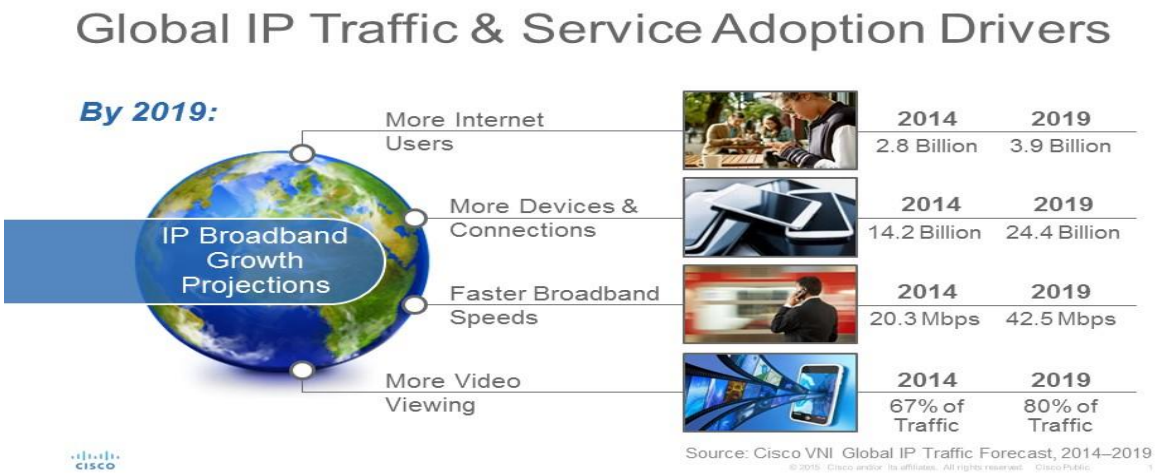
„ Big Data is like Sex in High School – Lots of people are talking about it, but few are having it."

(Eric Hansen,)

Úvod

Žijeme v dobe, kde sa všetko naokolo nás riadi informáciami, ktoré obsahujú čoraz viac dát. Informácie a dáta sa stávajú významnými komponentami úspechu všetkých subjektov na trhu. Jednou z hlavných potrieb businessu je potreba dostať čo najrýchlejšie a čo najjednoduchšie tie správne informácie o správnej osobe, na správne miesto a v správnom čase. Množstvo dát, ktoré ako spoločnosť každý deň generujeme a s ktorými sa musíme stretávať je vcelku nepredstaviteľné. Podľa autora O'Reilly približne ide o:

- každý deň viac než 144,8 miliárd emailov,
- každú minútu pribudne na server Youtube 72 hodín videa,
- každú minútu spracuje vyhľadávač Google cez 2 milióny vyhľadávacích dotazov,
- každú sekundu generuje urýchľovač častíc v CERNu 40 terabajtov dát.



Obrázok č. 1: predpokladaný nárast online dát od r. 2014 do r. 2019

Zdroj : http://www.cisco.com/

Motiváciou pre vypracovanie diplomovej práce na tému „veľkých dát" je okrem prezentovania vlastných výsledkov a poznatkov aj mediálny šum okolo technológie pre spracovanie problematiky Big Data.

Samotná práca sa skladá z troch hlavných častí a viacerých podkapitol.

V prvej kapitole sú objasnené východiská a jednotlivé pojmy potrebné pre pochopenie skúmanej problematiky. V rámci kapitoly sú objasnené definície skúmaných systémov. V tejto kapitole sú tiež uvedené konkrétne poznatky pre lepšie pochopenie problematiky Big Data – veľkých objemov dát.

Druhá kapitola je spracovaná tak, aby poskytla charakteristiku nami vybraného nástroja pre spracovanie a analýzu veľkého objemu dát a taktiež popisujeme metódy spracovania dát.

V poslednej kapitole implementujeme jednu nami zvolenú metódu z mnohých možných technologických riešení problematiky v oblasti Big Data. Táto prípadová štúdia sa skladá z viacerých častí podľa objektu našej analýzy.

1 Súčasný stav problematiky doma a v zahraničí

1.1 Big Data

„ You can have data without information, but you cannot have information without data.“

(Daniel Keys Moran)

Obrázok č. 2: znázornenie schémy Big Data

Zdroj : http://www.informatieplaats.nl/big-data/

1.1.1 Pojem „Big Data“

„Technológie Big Data opisujú novú generáciu technológií a architektúr, ktoré sú navrhnuté tak, aby ekonomicky extrahovali hodnotu z veľmi veľkých objemov údajov, umožňujúc vysokú rýchlosť zachytenia, objavovanie alebo analýzu“ (Minelly M., 2013, s.12). Táto definícia naznačuje, že Big Data je súbor technológií, ktoré majú za cieľ správu a analýzu veľkého množstva prevažne neštruktúrovaných dát pre získanie výsledkov, ktoré by sa dali finančne zhodnotiť.

Pre orientáciu v tejto záplave informácií sú už dlhšiu dobu využívané rôzne analytické nástroje s cieľom uľahčiť rozhodovanie, zrýchliť procesy, zviditeľniť určité extrémne hodnoty v dátach a podobne. Tieto nástroje sa označujú ako Business Intelligence a vedia pracovať s takzvanými štruktúrovanými dátami (ako sú dáta uložené v databázových

systémoch) a so záplavou nových druhov dát, neštruktúrovaných informácií, si nevedia poradiť.

Druhou - s prvou veľmi úzko zviazanou výzvou businessu je jeho nikdy nekončiaca potreba dokonale porozumieť zákazníkovi a dlhodobo ponúkať služby a produkty, o ktoré naozaj javí záujem a ktoré uspokojujú jeho potreby. Zároveň je potrebné s daným zákazníkom budovať vzťah.

Riešenie oboch týchto výziev vyžaduje zapojenie princípov a technológií, ktoré tu doposiaľ neboli a ktoré de facto vytvárajú nové odvetvie Big Analytics. Títo špecialisti spracovávajú obrovské množstvá rôznych dát s dôrazom na neštruktúrované dáta (predovšetkým v textovej podobe). Využívanie týchto pokročilých analytických metód pre lepšie poznanie a pochopenie nielen vlastného businessu, dodávateľov, či zákazníkov ale aj vzťahov medzi subjektmi na trhu, kde daná organizácia funguje, je v súčasnej dobe prakticky na začiatku. Ich implementáciou sa dá získať silná konkurenčná výhoda. Avšak tento stav nemusí trvať dlho a veľmi rýchlo sa blíži doba, kedy budú tieto praktiky nutnosťou pre prežitie na silne konkurenčnom globálnom trhu. Organizácia, ktorá premešká túto dobu, bude mať veľmi ťažkú pozíciu dohnať ostatných hráčov na trhu, pričom sa jej to nemusí ani podariť.

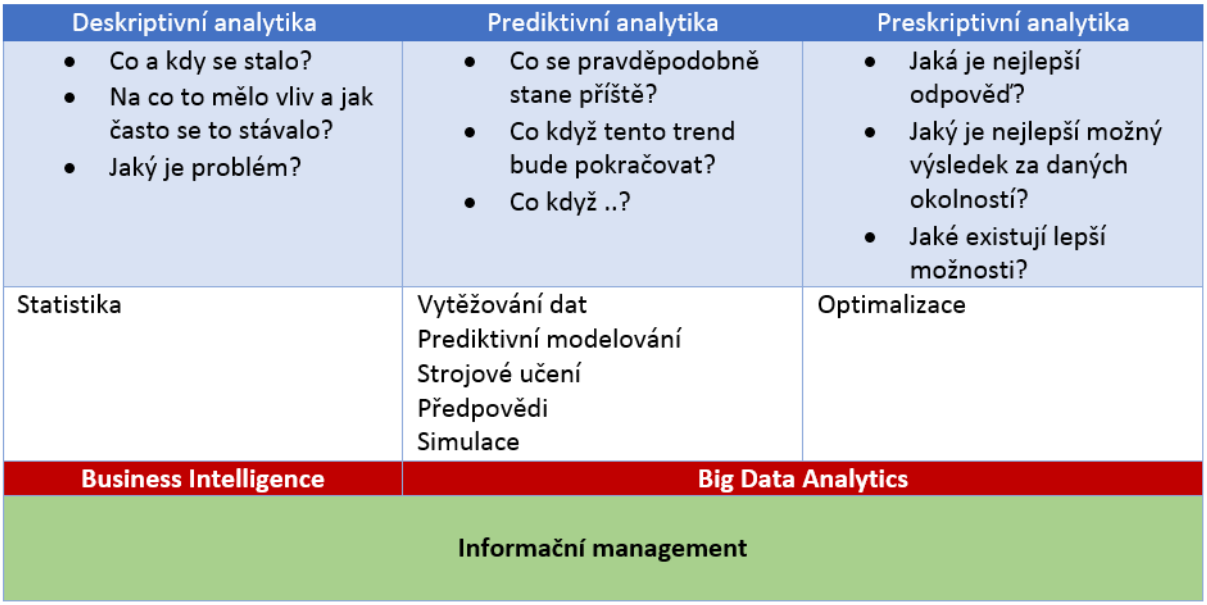
Big Data sú veľký pojem dnešnej doby, ktoré menia premýšľanie ľudí nad tým, čo sú vlastne dáta, čo sa z nich dá získať, a ako sa dajú tieto informácie využiť pre ďalší rast organizácie. Big Data pritom nie sú aplikácia alebo produkt, ktorý sa dá kúpiť, implementovať a hneď používať, je to platforma, filozofia a celý nový spôsob uvažovania.

K pokročilej analýze dát je pravdaže potreba dostatočného množstva dát v dostatočnej kvalite. V dátach musia byť zaujímavé informácie, ktoré z nich môžeme vyťažiť a následne použiť.

"Big Data sú charakteristické veľkými objemami dát, vysokou rýchlosťou spracovania alebo vysokou škálovateľnosťou informačných prostriedkov, ktoré vyžadujú nové formy spracovania, umožňujúce lepšie rozhodovanie, zisťovanie poznatkov a optimalizáciu procesov" (Minelly M., 2013, s. 52). Za Big Data považujeme teda také dáta, ktoré sa nedajú ukladať ani spracovávať tradičnými technológiami. Tieto dáta sú tak kapacitne náročné, že ich nemôžme uložiť na jeden disk, a je teda potrebné ich uložiť na viac pamäťových médií. Big Data sú zároveň brané ako večné dáta, takže nie je nutné žiadne dáta mazať iba pridávať nové. Čo je raz v dátach zachytené, ostáva v nich navždy. Preto je hlavná požiadavka na rozšírenie tohto úložného priestoru za chodu a podľa potreby (teoreticky až do nekonečna). Výpočty nad týmito dátami sa musia uskutočňovať distribuovane.

Koncept Big Data mení i samotný postup výpočtu, kde sa už nepresúvajú dáta na jedno miesto kvôli výpočtu, ale výpočet sa pomocou špeciálnych algoritmov Map & Reduce distribuuje bližšie k dátam. Výsledok sa následne dosiahne postupným redukovaním medzivýsledkov. Práve paralelné spracovanie a prístup k informáciám umožňuje riešeniam postaveným na tomto princípe o dosť vyššiu rýchlosť a robustnosť oproti štandardným systémom (Sathi A. 2012).

Rozdiel medzi týmto prístupom a bežnou Business Intelligence znázorňuje nasledujúci obrázok č. 3.



Obrázok č. 3: rozdiel medzi business intelligence a big data analytics

Zdroj : https://www.beapp.sk/aplikacie/7-qlik

Ďalším dôležitým benefitom konceptu Big Data je fakt, že celé riešenie sa dá zostaviť iba s využitím komoditného hardware riešenia či servermi. Napríklad ide o spoločnosti ako Google alebo Facebook, ktoré majú svoje dátové centrá zostavené z relatívne lacných počítačov zapojených v obrovskom počte paralelne, čo im dodáva potrebný výkon. Pri výpadku jedného uzla ho môžeme ľahko nahradiť iným, ktorý prevezme jeho prácu.

Hlavný rozdiel oproti bežným analytickým prístupom je pri samotnom kladení dotazov. Big Data, tým, že uchovávajú všetky dostupné informácie v ich pôvodnej podobe bez ohľadu na formát a štruktúru, umožňujú zadávať akékoľvek otázky, a to aj také, ktoré v dobe zhromažďovania dát ešte neboli známe, alebo sa nevedelo, že ich jedného dňa niekto bude potrebovať zodpovedať. To je presný opak tradičnej databázovej štruktúry, ktorá svojou pevne definovanou štruktúrou z princípu obmedzuje množinu dotazov. Pokiaľ sa chce

používateľ na takto uložené dáta pozrieť novým pohľadom, z iného uhla, ktorý nebol v pláne pri pôvodnom návrhu databázy, má častokrát problém, alebo sa musí štruktúra databázy upraviť a dáta do nej spätne prekonvertovať.

Je treba podotknúť, že Big Data nemenia nič na princípoch analytickej práce, iba v niektorých jej aspektoch podstatne rozširujú možnosti :

- Využitie všetkých dostupných dát pre vytvorenie modelu (predtým iba vzorka).
- Kombinácia viac analytických modelov a techník s cieľom zlepšiť a spresniť výsledok.
- Samo učiace algoritmy, ktoré sa učia nové získané informácie, a ďalej tak spresňujú svoje výsledky.
- Využitie prediktívnych modelov prakticky v reálnom čase.

1.1.2 Súčasnosť

V roku 2009 sa objavil nový vírus chrípky označovaný ako H1N1, ktorý kombinoval prvky vírusov spôsobujúcich vtáčiu a prasaciu chrípku a rýchlo sa šíril. Po niekoľkých týždňoch sa pracovníci hygienických služieb začali obávať príchodu nebezpečnej pandémie. Proti novému vírusu nebola rýchlo dostupná žiadna vakcína a preto zdravotníci mohli iba dúfať, že sa im podarí postup nákazy spomaliť. Na to by ale potrebovali vedieť, kde sa už choroba vyskytla.

Agentúra CDC (Centers for Disease Control and Prevention), ktorá je súčasťou amerického ministerstva zdravotníctva, požiadala lekárov, aby poskytli informácie o nových prípadoch chrípky. Obrázok o pandémii, ktorý mohla agentúra vytvoriť z týchto hlásení, bol však vždy o jeden či dva týždne oneskorený. Tak sa mohli ľudia obrátiť na lekárov až po niekoľkých dňoch problémov. Kvôli neaktuálnym údajom bola agentúra v najkritickejších dňoch úplne paralyzovaná.

Niekoľko dní pred tým, ako sa vírus H1N1 dostal do novinových titulkov, náhodou odborníci z internetového giganta Google publikovali vo vedeckom časopise Nature zaujímavý článok. Autori v ňom vysvetlili, ako môže spoločnosť Google predpovedať šírenie chrípky v USA, a to nielen na celoštátnej úrovni ale aj v rôznych oblastiach a dokonca aj v jednotlivých štátoch. Spoločnosť to dokázala sledovaním toho, čo ľudia vyhľadávali na internete. Vďaka tomu, že Google spracuje každý deň viac než tri miliardy vyhľadávajúcich dotazov a všetky ukladá, má na spracovanie dostatok dát. Spoločnosť zhromaždila 50 miliónov vyhľadávajúcich termínov, ktoré Američania najčastejšie

vyhľadávali a porovnala tento zoznam s dátami agentúry CDC o šírení chrípky medzi rokmi 2003 a 2008. Cieľom bolo identifikovať oblasti zasiahnuté vírusom chrípky podľa slov zadávaných do internetového vyhľadávača. O podobnú analýzu internetových vyhľadávajúcich termínov sa pokúsili aj iní, ale nikto nemal k dispozícií toľko dát, výpočtového výkonu a ani štatistických znalostí ako spoločnosť Google.

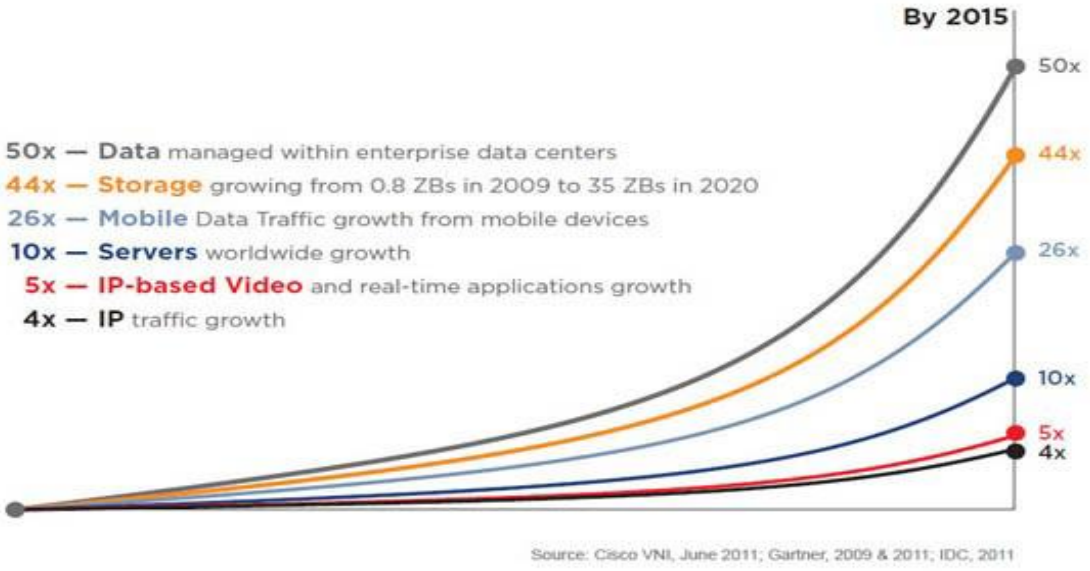
Takže spoločnosť Google zostavila program, ktorý hľadal koreláciu medzi frekvenciou určitých vyhľadávajúcich dotazov a šírením chrípky v čase a priestore. Pri testovaní vyhľadávacích dotazov spracovali celkom neuveriteľných 450 miliónov rôznych matematických modelov a porovnávali pritom svoje predpovede so skutočnými prípadmi chrípky, ako ich zachytila agentúra CDC v rokoch 2007 a 2008. Ukázalo sa, že ich software našiel kombináciu 45 vyhľadávajúcich termínov, ktoré po zahrnutí do jedného matematického modelu vykazovali silnú koreláciu medzi svojimi predpoveďami a oficiálnymi celoštátnymi číslami. Podobne ako agentúra CDC dokázal tento model určiť, kam sa chrípka rozšírila. Na rozdiel od dát CDC mohol však tieto informácie poskytovať takmer v reálnom čase a nie o týždeň alebo dva týždne neskôr ako dáta od CDC.
Keď sa teda v roku 2009 objavila epidémia vírusu H1N1, ukázalo sa, že systém spoločnosti Google je užitočnejším a včasnejším indikátorom než vládne štatistiky.
Metóda spoločnosti Google sa prekvapivo zaobíde bez spolupráce lekárskych ordinácií. Miesto toho je založená na „veľkých dátach“ (Big Data), vďaka ktorým môže spoločnosť inovatívnymi spôsobmi spracovávať informácie a získavať užitočné a cenné poznatky o výrobkoch a službách. Preto starostlivosť o verejné zdravie je iba jednou z mnoho oblastí, ktoré Big Data zásadne ovplyvňujú. Veľké dáta menia podobu celých podnikových oblastí (Mayer-Schonberger V. – Cukier K. 2014).

