EKONOMICKÁ UNIVERZITA V BRATISLAVE

FAKULTA HOSPODÁRSKEJ INFORMATIKY

Evidenčné číslo: 103006/I/2019/36086129772468996

PROBLEMATIKA BIG DATA V SÚČASNEJ AKTUÁRSKEJ PRAXI

Diplomová práca

2019 **Bc. Ján Masár**

EKONOMICKÁ UNIVERZITA V BRATISLAVE

FAKULTA HOSPODÁRSKEJ INFORMATIKY

PROBLEMATIKA BIG DATA V SÚČASNEJ AKTUÁRSKEJ PRAXI

Diplomová práca

Študijný program: Informačný manažment

Študijný odbor: Kvantitatívne metódy v ekonómií

Školiace pracovisko: Katedra matematiky a aktuárstva

Vedúci záverečnej práce: Ing. Michal Páleš, PhD.

Bratislava 2019 **Bc. Ján Masár**

Poďakovanie

Týmto sa chcem poďakovať vedúcemu mojej diplomovej práce Ing. Michalovi Pálešovi, PhD., za jeho odborné rady, usmernenia, ochotu, pomoc a vrelý prístup počas písania tejto diplomovej práce. Veľká vďaka takisto patrí spoločnosti Datavard s.r.o., ktorá poskytla hardvér a infraštruktúru na účely tejto práce, menovite sa chcem poďakovať IT oddeleniu tejto spoločnosti.

ABSTRAKT

MASÁR, Bc. Ján: *Problematika Big Data v súčasnej aktuárskej praxi*. – Ekonomická univerzita v Bratislave. Fakulta Hospodárskej informatiky; Katedra matematiky a aktuárstva. – Vedúci záverečnej práce: Ing. Michal Páleš PhD. – Bratislava, FHI 69 s.

Hlavným cieľom tejto práce je preskúmať problematiku použitia Big Data v aktuárstve a vytvorenia manuálu na prepojenie jazyka R a Apache Hadoop. Prepojenie oboch platforiem je opísané postupným vykonávaním krokov na strane R a aj Hadoopu. Použité príkazy sú opísané a doplnené o výstupy z konzoly. Práca je rozdelená do 4 kapitol. V prvej kapitole je charakterizovaná Data Science, Big Data vo všeobecnosti a aj vo vzťahu k poistnému trhu. Ďalej táto kapitola charakterizuje R ako nástroj Data Science a Apache Hadoop. Druhá kapitola obsahuje hlavný cieľ a čiastkové ciele. V tretej kapitole sa je popísaná metodika práce a metódy skúmania. Záverečná kapitola obsahuje manuál na prepojenie R a Apache Hadoop, ako aj porovnanie Hadoop User Experience ako platformy na spracovanie Big Data.

Kľúčové slová: aktuárstvo, Big Data, Data Science, Hadoop, integrácia platforiem, Jazyk R, poisťovníctvo

ABSTRAKT

MASÁR, Bc. Ján: *Big Data and actuarial science*. – University of Economics in Bratislava. Faculty of Economic Informatics; Department of Mathematics and Actuarial Science. – Supervisor of the thesis: Ing. Michal Páleš PhD. – Bratislava, FHI 69 p.

The main aim of this thesis is to explore the possibilities of using Big Data in actuarial science and creation of manual for connection R language and Apache Hadoop. Integration of both platform is described using step-by-step command execution on the side of R and Hadoop. Used command are described and the output from console is attached. Thesis is divided to four chapters. Data Science, Big Data in general and in relationship to insurance market is characterized in first chapter. Second chapter contains the main goal of the work. The third chapter described the methodology of work and methods of investgation. The final chapter contains a manual for R and Apache Hadoop as well as a comparison of Hadoop User Experience as a Big Data processing platform.

Keywords: actuarial science, Big Data, Data Science, Hadoop, integration of platforms, insruance, R language

OBSAH

Úvod 6
1. Súčasný stav riešenej problematiky doma a v zahraničí 7
1.1 Aktuárska veda 7
1.2 Data Science 8
1.3 Big Data 11
1.4 Big Data a poistný trh 16
1.5 Jazyk R a Data Science 26
1.6 Apache Hadoop 28
2. Cieľ práce 30
3. Metodika práce a metódy skúmania 31
4. Výsledky práce a diskusia 34
4.1 Nastavenie a konfigurácia prostredí R a Hadoop 34
4.2 Načítanie a uloženie dát z Hadoop 41
4.3 Operácie s dátami 45
4.4 Spracovanie údajov pomocou Hadoop User Experience 53
4.5 Porovnanie jednotlivých prístupov 63
Záver 65
Zoznam použitej literatúry 67
Prílohy 69

Úvod

V súčasnosti zažíva používanie Big Data veľký boom v rôznych odvetviach. Žijeme v dobe, keď je táto technológia masívne používaná spoločnosťami na každodennej báze a často si ich použitie ani neuvedomujeme. Použitie Big Data nie je len doménou technologických firiem, ale vo veľkom ich využívajú aj banky, štátne inštitúcie či poisťovne. V čase, keď sú tieto dáta považované za štvrtý výrobný faktor, ich spoločnosti používajú viac ako kedykoľvek predtým. Získavajú tým nielen prehľad o internom prostredí firmy, ale najmä o svojich zákazníkoch, ich správaní a aktuálnych trendoch. Správne použitie týchto dát znamená konkurenčnú výhodu a môže mať vplyv na budúcnosť firmy ako takej.

Aby sa tieto veľké a neustále meniace objemy dát dokázali efektívne analyzovať, sú potrebné nové prístupy a platformy. Veda, ktorá sa zaoberá získavaním pridanej hodnoty z týchto dát sa nazýva Data Science. Ide o jedno za najperspektívnejších odvetví v súčasnosti, ktoré sa neustále rozvíja veľkou rýchlosťou. Zaoberá sa analýzou, čistením, spracovaním dát a získavaniu nových poznatkov, ktoré spoločnostiam umožňujú prijímať rýchlejšie a lepšie rozhodnutia. Poistný trh a spoločnosti pôsobiace na ňom, si tieto skutočnosti veľmi dobre uvedomujú, a preto hľadajú stále nové možnosti ako skvalitniť existujúce a už implementované produkty. Pri aktuálnom trende je možné považovať niektorých aktuárov v poisťovniach aj za dátových vedcov. Ich úlohou je neustále sledovať vývoj priebiehajúceho stavu a v prípade zmien v dostatočnom predstihu informovať a prispôsobiť sa danej situácií. S novými možnosťami, ktoré Big Data prinášajú sa otvárajú nové možnosti aj aktuárom. Nové zdroje údajov, rastúci objem dát či meniace sa správanie zákazníkov – toto všetko sú výzvy, ktorým aktuár čelí. Dôležitým faktorom pri konkurenčnom súboji je používanie Big Data a spôsob ich spracovania. Existuje viacero možností, ako poskytnúť aktuárom prístup do týchto „veľkých dát“ a tým umožniť ich použitie v aktuárskej praxi. Jednu z takýchto možností predstavuje aj integrácia jazyka R a populárnej Big Data platformy Apache Hadoop.

1. Súčasný stav riešenej problematiky doma a v zahraničí

1.1 Aktuárska veda

Aktuárstvo ako vedné odvetvie je úzko späté s poisťovníctvom. Ide o disciplínu, ktorá sa zaoberá hodnotením rizika najmä v poisťovníctve a finančníctve, pomocou matematických a štatistických metód. Aktuárstvo zahŕňa množstvo vzájomne súvisiacich predmetov ako napríklad matematika, teória pravdepodobnosti, štatistika, financie, ekonómia a informatika. Špecialisti venujúci sa tejto oblasti sa nazývajú aktuári. Ide o kvalifikovaných odborníkov, ktorí musia absolvovať viacero certifikácií a akreditácií.

Ako už bolo povedané, ich hlavnou pracovnou náplňou je štúdium rizika. Aktuár by mal byť schopný riziko predpovedať, zhodnotiť a kontrolovať. Na základe týchto poznatkov funguje princíp poisťovníctva – spoločnosti sa spoliehajú na *risk manažment*, ktorý je dôležitou súčasťou ich stratégií. Poznanie rizika umožňuje poisťovniam rozhodovať sa lepšie tak, aby sa budúcnosti dokázali vyhnúť nepriaznivým situácia a zaistili finančnú stabilitu. V poisťovniach aktuári pomáhajú určovať ceny produktov (životné a aj neživotné poistenie), modelujú nové produkty a analyzujú aktuálny finančný stav. Cieľom týchto analýz je zabezpečiť likviditu poisťovne a stabilný cash flow.

Aktuár pri svojej práci analyzuje všetky dostupné a relevantné dáta – historické, ale aj *real-time* (v reálnom čase). Následne pomocou rozličných techník získava požadované výsledky. Po získaní relevantných výsledkov ich aktuár prezentuje a podieľa sa na tvorbe stratégie spoločnosti. Pri svojej práci aktuári používajú rôzne matematické a štatistické programy, či programovacie jazyky. Vďaka rýchlosti rozvoja technológie v poslednom desaťročí, aj prešlo aj toto odvetvie významnými zmenami. Nové zdroje dát si vyžadujú efektívnejšie metódy spracovania, čo predstavuje výzvu pre aktuárov. Z tohto dôvodu sa v tejto profesii kladie čoraz väčší dôraz na to, aby aktuár ovládal potrebné informačné technológie ako napríklad princípy databáz, programovacie jazyky či efektívne algoritmy na analýzu dát.

V súčasnej dobe je hlavným cieľom aktuárov vytvárať presnejšie a spoľahlivejšie predpovede a modely. Dôraz sa kladie najmä na nájdenie rovnováhy medzi finančným úspechom a akceptovateľným rizikom. Dosiahnutie tejto rovnováhy predstavuje pre poisťovne ideálny stav z hľadiska ich dlhodobej stability.

1.2 Data Science

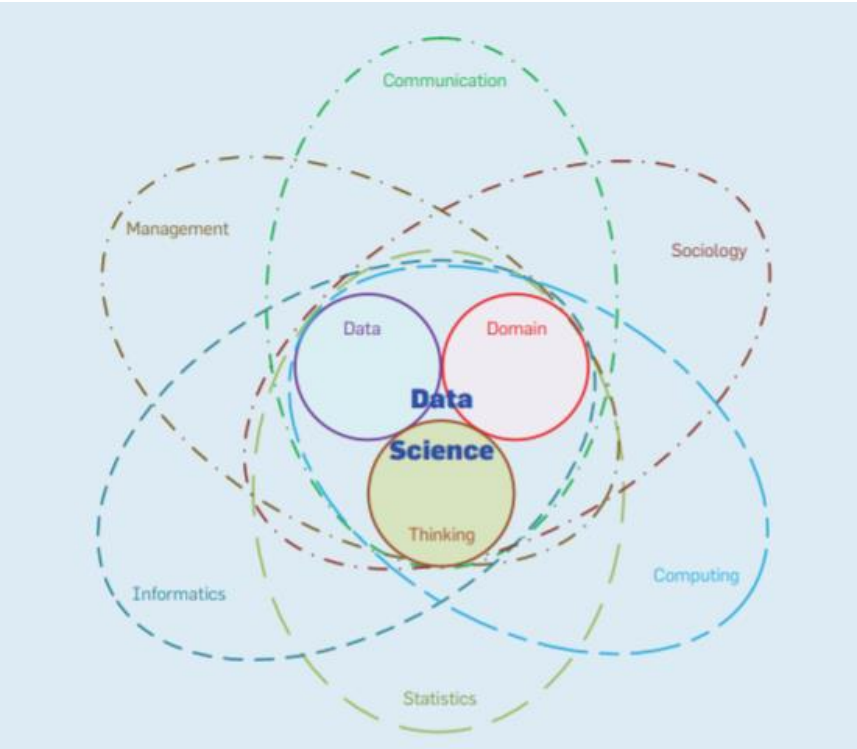
Toto slovné spojenie sa v posledných rokoch stáva čoraz populárnejším, no najmä za uplynulých pár mesiacov môžeme pozorovať stále rastúci záujem o Data Science. Tento pojem už nie je predmetom diskusií výhradne odbornej verejnosti v oblasti informačných technológií, získal si popularitu aj na oddeleniach marketingu, predaja, či je úspešným nástrojom pri prijímaní rozhodnutí vrcholného manažmentu.

V doslovnom preklade ide o takzvanú *dátovú vedu* – účelom tejto počítačovej vedy je extrahovať zmysluplné poznatky z dát a následne ich ďalej efektívne využívať tak, aby nám priniesli určitú hodnotu[1]. Avšak, vnímať Data Science výlučne ako počítačovú vedu by nebolo celistvé. Odborníci na túto problematiku sa zhodujú, že ide o multidisciplinárny vedný odbor, ktorý využíva relevantné poznatky z oblastí informatiky, štatistiky, komunikácie, sociológie a manažmentu na analýzu dát. Tieto poznatky vznikajú analýzou dát, často hovoríme o Big Data, pričom dáta pochádzajú z rôznych informačných zdrojov a platforiem. Ide o dáta štruktúrované aj neštruktúrované, formalizované aj neformalizované a pochádzajú z rôznych zdrojov – z firemných databáz, zo senzorov, z internetu či prieskumov. Na základe uvedomenia si týchto skutočností si môžeme utvoriť predstavu o komplexnosti a náročnosti tejto vednej disciplíny. Rozmach data science priamo súvisí s rozmachom Big Data, nakoľko si spoločnosti začali uvedomovať, že síce disponujú obrovským množstvom dát, no bez efektívnej analýzy a využitia týchto dát spoločnosť nemá žiadnu pridanú hodnotu z týchto dát. Uvedomenie si tejto skutočnosti viedlo spoločnosti k rozhodnutiu investovať značné finančné prostriedky, práve na získanie tejto pridanej hodnoty.. O dátach sa v súčasnej dobe hovorí ako o štvrtom výrobnom faktore – popri tradičných troch práci, pôde a kapitáli. Spoločnosti praktizujú tzv. *data driven decision making (DDDM)* – čo predstavuje rozhodovanie sa na základe údajov, ktoré máme podložené dátami radšej, ako spoliehanie sa na intuíciu[2]. Napríklad, marketingový manažér sa nebude spoliehať na svoje skúsenosti a odhad pri výbere typu reklamy, ale využije poznatky zistené na základe analýzy dát ako zákazníci reagujú na daný typ reklamy. Tieto metódy umožňujú spoločnostiam lepšie porozumieť tomu, čo zákazník preferuje a očakáva. Spoločnosti

[1] PEARSON, Lillian. *Data Science for Dummies*. Hoboken : John Wiley & Sonc, Inc., 2017. 384 s. ISBN 978-1-119-32763-9

[2] PROVOST, Foster – FAWCETT, Tom. *Data Science And Its Relationship to Big Data and Data-Driven Decision Making* [online]. 2013. [cit. 18.01.2019. Dostupné na: https://www.liebert-pub.com/doi/pdfplus/10.1089/big.2013.1508

sa ďalej zameriavajú na to, podľa čoho sa rozhoduje a prispôsobujú sa mu, čo im umožňuje získať konkurenčnú výhodu.



Obrázok 1.1: Disciplíny v Data Science
Zdroj:[5]

Špecialista, ktorý sa zaoberá data science sa nazýva Data Scientist – *dátový vedec.* Tento termín sa objavuje od roku 2008, odkedy sa potrebe analýzy a spracovania dát začal klásť čoraz väčší význam[3]. Ako môžeme vidieť na obrázku 1.1, výkon tohto povolania si vyžaduje znalosť mnohých vedeckých disciplín. Zároveň je však v súčasnosti považované za jedno z najžiadanejších povolaní, v USA bola označená za najatraktívnejšiu štyri roky po sebe[4]. Data Scientist by mal okrem znalostí technológií disponovať nasledujúcimi schopnosťami: poznať špecifiká daného priemyslu, objasniť získané poznatky, rozumieť aktuálnej firemnej stratégii a podieľať sa na jej tvorbe a svojou prirodzenou zvedavosťou hľadať odpovede na aktuálne otázky. Z technologického hľadiska by mal dátový vedec poznať nasledujúce oblasti: rôzne programovacie jazyky (*R, Python*), technologické platformy (*Apache Hadoop, non-SQL databázy*) a sledovať aktuálne technologické trendy (*cloud computing, reporting – Tableau*).

[3] BERKLEY. *What is Data science* [online]. UC Berkley School of Information, 2018. [cit. 18.01.2019].Dostupné na: https://datascience.berkeley.edu/about/what-is-data-science/
[4]GLASDOOR. *50 Best Jobs in America* [online].2018. [cit. 18.01.2019]. Dostupné na: https://www.glassdoor.com/List/Best-Jobs-in-America-2018-LST_KQ0,25.htm

Z popisu náplne práce data science a profilu samotného dátového vedca vieme, že medzi jeho hlavné schopnosti a znalosti určite patrí štatistika, dátové analýzy, tvorba predikčných modelov, programovanie a hľadanie súvislostí v dátach a následné odporúčania a prezentácie jeho znalostí. V oblasti poisťovníctva existuje podobná práca – a to práca aktuára. Aktuár, podobne ako dátový vedec, takisto musí ovládať oblasti ako matematika, štatistika, dátové analýzy a tvorbu predikčných modelov. Môžeme teda nájsť podobnosť medzi týmito dvoma povolaniami. Čím viac dát dostanú, tým presnejšie ich výsledky a predikčné modely budú. Najmä s rozmachom Big Data, tieto dve profesie budú čoraz viac prepojené – potenciál Big Data v oblasti poisťovníctva a aktuárstva rozhodne existuje. Napriek tomu, stále tu môžeme pozorovať rozdiely medzi nimi – kým aktuár je viac zameraný na dve oblasti ako financie, investície a poistenie, naproti tomu data science skúma dáta a vzťahy medzi nimi vo všeobecnosti (hlavné rozdiely medzi povolaniami sú uvedené v tabuľke 1.1). Dátový vedec je viac programátorsky zameraný a medzi jeho znalosti patria aj rôzne no-SQL databázy, Hadoop či práca s neštruktúrovanými dátami.

Tabuľka 1.1 : **Porovnanie akutárstva a Data Science**

Aktuárska veda	**Data Science**
Zameranie sa poistné produkty, finančné modely, investície a poisťovníctvo.	Zameranie na dáta vo všeobecnosti, vzťahy medzi nimi, analýzy, nezávisle od priemyslu.
Finančné modely a odporúčania, riadenie rizík, dizajn produktov	Databázy, data mining, machine learning a dátove vizualizácie, neurónové siete
Špecifické vzdelanie	Vzdelanie v oblasti IT
Nástroje SAS, Excel, R, Visual Basics, SQL..	Nástroje R, Python, Hadoop, C++...

Zdroj: vlastné spracovanie

Aktuári oproti dátovým vedcom majú výhodu, že pracujú v oblasti, ktorá im je dôverne známa a poznajú ju, ide o špecifickú oblasť v oblasti poisťovníctva. Takisto, práca aktuárov je regulovaná nariadeniami, predpismi a na svoju činnosť musia mať licenciu. Aktuár môže byť vo svojej profesii aj dobrým dátovým vedcom, no opačne to až tak neplatí z dôvodu regulácií. Vďaka aktuálnemu trendu rastu Big Data a rastúcej potrebe spoločností hľadať súvislosti v dátach a odhaľovať nové skutočnosti, tieto dve profesie konvergujú. Môžeme aj predpokladať, že aktuári v priebehu rokov budú vykonávať aj úlohu dátových vedcov.

1.3 Big Data

Pojem *Big Data* sa používa v súčasnej dobre čoraz častejšie, pričom sa teší veľkej obľube nielen v IT svete, ale aj iných odvetviach. Big Data dnes vo veľkom využívajú automobilky, finančné burzy, výrobné podniky a v neposlednom rade našli uplatnenie aj na poistnom trhu. V slovníkoch Oxfordskej univerzity – *Oxford English Dictionaries* – sa môžeme stretnúť s nasledujúcou definíciou tohto termínu: *extrémne veľké dátové sety, ktoré môžu byť analyzované pomocou počítačových systémov na odhalenie vzorov, trendov a asociácií, najmä v oblasti ľudského správania a interakcie*[5]. Iná, populárna a pomerne rozšírená definícia hovorí, že ide o dátové sety, ktoré sú tak rozsiahle a komplexné, že tradičné metódy spracovania dát sú neefektívne alebo nemožné. Najznámejšiu a dodnes zaužívanú metódu charakterizujúcu Big Data zaviedol v roku 2001 Douglas Laney – viceprezident analytického tímu firmy Gartner. Ide o známu teóriu 3Vs (troch véčok)[6]. Patria sem Volume, Velocity a Variety.