1.1.3 Dôvody pre využívanie technológie Big Data

Ako základné dôvody pre využívanie technológie Big Data môžeme spomenúť princíp čiernej skrinky a presnosť.

Princíp čiernej skrinky nám hovorí, že na trhu sa nachádza množstvo dostupných platforiem, ktoré nám umožňujú spracovávať veľké množstvo neštruktúrovaných dát konzervatívnym spôsobom. Tieto systémy sú náročné najmä na počiatočnú implementáciu a vyladenie. Technológia Big Data sa nezaujíma o vnútorné vzťahy medzi dátami a neanalyzuje spôsob získavania dát, jednoducho tieto komplikované kroky implementácie abstrahuje, takže celý proces implementácie výrazne zjednodušuje.

Keďže Big Data technológia rieši spracovávanie veľkých objemov dát, nie je možné presne analyzovať každý dátový súbor. Konvenčné spôsoby sa snažili pri spracovávaní dát chybné hodnoty eliminovať alebo opraviť. Naopak Big Data do analýzy zahŕňajú chybné hodnoty, takže očakávaný výsledok pri spracovávaní nie je úplne bezchybný. Na druhej strane výsledok analýzy je takmer totožný so skutočnosťou.

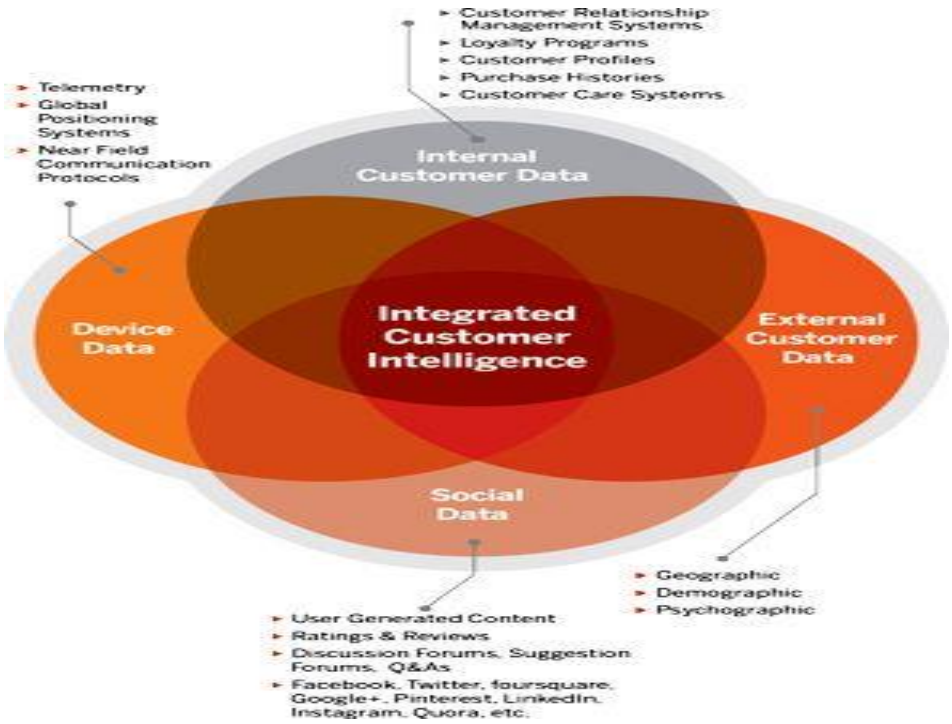


Obrázok č. 4: Trendové krivky objemu dát v čase

Zdroj: http://www.cisco.com/

1.1.4 Customer Intelligence

Základnou otázkou Customer Intelligence v problematike Big Data všeobecne je „Aký druh informácií o klientoch Big Data obsahujú a ako sa dajú využiť?“ (Novotný O., Pour J., Slánský D., 2005, s. 35). Je dôležité mať dáta, ale dôležitejšie je vedieť, čo s nimi môžeme vykonávať, aké výstupy z nich môžeme očakávať a ako tieto výstupy môžu byť využité v businesse.



Obrázok č. 5: Big Data v oblasti customer intelligence

Zdroj : https://sk.pinterest.com/pin/560064903633213973/

Typy dát

Big Data nám umožňujú zahrnúť do analytického procesu akékoľvek dáta bez ohľadu na formát, takže musíme myslieť vo veľkom. Ako príklad nových zdrojov informácií môžeme uviesť nasledujúci zoznam:

- internetové dáta – sociálne médiá, webové diskusie, články, clickstream,
- primárny výskum – ankety, dotazníky, pozorovania, experimenty,
- sekundárny výskum – business dáta, dáta o konkurencií a trhu, reporty,
- obrazové dáta – obraz, satelitné snímky, video, sledovanie,
- geolokačné informácie – mobilné dáta, GPS,
- dáta zo zariadení – senzory.

Príklad využitia dát

Zaujímavým príkladom využitia platformy Big Data môže byť napr. sledovanie „clikstreamu“, teda detailného záznamu pohybu myši napríklad v elektronickom obchode. Ďalej môžeme uviesť ako príklady využitia platformy Big Data zdravotníctvo, vyšetrovanie alebo telekomunikácie, ktorej problematike sa budeme venovať aj v tejto práci.

1.1.5 Korelácia

Vo svete malých dát je korelácia medzi údajmi užitočná ale v prípade veľkých dát je naozaj neoceniteľná. Prostredníctvom nej môžeme získavať nové poznatky ľahšie, rýchlejšie a jasnejšie než doteraz.

Princíp korelácie spočíva v tom, že kvantifikuje statický vzťah medzi dvomi dátovými hodnotami. Silná korelácia ukazuje, že keď sa mení jedna dátová hodnota, s vysokou pravdepodobnosťou sa bude meniť aj druhá hodnota. So silnými koreláciami sme sa stretli pri systéme Google Flu Trends: čím viac ľudí v určitej lokalite vyhľadáva pomocou Google isté termíny, tím viac ľudí v danom mieste má chrípku. Slabá korelácia naopak ukazuje, že pri zmene jednej dátovej hodnoty sa druhá príliš nemení. Keby sme spočítali koreláciu dĺžky vlasov a subjektívneho šťastia, zistili by sme, že dĺžka vlasov nám toho o spokojnosti danej osoby moc nepovie. Vďaka koreláciám nemusíme pri analýze daného javu skúmať jeho vnútorné princípy, ale naopak môžeme len identifikovať užitočný zástupný jav. Je treba podotknúť, že ani silné korelácie nebývajú dokonalé. Je možné, že sa dva javy správajú podobne iba zhodou okolností. Korelácia nám nedáva žiadnu istotu ale pravdepodobnosť s akou sa dané dva javy navzájom ovplyvňujú. Pokiaľ je však korelácia silná, znamená to, že vzťah je dosť pravdepodobný.

Vďaka tomu, že nám korelácia umožňuje identifikovať veľmi dobrú náhradu za určitý jav, pomáha nám zachytiť prítomnosť a predpovedať budúcnosť: v prípade ak jav A nastáva často spolu s javom B, môžeme sledovať výskyt javu B a predpovedať, že dôjde k javu A. Pomocou javu B ako zástupcu môžeme zaznamenať, čo sa pravdepodobne deje s javom A, aj keď jav A nedokážeme pozorovať ani merať priamo. Dôležité je, že nám tento postup umožňuje odhadovať, čo sa s javom A stane v budúcnosti. Pomocou korelácie samozrejme nedokážeme predpovedať budúcnosť s istotou ale len s určitou pravdepodobnosťou. Avšak táto možnosť je mimoriadne cenná (Mayer-Schonberger V. – Cukier K. 2014).

Využitie korelácie

Opodstatnenosť dôležitosti využitia korelácie si môžeme ukázať na príklade firmy, ktorá v roku 1997 začínala predávať knihy cez internet. Len za dva roky svojej existencie firma dosiahla značný úspech. Obchod mal adresu Amazon.com. História spoločnosti Amazon je síce pomerne známa, ale málo ľudí vie, že obsah na ich stránke pôvodne tvorili ľudia. Editori a kritici hodnotili a vyberali tituly, ktoré následne Amazon propagoval na svojich stránkach. Boli zodpovední zato, čomu sa hovorilo „štýl Amazonu“ a čo spoločnosť

pokladala za jedno zo svojich kľúčových aktív a svoju konkurenčnú výhodu. Vtedajší článok v novinách Wall Street Journal ich označil za najvplyvnejších literárnych kritikov v celej zemi, pretože na základe ich článkov sa predávalo množstvo kníh.

Následne sa však Jeff Bezos, zakladateľ a CEO Amazonu začal pohrávať zo zásadnou myšlienkou: čo keby spoločnosť dokázala zákazníkom odporučiť konkrétne knihy na základe ich individuálnych nákupných preferencií? Od začiatku svojej existencie Amazon zaznamenával množstvo údajov o svojich zákazníkoch: čo kupujú, na ktoré knihy sa iba pozreli, ale nekúpili ich, ako dlho si ich prezerali a ktoré knihy si dali do košíka spoločne. Dát bolo toľko, že ich Amazon najskôr spracovával konvenčným spôsobom a to tak, že vybral z nich vzorku a tú potom analyzoval, aby našiel podobnosti medzi zákazníkmi. Výsledné odporučenia neboli príliš presné. Keď ste si kúpili knihu o Poľsku, systém vás zaplavil publikáciami o strednej Európe. Po nákupe knihy o malých deťoch ste dostávali ponuky na ďalšie knihy rovnakého druhu. Systém spravidla navrhoval malé odchýlky od predchádzajúcich nákupov.

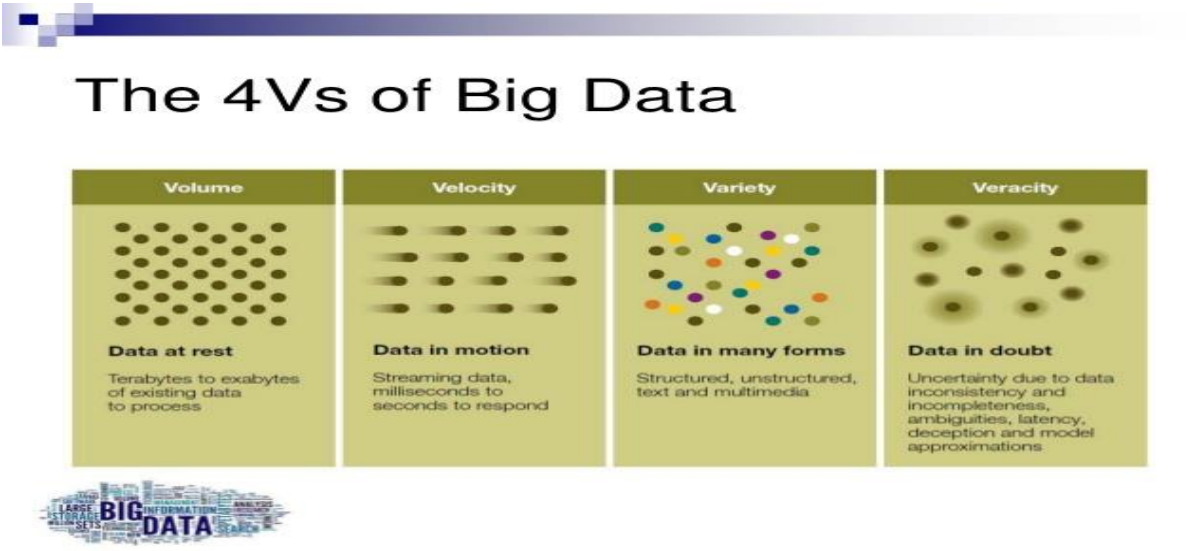
Riešenie objavil Greg Linden. Uvedomil si, že systém odporúčaní nemusí porovnávať používateľov s inými ľuďmi, čo bol technicky náročný proces. Stačilo nájsť asociáciu medzi samotnými produktami. V roku 1998 Linden a jeho kolegovia podali žiadosť o patent na kolaboratívne filtrovanie medzi položkami, ako sa táto metóda označuje. Táto zmena prístupu mala zásadný dopad.

Vzhľadom nato, že výpočty bolo možné uskutočniť vopred, odporúčanie bolo bleskovo rýchle. Metóda bola taktiež všestranná a bolo ju možné nasadiť v rôznych kategóriách. Keď sa Amazon pustil do predaja iných produktov ako kníh, dokázal rovnakým spôsobom odporučiť napr. filmy a rôzne iné produkty, ktoré mal vo svojom portfóliu. A odporúčania boli omnoho lepšie ako predtým, pretože systém spracovával všetky dáta. Takže sa musela spoločnosť rozhodnúť čo sa bude zobrazovať na webe. Počítačom generovaný obsah typu osobných odporúčaní a zoznamov bestsellerov, alebo recenzie napísané internými redaktormi Amazonu? Čo hovoria dáta alebo kritici? Túto bitku bojovali počítačové myši s ľuďmi.

Keď Amazon uskutočnil pokus a porovnal predaje, ktoré redaktori dosiahli, s predajmi, ktoré zabezpečil počítačom generovaný obsah, výsledky boli úplne odlišné. Materiál vytvorený na základe dát generoval omnoho vyšší objem predaja. Počítač nemusí vedieť, prečo čitateľ Ernesta Hemingwaya môže mať záujem o knihu F. Scotta Fitzgeralda. Ale na tom nezáleží. Nakoniec bol tím redaktorov rozpustený. Odhaduje sa, že v súčasnosti tretinu všetkých predajov spoločnosti Amazon zabezpečujú ich systémy odporučení

a používateľského prispôsobenia. Vďaka týmto systémom Amazon vyradil z hry mnoho svojich konkurentov – nielen veľké kníhkupectvá a hudobné predajne, ale taktiež miestne kníhkupectvá, ktoré sa domnievali, že je pred technologickými novinkami ochráni ich osobný prístup. Tisíce webov, ktoré nasledujú prístup Amazonu, dokážu odporučiť produkty, obsah, priateľov a skupiny, bez toho aby autori týchto systémov vedeli, prečo majú užívatelia takéto záujmy (Mayer-Schonberger V. – Cukier K. 2014).

1.2 Základné charakteristiky



Obrázok č. 6: Základné charakteristiky „veľkých dát“

Zdroj:https://www.promptcloud.com/blog/The-4-Vs-of-Big-Data-for-Yielding-Invaluable-Gems-of-Information

Napriek tomu, že ohľadne presnej definície Big Data existuje množstvo dohadov, ustálili sa niektoré charakteristiky, ktoré majú všetky Big Data spoločné, ktoré ich odlišujú od tých bežných, zvládnuteľných dát. Väčšina zdrojov uvádza charakteristiky Volume (teda množstvo dát), Velocity (rýchlosť akou dáta pribúdajú) a Variety (rôznu podobu a formát dát). K týmto trom základným charakteristikám (3V) spoločnosť IBM identifikovala ešte jednu - Veracity, charakteristiku, ktorá predovšetkým rieši úplnosť a dôveryhodnosť spracovávaných dát. Ďalšie charakteristiky môžu postupne pribúdať, čo je dôkazom toho, že proces poznávania Big Data je ešte len na začiatku (O’Reilly. 2012).

1.2.1 Volume – Objem

Charakteristika objemu reprezentuje celkové množstvo dostupných dát. To môže byť tvorené historickými záznamami zákazníka a všetkými jeho transakciami, dátami zo sociálnych sietí, internetu celkovo a pod. Nemôžeme zabudnúť na fakt, že veľké percento tohto objemu je iba šum, čo znamená, že ide o neužitočné informácie, ktoré neprinášajú žiadnu reálnu hodnotu a ktoré potenciálne môžu skresliť výsledky.

Klasický doterajší prístup pri analytickej práci nad dátami bol pomocou vzorkovacích algoritmov vybrať určitú reprezentatívnu podmnožinu dát a na nej testovať určité hypotézy, či robiť štatistiky. Z takto získaných výsledkov boli často odvodené výsledky (aký produkt si zákazník nabudúce kúpi a podobne). Kľúčovým bodom v tomto štandardnom procese je pritom už samotný výber vzorky, čo predstavuje zložitú operáciu. Je vysoko pravdepodobné, že takmer každé vzorkovanie zavedie určitú chybu alebo skreslenie, ktoré môže ovplyvniť s veľkou pravdepodobnosťou výsledok. Rozdiel v prístupe Big Data v tejto oblasti je, že dokáže spracovať všetky dáta bez ohľadu na ich množstvo pri čom si môžeme byť istí, že výsledok je naozaj najlepší možný a zodpovedá realite.