- **Volume** (*Objem*) – charakterizuje nielen veľkosť zbieraných dát, ale aj detailnosť, s ktorou sú tieto dáta zaznamenávané, či obdobie ich zachovania. Ide o veľmi rýchlo rastúcu dimenziu.
- **Velocity** (*Rýchlosť*) – hovoríme nielen o rýchlosti, ktorou sú dáta vytvárané, ale aj o frekvencii ich zberu a následného spracovania. Čím rýchlejšie sa dokážu dáta spracovať, tým lepšie – v ideálnom prípade v reálnom čase. Cieľom spoločností je tieto dáta využiť v momente, keď sú vytvorené a tak pružne reagovať na dopyty používateľov.
- **Variety** (*Rôznorodosť*) – opisuje typ dát, ktoré sú zbierané. Zdroje dát sú divergentné – tradičné dáta z relačných databáz, senzorové dáta, internet, vnútropodnikové dáta a aj externé zdroje. Forma dát a štruktúra sa navzájom líšia – textové, binárne, grafické – pričom môžu byť štruktúrované alebo neštruktúrované.

Postupom času boli k tejto pôvodnej teórií boli pridané ďalšie koncepty pre lepšie pochopenie Big data, najčastejšie sa stretávame so štvrtým V ako **Veracity** – pravdovravnosť. Táto

[5] OXFORD ENGLISH DICTIONARIES. *Big Data Definition,* [online]. 2019. [cit. 20.01.2019]. Dostupné na: https://en.oxforddictionaries.com/definition/big_data https://en.oxforddictionaries.com/definition/big_data
[6] LANEY, Douglas. *3D Data Management: Controlling Data Volume, Velocity and Variety* [online]. 2001. [cit. 20.01.2019]. Dostupné na https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

dimenzia nám opisuje **spoľahlivosť, kredibilitu, objektívnosť a čistotu dát**. Čím menej zavádzajúcich informácií dáta obsahujú, tým viac sa naň analytici a v konečnom dôsledku aj manažéri podnikov môžu spoľahnúť. Okrem spomínaných 4 V sa s odstupom času definujú aj ďalšie dimenzie Big Data ako napríklad *Value* (hodnota), *Volatility* (prchavosť) či *Vulnerability* (zraniteľnosť).

Vznik Big Data súvisí so stále dostupnejšími technológiami a to nielen s kvalitnejším, modernejším hardvérom a softvérom, ale aj vzostupom popularity internetu a pôsobeniu veľkých korporácií v tejto oblasti. Ľudstvo každým okamihom generuje viac dát ako kedykoľvek predtým. Spisovateľ a poradca v oblasti business stratégií Bernard Marr uvádza, že za posledné dva roky ľudstvo vytvorilo 90 percent všetkých svetových dát.[7] Spoločnosti ako Google alebo Facebook ako čelili výzve spracovania a uchovania dát ako jedny z prvých, nakoľko boli priamo „zasiahnuté“ touto problematikou. Pre potreby lepšej ilustrácie objemu dát uvádzame nasledovné fakty:

- **1,52 miliardy** ľudí sa denne prihlasuje na sociálnu sieť Facebook, v Európe sieť používa **307 milióna** používateľov, denne sa nahrá približne **300 miliónov** fotiek, každú minútu sa uverejní vyše **pol milióna** komentárov, **5** profilov vznikne každú sekundu, pričom **83 miliónov** profilov je falošných,[8]
- každú minútu roku 2018, používatelia Instagramu pridali **49 380** fotiek, Google spracoval **3 887 140** vyhľadávaní, Skype uskutočňuje **176 220** hovorov **a 4 333 560** videí na Youtube je pozretých[9],
- každý deň sa pošle a príjme v priemere **281 miliárd emailov**, pričom predpokladaný rast toho čísla je **4 percentá** ročne[10]

Spoločnosti Seagate a International Data Corporation (ďalej len IDC) vo svojej správe z novembra roku 2018 uvádzajú, že globálna datasféra (angl. global datasphere) narastie zo súčasných 33 zettabajtov na 175 zettabajtov v roku 2025 (obrázok 1.2) , pričom takmer polovica

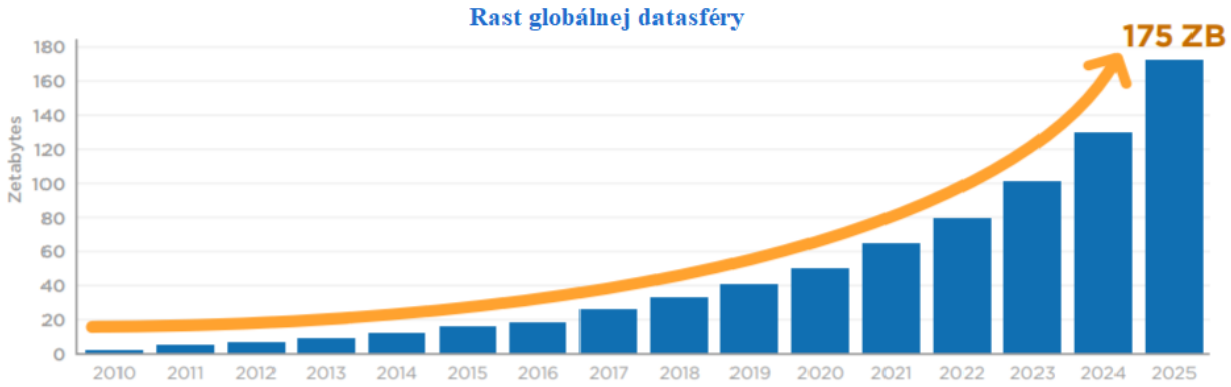
[7] MARR, Bernard. *How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read* [online]. 2018. [cit. 20.01.2019]. Dostupné na:https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/

[8]ZEPHORIA. *Top 15 Valuable Facebook Statistic*.[online]. 2019. [cit. 23.01.2019]. Dostupné na: https://zephoria.com/top-15-valuable-facebook-statistics/

[9] JAMES, Josh. *Data Never Sleeps 6.0* [online]. 2018. [cit. 23.01.2019]. Dostupné na: https://-www.domo.com/blog/data-never-sleeps-6/

[10] THE RADICATI GROUP, INC. *Email Statistics Report, 2018-2022* [online]. 2018. [cit. 23.01.2019].Dostupné na: https://www.radicati.com/wp/wp-content/uploads/2017/12/Email-Statistics-Report-2018-2022-Brochure.pdf

bude vo verejných cloudoch[11]. Pod pojmom globálna datasféra môžeme chápať celkový objem dát, ktorý ľudstvo vygeneruje a používa pre svoju činnosť.



Obrázok 1.2: Nárast globálnych dát do roku 2025
Zdroj: Seagate, 2018.

Ako môžeme vidieť na grafe vyššie, objem dát produkovaných každý rok sa prudko zvyšuje, preto s rozvojom Big Data prišli na rad dve dôležité otázky – a to, ako tieto dáta budú uchovávané či spracovávané a najmä, akú pridanú hodnotu dokážu priniesť. Analýza Big Data umožňuje spoločnostiam a podnikom túto hodnotu získať – ich spracovaním a vhodnou interpretáciou je možné optimalizovať výrobné procesy, zlepšiť segmentáciu zákazníkov a tým pádom využiť tzv. tailormade marketing, predpovedať trendy, detekovať podvody a mnoho ďalších. Touto analýzou sa zaoberá *data science* – čiže dátová veda, ktorá je bližšie predstavená v nasledujúcej kapitole tejto práce. Big Data v súčasnosti majú veľa praktických uplatnení a to napríklad v nasledujúcich oblastiach[12]:

- *analýza zákazníkov a marketing* – spoločnosti dokážu predikovať správanie sa zákazníkov a následne prispôsobovať svoju ponuku na základe trendov či preferencií.
- *šport* – lyžiarske strediská vďaka čipom v skipasoch dokážu nielen zabrániť podvodom, ale aj zistiť, ktoré zjazdovky a lanovky sú najviac využité a v akom čase či analyzovať vyťaženie strediska. National Football League (NFL), liga amerického futbalu, vyvinula vlastnú platformu aby svojim tímom uľahčila rozhodovanie vzhľadom na mnohé faktory, ako napríklad stav povrchu hracej plochy, počasie, individuálne výkony

[11]SEAGATE. *Digitalization of the World.* [online]. 2018. [cit. 23.01.2019]. Dostupné na: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf
[12]MARR, Bernard. *Big Data in Practice.* [online]. 2018. [cit. 23.01.2019]. Dostupné na: https://www.bernardmarr.com/default.asp?contentID=1076 6

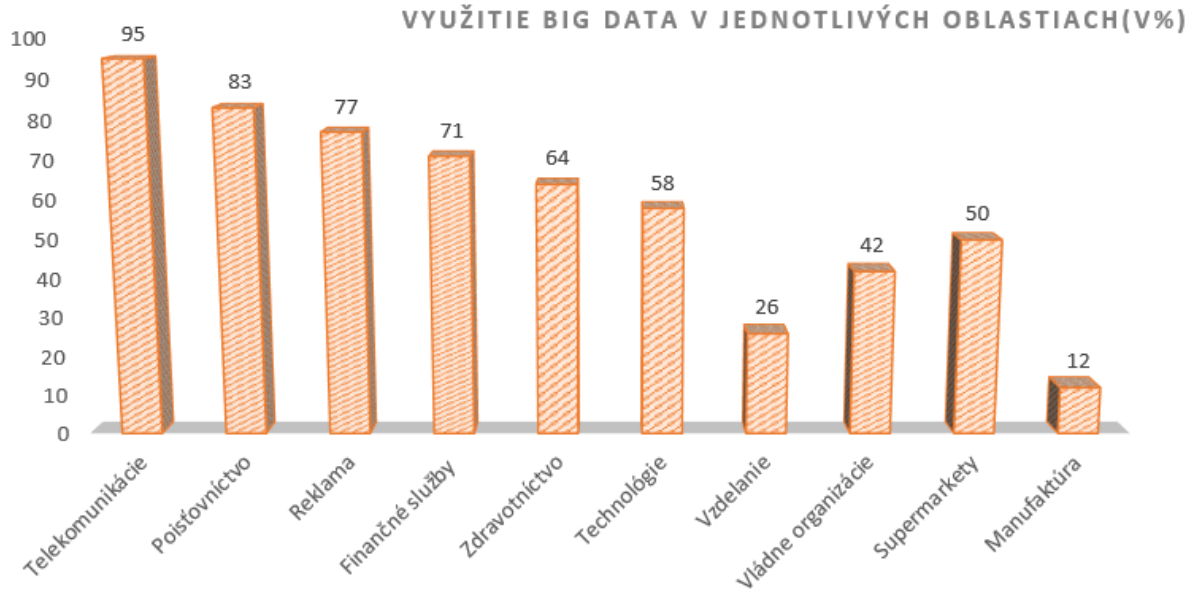
v reálnom čase. Týmto spôsobom tréner mužstva dokáže zvoliť ideálnu taktiku a zároveň redukovať množstvo úrazov.

- *optimalizácia zariadení, procesov a výkonov* – servery a počítače v datacentrách sú na základe dát o ich vyťaženosti a stave nastavené tak, aby sa čo najviac predĺžila ich životnosť. Americká Bank of a America pomocou analýzy dát zistila, že najefektívnejší pracovníci call centra sú tí, čo trávia pracovné prestávky spolu a zaviedla povinné skupinové prestávky. Celková výkonnosť následne stúpla o 23 percent.
- *zdravotníctvo* – pomocou senzorov a údajov zo smart zariadení je možné sledovať aktuálny zdravotný stav pacientov, čo urýchľuje dobu potrebnú na stanovenie diagnózy a predchádza chorobám. Big Data sa používajú napríklad na monitorovanie bábätiek – pomocou sledovania srdcového tepu a frekvencie dýchania, algoritmy dokážu zaznamenať anomálie a potenciálnu chorobu už 24 hodín predtým, ako sa dostavia prvé fyzické príznaky.
- *bezpečnosť* - štátne orgány dokážu predvídať a zabraňovať kriminálnej aktivite, bankové spoločnosti sú schopné odhaliť podvodné transakcie a poisťovne podvody pri poistnom plnení. V roku 2014 oddelenie polície v Chicagu posielalo policajné hliadky za jednotlivcami u ktorých bola pravdepodobnosť spáchania trestného činu najväčšia.
- *zefektívnenie chodu miest a obcí* – mestá dokážu na základe údajov zo senzorov a kamier optimalizovať dopravu, šetriť prevádzkové náklady, výdavky občanov a menia sa na tzv. smart cities, čiže múdre mestá. Mestu Los Angeles sa podarilo analýzou údajov o doprave znížiť zápchy v doprave o 16 percent, pričom počítačový systém v reálnom čase spravuje vyše 4 500 semaforov. Long Beach v Kalifornii pomocou monitorovania využitia závlahových systémov poskytlo spätnú väzbu obyvateľom a niektorí občanie zredukovali použitie vody na zavlažovanie až o 80 percent.

Spoločnosť Dresner Advisory Services vo svojej správe o využití Big Data medzi podnikmi uvádza, že v roku 2018 ich využívalo 59% spoločností, čo predstavuje nárast o 36 percent oproti roku 2015[13]. Z hľadiska regionálneho členenia vedie Latinská Amerika (76%), ďalej nasleduje oblasť EMEA (63%), Severná Amerika (57%) a v oblasti Ázie a Pacifiku je

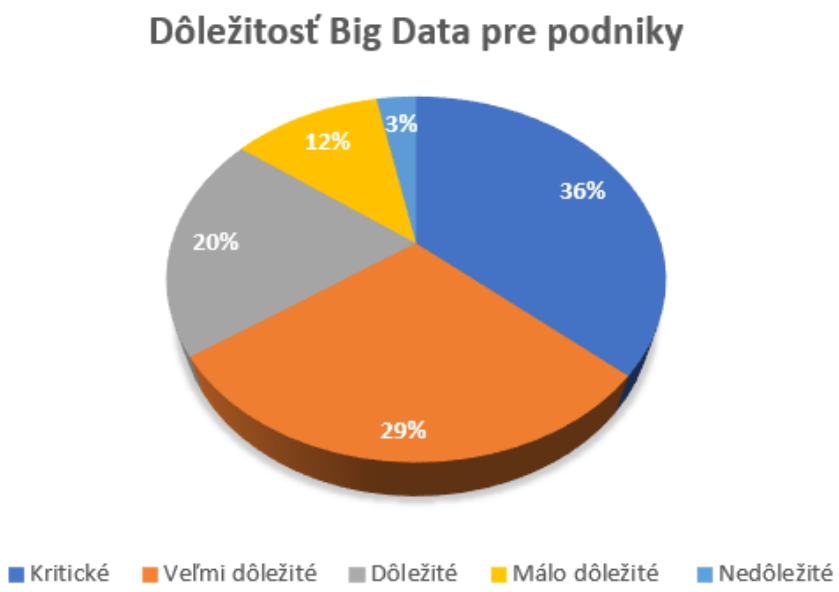
[13] COLUMBUS, Louis. *Big Data Analytics Adoption Soared In The Enterprise In 2018*. [online]. 2018. [cit. 24.01.2019]. Dostupné na: https://www.forbes.com/sites/louiscolumbus-/2018/12/23/big-data-analytics-adoption-soared-in-the-enterprise-in-2018/

využitie na úrovni 50 percent. V rámci jednotlivých odvetví dominujú telekomunikácie a v tesnom závese poisťovníctvo, významný podiel majú aj reklamné spoločnosti, finančné služby či zdravotníctvo(obrázok 1.3).



Obrázok 1.3: Využitie Big Data v jednotlivých oblastiach
Zdroj: vlastné spracovanie podľa Dresner Advisory Services

S rastúcim trendom používania Big Data a stále zvyšujúcou rýchlosťou ich adaptovania pri činnosti podnikov, sa ich použitie stáva čoraz viac signifikantným pre chod spoločností. Až 65 percent podnikov ich považuje za kritickú či veľmi dôležitú časť ich biznisu(obrázok 1.3). Na základe aktuálneho vývoja v tejto oblasti môžeme predpokladať, že Big Data budú v najbližších rokoch integrované čoraz viac a to najmä v oblastiach, kde majú malé zastúpenie. Odvetvia ako telekomunikácie, poisťovníctvo a aktuárstvo či technologické spoločnosti, ktoré ich do značnej miery používajú už v súčasnosti, sa budú venovať najmä spresneniu výsledkov dosiahnutých analýzou týchto dát a zavádzaním automatizovaných algoritmov a umelých inteligencií založených práve na Big Data. Kontinuálnym rastom miery integrácie a celkového objemu dát sa priamo úmerne aj zvyšuje presnosť a spoľahlivosť výsledkov získaných z týchto dát, čo spôsobí urýchlenie rozhodovania sa, prinesie nové produkty a umožní lepšie pochopenie zákazníka a jeho potrieb.



Obrázok 1.4: Dôležitosť Big Data pre podniky
Zdroj: vlastné spracovanie podľa Dresner Advisory Services.

1.4 Big Data a poistný trh

Poisťovníctvo ako odvetvie zohráva dôležitú úlohu v živote ľudí a firiem, nakoľko sa priamo zaoberá bezpečnosťou jednotlivcov a ich majetkov, pričom princíp fungovania tohto odvetvia je na založený na existencii rizika. Ďalším jeho špecifikom je fakt, že poisťovníctvo nevyrába ani nespravuje žiaden fyzický produkt. V takomto prípade sú dáta pre poisťovne jedným z ich najhodnotnejších aktív. Vo všeobecnosti môžeme konštatovať, že celý princíp fungovania poisťovníctva je založený na predikcii rôznych rizík. Vysoko kompetitívny charakter tohto odvetvia a neustále zmeny správania zákazníkov tvoria predpoklad pre neustále investície do nových, efektívnych a presných spôsobov ako predpovedať toto správanie a minimalizovať ich riziko. Všetky tieto skutočnosti naznačujú a vytvárajú predpoklady pre použitie Big Data v poisťovníctve. Ako sme mali možnosť vidieť na obrázku 1.2 v predchádzajúcej kapitole, toto odvetvie si už v súčasnosti veľkom osvojilo používanie Big Data pre svoju činnosť, pričom sa odhaduje, že zložená ročná miera rastu (CAGR) v najbližších troch rokoch bude na úrovni 14 percent[14].

[14] RESEARCH AND MARKETS. The *Big Data Market: 2018 - 2030 - Opportunities, Challenges, Strategies, Industry Verticals & Forecasts* [online]. 2018. [cit. 27.01.2019]. Dostupné na: https://www.researchandmarkets.com/reports/4564313/the-big-data-market-2018-2030-opportunities

Jednou z najpoužívanejších techník pri práci aktuára je prediktívne modelovanie a analyzovanie. Ide o metódu, ktorá umožňuje určiť, čo sa udeje v budúcnosti na základe porozumenia a merania historických údajov. Následne sa vytvárajú modely založené na rôznych vzťahoch medzi jednotlivými premennými, ktoré boli objavené v týchto údajoch[15]. Prediktívne modelovanie je kľúčovým nástrojom data science a Big Data, pričom poisťovníctvo a aktuárstvo predstavuje odvetvie, ktoré tieto techniky dokázali rýchlo adaptovať. Táto metóda sa používa v akutárstve najmä v nasledovných oblastiach:

- **posúdenie a hodnotenie rizík – risk assesment** : jedným z najdôležitejších oblastí pre poisťovne je stanovenie výšky poistného. Používa sa v životnom a neživotnom poistení a to napríklad pri automobiloch, zdravotnom a poistení domácností. Poistenie dokážu používať telekinematiku , IoT (Internet of Things) zariadenia či iné smart zariadenia na sledovanie zákazníkov a tým pádom predpovedať dané riziko. Pomocou prediktívneho modelovania spoločnosti dokážu identifikovať pravdepodobnosť havárie, odcudzenia automobilu alebo nastanie inej poistnej udalosti pomocou kombinácie behaviorálnych dát a iných externých faktorov(ako napr. počasie, stav vozovky či charateristiky obytnej štvrte. V životnom poistení zase fitness náramky či smart hodinky umožňujú sledovanie správania a zvykov zákazníkov, čo umožňuje určenie rizika na základe ich návykov.
- **odhalenie poistných podvodov** – **fraud detection**: dátová analýza spolu s kombináciou prediktívneho modelovania zlepšujú a umožňujú odhalenie kriminálnej aktivity. Sledovaním premenných v každom poistnom plnení a ich následných porovnaním s poistnými podvodmi dokážu poisťovne vytvoriť profil podozrivých nárokov na poistné. Pri odhalení zhody sa takýto prípad automaticky označí na ďalšie, podrobnejšie preskúmanie.
- **informácie o zákazníkoch** – **customer insights**: získavanie komplexného pochopenia správania, zvykov a potrieb zákazníkov z rôznych zdrojov je veľmi dôležité, aby poisťovne mohli predvídať budúce správanie, správne produkty a identifikovať správne segmentácie zákazníkov. Informácie získané z call centier, e-mailov, sociálnych médií užívateľských fór, či správanie zákazníkov na stránkach spoločností umožňuje poisťovateľom vytvorenie unikátneho profilu jednotlivého zákazníka. Získanie