Avšak táto charakteristika kladie na platformu Big Data vysoké požiadavky na pomer ceny a efektivity ukladania dát, ich bezpečnosti ale primárne rýchlosti prevedenia jednotlivých procesov (algoritmov). Pretože nebolo možné efektívne využiť technológiu ukladania dát, vznikol čisto pre potreby spracovania Big Data nový systém Hadoop, ktorý vychádza z pôvodného Google File Systému, na ktorého základe stojí väčšina súčasných open source ale aj komerčných Big Data riešení. Charakteristika objemu je ale vysoko subjektívna a nedá sa presne určiť, aké množstvo dát je považované za Big Data. Hranica medzi normálnym množstvom dát a „veľkými dátami“ je tenká a neustále sa pohybuje. Konečné číslo sa prakticky pre každú organizáciu odlišuje a je dané aj sektorom, v ktorom firma pôsobí, či hardwarovými a softwarovými nástrojmi, ktoré k ich analýze používa. Táto hranica sa bude neustále vyvíjať a posúvať s rozvojom technológií a možností analytických nástrojov, ktoré k ich analýze využíva. Dáta, ktoré sme v minulosti mohli klasifikovať ako Big Data, nie sú Big Data dnes a čo všeobecne považujeme za Big Data dnes, môže byť pomerne ľahko zvládnuteľný a rozumný objem dát za pár rokov. Všeobecne považujeme za Big Data tie dáta, ktorých objem je od niekoľko terabajtov po desiatky petabajtov (Mayer-Schonberger V. – Cukier K. 2014).

1.2.2 Velocity - Rýchlosť

Rýchlosť je myslená predovšetkým ako celková doba od vytvorenia novej informácie cez jej získanie, spracovanie až k samotnému rozhodnutiu. Aspekt rýchlosti ku spracovaniu dát pribudol hlavne v posledných rokoch a s rozvojom technológií bude nadobúdať stále väčší význam. Dnes by sa už bez tohto atribútu v určitých prípadoch nedalo zaobísť, pretože niektoré druhy businessu (automatizované systémy obchodovania na burze a pod.), ktoré sú nútené spracovávať informácie v reálnom čase a musia sa okamžite rozhodnúť o ďalšom kroku.

Reálnym príkladom môže byť situácia, kedy operátor sleduje polohu mobilného telefónu podľa jeho signálu. Jeho majiteľ -klient- akurát prechádza okolo obchodu so športovými potrebami, na ktoré mu predajca chce poskytnúť zľavu v prípade, že okamžite príde do predajne a niečo kúpi. V túto chvíľu sú možné dva scenáre. V prvom prípade operátor stihne zachytiť polohu zákazníka (ideálne, keď sa bude k predajni blížiť) a včas ho informovať o ponuke. V druhom prípade operátor túto aktivitu uskutoční až retrospektívne a o ponuke bude informovať s oneskorením. Potenciálny zákazník, ktorý však dostane túto ponuku oneskorene, sa ale do obchodu pravdepodobne nevráti a predajca stratí príležitosť predať svoj výrobok. Súčasné algoritmy pre automatické obchodovanie na burze pracujú v rozpätí milisekúnd a uzatvárajú obchody za dobu kratšiu, ako človek stihne kliknúť na myš. Milisekunda rozdielu pri prevádzaní príkazu v tomto prípade znamená zisk alebo stratu.

1.2.3 Variety - Rozmanitosť

Kľúčovým rozdielom Big Data je predovšetkým rozmanitosť informácií, ktorých sa celý koncept týka. Tradičné dáta, s ktorými sa pri bežnej činnosti pracuje, majú pevne stanovenú štruktúru (označujú sa pojmom štruktúrované). Typickým príkladom sú dáta ukladané do databázových štruktúr. Počas posledných rokov ale začal prevládať neštruktúrovaný obsah a semištrukturovaný obsah.

Neštruktúrovaný obsah môže mať veľa podôb a len ťažko by sa dal do detailu definovať. Do tejto kategórie zaraďujeme väčšinu moderných formátov od obrazových alebo zvukových dát, príspevkov na sociálnych sieťach, záznamy kliknutí na weboch, dáta dostupné na internete a ďalšie. Všetky tieto formáty majú spoločný fakt, že nemajú pevne stanovenú štruktúru. Často obsahujú väčšie množstvo textových informácií, ktoré nemusia byť vo forme zrozumiteľných viet ale môžu to byť rôzne logy zo zariadení, záznamy

pohybov mobilných telefónov, kódy, rôzne signály a pod. Práve logy sú príkladom informácií, ktoré sa často označujú ako semištrukturované.

Semištrukturované dáta môžu mať v niektorej časti určitú štruktúru (dajú sa predvídateľne čítať), ale väčšiu časť tvorí pravdaže neštruktúrovaný text.

Úlohou Big Data je všetky tieto informácie vyťažiť a zjednotiť do podoby vhodnej ďalšieho spracovania dát a ideálne oddeliť šum, či zbytočné údaje od tých dôležitých.

1.2.4 Veracity - Vierohodnosť

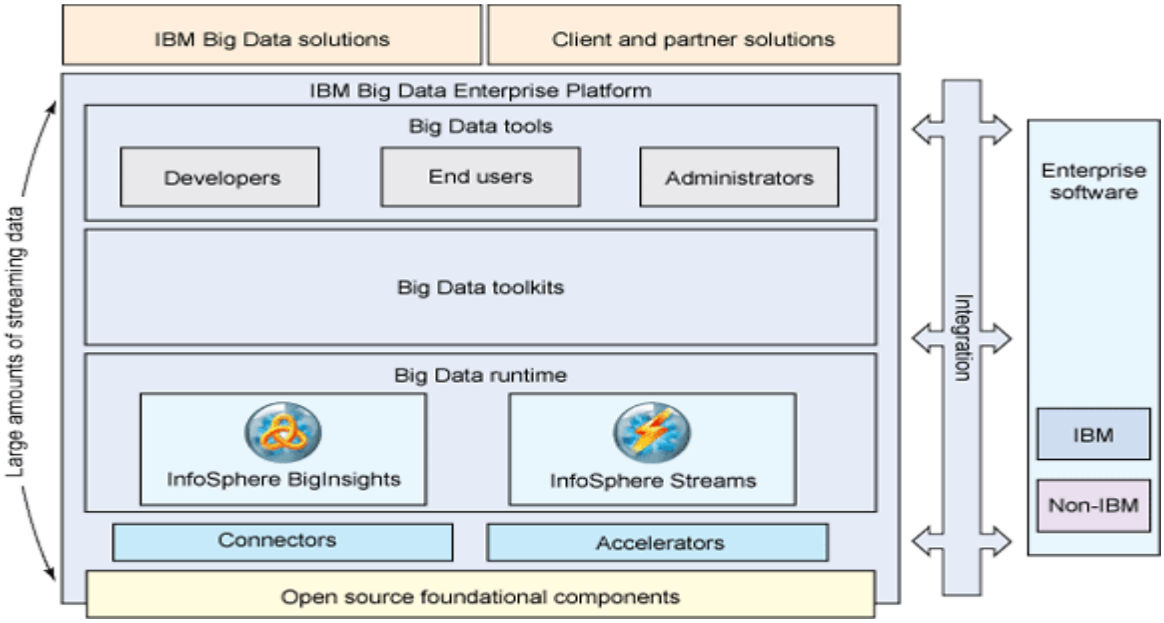
Spoločnosť IBM definovala ešte jednu Big Data charakteristiku, ktorá vznikla predovšetkým na základe skúsenosti z praxe. Táto charakteristika je dôležitá a jej význam rastie spolu s rastúcim množstvom dát (O'Reilly, 2012).

V tradičných databázových systémoch je venovaná značná pozornosť príprave a predspracovaní dát predtým, ako vstúpia do samotného systému a analýz. Systém následne predpokladá, že tieto dáta, ktoré sa sem dostali, sú vierohodné, vyčistené, kompletné, že záverom z nich získaných môžeme dôverovať. Pokiaľ berieme do úvahy v kontexte Big Data myšlienku, že dáta čerpáme z rôznych sociálnych sietí, webových článkov a iných podobných zdrojov, je logické, že dôveryhodnosť týchto dát môže z pohľadu analytika klesnúť. V skutočnosti ale dve zo spomínaných charakteristík Big Data – rýchlosť a rozmanitosť pracujú proti vierohodnosti. Pretože pokiaľ v reálnom čase spracúvame naozaj veľké množstvo dát z rôznych zdrojov, nie je na ich dôkladné čistenie alebo filtrovanie veľa času (aj z dôvodu rýchlosti hardwaru).

Avšak nemusí to znamenať nevýhodu ale je potreba si túto skutočnosť uvedomiť. Je potrebné rozlišovať dva pojmy, ktoré sa často v súvislosti s Big Data môžu pliesť – kauzalita a korelácia.

Kauzalita je príčinná súvislosť, zatiaľ čo korelácia znamená, že dva druhy dát spolu nejakým spôsobom úzko súvisia. Pomerne ľahko ide odhaliť koreláciu pomocou matematického aparátu, pričom na odhalenie kauzality je potrebné predovšetkým porozumieť kontextu Big Data. Môžeme uviesť príklad, kde došlo k skresleniu výstupov z dát, ktoré nastalo okolo hurikánu Sandy v USA na konci roka 2012. Medzi 27. októbrom a 1. novembrom 2012 o tomto hurikáne vzniklo viac ako 20 miliónov tweetov na sociálnej sieti Twitter. Ich spätná analýza ukázala, že najviac ich pochádzalo z Manhattanu (centrum New Yorku), čo by naznačovalo, že táto lokalita bola najviac zasiahnutá hurikánom. Skutočný dôvod je však úplne iný. Viac tweetov z danej oblasti bolo zrealizovaných

z dôvodu, že miestne obyvateľstvo malo viac mobilných telefónov a pripojenie na internet oproti viac postihnutým hurikánom, ktorým naviac vypadával prúd a mobilné siete (Mayer-Schonberger V. – Cukier K. 2014).



Obrázok č. 7: Návrh schémy platformy IBM pre spracovanie problematiky Big Data

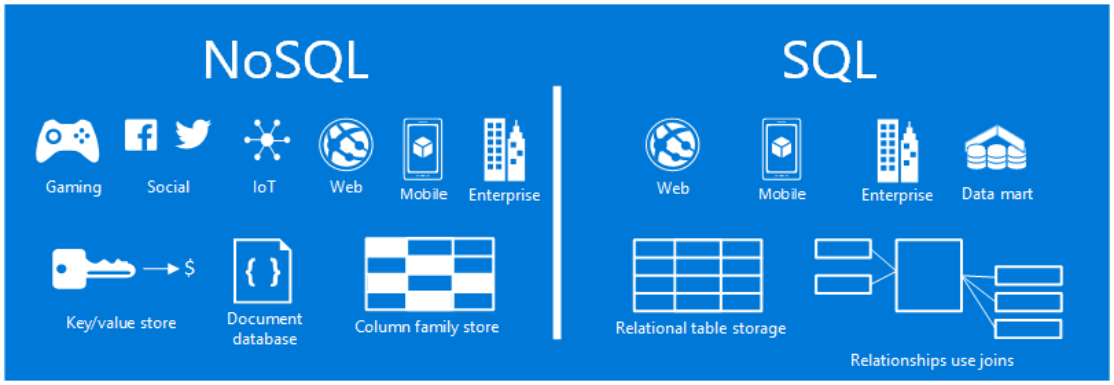
Zdroj : http://www.ibm.com

1.3 NoSQL

NoSQL v preklade znamená „nie len SQL“ (not only SQL). NoSQL nepredstavuje konkrétnu technológiu alebo riešenie, je všeobecným pojmom, v ktorom sú zahrnuté všetky riešenia, ktoré nie sú založené na báze relačného modelu.

Pojem NoSQL vznikol v internetových gigantoch Facebook, Google a Amazon, ktorí mali ťažkosti so spracovaním veľkých objemov dát tradičnými RDBMS.

Na základe nimi získaných poznatkov boli vytvorené ďalšie NoSQL databázy. Jednoduchý prehľad využiteľnosti NoSQL databáz a tradičných relačných databáz je znázornený na obrázku č. 8 (Kosek J. – Holubová I. – Minarik K. – Novák D. 2015).



Obrázok č. 8: Porovnanie platformy SQL a NoSQL

Zdroj : https://docs.microsoft.com/en-us/azure/documentdb/documentdb-nosql-vs-sql

Spracovávané dáta v NoSQL databázach môžu byť neštrukturované, štrukturované alebo semištrukturované. Tieto databázy využívajú schopnosť načítať a ukladať veľké množstvo dát bez závislosti na vzájomných vzťahoch.V tomto prípade sú dáta uložené redudantným spôsobom na niekoľkých serveroch. Týmto spôsobom je prakticky možné ľahko škálovať systém, pridaním ďalších serverov, čím vzniká tolerancia k prípadným výpadkom.

Rozdiely medzi NoSQL dátovým modelom a relačným modelom sú :

- V relačnom dátovom modeli je dôraz kladený na štruktúru dostupných dát – Aké odpovede mám ?
- V NoSQL modeli je dôležitá forma prístupu k dátam – Aké otázky mám?
- NoSQL model vyžaduje hlbšie znalosti štruktúr a algoritmov ako tradičné relačné databázy.
- Relačné databázy nie sú vhodné pre grafické alebo hierarchické modely.

1.3.1 Základné princípy NoSQL modelingu

Prvým pojmom je denormalizácia, v ktorej ide o kopírovanie dát do viacerých tabuliek alebo dokumentov s cieľom uľahčiť vyhľadávací proces.

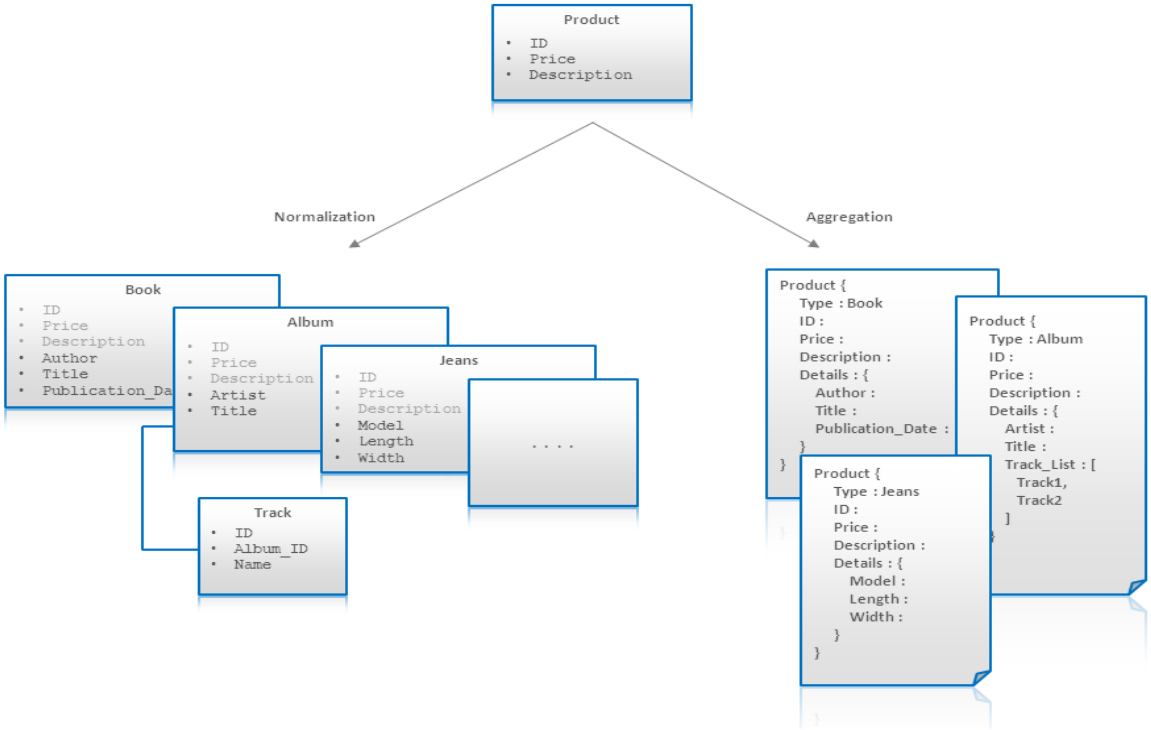
V prípade, že hovoríme o agregácii, hovoríme o zhlukovaní dát. Všetky NoSQL modely používajú tento model.

Úložisko kľúč – databázy grafov (Graph databases) a hodnota (key-value stores) – prakticky nemajú žiadne obmedzenia na dátové hodnoty. Tieto hodnoty môžu byť ukladané

v ľubovoľnom formáte pod jedným kľúčom. To môže byť napr. email, správy, meno pod kľúčom UserID vo formáte UserID_email, UserID_spravy, UserID_meno.

Big Table – podporuje zlučovanie viacerých stĺpcov.

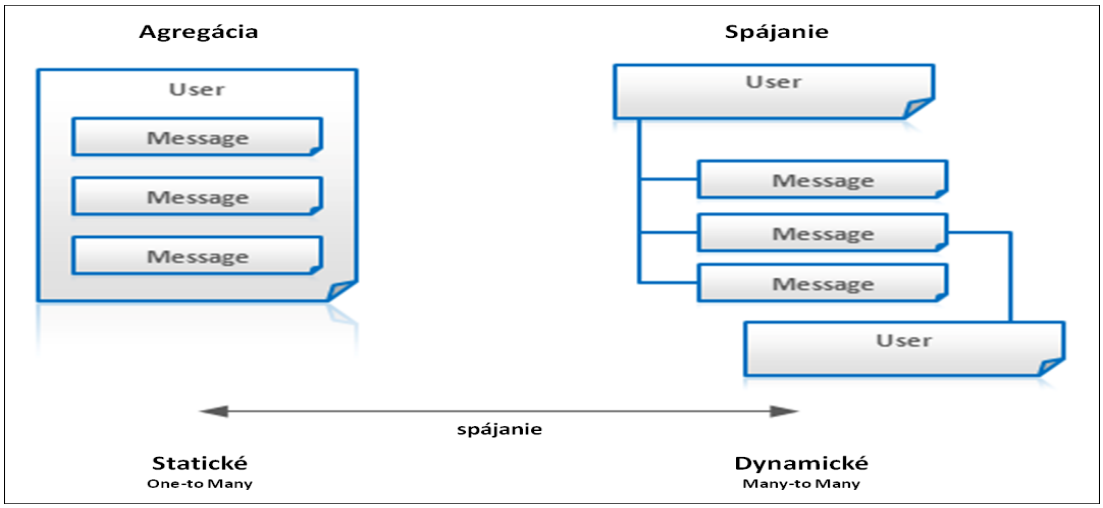
Dokumentovo – orientované – ukladanie dokumentov podľa vopred používateľom zadefinovanej schémy (Kosek J. – Holubová I. – Minarik K. – Novák D. 2015).