[15] MARR, Bernard. *How Big Data is Changing Insurance Forever* [online]. 2015. [cit. 27.01.2019]. Dostupné na: https://www.forbes.com/sites/bernardmarr/2015/12/16/how-big-data-is-changing-the-insurance-industry-forever/

prehľadu o zákazníkoch pomocou Big Data poskytuje predpovede nielen o tom, kedy zákazník pravdepodobne odíde ku konkurencii alebo formuje politiku zákazníka. To môže pomôcť poisťovateľom rozvíjať dôverné vzťahy a integrovať zákazníkov správnym spôsobom a s presnými informáciami. Výsledkom týchto poznatkov je, že sú dosiahnuté pozitívne výsledky ako napríklad riešenie problémov zákazníkov v reálnom čase so správnym prístupom a zlepšenie vzájomných vzťahov.

- **marketing**: poisťovatelia používajú prediktívnu analýzu na trhu pre spotrebiteľov. Spoločnosti môžu sledovať správanie spotrebiteľov v rôznych formách a vytvárať cielené reklamy, zbierať informácie pomocou cookies a iných mechanizmov či budovať modely „sklonu ku kúpe“ aby sa mohli zamerať na tých, u ktorých je pravdepodobnejšie že si kúpia produkt. Tieto aktivity môžu pomôcť k zníženiu nákladov na reklamu, čo vedie k zníženiu celkových nákladov a tieto marketingové fondy prerozdeliť na iné účely.
- **vývoj produktov**: prediktívna analýza umožňuje nájsť nové oblasti pôsobenia a navrhnúť nové produkty šité na mieru či zlepšiť už existujúce. Poisťovne následne môžu na trh dodať lepšie produkty založené na rozbore predchádzajúcich údajov o poistení, zdravotných záznamov, spôsobu riadenia vozidla a životného štýlu.
- **automatizácia procesov**: poisťovatelia používali automatizáciu najmä pri jednoduchých a opakujúcich sa úkonoch, ktoré nevyžadovali ľudskú interakciu ako napríklad kontrola povinných polí či formátu dát. S nárastom Big Data je možné automatizovať aj komplexnejšie a zložitejšie úkony ako upisovanie úverov, posúdenie majetku, overovanie nároku na plnenie či odhalenie podvodov. Posunom k inteligentnejšej automatizácií môžu spoločnosti ušetriť obrovské množstvo času a peňazí pomocou strojového učenia a samozrejme, prediktívnej analýzy.
- **data driven decision making – DDDR**: prediktívne modelovanie dokáže napodobniť ľudské rozhodovanie na vytváranie nových pravidiel, ktoré sú lepšie ako tie používané predtým. Tým pádom sú získané výsledky mapované rýchlejšie a s väčšou presnosťou. Toto modelovanie prináša nesporné výhody, ale aj určité obmedzenia – modelovanie ľudského správania môže spôsobiť, že v algoritmoch ostanú zachované ľudské predsudky.

Spomínané techniky je možné využiť v životnom aj neživotnom poistení. Rozvoj Big Data umožňuje nielen použitie týchto techník, ale prinášajú poisťovniam a viac údajov, čo napomáha skvalitňovať a ich produkty a mať lepší prehľad o zákazníkov.

V oblasti životného poistenia patria medzi tradičné zdroje dát najmä interné záznamy poisťovní, kde patrí napríklad vek poistenca, pohlavie, údaje o zdravotnom stave či sociálnom statuse a iné. Nové, rozvíjajúce sa dáta pre životné poistenie zahŕňajú:

- **dáta zozbierané z marketinového oddelenia**, ktoré zahŕňajú informácie o preferenciách spotrebiteľa.
- **Verejné záznamy, kriminálna história, demografické údaje, genetické informácie, elektronické lekárske záznamy, história predpisov liekov, životný štýl a dáta z nosičov ako sú napríklad smart hodinky a fitness náramky**. Niektoré z týchto údajov sa používajú pre uzatvorením poistnej zmluvy, iné počas a po uzavretí ako určitá forma kontroly.
- **Dáta zo sociálnych sietí** ako napríklad Facebook, Instagram či Snapchat. Tieto dáta môžu odhaliť napríklad fajčenie alebo užívanie alkoholu, či iné dôležité informácie o zdravotnom stave poistenca.
- **Údaje o príjme a majetku**, čo umožňuje lepšie zaradenie do jednotlivých rizikových skupín a marketing.

Medzi často používané dáta v tejto oblasti patria najmä dáta zo smart zariadení – hodinky a fitness náramky. Vďaka dátam nazbieraným pomocou týchto nosičov poisťovne majú komplexný prehľad o životnom štýle a návykoch daných osôb. Priekopníkom v tejto oblasti je poisťovňa John Hancock, ktorá ako prvá zaviedla zľavu na poistnom pre používateľov fitness náramkov Fitbit. Ide o náramok, ktorý je prepojený s internetom a odosiela údaje o fyzickej aktivite zákazníka, tento náramok svojím klientom zadarmo zabezpečí poisťovňa. Spoločnosť uvádza, že klienti zapojení do tohto programu môžu získať zľavu na životnom poistení až do výšky 15 percent[16]. Dáta sú pravidelne odosielané cez internet a používatelia tohto programu dostávajú body za svoju fyzickú aktivitu. Následne môžu túto aktivitu sledovať a získané body premeniť na rôzne zľavy a iné odmeny. Prezident pre finančné služby tejto poisťovne, Craig Bromley povedal, že spoločnosť verí, že táto ponuka zlepší životné poistenie pre nové kategórie

[16] MEARIAN, Lucas. *Insurance Company Now Offers Discounts* [online]. 2015. [cit. 28.01.2019]. Dostupné na: https://www.computerworld.com/article/2911594/insurance-company-now-offers-discounts-if-you-let-it-track-your-fitbit.html

spotrebiteľov a oživí celú kategóriu. Podľa prieskumu spoločnosti Accenture, tretina poisťovní v Severnej Amerike má vo svojom portfóliu podobný produkt, založený na sledovaní fyzickej aktivity spotrebiteľa[17].

Nové zdroje dát sa objavujú aj v neživotnom prostredí – pri poistení majetku či úrazovom poistení. Kým ku tradičným zdrojom dát v tejto oblasti patrili napríklad údaje o vodičovi, počet rokov skúseností, typ vozidla, počet najazdených kilometrov ročne(poistenie automobilov), typ konštrukcie budovy, geografická lokalita, druh zabezpečenia budov proti ohňu a vlámaniu(poistenie majetku) či údaje o oblasti podnikania. Nové zdroje dát v tejto oblasti sú:

- **Údaje o počasí, kriminalite, hustota obyvateľstva, hustota premávky, informácie zo sčítania obyvateľstva** – pre všetky druhy poistenia.
- **Telematické zariadenia v automobiloch**, ktoré uľahčujú získať detailné informácie o správaní vodiča Tieto údaje sa používajú pri poistení automobilov pre firmy a jednotlivcov.
- **Mobilné dáta**, dáta zo zariadení IoT a iných technológií napríklad z inteligentných domov – najmä pri poistení majetku a podnikania.
- **Dáta zo smart zariadení, mobilov a fitness náramkov, elektronické zdravotné záznamy, záznamy o činnosti na internete** ako napríklad história hľadaných výrazov – pri úrazovom poistení. Do tejto kategórie patria aj údaje o genofonde zákazníka a genetických preddispozíciach na základe analýzy DNA, čo umožňuje genetické profilovanie.

Používanie nových dátových zdrojov v neživotnom poistení nielen zlepšuje segmentáciu zákazníkov a otvára nové možnosti poisťovniam, vďaka možnosti spracovania dát sú v reálnom čase dosiahnuté presnejšie výsledky. Až do rozvoja technológie Big Data sa aktuárstvo pri výpočtoch spoliehalo najmä na historické dáta. Klimatické zmeny sa v súčasnosti dejú rýchlejšie ako kedykoľvek predtým a globálna klíma má odlišný charakter ako pred rokmi. Tieto zmeny predstavujú nové riziko pre poisťovne a preto je dôležité ich identifikovať a odhadnúť aktuálne trendy. Pomocou Big Data môžu poisťovne tieto skutočnosti odhaliť

[17] MARR, Bernard. *How Big Data is Changing Insurance Forever* [online]. 2015. [cit. 27.01.2019]. Dostupné na: https://www.forbes.com/sites/bernardmarr/2015/12/16/how-big-data-is-changing-the-insurance-industry-forever/

vďaka použitiu veľkého množstva senzorov, ktoré zachytávajú informácie o počasí, barometrickom tlaku, teplote, zmeny v prúdeniach vzduchu a podobne.

Dôležitým aspektom pri použití dát z nových zdrojov je ich kvalita. O kvalite týchto údajov pojednáva aj smernica EÚ s názvom Solventnosť II. Ide o smernicu, ktorá sa sťahuje na všetky poistné spoločnosti, ktoré pôsobia v Európskej únií a umožňuje lepšie pokryť riziká, ktorým musia poisťovne čeliť. Toto nové nariadenie vyžaduje posúdenie budúceho vývoja, pretože práve ten dokáže ovplyvniť finančnú situáciu poisťovateľa. Toto nariadenie definuje tri štandardy pre dátovú kvalitu podľa článku 82 a to nasledovné:

- **presnosť,**
- **úplnosť,**
- **kompletnosť.**

Dodržiavanie týchto štandardov eliminuje riziko zlej interpretácie údajov, čo môže viesť k nesprávnym rozhodnutiam a nedorozumeniam. Použitie nových, často neštruktúrovaných údajov prináša nové výzvy pre spracovanie týchto dát. Solventnosť II má za úlohu zabezpečiť, aby všetky údaje boli presné, vhodné a úplné.