Obrázok č. 9: Porovnanie agregácie a normalizácie

Zdroj : vlastná tvorba

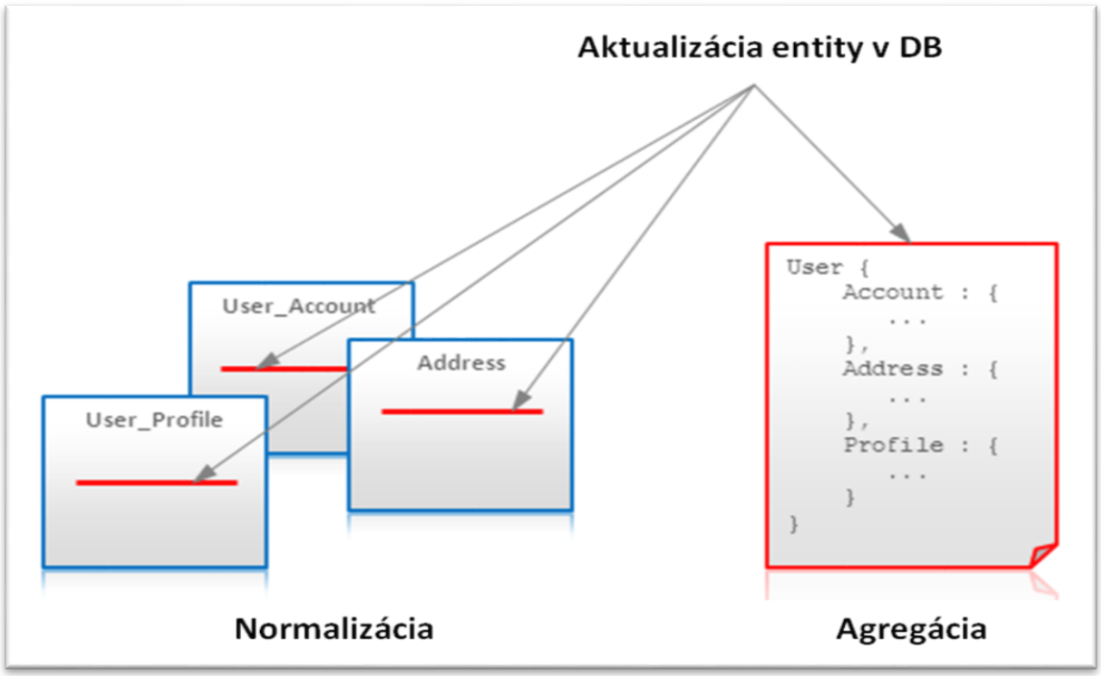
Aplikácia bočného spájania – v NoSQL databázach je zriedkavé spájať dva alebo viacej riadkov do jedného riadku alebo tabuľky. V NoSQL databázach je možné spájanie nahradiť denormalizáciou alebo agregáciou, ale nie vždy to je možné. V takýchto prípadoch je potreba použiť aplikáciu bočného spájania hlavne, keď je entita často modifikovaná pri spájaní veľkého množstva vzťahov. Príklad aplikácie bočného spájania a agregácie je znázornený na obrázku č. 10, kde user je entita, pri ktorej dochádza pomerne často k modifikácii (prichádzajúce nové správy).



Obrázok č. 10: Porovnanie agregácie a spájania

Zdroj : vlastná tvorba

Atomická agregácia – prevažná väčšina NoSQL riešení má limitovaný transakčnú podporu.



Obrázok č. 11: Porovnanie normalizácie a atomickej agregácie

Zdroj : vlastná tvorba

Rozdiel v aktualizácii medzi NoSQL (agregácia) a relačnou databázou (normalizácia) je zobrazený na obrázku č. 11. V relačnej databáze je zvyčajne potrebné vykonať aktualizáciu v niekoľkých tabuľkách, pričom NoSQL dovoľuje uložiť aktualizáciu ako jeden riadok, kľúč alebo dokument pomocou atomickej agregácie.

Kľúče triedenia – jednou z najväčších výhod úložiska kľúč – hodnota je, že dáta sa môžu nachádzať a byť rozdelené na niekoľkých serveroch len za pomoci zatrieďovacieho kľúča.

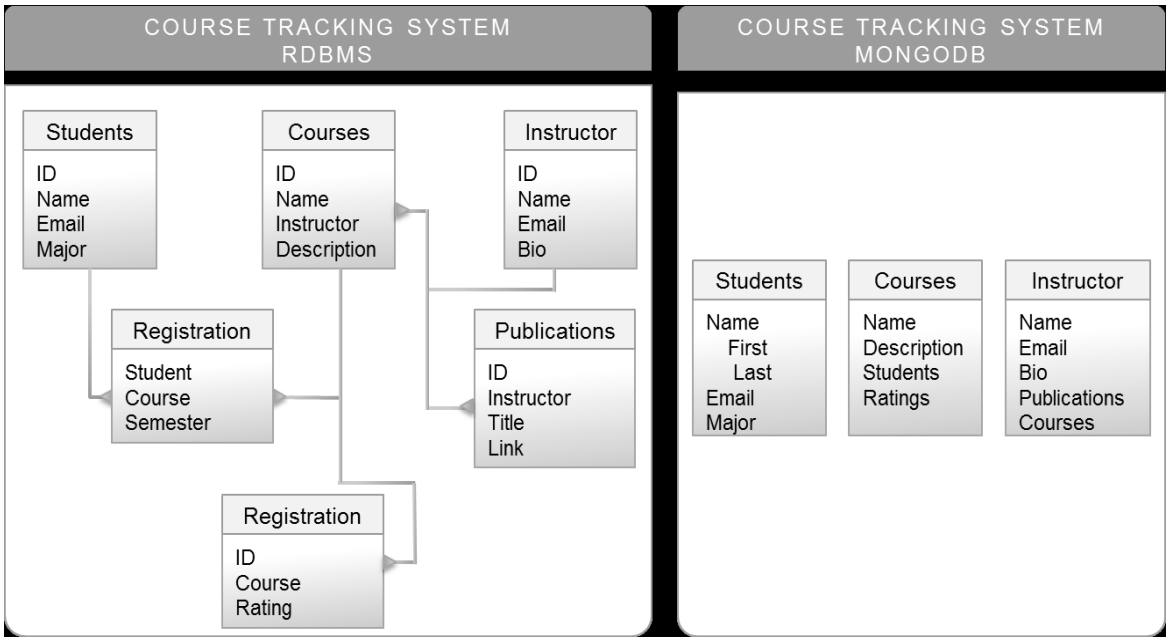
1.3.2 Porovnanie niektorých NoSQL databáz

MongoDB

MongoDB je dokumentovo orientovaná, voľne dostupná databáza. Záznamy v databáze sú vnímané ako dokumenty s vlastným ID a operácie nad nimi sú atomické. Dokumenty sa ukladajú do zbierok. Každá zbierka môže obsahovať ľubovoľné množstvo dokumentov. V tejto databáze sa ukladajú štruktúrované dáta JSON (Java Script Object Notation), ktoré predstavujú dokumenty s dynamickými schémami. Dáta sú ukladané a navonok prezentované formátom BSON, čo znamená prevod JSON do binárnej podoby. Každé políčko môže obsahovať jednu alebo viacej hodnôt, takže dátový model je flexibilný. MongoDb podporuje:

- Ad hoc dotazy – vlastný dotazovací jazyk na štýl SQL.
- MapReduce – implementovaný pre využívanie hromadných operácií nad dátami a využíva JavaScript.
- Indexovanie – každé pole môže byť indexované.
- Replikácia – MongoDB podporuje master-slave replikáciu. Master číta a zapisuje dáta, následne Slave tieto dáta kopíruje z Master, pričom dáta je možné použiť len na zálohovanie alebo čítanie.
- Load balancing MongoDB – podpora horizontálneho škálovania. Zvolí sa fragment – kľúč, ktorý určí spôsob ako budú dáta distribuované. Dáta sú distribuované a rozdelené na viac fragmentov.

MongoDB môže bežať naraz na viacerých serveroch, kvôli vyváženiu zaťaženia.



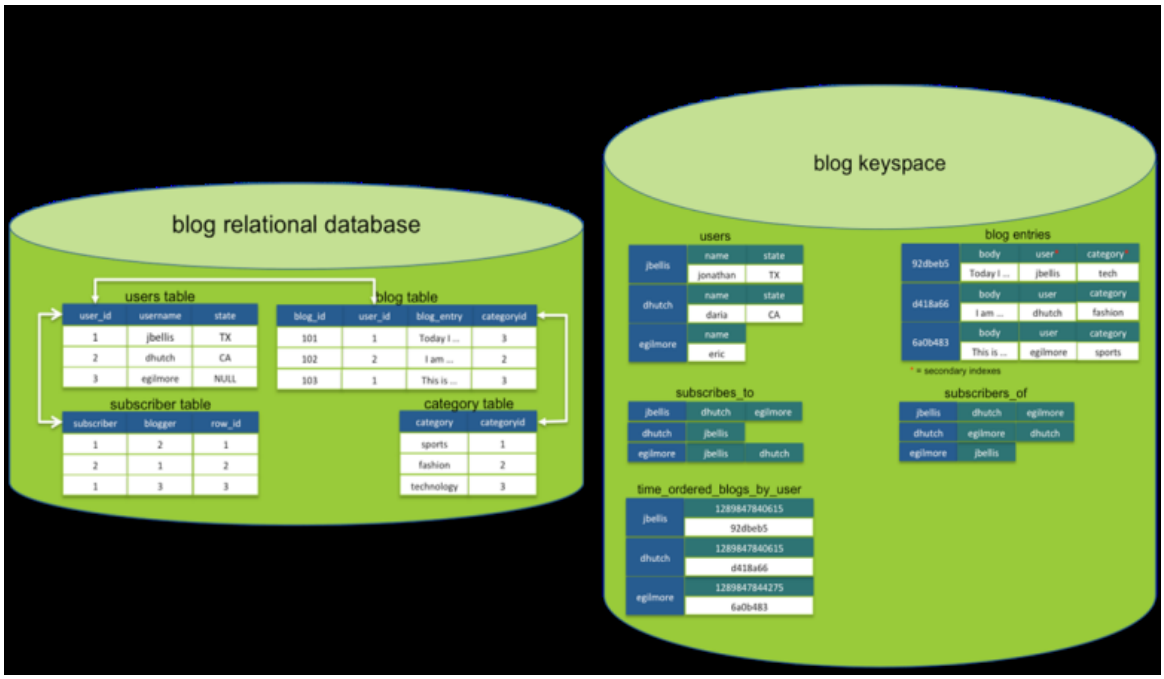
Obrázok č. 12: Porovnanie MongoDB a tradičnej RDBMS

Zdroj : http://software.danielwatrous.com/mongodb-boise-code-camp-2012/

Cassandra

Cassandra je postavená na Google Bigtable a pôvodne išlo o interný projekt Facebooku, z ktorého sa postupne stal voľne dostupný projekt, podporujúci Apache Foundation. V databáze Cassandra sa dáta ukladajú ako viacrozmerné mapy. Každý jeden dátový záznam obsahuje určité hodnoty, ktorým je priradený jednotlivý kľúč. V každej rodine stĺpcov sa nachádza mapa dát. Dáta sa ukladajú do riadkov, ktoré obsahujú viacej stĺpcov a sú zlúčené s kľúčom riadka. Obsahom každej bunky okrem samotnej hodnoty je aj časová známka, ktorá slúži na odhalenie prípadného konfliktu pri zápise.

Databáza Cassandra je určená na spracovávanie veľkých objemov dát cez viac uzlov, bez akéhokoľvek bodu zlyhania. V prípade, že nastane problém zlyhania, Cassandra ho rieši pomocou peer-to-peer distribuovaného systému, v ktorom sú uzly rovnaké a dáta sú medzi ne rozdelené.



Obrázok č. 13: Porovnanie CassandraDB a tradičnej RDBMS

Zdroj : http://docs.datastax.com/en/archived/cassandra/0.8/docs/

1.4 Technológie

Koncept Big Data by ale nefungoval bez potrebného technologického zázemia, na ktoré pravdaže tradičné nástroje už nestačia. Možností zvládnutia Big Data je stále viac a viac, najvýznamnejšia z nich je bezpochybne technológia Hadoop a z nej odvodené ďalšie nástroje.

1.4.1 Apache Hadoop

Ide o open-source softwarový produkt, ktorý je odvodený z Google File System. Aj keď môže byť používaný iba na jednom počítači, jeho ozajstná sila spočíva v schopnosti škálovať dáta na stovky, či tisíce počítačov, každý s niekoľkými procesorovými jadrami. Prvé plné funkcie sa na trhu objavili na konci roka 2007. Od samotného začiatku vývoja technológie Hadoop až do súčasnosti bolo vydaných takmer 50 releasov, ktoré postupne obohacovali a doplňovali jeho funkcionalitu.

Hadoop je primárne navrhnutý pre linuxové platformy tak, aby efektívne spracoval údaje o veľkosti niekoľko stoviek gigabajtov, terabajtov až petabajtov. Hadoop preto disponuje distribuovaným súborovým systémom, ktorý rozdelí objemom veľké vstupné dáta a odosiela podstatne menšie časti (bloky) originálnych dát do niekoľko počítačov v clusteri (skupina

počítačov), kde následne dochádza k paralelnému spracovávaniu prijatých dát a tým k maximálne efektívnemu výpočtu zadanej úlohy.

Ako príklad môžeme uviesť spoločnosť s tridsať miliónmi zákazníkov, ktorá získava údaje z dátových skladov, reportov a záznamov call centier pre vylepšovanie svojich služieb. Operácia, ktorá kedysi trvala týždeň, zaberie použitím novej technológie iba pár hodín. V technológií Hadoop sú zabudované ďalšie softwarové komponenty pre prácu s veľkým objemom dát (White T. 2012).

S rastúcim počtom počítačov rastie aj pravdepodobnosť výskytu chýb. Príkladom prekážok, ktoré sa môžu v distribuovanom prostredí objaviť, môžu byť rôzne verzie klientskeho softwaru alebo dočasná strata pripojenia k sieti. Okrem obavy možných chýb a teda akceptovania niekoľko druhov problémov je potrebné zabezpečiť dostatočne výkonné zdroje. Predovšetkým ide o procesor, pamäť a pevný disk. Hadoop je teda navrhnutý tak, aby efektívne spracoval veľké množstvo informácií vytvorením paralelne pracujúcej komunity počítačov, vzhľadom k tomu, že hypoteticky počítač s tisícom jadier by bol mnohonásobne drahší ako tisíc počítačov s jedným jadrom.

Hadoop a jeho výhody oproti ostatným metódam :

- Jednoduchý programovací model – umožňuje rýchlo písať a testovať distribuované systémy.
- Automatická distribúcia dát – efektívne využíva základné podobnosti procesorových jadier.

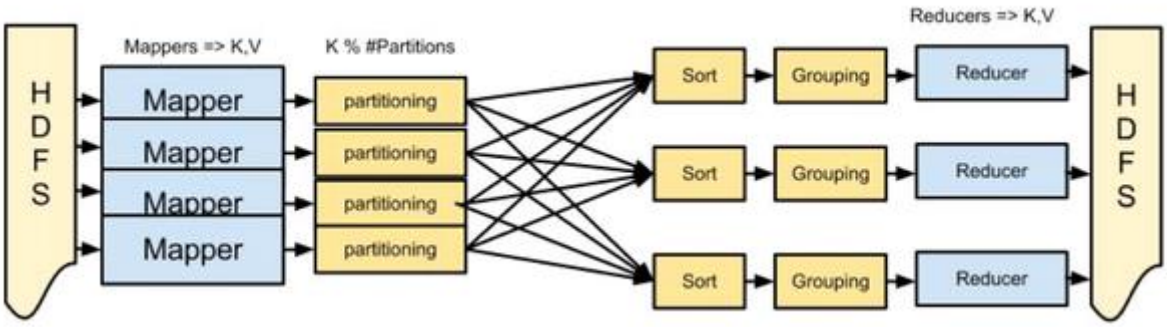
1.4.2 HDFS – Hadoop Distributed File System

Ide o distribuovaný súborový systém navrhnutý k uloženiu a spracovaniu veľmi veľkého objemu dát, ktorý poskytuje klientovi svojou vysokou priepustnosťou prístup ku všetkým uloženým informáciám. Aby bola zvýšená odolnosť voči chybám a možná práca paralelne spusteným aplikáciám, tak sú súbory ukladané redundantným spôsobom.

HDFS je založený na architektúre Google File System (GFS) a je navrhnutý tak, aby bol odolný voči čo najviac problémom. Patria mu tieto vlastnosti :

- Poskytnutie rýchleho, škálovateľného a stáleho prístupu k veľkému množstvu informácií.

- Šírenie dát v rámci veľkého počtu strojov (uzlov) zoskupených do jednotlivých clusterov.
- Cluster odolný voči kompletnému zlyhaniu niekoľkých strojov, bez toho, aby stratil informácie alebo výrazne spomalil výkon.



Obrázok č. 14 : Platforma HDFS

Zdroj: (http://www.bigdatablog.sk/mapreduce/)

Funkcie systému Hadoop

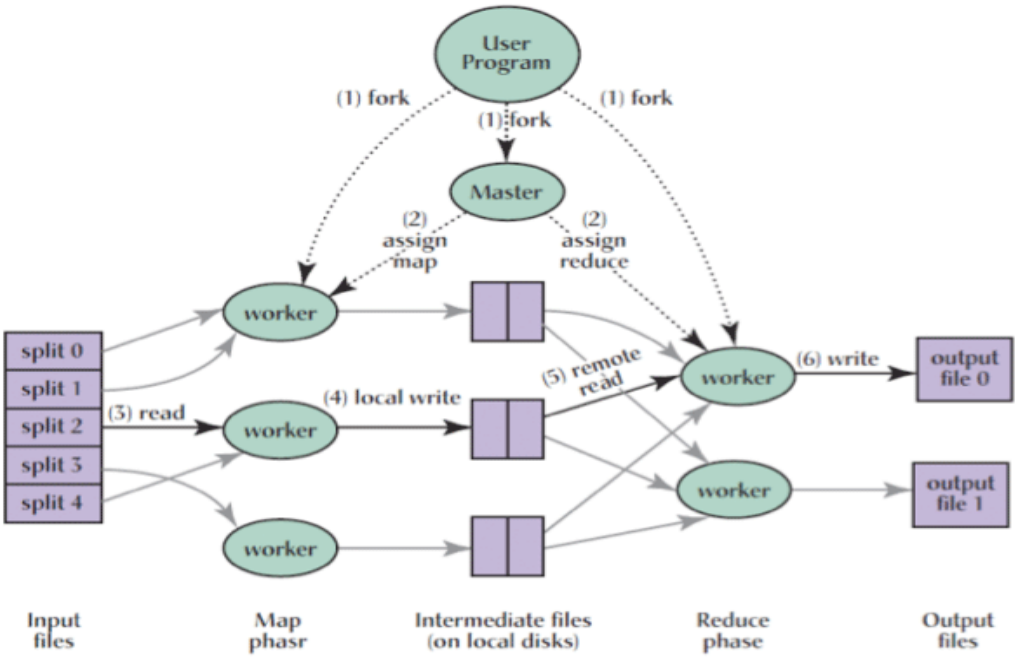
Systém Hadoop je primárne zložený z dvoch základných celkov. Pre ukladanie veľkého objemu dát využíva distribuovanú súborovú databázu a k paralelnému spúšťaniu výpočtových úloh nad uloženými dátami slúži programový rámec MapReduce. Na ukladanie dát do databázy a správu dohliada master uzol NameNode, zatiaľ čo daemon JobTracker má za úlohu spracovanie dát. Ich príkazy vykonávajú podriadené slave uzly DataNodes pre NameNode a podriadené slave deamons TaskTrackers pre JobTracker. Hadoop preferuje rýchlosť spracovania, ale aj veľký stupeň odolnosti proti chybám. Úlohové uzly TaskTrackers sú v neustálej komunikácii s hlavným uzlom JobTracker, ktorý má prehľad o rozdelení úloh a právomoc k prípadnému reštartu uzlu, či k vyradeniu v prípade jeho zlyhania.

1.4.3 MapReduce

MapReduce je programovací model navrhnutý pre spracovávanie veľkého objemu dát paralelným spôsobom. Podľa presne definovaných štýlov spúšťa aplikácie, ktoré bežia v Hadoop prostredí. Vstupné dáta sú pomocou MapReduce funkciami transformované na výstupné dátové prvky. Pri spracovávaní dát programovacím modelom MapReduce sú teda použité dve na seba nadväzujúce metódy a to Map a Reduce.

Funkcia fázy Map spočíva v postupnom transformovaní každého zo vstupných dátových prvkov na výstupné dátové prvky. Ako príklad užitočného využitia metódy Map môžeme uviesť funkciu toUpper (str), ktorá vytvorí zo vstupného textového reťazca nový text s čisto veľkými písmenami.

Ďalšia fáza Reduce spojuje vstupné dátové prvky v jednu výstupnú hodnotu podľa zadaného parametra. Bežným využitím metódy Reduce je sumarizácia vstupných dát, kde funkcia Reduce vráti celkový súčet zadaných hodnôt v podobe jedného čísla.

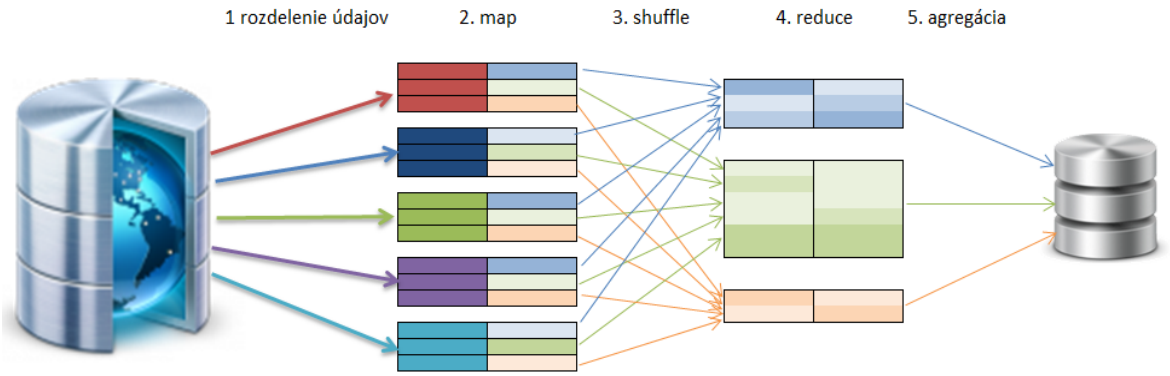


Obrázok č. 15: Schéma spracovávanej úlohy programového modelu MapReduce

Zdroj: (MapReduce: Simplifed data processing on large clusters, 2008, s. 3)

Spojenie metódy Map a Reduce

Architektúra MapReduce využíva obe metódy a rozširuje ich možnosť pre spracovávanie veľkého množstva dát. Funkcie Map a Reduce nepracujú iba s hodnotami ale vždy s dvojicou hodnota – kľúč. Každá z obidvoch funkcií berie ako vstup dvojicu hodnota – kľúč a poskytuje ju rovno ako výstup. Funkcia Reduce zjednocuje vstupné prvky do jednej výstupnej hodnoty. Obvykle sa nachádza na výstupe z modelu MapReduce viac výstupných prvkov podľa definovaných kľúčov. Všetky hodnoty s rovnakým kľúčom sú agregované nezávisle na ostatných a sú súčasťou konečného výstupu modelu MapReduce.

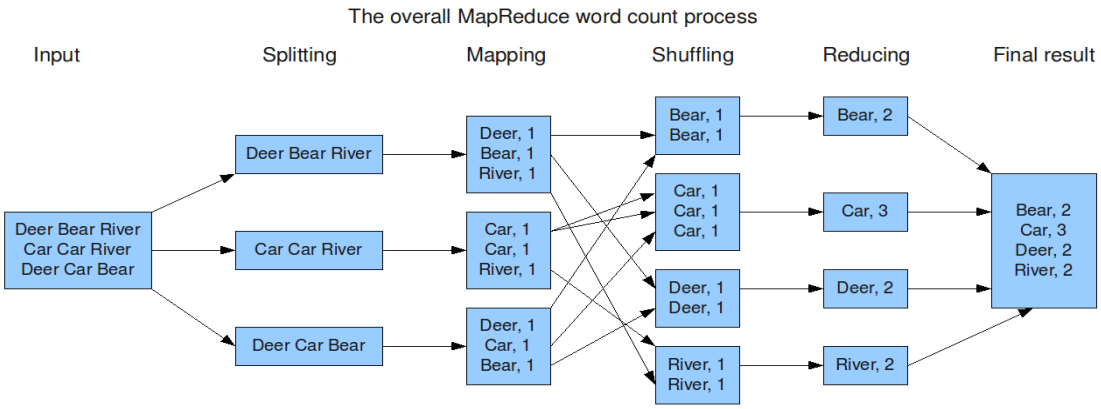


Obrázok č. 16: Spojenie modelu Map a Reduce

Zdroj: (http://www.bigdatablog.sk/mapreduce/)

Dátový tok modelu MapReduce

Vstupná vrstva vchádzajúca do modelu MapRaduce obyčajne pochádza priamo zo súborov načítaných do HDFS a je rovnomerne rozdelená na všetky uzly v clusteri. Spustenie programu MapReduce preto vyvolá zahájenie map úloh všetkých uzlov. Každý vstupný súbor je spracovávaný ľubovoľným uzlom. Následne sa prechádza k jedinému komunikačnému kroku v modele MapReduce. Nedochádza k žiadnej vzájomnej výmene informácií medzi map úlohami, ako aj medzi Reduce úlohami, ale dochádza k výmene vzniknutých dvojíc kľúč-hodnota, aby každá inštancia funkcie Reduce obdržala všetky hodnoty s rovnakým kľúčom. Pokiaľ dôjde k zlyhaniu niektorého uzla v clusteri, spracovávaná úloha sa po reštarte spustí znovu.



Obrázok č. 17: Dátový tok modelu MapReduce

Zdroj : https://www.pcrevue.sk/a/Hadoop---overena-sucast-rieseni-na-big-data

1.4.4 Programovacie jazyky v spolupráci so systémom Hadoop

Hadoop systém je vytvorený v programovacom jazyku Java, preto v prvom rade spúšťa programy Map a Reduce s kódom napísaným v Jave. Okrem toho Hadoop prichádza s adaptačnými vrstvami pipes a streaming, ktoré umožňujú rámcu MapReduce spúšťanie kódu napísaného v inom programovacom jazyku. Pipes sa dá označiť za knižnicu, vďaka ktorej môžeme používať programovací jazyk C++. Streaming obsahuje aplication programming interface (API) niekoľko ďalších programovacích jazykov, medzi ktoré patria napr. Bash, Perl alebo Python. Pre spracovávanie veľkého objemu priestorových dát a implementáciu Map a Reduce sa využíva práve jazyk Python.

Python pre dátovú analytiku

Pre veľa používateľov je Python ľahko pochopiteľný, prehľadný a obľúbený programovací jazyk. Od prvého nasadenia v roku 1991 sa Python začlenil k najpopulárnejším dynamickým programovacím jazykom ako sú Perl, Ruby a iné. Niektoré jazyky sú často nazývané skriptovacie jazyky, ktoré môžu byť využívané na písanie rýchlych, malých programov alebo skriptov. Python sa vyznačuje rozsiahlou, aktívnou vedeckou výpočtovou komunitou.

Pre dátové analýzy, interaktívne výpočty a dátové vizualizácie sa Python podobá rôznym open source a komerčným programovacím jazykom a nástrojom ako sú R, MATLAB, SAS, Stata a iné. Python implementuje knižnice (primárne pandas) podporujúce nástroje pre manipuláciu s dátami. Môžeme povedať, že Python je jednoduchý jazyk a výborná voľba čo sa týka otázky pre zostavenie data-science aplikácie (McKinney. 2013).

1.5 Vizualizácia

Hovorí sa, že lepšie je raz vidieť ako stokrát počuť. Taktiež to platí pre písaný text. Ľudská pamäť a ľudský mozog spracúvajú obrázky lepšie a oveľa rýchlejšie ako hovorené slovo a text. Kvôli tomu sa analytici pôsobiaci v problematike „veľkých dát“ snažia svoje výstupy práce prezentovať najmä vizuálne. Hlavným cieľom musí byť informatívna stránka danej vizualizácie. Kľúčovou podmienkou je, aby nám poskytla čo najväčšie množstvo plnohodnotných informácií, ktoré potrebujeme pochopiť. Preto musí byť vizualizácia jasná a priama a v neposlednom rade musí byť estetická, aby používateľ dospel rýchlejšie k potrebnému záveru. Pri zostavovaní vizualizácie je podstatným faktorom voľba parametrov ako sú veľkosť, pozícia, farba a podobne, ktoré slúžia ako návod. Správny výber

parametrov okrem toho, že prispieva k estetickému zážitku, napomáha ku bezproblémovému prezentovaniu informácií, ktoré potrebujeme ukázať. Pomocou parametrov ukazujeme, čo je dôležité a na čo sa sústrediť.

1.5.1 Proces vizualizácie

Prvý krok vizualizácie spočíva v zistení jednotlivých požiadaviek zúčastnených strán. Keď presne poznáme predmet záujmu a aké otázky je potrebné zodpovedať, vieme určiť level a typ analýzy. Na začiatok je potrebné získať podľa možností čo možno najkvalitnejšie a najrelevantnejšie dáta. Množstvo a kvalita dát závisia od dostupného rozpočtu a času. Dáta sú spravidla poskytované od zúčastnených strán, ale taktiež môžu byť zakúpené od hlavných poskytovateľov alebo získané z vlastných zdrojov. Získané dáta sú spravidla uložené v tabuľkovom formáte, v textovom formáte alebo sa nachádzajú v databáze. Nie vždy sú tieto dáta pripravené na konkrétnu analýzu. Z toho dôvodu je často potrebná ich údržba pred tým, ako začneme so samotnou analýzou, aby všetky dáta mali konzistentný formát. Ak dáta pochádzajú z viacerých zdrojov, je dosť možné, že bude potrebné prekonvertovať niektoré jednotky. Napríklad americké a európske dáta môžu byť rozdielne, čo sa týka metrického systému alebo spôsobu zápisu dátumu a času a meny, ktorú bude potrebné prekonvertovať. Keď dostaneme dáta do požadovanej podoby, musíme vyfiltrovať poškodené a nepoužiteľné dáta, aby sme dostali čisté a konzistentné dáta. Dáta vložíme do zvolenej vizualizácie a doplníme, titulky, legendu a všetko potrebné pre efektivitu zobrazenia a sprehľadnenie. Ako posledný krok budeme prezentovať výsledok a umožníme interakciu z vizualizáciou. Zviditeľnenie jednotlivých prvkov, zmena farby, zoom, výber premenných, zväčšovanie, toto všetko umožňuje človeku, ktorý príde do kontaktu z vizualizáciou vylepšiť jeho dojem (Fry B. 2007).

2 Cieľ práce, metodika práce a metódy skúmania

2.1 Cieľ práce

Cieľom práce je naštudovať problematiku Big Data a získané poznatky demonštrovať na riešení v praxi. Na dosiahnutie tohto cieľa treba dosiahnuť rad podcieľov, ktorými sú hlavne:

- Charakteristika spracovania témy Big Data so zameraním na podrobný popis platformy Big Data, ich vlastností a špecifikácií.
- Uviesť základné princípy ich modelovania a technológií spracovania.
- Porovnať jednotlivé technológie, systémy spracovania, ich odlišnosti, výhody a nevýhody.
- Na základe týchto získaných znalostí, vypracovať projekt, týkajúci sa problematiky Big Data v sektore telekomunikácií.
- Na vybranej vzorke dát z praxe uskutočniť analýzu. Analýzu budeme vykonávať na základe vopred určených kategórií s cieľom demonštrovať možnosti spracovania Big Data pre potreby rozhodovania.

2.2 Metodika práce a metódy skúmania

Na základe preštudovania danej problematiky v prvej časti práce sa zaoberáme základnými problémami, charakteristikami a technikami v oblasti Big Data.

Pri spracovaní budeme primárne pracovať s jazykom Python, v ktorom využívame na prácu s dátami hlavne package pandas a flask. Použijeme tiež jazyk JavaScript, pomocou ktorého zabezpečíme vizualizáciu výsledkov analýzy dát. Jazyk HTML využijeme na komunikáciu so serverom, na ktorom sa bude vizualizácia danej analýzy nachádzať.

Na vybranej vzorke dát uskutočníme pomocou vyššie spomenutých nástrojov spracovanie. Získané výsledky budeme interpretovať vzhľadom na danú oblasť rozhodovania. Samozrejme, že v praxi sa môžu vyskytnúť rôzne iné problémy, ale cieľ a rozsah práce nevyžaduje sa venovať každému jednému aspektu tejto problematiky v plnom rozsahu.

Ďalším dôležitým faktorom sú prax a skúsenosti, preto sme sa rozhodli danú problematiku prediskutovať s odborníkom[1] pôsobiacim v problematike Big Data z oblasti

[1] Jozef Bilý – Big Data Business Architect

telekomunikácií. Týmto spôsobom sme mohli získať poznatky, ku ktorým by sme sa len ťažko dostali a v nasledujúcej kapitole ich krok za krokom budeme spájať s teoretickými poznatkami z predošlej kapitoly.

Na základe týchto získaných znalostí, bude v nasledujúcej časti vypracovaný projekt, týkajúci sa problematiky Big Data v sektore telekomunikácií. Spracovanie „veľkých dát“ budeme demonštrovať na konkrétnom príklade zo získaných dát za mesiac január, február a marec r. 2017.

3 Výsledky práce a diskusia

V tejto časti detailne popisujeme praktický prístup k spracovaniu a reálnemu využitiu „veľkých dát“. Pomocou zvolenej technológie budeme analyzovať dáta, ktoré sme získali z odvetvia telekomunikácií. Pre prezentáciu výsledkov vytvoríme jednoduchú web aplikáciu, kde budú prezentované výsledky analýzy a kde preukážeme, že aj tak komplikovaná problematika, akou analýza rozsiahlych klientskych dát bezpochyby je, sa dá vizualizovať prehľadne a zrozumiteľne na jednej obrazovke tak, aby bežný používateľ okamžite pochopil danú tému.

3.1 Príprava dát

Pri spracovaní platformy Big Data na komerčné účely prestáva byť konfrontácia osobných údajov ako nového platidla digitálnej ekonomiky iba frázou a stáva sa skutočnosťou. Komerčný potenciál Big Data je neuveriteľný, čoho dôkazom je fakt, že korporácie, ktoré stoja za najúspešnejšími sociálnymi sieťami a internetovými vyhľadávačmi sú súčasnými víťazmi na stupňoch globálnej ekonomiky. Existuje veľa príkladov rôznych analýz datasetov, ktoré nebudú generovať a ani využívať žiadne osobné údaje (napr. údaje o celosvetovej klíme a počasí, údaje o hustote dopravy, či GPS dáta lodných kontajnerov určených na prevoz tovaru apod.), ale zároveň existuje aj mnoho datasetov využívaných v rámci analýzy tzv. „veľkých dát“, ku ktorým je potrebné pristupovať ako k osobným údajom (napr. dáta zo zdravotníckych monitorovacích prístrojov priradených ku konkrétnemu pacientovi, lokalizačné a prevádzkové údaje spracúvané mobilnými operátormi, či rozsiahle dáta analyzujúce spotrebiteľské správanie zákazníka a alebo člena vernostného programu). Ak takýto prevádzkovateľ pri analýzach Big Data využije externých partnerov (napr. HR agentúry, špecializované marketingové spoločnosti alebo nové typy technologicko-analyticko-poradenských organizácií schopných analyzovať Big Data pomocou vlastných unikátnych algoritmov a prevádzkovateľov rôznych elektronických úverových registrov), ktorí vykonajú profilovanie dotknutej osoby de facto namiesto prevádzkovateľa, tak tieto subjekty budú mať spravidla postavenie sprostredkovateľov a tiež budú musieť pri zabezpečovaní ochrany osobných údajov splniť časť regulačných povinností. Sprostredkovateľ a poskytovateľ zároveň budú musieť takéto spracovávanie osobných údajov dostatočne právne legitimizovať takým spôsobom, ktorý presne stanovuje reguláciu definovania požiadaviek na obsahové náležitosti súvisiaceho

zmluvného vzťahu (aktuálne § 8 ods. 4 Zákona o ochrane osobných údajov, od 25. mája 2018 článok 28 ods. 3 GDPR[2]).