Platí však, že v každom systéme sa nachádzajú chybné údaje a to buď z dôvodu ľudského zlyhanie (preklep, nesprávne pole...), ale z dôvodu inkonzistencie informačných systémov podniku. V prípade Big Data je logické, že dáta budú obsahovať viac chýb, nakoľko ide o agregované dáta prichádzajúce z rôznych zdrojov. Pri transformácií týchto údajov môže dôjsť k chybám spôsobeným rozdielnou technológiou či zle nastaveným procesom. Preto je dôležité, aby si jednotlivé spoločnosti definovali minimálnu kvalitu dát, najmä pri spracovaní a analýze týchto údajov. V poistnom priemysle sa využívajú najmä aktuárske, finančné, majetkové a rizikové údaje, ktoré sú podľa normy Solventnosť II klasifikované ako analytické údaje. Tieto údaje sa od tradičných dát používaných v poisťovniach líšia napríklad vyššou úrovňou granularity.

Ďalším, rozmáhajúcim sa trendom v oblasti neživotného poistenia motorových vozidiel predstavuje tzv. *Usage based insurance* (UBI), známe aj ako *Pay As you Drive* (PAYD) alebo *Pay How you Drive* (PHYD), čiže poistenie založené na používaní alebo plať ako jazdíš. Základnou myšlienkou UBI je, že správanie vodiča je monitorované priamo počas toho, kedy dotyčná osoba jazdí, čo umožňuje poisťovateľom získať komplexnú predstavu o spôsobe jazdy a tým pádom poskytovať individuálny prístup ku klientom. UBI je založené na telematike, ide o vedu, ktorá sa zaoberá diaľkovým prenosom a spracovaním dát. Najčastejšie používanou v

praxi je dopravná telematika, kde ide o prenos údajov o vozidle do vzdialeného počítača, pričom sa využíva technológia GPS. Medzi najčastejšie zbierané údaje patrí počet kilometrov, poloha, rýchlosť, indikácia rýchleho zrýchlenia a spomalenia, použitie smeroviek, spôsob prechodu zákrut, použitie airbagov a iné. Úroveň zozbieraných údajov vo všeobecnosti závisí od typu použitej technológie a ochoty poistencov odosielať osobné údaje. Poisťovňa následne tieto údaje analyzuje a určí výšku poistného. Napríklad vodič často jazdiaci dlhé trasy jazdiaci vo vysokých rýchlostiach bude mať vyššie poistné ako vodič jazdiaci krátke trasy pri nízkych rýchlostiach, Spôsob zberu týchto údajov sa neustále vyvíja - v začiatkoch(približne od roku 2005) sa do automobilov montovali prídavné zariadenia, ekvivalenty čiernych skriniek v leteckom priemysle, následne sa používali zariadenia na princípe USB, ktoré sa vložili do auta a v súčasnosti prevládajú mobilné aplikácie.

Tento princíp v dnešnej dobre už používajú mnohé poisťovne, ako napríklad All-state, ‚Metromile, Root Insurance (USA), AIOI (Japonsko), Generali (EÚ), Wubi (Česká republika) či Allianz (Nemecko). Použitie tohto typu poistenia prináša poistencom viacero benefitov, najmä tým, čo ročne najazdia menej kilometrov – namiesto fixnej platby platia len za prejdené kilometre. Americká spoločnosť Metromile na svojich stránkach uvádza úsporu až 611 dolárov pri prejdení 5000 míľ (8046km) ročne, vo všeobecnosti dokážu vodiči s nízkym počtom kilometrov ušetriť až 50 percent za rok[18]. Implementácia a využitie telematiky prináša do tohto segmentu poistenia veľa výhod. Poisťovne získavajú komplexný obraz o spôsobe jazdy, správaní sa zákazníkov a reálnom využití vozidiel a z týchto údajov môžu využiť aj jednotlivé podniky a to prehľadom o využití firemných vozidiel a správania a zamestnancov k nim. Zákazníkov zase láka vidina úspory na poistnom – okrem príležitostných vodičov získavajú benefit aj mladí vodiči, nakoľko pri klasickom spôsobe poistenia často platia vyššie poistné z dôvodu ich nízkeho veku a skúseností. Asociácia GSMA združujúca mobilných operátorov vo svojej správe uvádza, že zavádzanie telematiky sa do roku 2015 zdvojnásobí, pričom rozlišuje 3 spôsoby implementácie v tejto oblasti a to:

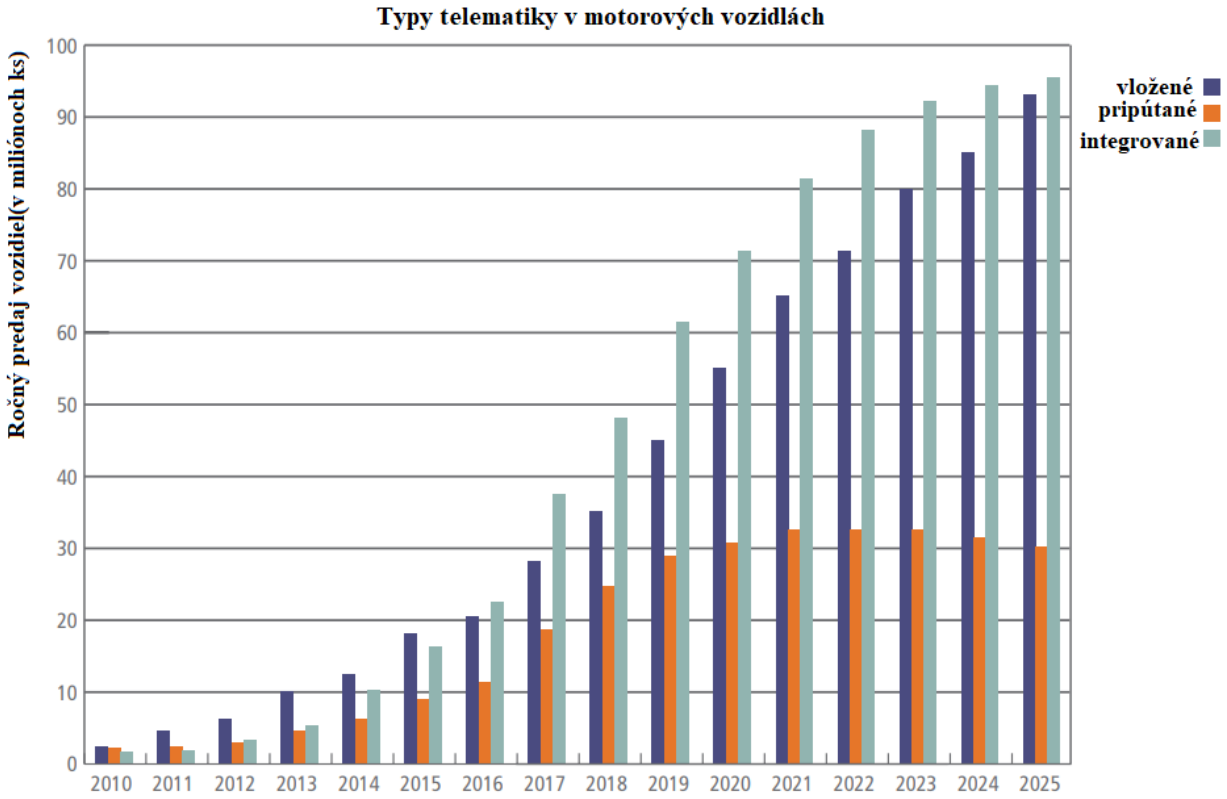
- **vložené** – ide o inštaláciu prídavného zariadenia do motorového vozidla. Tieto zariadenia poskytujú poistné spoločnosti a zber údajov sa uskutočňuje pomocou týchto

[18] FINANCIAL WEB. *Pros and Cons of Pay as You Drive Insurance.* [online]. 2016. [cit. 02.02.2019]. Dostupné na: https://www.finweb.com/insurance/pros-and-cons-of-pay-as-you-drive-insurance.html

mechanizmov. V súčasnosti ide o malé skrinky, ktoré sú namontované fixne do vozidla alebo zapojené do palubnej dosky.

- **pripútané** - vodič vozidla používa mobilný telefón na odosielanie informácií pomocou mobilných dát, wifi či Bluetooth. Nevýhodou tohto typu je že je závislé na mobilnom zariadení, čo môže spôsobiť vyššie účty za telefón a takisto stratu dát v prípade výpadku mobilnej siete či zariadenia
- **integrované** – aplikácie sú priamo integrované do automobilov, ide o inteligentné automobily, ktoré obsahujú internetovú komunikáciu, pričom odpadá nutnosť použitia mobilného telefónu používateľa – systémy ako Apple CarPlay.

GSMA predpokladá, že práve integrované riešenia budú v budúcnosti dominovať(obrázok 1.4)



Obrázok 1.5: Použitie rôznych typov telematiky v motorových vozidlách
Zdroj: https://www.gsma.com/iot/wp-content/uploads/2012/03/gsma2025everycarconnected.pdf

Použitie telematiky so sebou takisto prináša aj určité nevýhody, plynúce najmä z nutnosti zdieľania dát s poisťovateľom. V prípade inštalácie zariadenia alebo použitia mobilnej aplikácie na zber týchto údajov, poisťovňa dostane všetky relevantné informácie, čo veľa zákazníkov považuje za narušenie ich súkromia, nakoľko sa odosielajú údaje o polohe,

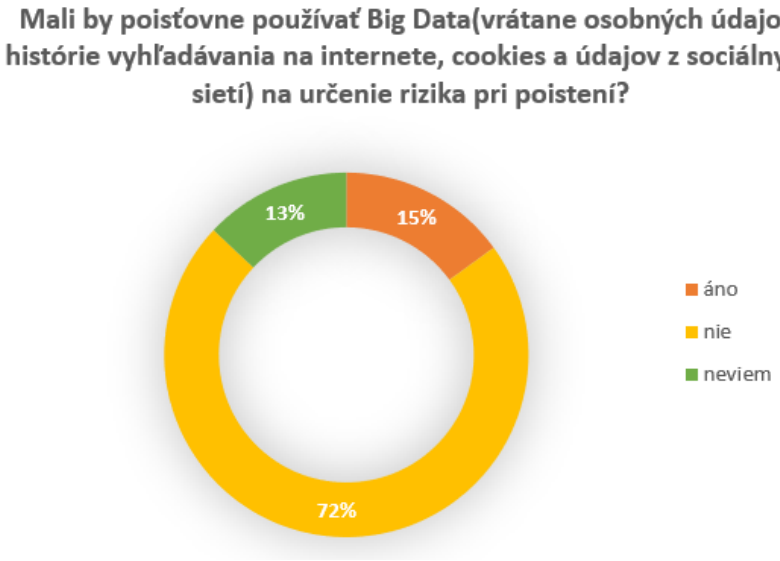
spôsobe jazdy a rýchlosti. Analýza týchto dát nemusí priniesť výhody, práve naopak – pre určitých zákazníkov môže viesť k zvýšeniu poistného(napríklad v prípade agresívneho štýlu jazdy) či intenzívneho používania motorového vozidla. Vo všeobecnosti PHYD poistenie viac oplatí vodičom s nízkym počtom najazdených kilometrov ročne.

Použitie Big Data v aktuárstve a poisťovníctve nesporne prináša veľa výhod a spoločnosti získavajú lepší prehľad o produktoch či zákazníkoch – avšak treba brať do úvahy aj etický aspekt použitia týchto údajov. Aktuári majú povinnosť vyplývajúce z ich profesionalizmu dodržiavať povesť poistno-matematickej profesie a nesú zodpovednosť voči verejnosti v novovznikajúcej oblasti Big Data. Dôležitou súčasťou je dodržiavanie regulácií a zákona. V mnohých situáciách majú aktuári jedinečný prehľad o výsledkoch a dôsledkoch použitiach týchto dát a musia byť ochotný s schopný tieto poznatky kľúčovým zainteresovaným stranám ako napríklad regulačné úrady, auditori, verejnosť či exekutíve spoločnosti. Vysvetlenie svojich záverov je jedným z kľúčových atribútov poistných matematikov. Odborný posudok aktuárov je rozhodujúci pri využívaní Big Data v poisťovniach. Tento posudok sa týka nielen pridanej hodnoty, ktorú Big Data prinášajú, ale aj etickosti ich využitia, pričom treba dodržiavať všetky regulácie tohto odvetvia a príslušné zákony jednotlivých krajín – napríklad v krajinách Európskej únie ochranu osobných údajov všeobecné nariadenie GDPR (General Data Protection Regulation). Toto všeobecne záväzné nariadenie vstúpilo do platnosti v roku 2018, pričom povinnosť dodržiavať ho majú všetky spoločnosti a organizácie spracovávajú osobné údaje a pôsobia v krajinách EÚ28, poisťovne nevynímajúc. Spoločnosti, ktoré nesplnia podmienky tohto nariadenia alebo nezabezpečia bezpečnosť osobných údajov môžu dostať pokuty až do výšky 20 miliónov eur, prípadne 4% z obratu[19]. Nové dáta ako údaje z fitness hodiniek, zdravotnom stave, sociálnych sietí, automobilov a iných zdrojov predošle uvádzaných sa dajú ľahko zneužiť. Odhalenie zneužitia údajov môže viesť k podniknutiu právnych krokov voči poisťovniam, pokutám, zákazu činnosti a poškodeniu dobrého mena spoločnosti. Je preto na individuálnom posúdení aktuára, vedenia spoločnosti a iných zúčastnených strán, ako a do akej miery budú tieto nové poznatky použité.

Internetový portál Lendedu v roku 2018 podnikol prieskum verejnej mienky odľadom použitia Big Data v poisťovníctve. V prieskume odpovedalo 1000 respondentov, pričom iba 15 percent opýtaných si myslí, že poisťovne by mali používať Big Data (vrátane osobných údajov,

[19] Viac informácií na: https://gdpr-slovensko.sk/co-je-gdpr/

histórie vyhľadávania na internete, cookies a údajov z sociálnych sietí) na určenie rizika pri poistení(viď obrázok 1.5).



Obrázok 1.9:Názor verejnosti na používanie Big Data v poistnom odvetví
Zdroj: vlastné spracovanie na základe Lendedu

V prieskume takisto 55 percent opýtaných uviedlo, že považuje používanie Big Data v poistení za rovnakú hrozbu ako používanie týchto osobných dát technologickými spoločnosťami ako napríklad Facebook, ktoré tieto dáta používajú na cielený marketing a reklamu. Majorita opýtaných je aj proti možnosti inštalácie biometrického zariadenia do ich tela, aj keď by to znamenalo nižšiu cenu poistného, jednak úrazového a aj životného.

Big Data je nepochybne nástrojom, ktorý prináša a v budúcnosti prinesie veľké množstvo pozitívnych zmien v poisťovníctve a to vo forme lepšieho zákazníckeho servisu, pochopenia jeho potrieb, zníženie celkových škôd spôsobných podvodmi. Je to aj odvetvie, ktoré prináša jedinečný súbor výziev, no zároveň so sebou nesie obavy o súkromie poistencov. Samotné poisťovne by sa mali snažiť s ich využitím na poskytovanie lepšej hodnoty pre zákazníka, za dodržania regulácií a etických kódexov, aby túto pridanú hodnotu vnímali ich zákazníci pozitívnom slova zmysle. V prípade dodržania všetkých morálnych zásad a zákonov, Big Data prinesú v budúcnosti veľké zmeny do tohto odvetvia a ich použitie a dôležitosť bude v priebehu najbližších rokov narastať.

1.5 Jazyk R a Data Science

R ako programovací jazyk je využívaný najmä na štatistické výpočty, dátové analýzy a grafickú interpretáciu dát. Bol vytvorený v 90-tych rokoch minulého storočia Rossom Ihaka a Robertom Gentlemanom. Ich primárnou úlohou bolo vytvoriť výkonnú platformu pre štatistické účely, ktorá sa zameriava na čistenie, analýzu a reprezentáciu dát. Vo svojich začiatkoch R ako programovací jazyk nezískal veľkú popularitu, tá prišla až o takmer 20 rokov neskôr. V dnešnej dobe hrá jazyk R dôležitú úlohu a svoj potenciál naplno využil s príchodom Big Data a rozmachom Data Science. Spoločnosť Burtch Works vo svojej ankete medzi dátovými vedcami a analytikmi z roku 2018 uvádza, že 33 percent opýtaných preferuje R pred programovacím jazykom Python a softvérom SAS[20]. Stránka KDNuggets, zameraná na trendy v oblasti Big Data a Data Science, ho uvádza ako tretí najpopulárnejší nástroj používaný medzi profesionálmi zaoberajúcimi sa touto tematikou[21]. Tieto čísla nám dokazujú, akú popularitu má jazyk R v súčasnej dobe.

Od svojho vytvorenia prešlo R dlhým a postupným vývojom než sa dostalo do podoby, v ktorej ho poznáme dnes. K jeho popularite nemalou časťou prispelo aj R Studio – ide o grafický front-end, ktorý je rozšírením R a ponúka príjemné používateľské rozhranie. Vďaka jeho open source charakteru sa R významnou mierou rozvíja v podobe dodatočných balíčkov resp. knižníc, ktoré boli v priebehu rokov vytvorené. Tieto knižnice rozširujú funkcionalitu jazyka a sú voľne dostupné v centrálnom repozitári CRAN (Comprehensive R Archive Network). V súčasnej dobe má tento repozitár vyše 14 tisíc knižníc. Okrem knižníc v tomto repozitári, je možné stiahnuť a nainštalovať aj knižnice ktoré sú voľne dostupné. To však so sebou prináša riziko – pri týchto knižniciach nie je žiadna záruka kompatibility s najnovšou verziou jazyka R.

Za úspechom R v oblasti spracovania dát a ich analýze stojí viacero faktorov. Okrem dôvodov spomínaných vyššie sú to najmä:

- **Akademické zázemie** – R patrí medzi veľmi populárne jazyky na univerzitách a prácach výskumníkov. Často je vyučované na univerzitách a predstavuje pomyslený „odrazový

[20] NEW GEN APPS. *6 Reasons Why Choose R Programming For Data Science Projects* [online]. 2017. [cit. 02.02.2019]. Dostupné na: https://www.newgenapps.com/blog/6-reasons-why-choose-r-programming-for-data-science-projects

[21] PIATETSKY, Gregory. *19th Annual KDnuggets Sofware Poll* [online]. 2018. [cit. 02.02.2019].Dostupné na: https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html

mostík“ k Data Science. To prirodzene znamená, že pokiaľ študenti počas štúdia budú používať R tak môžeme predpokladať, že ho budú používať aj následne v praxi.

- **Úprava dát** – čistenie a rôzne iné operácie s dátovými setmi hrajú dôležitú úlohu pri Data Science – čím kvalitnejšie dáta máme k dispozícií, tým kvalitnejší bude aj výsledkov spracovania týchto údajov. Tento proces je často časovo náročný. R ponúka viacero knižníc, ktoré nám umožňujú efektívnu manipuláciu a úpravu dát, napr. knižnice *dplyr, data.table* či *readr*.
- **Rozsiahle možnosti vizualizácie** – pri analýze dát je dôležité nielen získať požadované výsledky, ale ja ich názorne interpretovať. R ponúka množstvo knižníc, ktoré umožňujú kvalitné spracovanie grafických výstupov.
- **Špecifickosť** – primárnou úlohou R je štatistická analýza a práca s dátovými setmi. Takmer všetky knižnice dostupné pre R majú spoločný cieľ – detailnejšiu, jednoduchšiu a efektívnejšiu analýzu dát. Tieto predpoklady robia z R výborného pomocníka pri Data Science.
- **Machine learning** – dátový vedec pri svojom výskume môže použiť rôzne algoritmy a postupy na automatizáciu určitých úloh a predikcií. R ponúka knižnice zamerané presne na túto oblasť, ktoré sa napríklad dokážu vysporiadať s chýbajúcimi hodnotami, vytvoriť rozhodovacie stromy, deliť dáta na partície a mnohé iné.
- **Dostupnosť** – R je open source, ktokoľvek si ho môže zadarmo vyskúšať. Môže byť použité pri projektoch rôznej veľkosti. Množstvo developerov pracuje na neustálom vylepšovaní aktuálnych knižníc a vývoji nových. Podporuje aj množstvo platforiem, ako napríklad Linux/Unix, Windows, MacOS a iné.