3.2 Výber dát

Dáta pre túto prácu sme čerpali zo sektora telekomunikácií, kde je problematika Big Data dosť rozšírený a využívaný pojem. Pre túto prácu sme vybrali agregované dáta zaznamenané za mesiace január, február a marec r. 2017. Z týchto dát dokážeme zistiť všetky potrebné informácie o stave telefonovania na území Slovenskej republiky. Tieto dáta obsahujú informácie ako sú vek zákazníka, kraj, kde je hovor uskutočnený, značka mobilného telefónu zákazníka, dĺžka hovorov a prípadné zlyhanie hovorov z technickej príčiny.

3.2.1 Sample dáta a realita

Pre túto diplomovú prácu bola vytvorená dátová vzorka, ktorá nie je reálna, no v niektorých hodnotách nie je ďaleko od reality ľubovoľného mobilného operátora na Slovensku.

Keďže sme pracovali s agregovanými dátami, v tejto kapitole popíšeme, ako sa vytvára podobný pohľad v reálnom telekomunikačnom svete.

Zdrojom pre takýto pohľad sú detailné informácie o hlasovej, sms a dátovej prevádzke, ktoré v telekomunikačnom svete poznáme ako CDR (call detailed records), ktoré si v krátkosti popíšeme. Realita je omnoho komplexnejšia, ale pre pochopenie nášho výstupu sa sústredíme na najdôležitejšie atribúty:

- **hlasové CDR** –
  - telefónne číslo volajúceho,
  - telefónne číslo volaného,
  - presný čas,
  - typ hlasového hovoru napr. v rámci siete operátora; zahraničný operátor; roaming do cudzej siete,
  - dĺžka trvania hovoru v sekundách,

[2] Nariadenie Európskej únie o ochrane fyzických osôb a ich údajov

- geolokalizácia – na ktorej geo bunke bol hovor uskutočnený,
- IMEI volajúceho - International Mobile Equipment Identity; jedinečný kód používaný k identifikácii každého mobilného telefónu,
- informácia o padnutom alebo dokončenom hovore,
- typ siete – 2G; 3G; 4G,
- atď.

➢ ***SMS CDR –***

- telefónne číslo odosielateľa,
- telefónne číslo prijímateľa,
- presný čas,
- typ sms smeru napr. v rámci siete operátora; zahraničný operátor; roaming do cudzej siete,
- počet znakov,
- IMEI volajúceho - International Mobile Equipment Identity; jedinečný kód používaný k identifikácii každého mobilného telefónu,
- typ siete – 2G; 3G, 4G,
- atď.

➢ ***Dátové CDR –***

- telefónne číslo,
- presný čas,
- počet kB uplink,
- počet kB downlink,
- typ siete – 2G; 3G; 4G,
- atď.

V tejto práci sme sa sústredili na hlasové CDR. Aby bolo možné vytvoriť takúto agregáciu, je nutné spojiť detailné hlasové CDR s informáciami o zákazníkovi, ktorých je niekoľko terabajtov až petabajtov. Z tohto spárovania cez telefónne číslo sa dopracujeme k dátam ako napr.: vek, segment (podnikatelia vs. privátni zákazníci). V reálnom telekomunikačnom svete sa analyzuje omnoho viac atribútov zákazníka (napr. paušál, jeho história, dĺžka kontraktu, atď.).

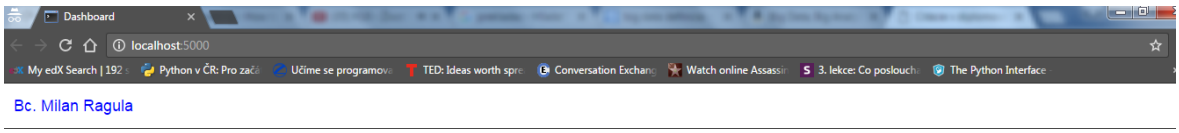
Hlasové CDR je potrebné spojiť aj s GEO dátami, aby bolo možné získať lokalizačné informácie napr. ulica, mesto, atď.

A na záver, z IMEI je možné získať informáciu o type a rade telefónu napr. iPhone, Samsung S8, atď., a jeho parametroch.
Tieto dáta spracujeme a následne uložíme do formátu JSON (JavaScript Object Notation), kde s nimi budeme pracovať.

Ako sme už spomínali, dáta budeme analyzovať podľa rôznych kategórií. Prvú kategóriu, ktorú budeme analyzovať, je lokalita volajúceho zákazníka a toho, čo prijíma hovor. Dôležitým faktorom našej analýzy je zistiť, v akých miestach najčastejšie tieto hovory zlyhávajú. Ďalšou kategóriou, ktorú budeme analyzovať, je vek zákazníkov, ktorý sa pohybuje od používateľov vo veku 18 rokov až po obyvateľov starších ako 65 rokov. Pre lepšie pochopenie požiadaviek a preferencií spotrebiteľa taktiež potrebujeme analyzovať, ktoré mobilné telefóny zákazníci preferujú najčastejšie pri telefonickom kontakte. V neposlednom rade budeme analyzovať zákazníkov, ktorých zaradíme do segmentu buď privátneho, ktorý znamená bežné používanie alebo business segment, v ktorom sa nachádzajú zákazníci využívajúci tzv. podnikateľské balíky služieb.

3.2.2 Mapa Slovenskej republiky

Aby sme sa bez problémov dostali k vizualizácii spracovanej analýzy, vytvorili sme jednoduchú web aplikáciu na serveri localhost, kde spustíme predmetnú analýzu.

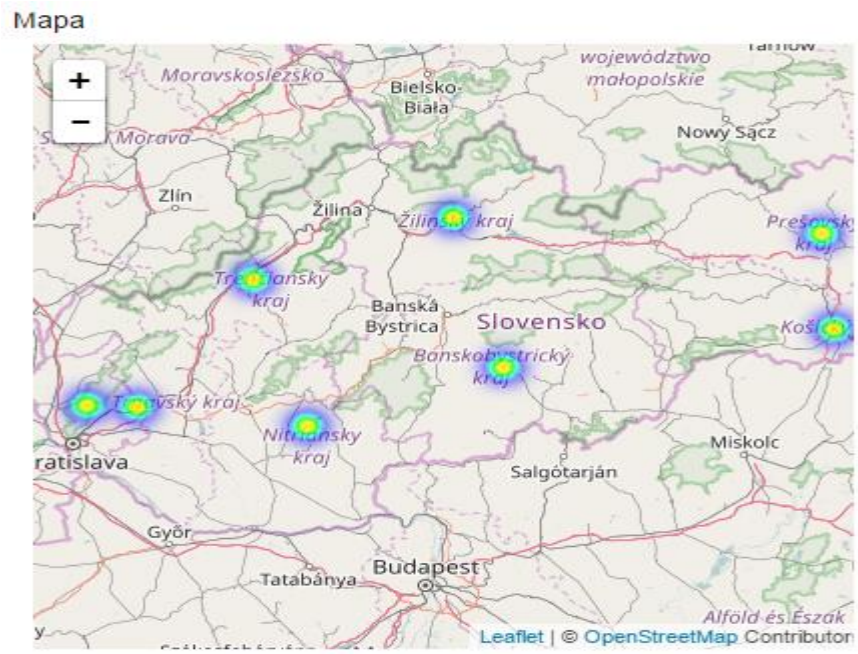


Obrázok č. 18: Dashboard vizualizácie

Zdroj : vlastná tvorba

Pre prehľadnejšie a lepšie pochopiteľné zobrazenie analyzovaných hovorov sme využili interaktívnu mapu Slovenskej republiky, prostredníctvom Google Api. Pretože ide o agregované dáta, hovory sme spracovávali za jednotlivých 8 krajov (Bratislavský kraj, Trnavský kraj, Nitriansky kraj, Trenčiansky kraj, Žilinský kraj, Banskobystrický kraj, Prešovský kraj a Košický kraj), ktoré sme zvýraznili na mape.

Graf č. 1 :Mapa Slovenskej republiky v interakcii s dátami



Zdroj : Google APIs

3.3 Analýza dát

Ďalším krokom po zbere dát je už samotná analýza. Na analýzu našich dát sme aplikovali nami už vopred definované kategórie, ktoré si rozoberieme postupne krok za krokom a interpretujeme výsledok ku ktorému sme dospeli. Tieto dáta sme uložili do formátu JSON (JavaScript object notation), ktorý je zobrazený na obrázku č. 19.

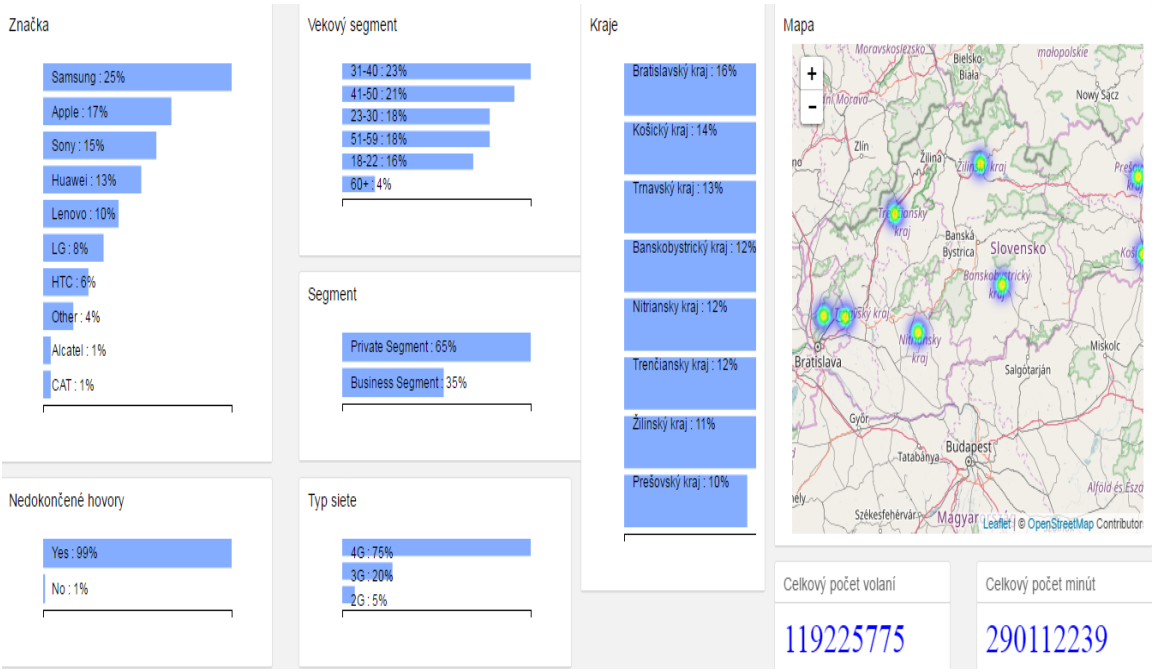
Obrázok č. 19: Vzorka dát uložených vo formáte JSON

Zdroj : vlastná tvorba

3.3.1 Celkový počet hovorov

Ako prvé pri spustení vizualizácie analýzy vidíme celkový počet hovorov a celkový počet minút pretelefonovaných v mesiacoch január, február a marec v r. 2017. Za uvedené 3 mesiace bol celkový počet hovorov vyše 119 miliónov, čo predstavuje vyše 290 miliónov minút prevolaných zákazníkmi. Hovory uskutočnené v tomto období boli približne rovnako rozložené po celom území Slovenskej republiky.

Graf č. 2: Zobrazenie celkového počtu hovorov

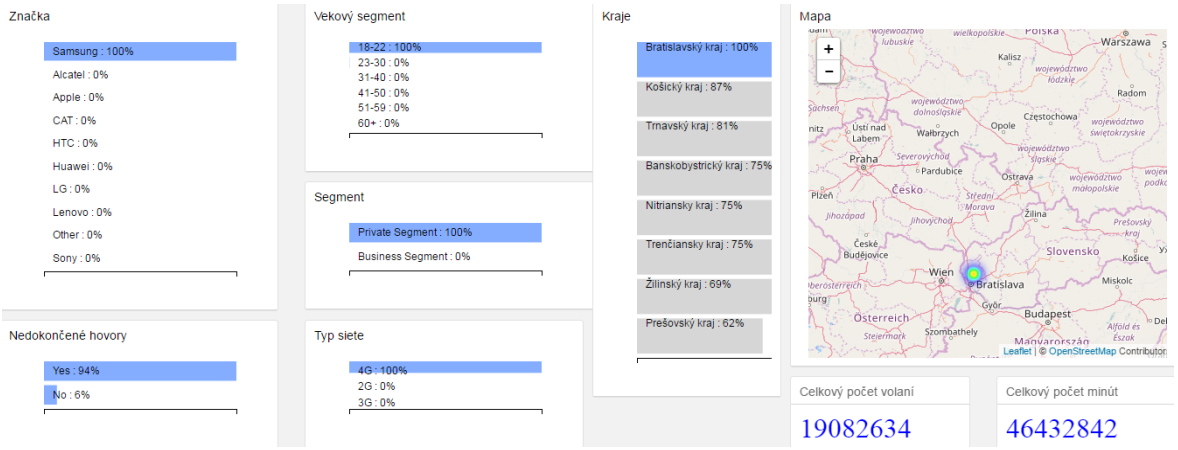


Zdroj : vlastná tvorba

3.3.2 Analýza dát podľa krajov

Z grafu č. 3 vidíme, že najviac hovorov bolo zaznamenaných z Bratislavského kraja vo veku od 18 do 22 rokov. Hovory boli uskutočnené z mobilného telefónu značky Samsung, pričom zlyhalo 6 % týchto hovorov, čo znamená, že z 19 082 634 hovorov zlyhalo približne 1 200 000 hovorov. Okrem toho môžeme z grafu č. 3 vyčítať, že v Bratislavskom kraji je silné pokrytie 4G siete (štvrtá generácia technológií mobilnej komunikácie).

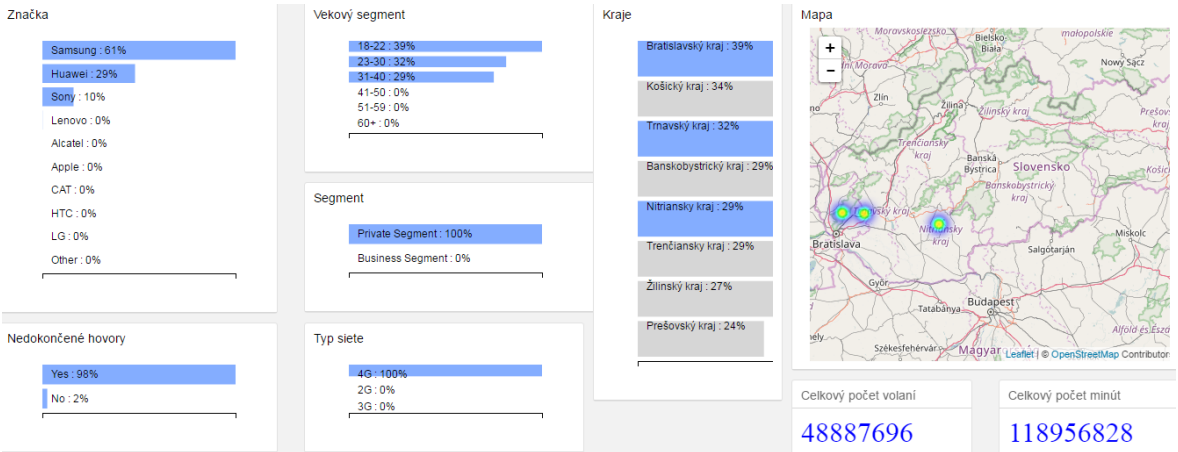
Graf č. 3: Analýza dát vzhľadom na hovory uskutočnené z Bratislavského kraja



Zdroj : vlastná tvorba

Taktiež je pre každú spoločnosť dôležité poznať svoju zákaznícku základňu, ktorá v tomto prípade zaznamenala najviac hovorov na juhu Slovenska. V tomto kraji sú zákazníci vo veku do 40 rokov a spoločnosť ponúka pokrytie siete vo forme 4G siete. Zákazníci na juhu Slovenska využívajú dané služby pre svoje súkromné účely.

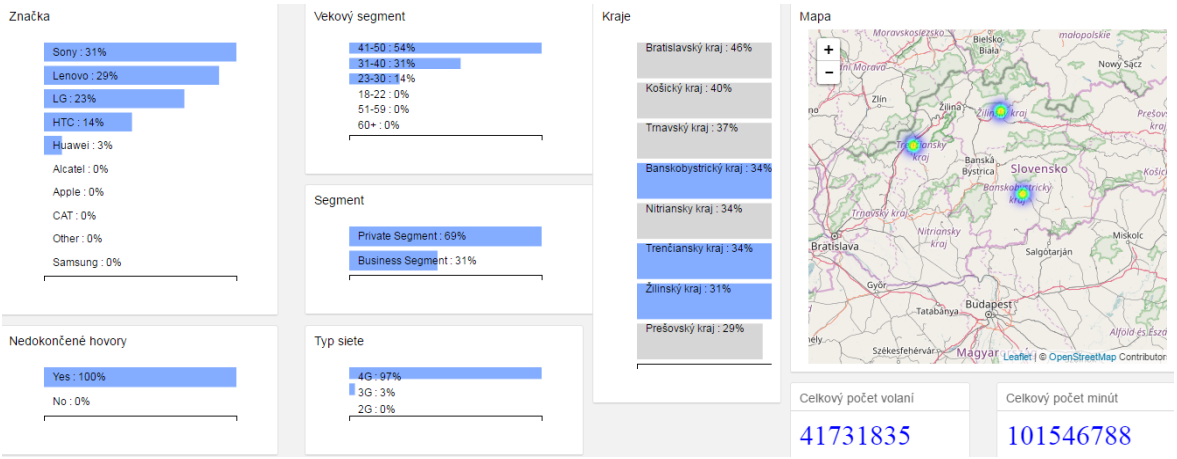
Graf č. 4: Analýza dát vzhľadom na hovory uskutočnené z južného Slovenska



Zdroj : vlastná tvorba

Na západe, severe a strede Slovenskej republiky je možnosť využívania aj 3G siete a z grafu č. 5 môžeme vidieť, že 30 % hovorov z tohto územia bolo uskutočnených prostredníctvom podnikateľských produktov. Na tomto území je veľmi dobré pokrytie siete, keďže tu nebol zaznamenaný žiadny hovor, ktorý by zlyhal. Zákazníka základňa na tomto území je od 20 do 50 rokov a najviac tu zákazníci preferujú značky mobilných telefónov ako sú Sony, Lenovo a LG.

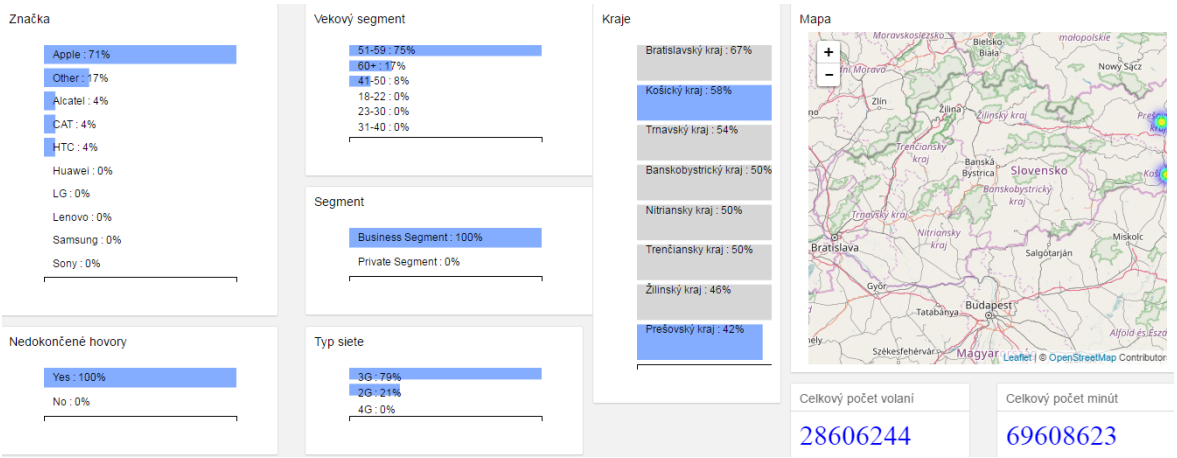
Graf č. 5: Analýza dát vzhľadom na hovory uskutočnené zo severného a stredného Slovenska



Zdroj : vlastná tvorba

Na východe Slovenskej republiky 4G pokrytie siete nie je dostupné, napriek tomu najviac hovorov prostredníctvom podnikateľských produktov bolo práve z tohto kraja, pričom až 70 % uskutočnených hovorov bolo zo smartfónu Apple. V tomto kraji mobilné služby využívali prevažne starší zákazníci.

Graf č. 6: Analýza dát vzhľadom na hovory uskutočnené z východného Slovenska

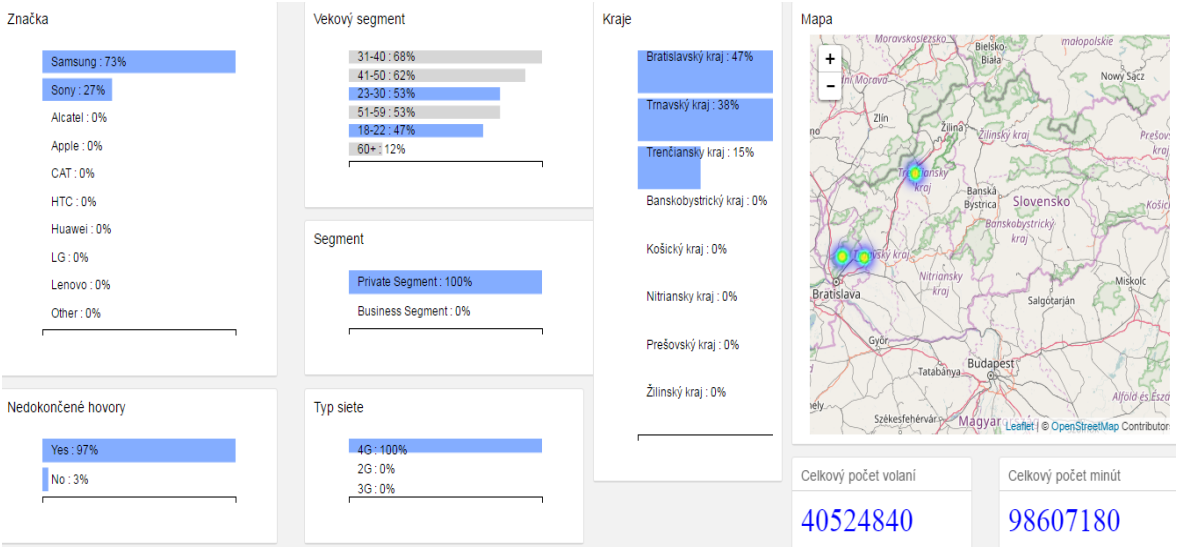


Zdroj : vlastná tvorba

3.3.3 Analýza dát podľa vekového segmentu

Ďalším dôležitým aspektom lepšieho poznania zákazníckej základne je poznanie potrieb zákazníka vyplývajúcich z jeho veku. Vek zákazníkov sme roztriedili do viacerých skupín. Z grafu č. 2 môžeme vidieť, že najviac hovorov uskutočnili zákazníci vo veku od 31 do 40 rokov. Zákazníci do 30 rokov sa nachádzali v Bratislavskom, Trenčianskom a Trnavskom kraji, pričom využívali telefónne zariadenia značky Samsung a Sony. Taktiež môžeme vidieť, že využívali služby spoločnosti na súkromné účely a z toho 3 % hovorov nebolo dokončených z neznámych technických príčin.

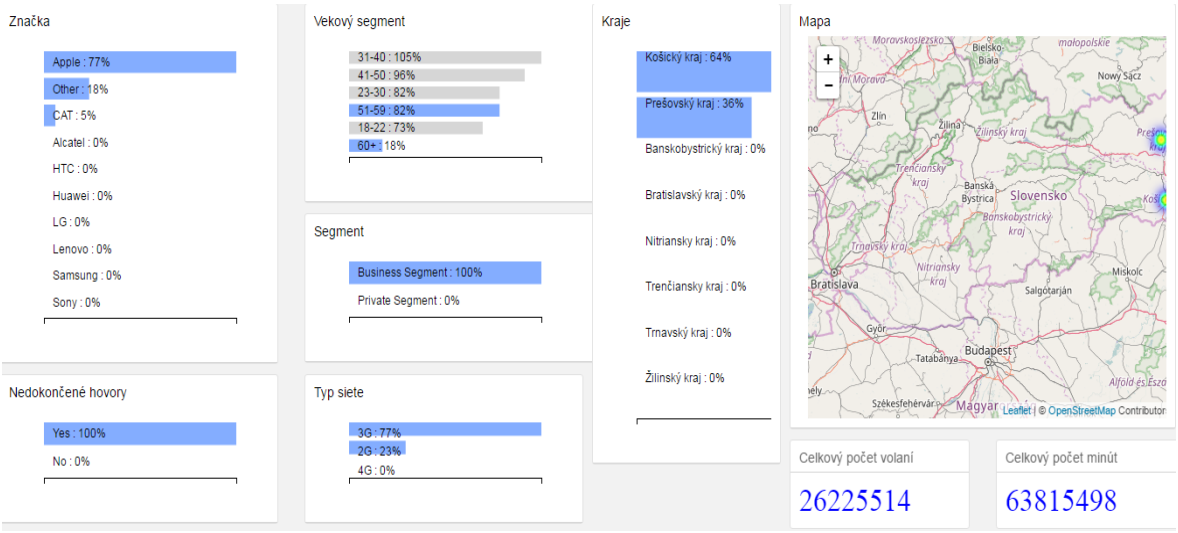
Graf č. 7: Analýza dát vzhľadom na zákazníkov vo veku od 18 do 30 rokov



Zdroj : vlastná tvorba

Naopak staršia generácia zákazníkov sa nachádzala na východe Slovenskej republiky, kde nebol zaznamenaný žiadny problém z dokončením hovorov a najviac z nich bolo sprostredkovaných cez telefón značky Apple. Z grafu č. 8 je zrejmé, že mladšia generácia klientov sa nachádza bližšie k nášmu hlavnému mestu a naopak starší zákazníci sú prevažne na východe Slovenska.

Graf č. 8: Analýza dát vzhľadom na zákazníkov vo veku od 51 do 60 a viac rokov

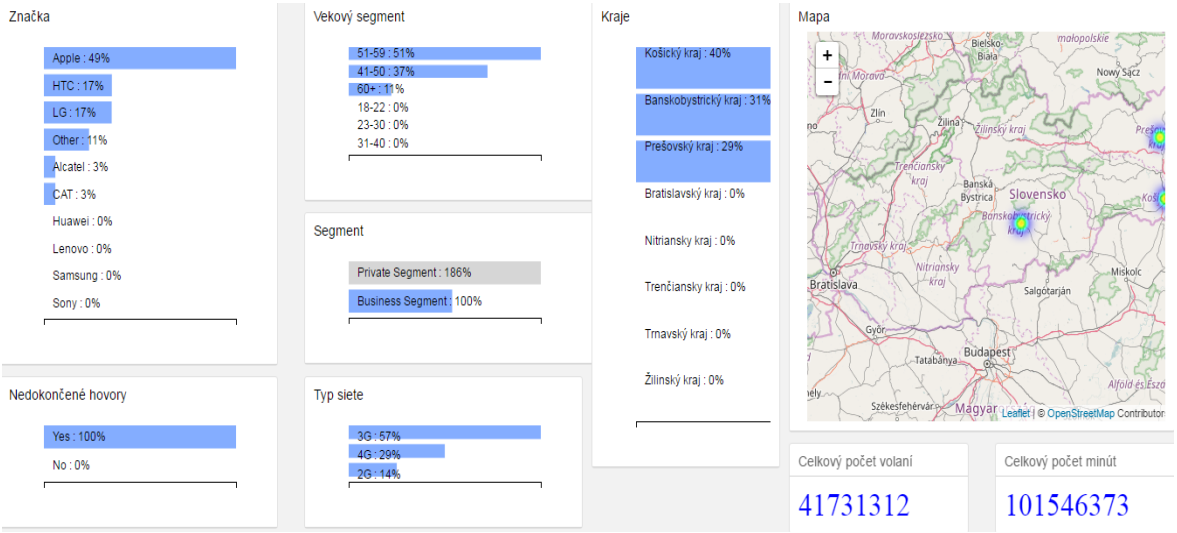


Zdroj : vlastná tvorba

3.3.4 Analýza dát podľa segmentu

Spoločnosti v tejto sfére ponúkajú pre svojich zákazníkov rôzne možnosti využívania ich služieb, ktoré môžeme zhrnúť do dvoch kategórií a to private segment, kde sa nachádzajú zákazníci využívajúci telekomunikačné služby pre vlastné účely a business segment, kde sú zahrnutí zákazníci, ktorí prečerpajú väčšie množstvo minút, čo znamená, že spoločnosť im vie ponúknuť výhodnejšie tzv. podnikateľské produkty. Z grafu č. 2 je zrejmé, že z týchto podnikateľských produktov bolo zaznamenaných 35 % z celkového počtu hovorov. Pokiaľ berieme len hovory z podnikateľských produktov, tak konštatujeme, že ich bolo o 83 % menej ako hovorov uskutočnených z bežných produktov. V tomto období boli všetky prevolané minúty zaznamenané len z východného územia Slovenskej republiky.

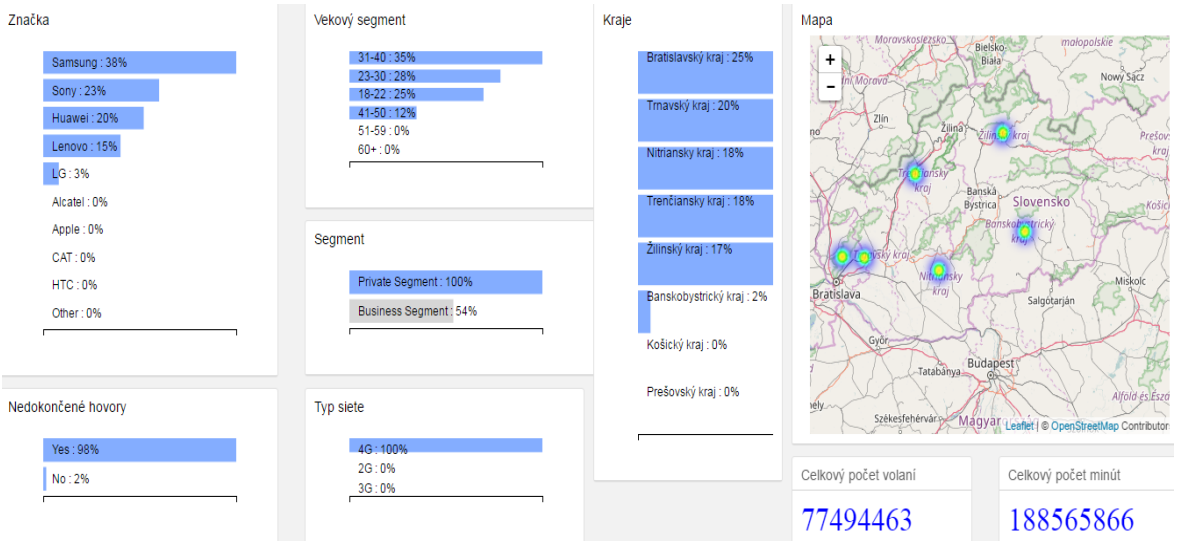
Graf č. 9: Analýza dát vzhľadom na Business segment



Zdroj : vlastná tvorba

Privátni zákazníci za sledované obdobie sa nachádzali okrem východného Slovenska na celom území, kde bolo uskutočnených vyše 77 miliónov hovorov. Na grafe č. 10 vidíme, že privátne služby využívajú ľudia do 50 rokov.

Graf č. 10 : Analýza dát vzhľadom na Private segment

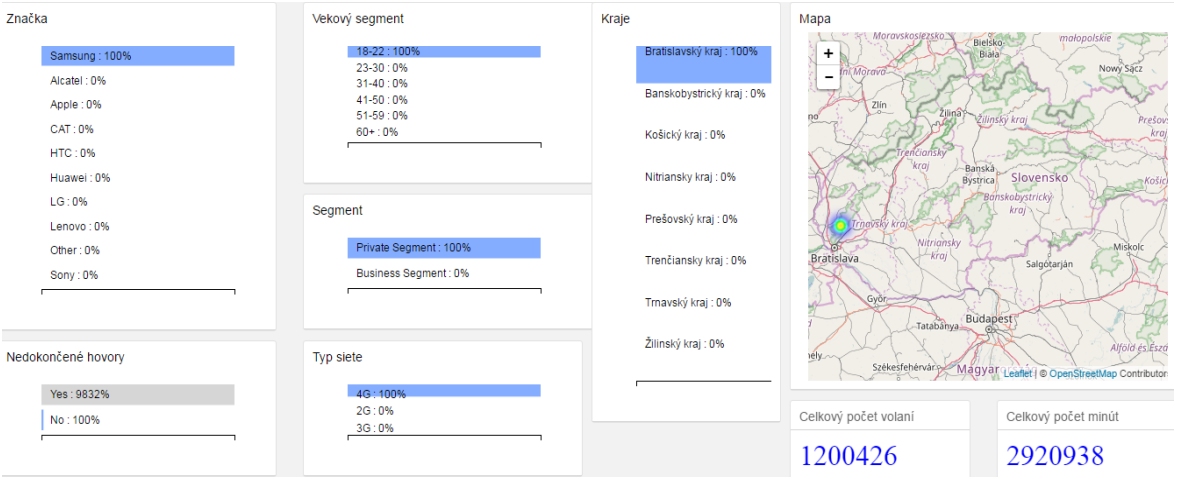


Zdroj : vlastná tvorba

3.3.5 Analýza dát podľa zlyhaných hovorov

V týchto mesiacoch sme zaznamenali zlyhanie siete v Bratislavskom kraji, v ktorom zlyhalo za hodnotené obdobie vyše 1 milión hovorov, pričom všetky boli uskutočnené z mobilného telefónu Samsung, zákazníkmi od 18 do 22 rokov.

Graf č. 11: Analýza dát vzhľadom na zlyhané hovory



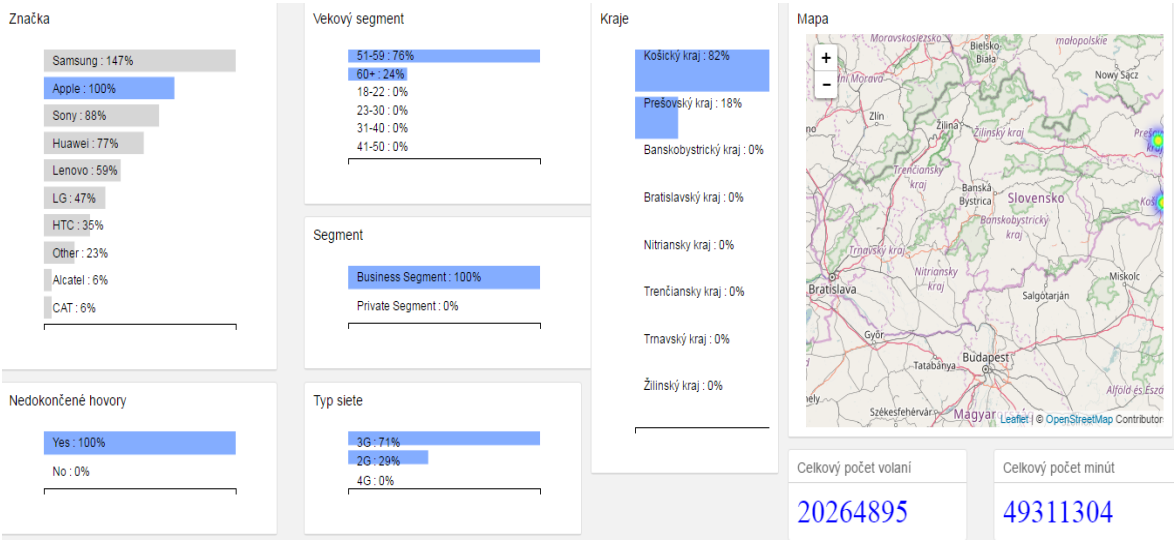
Zdroj : vlastná tvorba

3.3.6 Analýza dát podľa mobilných telefónov

Spoločnosti telekomunikácií už dlhšiu dobu ponúkajú k jednotlivým produktom aj mobilné telefóny, preto je pre každú spoločnosť vhodné vedieť o aké zariadenie je momentálne najväčší záujem. Telefóny, ktoré nás najviac zaujímali pre túto analýzu, sú zobrazené na grafe č. 12.

Z grafu vidíme, že zákazníci preferujú značku Apple na podnikateľské účely v rámci východného Slovenska, pričom sú to starší zákazníci.

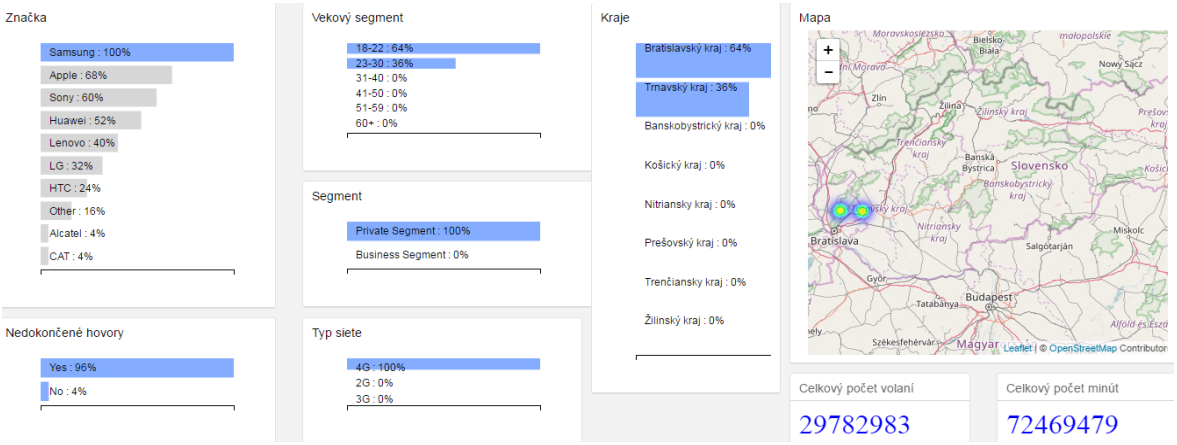
Graf č. 12: Analýza dát vzhľadom na mobilný telefón Apple



Zdroj : vlastná tvorba

Naopak čisto len na súkromné účely sa využíval telefón značky Samsung, ktorý preferovali zákazníci do 30 rokov v sieti 4G, pričom z tohto počtu bolo zaznamenaných 4 % zlyhaní hovorov. Z tohto mobilného telefónu bolo uskutočnených najviac hovorov v rámci Slovenskej republiky za sledované obdobie.

Graf č. 13: Analýza dát vzhľadom na mobilný telefón Samsung

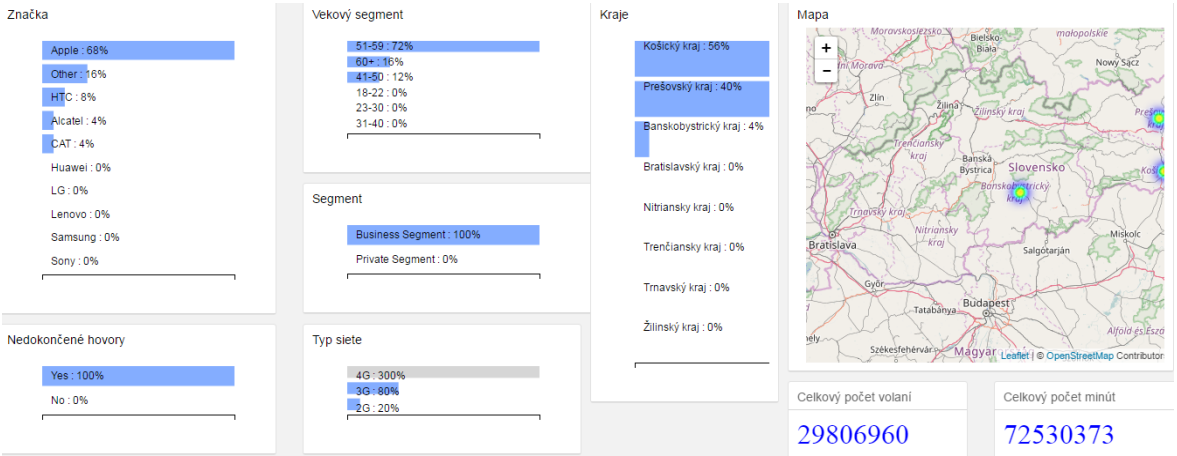


Zdroj : vlastná tvorba

3.3.7 Analýza dát podľa typu siete

Spoločnosti ponúkajú možnosti telefonovania zo všetkých možných dostupných sietí. Ako vidíme z grafu č. 14 na východnom Slovensku nie je možné momentálne využívanie 4G siete, ale len siete 2G (druhá generácia technológií mobilnej komunikácie) a siete 3G (tretia generácia technológií mobilnej komunikácie), napriek tomu všetky hovory boli uskutočnené prostredníctvom podnikateľských produktov.

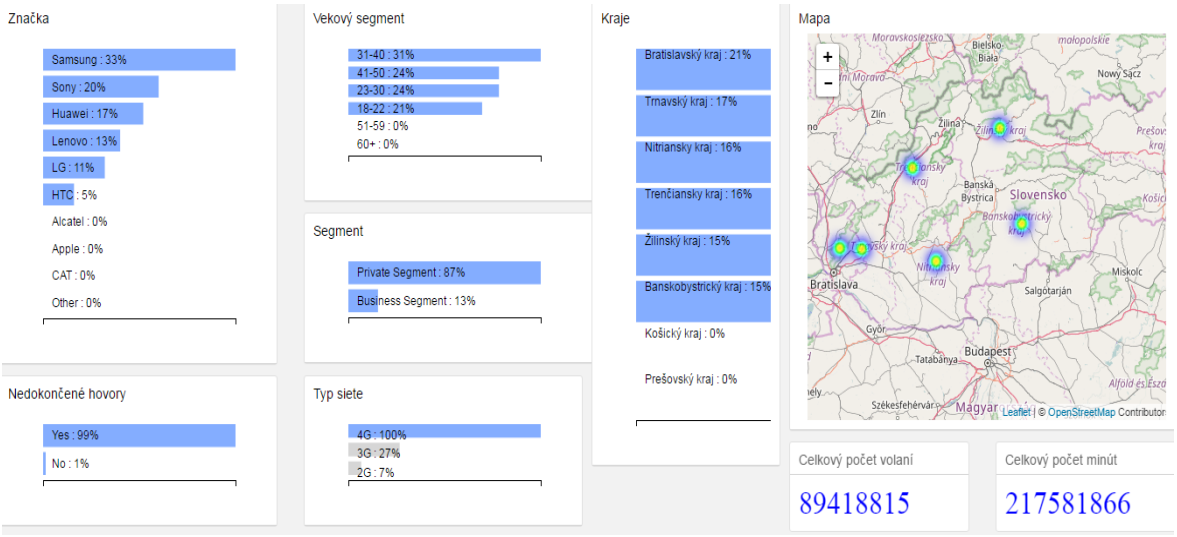
Graf č. 14: Analýza dát vzhľadom na siete 2G a 3G



Zdroj : vlastná tvorba

Keď, sa pozrieme na zastúpenie siete 4G (štvrtá generácia technológií mobilnej komunikácie) vidíme, že telekomunikačné spoločnosti ňou pokrývajú väčšinu územia Slovenskej republiky, okrem krajov, ktoré sa nachádzajú na východnom Slovensku.

Graf č. 15: Analýza dát vzhľadom na sieť 4G



Zdroj : vlastná tvorba

3.4 Výsledky analýzy

Z našej vizualizácie sme si ako prvé mohli všimnúť, že celkový počet hovorov a minút za mesiace hodnotené obdobie r. 2017 predstavuje obrovské množstvo dát, ktoré musia spoločnosti spracovávať a na základe týchto výstupov sa rozhodovať o ďalšej stratégii spoločnosti.

Z analýzy sme zistili, že len 1 % všetkých hovorov boli hovory, ktoré neboli nejakým spôsobom dokončené a to v Bratislavskom kraji, no aj tak predstavovali počet vyše 1 200 000 hovorov. Taktiež z Bratislavského kraja bolo uskutočnených najviac hovorov v rámci Slovenskej republiky. Zistili sme, že telefóny, ktoré sa nachádzajú v drahšej cenovej kategórii, ako je značka Apple, prevažne používajú starší zákazníci nad 50 rokov z východného Slovenska. Naopak, mladší zákazníci preferujú značku Samsung v okolí Bratislavského kraja. Taktiež sme zistili, že spoločnosť nemá pokrytie siete 4G na východe Slovenska, pričom je tam prevažná väčšina zákazníkov využívajúcich podnikateľské produkty. Okrem nami zobrazených grafov, je možné si pozrieť ľubovoľnú kombináciu daných kategórií, pri analýze hovorov, vzhľadom na zistenie potrebných faktov pri hodnotení zvoleného obdobia ale pri predikcii očakávaného vývoja dát v telekomunikačnom

sektore. Ako sme už uviedli na začiatku tejto kapitoly, dáta, s ktorými sme pracovali, nie sú reálne ale nemajú ďaleko od reality a napriek tomu je možné nami uskutočnenú analýzu implementovať aj do reálneho sveta na reálne dáta.

3.5 Možnosti využitia výstupov dosiahnutých výsledkov

Spracovanie, spájanie, analyzovanie dát a vytváranie akcií napr. kampane, optimalizácia siete atď., je v reálnom telekomunikačnom svete veľmi zaujímavá a komplexná práca, ktorá zamestnáva desiatky pracovných miest na slovenskom trhu. Zároveň podlieha prísnym normám o ochrane informácii a dát ako aj ďalších nariadení Európskej únie.

Výstup, s ktorým pracujeme v diplomovej práci je možné využiť pre nasledovné prípady:

- Optimalizácia siete – informácia o spadnutých hovoroch so zacielením na detailnejšiu lokalizáciu napr. ulica pomôže k plánovaní posilnenia siete v lokalitách, kde sú časté výpadky hovorov.
- Externá monetizácia a open data – sumárne dáta je možné poskytnúť pre univerzity, štatistický úrad, MHD alebo monetizovať na trhu.
- Marketingové kampane - typickým výstupom analytickej práce je cielenie kampaní pre predaj služieb, balíkov a ďalších produktov.

4 Záver

Každá spoločnosť, ktorá sa chce presadiť na trhu a udržať si stabilnú pozíciu, sa musí vedieť vysporiadať s rýchlo narastajúcim objemom dát. Musí byť schopná dáta čo najlepšie a najefektívnejšie spracovať a získať tak napr. možnú výhodu pred konkurenciou. Problematika Big Data je relatívne mladý no rýchlo sa rozvíjajúci pojem hlavne v odvetviach, kde sa kladie dôraz na poznatky a kde nie je možné využívať tradičné relačné databázy.

V prvej kapitole sme sa zamerali na porozumenie platforme Big Data z teoretického hľadiska. Dôležitou súčasťou pochopenia tejto problematiky je poznanie štyroch základných charakteristík, ktorými sa „veľké dáta“ vyznačujú, a ktoré musia dáta spĺňať, aby sme ich vôbec mohli nazvať „veľkými dátami“. Ďalším dôležitým faktorom pri spracovávaní dát je ich samotné poznanie, podľa toho či sú štruktúrované alebo neštruktúrované. Taktiež sme si opísali možné spôsoby spracovania štruktúrovaných, neštrukturovaných a semištrukturovaných dát a uviedli sme porovnanie tradičných relačných databáz a NoSQL databáz. Pre pracovanie s „veľkými dátami“ sme si rozobrali softvér Hadoop a jeho súčasť MapReduce.

V druhej kapitole sme si uviedli cieľ práce a metódy práce, ktoré nám slúžili na dopracovanie sa k stanovenému cieľu.

V tretej kapitole sme sa venovali samotnej analýze dát. Problematika Big Data je podstatne rozšírená najmä v oblasti telekomunikácií. Na základe zozbieraných agregovaných dát sme vytvorili jednoduchú webovú aplikáciu, kde sme analyzovali hovory zákazníkov za mesiace január, február a marec 2017. Ukázali sme si na vzorke dát možnosť spracovania problematiky Big Data a ich využitie.

Daný výstup diplomovej práce je možné využiť v reálnom svete napr. na optimalizáciu siete alebo pre rôzne marketingové kampane ako aj pre lepšie pochopenie správania sa zákazníkov a ich potreby.

Dokumentácia – Bibliografické odkazy

[1] Mayer-Schonberger V. – Cukier K. 2014. *BIG DATA – Revolúcia, ktorá zmení spôsob ako žijeme, pracujeme a myslíme.* 1. vyd. Praha : Albatros Media a. s. 2014. 248 s. ISBN 978-80-251-4119-9

[2] Mohanty S. – Jagadeesh M. – Srivatsa H. 2013. *Big Data Imperatives: Enterprise Big Data Warehouse, BI Implementations and Analytics (The Expert's Voice).* 1 vyd. New York: Springer Science + Business Media. 2013. 268 p. ISBN 978-1-4302-4873-6

[3] White T. 2012. *Hadoop: The Definitive Guide*. 3. vyd. *O'Reilly Media / Yahoo Press.* 2012. 688 p. ISBN 978-1-4493-1152-0

[4] McKinney. 2013. *Python for Data analysis.* 2. vyd. O'Reilly Media. 2013. 433 p. ISBN 978-1-449-31979-3

[5] O'Reilly. 2012. *Big Data Now.* 1. vyd. O'Reilly Media. 2012. 119 p. ISBN 978-1-449-35671-2

[6] Sathi A. 2012. Big Data Analytics. 1. vyd. MC Press. 2012. 96 p. ISBN 978-1-583-47380-1

[7] Novotný O. – Pour J. – Slánský D. 2005. Business Intelligence. Jak využít bohatství ve vašich datech. 1. vyd. Vydavateľstvo : Grada. 2005. 192 s. ISBN 802-4-710-943

[8] Sharda R. – Delen D. – Turban E. 2014. Businesss Intelligence and Analytics: Systems for Decision Support. 10. vyd. Prentice Hall. 2014. 1188 p. ISBN 978-0-133-05090-5

[9] Kosek J. – Holubová I. – Minarik K. – Novák D. 2015. Big Data a NoSQL databáze. 1. vyd. Vydavateľstvo : Grada. 2015. 288 s. ISBN 978-8-024-75466-6

[10] Fry B. 2007. Visualizing Data. 1. vyd. O´Reilly Media, Inc.2007. 384 p. ISBN 978-0-596-51455-6

[11] Gröpl T. 2015. HTML, CSS a JavaScript. 1. vyd. Vydavateľstvo: BEN - technická literatura. 2015. 160 p. ISBN 80-7300-099-7

[12] Pacáková V. a kolektív. 2003. Štatistika pre ekonómov. 1. vyd. Vydavateľstvo: Wolters Kluwer (Iura Edition). 2003. 358 s. ISBN 80-8904-774-2

[13] Kelly S. 2006. Customer Intelligence From Data to Dialogue. 1. vyd. Vydavateľstvo: John Wiley & Sons. 2006. 276 p. ISBN 978-0-470-01858-3

[14] Minelly M. – Chambers M. – Dhiraj A. 2013. Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses. 1. vyd. Vydavateľstvo: John Wiley & Sons. 2013. 179 p. ISBN 978-1-118-23915-5

Elektronické dokumenty - Monografie

[1] Grobelnik M. 2012. Big – Data tutorial. [online]. Ljubljana : Jozef Stefan Institute. 2012. 40 p. [cit. 20.3.2017]. Dostupné na internete: http://www.planetdata.eu/sites/default/files/presentations/Big_Data_Tutorial_part4.pdf

[2] Raj P. 2014. Handbook of Research on Cloud Infrastructures for Big Data Analytics. [online]. India: IBM India Pvt Ltd. 2014. 420 p. [cit. 01.4.2017]. Dostupné na internete: https://books.google.sk/books?id=m95GAwAAQBAJ&pg=PA416&lpg=PA416&dq=big+data+pdf&source=bl&ots=VwBX0G69jr&sig=veNJGffXs_LuVzbxczE2N6ODvOc&hl=sk&sa=X&ved=0ahUKEwjPp_3bnrPTAhWHLlAKHYAQClo4HhDoAQgxMAI#v=onepage&q=big%20data%20pdf&f=false

[3] Autor: Neznámy. 2017. Top Hadoop Terms You Need to Know. In bigdata analyticsnews. [online]. vol. 3, no. 3. [cit. 25.03.2017]. Dostupné na internete: http://bigdataanalyticsnews.com/top-hadoop-terms-you-need-to-know/

[4] Jeffrey D. - Sanjay G. 2004. MapReduce: Simplified data processing on large clusters. Google, Inc. [online]. [cit. 25.02.2017]. Dostupné na internete: https://static.googleusercontent.com/media/research.google.com/sk//archive/mapreduce-osdi04.pdf

Zoznam príloh

Príloha A : Zdrojový kód pre analýzu dát v jazyku Python

Príloha B : Zdrojový kód pre vizualizáciu dát v jazyku JavaScript

Príloha C : Zdrojový kód pre komunikáciu so serverom v jazyku HTML