Jazyk R umožňuje používanie širokej škály štatistických a grafických techník, ako napríklad lineárne modelovanie, analýza časových radov, štatistické testovanie, klastrovanie a podobne. Všetky tieto skutočnosti robia z neho robia ideálnu voľbu pre Data Science, analýzu Big Data a machine learning. Aktuár vo svojej praxi tiež používa spomenuté techniky, preto nie je prekvapením, že R je populárnym nástrojom aj v tejto oblasti.

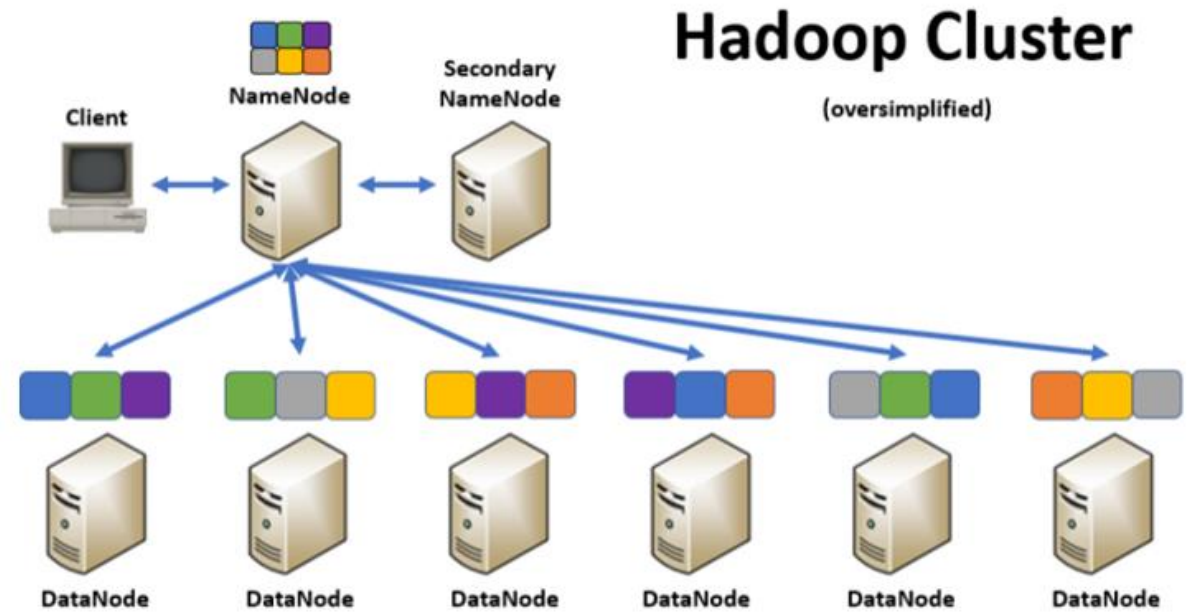
1.6 Apache Hadoop

Hadoop predstavuje open source projekt od spoločnosti Apache. Ide o softvérový framework, ktorý dokáže spracovať rozsiahle súbory v rámci počítačového klastra pomocou distribuovaného spracovania dát. Jednou z kľúčových charakteristík Hadoopu je jeho škálovateľnosť – dokáže pôsobiť na jednom serveri, ale aj na tisícoch počítačov. V súčasnosti ide o vedúcu platformu na spracovanie Big Data.

Obsahuje viacero komponentov, ktoré zabezpečujú fungovanie klastra. Základné komponenty sú:

- **HDFS –** Hadoop Distributed File System, distribuovaný súborový systém používaný na spracovanie veľkých dát pozdĺž klastra. Je určený na beh na komoditnom hardvéri a vysoko odolným voči jeho zlyhaniam. Dáta uložené v ňom sú uložené v blokoch, ktoré sú následne replikované. Je napísaný v jazyku Java.
- **MapReduce** – programový model určený na spracovanie veľkých dát. Je navrhnutý tak, aby poskytoval spracované výsledky v prijateľnom čase. Pôvodne bol vyvinutý spoločnosťou Google.
- **YARN** – Yet Another Resource Negotiator, doslova ďalší vyjednávač zdrojov. Jeho úlohou je efektívna alokácia zdrojov klastra na vykonávanie jednotlivých úloh.

Apache Hadoop je založený na master/slave architektúre. Klaster pozostáva z viacerých serverov/počítačov, ktoré nazývame nódy respektíve uzly. Zvyčajne obsahuje aspoň jednu NameNode – hlavný server, ktorý riadi a rozdeľuje úlohy ostatným, podriadeným serverom, ktoré nazývame DataNody. Týchto serverov môže byť v klastri ľubovoľný počet, pričom sú všetky riadené NameNodou. Dáta uložené v HDFS naprieč klastrom sú replikované – to koľko krát majú byť replikované, určuje replikačný faktor. To zabezpečí, že v prípade zlyhania jednej z DataNodes nedochádza k strate dát. O tom, kde sú jednotlivé bloky uložené ma prehľad NameNode. Ak zlyhá jedna z DataNode, nastane automatický replikácia dát medzi ostatné, stále dostupné uzly. Tým pádom sa zabezpečí dodržanie replikačného faktora. Informácie o tom, kde sú dáta v HDFS uložené má k dispozícií NameNode v svojej pamäti RAM. Aby sa predišlo jej zlyhaniu, klaster často obsahuje aj sekundárnu NameNode. Tá prevezme úlohu riadenia klastra v prípade výpadku primárnej a zabezpečí chod klastra.



Obrázok 1.6: Architektúra Hadoop klastra
Zdroj: [25]

V závislosti od replikačného faktora dokáže klaster fungovať aj pri výpadku viacerých nódov naraz. Po výmene chybných DataNodes a nahradení ich novými, sú na tieto serveri dáta znova automaticky replikované. To, že nastane niektorej z nódov zlyhanie je očakávané, nie je dôležité či nastane, ale kedy. Tento princíp fungovania klastra nám nielen eliminuje hrozbu straty dát uložených v HDFS a potrebu záloh, ale aj ponúka značnú výpočtovú silu na spracovanie týchto dát.

Existuje viacero distribúcií Hadoopu, najväčšími hráčmi na trhu sú spoločnosti Cloudera, Hortonworks a MapR. Všetky uvedené ponúkajú svoje vlastné distribúcie a pokročilé nástroje na manažovanie Hadoop klastrov. Najväčším podielom na súčasnom trhu disponuje Cloudera, ktorá niekedy býva označovaná za prvú komerčnú Hadoop spoločnosť. Ponúkajú vlastný produkt pod názvom Cloudera´s Distribution Including Apache Hadoop (CDH), ktorý okrem už spomínaných komponentov používa množstvo iných nástrojov, ktoré pomáhajú k efektívnemu spracovaniu veľkých dát ako napríklad Apache Hive, HBase, Impala, Hue a iné. Dôležitým prvkom CDH je Cloudera Manager – centralizovaný správca celého klastra, v ktorom je možné spravovať, monitorovať a nastavovať jednotlivé služby Apache Hadoop.

2. Cieľ práce

Hlavným cieľom našej diplomovej práce je analyzovať problematiku Big Data v aktuárstve a vytvoriť manuál na prepojenie jazyka R a Apache Hadoop, ako vedúcej platformy na spracovanie Big Data. Vytvorený manuál bude obsahovať kroky potrebné na integráciu týchto platforiem na oboch stranách.

Popri hlavnom ciele sme si stanovili aj vedľajšie ciele, ktoré sa budeme snažiť v našej práci naplniť:

- hodnotenie súčasného stavu integrácie Big Data v odvetví,
- oboznámenie s technológiou Big Data a Apache Hadoop,
- pochopenie princípov fungovania vybraných služieb Apache Hadoop,
- vysvetlenie funkcionality vybraných balíčkov zabezpečujúcich konektivitu týchto platforiem,
- preskúmanie možnosti jazyka R na spracovanie dát pochádzajúcich z Big Data
- ukázať proces načítania dát do pracovného prostredia jazyka R a Hadoop User Experience (HUE),
- porovnanie vybraných prístupov spracovania Big Data.

Po prečítaní tejto práce a vykonaní praktických činností v tejto práci by mal byť čitateľ schopný nastaviť pracovné prostredie jazyka R pre komunikáciu s Apache Hadoop, čítať, zapisovať a manipulovať s dátami, ktoré sú uložené v Hadoope. Čitateľ získa prehľad aj o vybraných službách technológie Hadoop, princípoch ich fungovania a možnostiach použitia integrovaných nástrojov Hadoopu na spracovanie Big Data.

3. Metodika práce a metódy skúmania

Teoretická časť našej bola rozdelená do 5 okruhov. V prvom sme sa venovali vymedzeniu pojmu Big Data so zreteľom na súčasné trendy a analýze ich uplatnenia v rôznych oblastiach. V druhej časti sme syntetizovali perspektívy použitia Big Data v aktuárstve. V tretej časti sme definovali Data science a použili metódu analógie vo vzťahu medzi data science a aktuárstvom. Štvrtý okruh je venovaný jazyku R a jeho použití v aktuárstve. Posledná časť predstavuje Apache Hadoop ako platformu spracovania Big Data a charakterizuje jeho architektúru a základné služby. Z uvádzaného teoretického rámca sme si na základe princípu dedukcie vytýčili ciele práce, ktoré sa stali predmetom nášho skúmania.

Východiskom našej práce bolo nadviazanie spojenia medzi prostrediami R a Hadoop. To sprevádzalo rad krokov potrebných na vytvorenie konekcie na strane Hadoopu, jazyka R a R Studio servera. Ako platformu pre spracovanie Big Data v tejto práci sme si zvolili produkt Cloudera Quickstart. Ide o funkčnú simulácia skutočného Hadoop klastra na jednom počítači. Obsahuje všetky potrebné služby a služby funkcionality skutočného, distribuovaného Hadoop klastra, ktoré sú potrebné pre účely tejto práce. Tento produkt sme si vybrali z dôvodu jeho komplexnosti a jednoduchosti – inštalácia klastra s niekoľkými počítačmi by bol zdĺhavá a vyžadovala niekoľko strojov. Všetky postupy uvedené v tejto práci sú platné aj pre akýkoľvek klaster, použitie tohto "demo" produktu nemá žiaden vplyv na funkcionalitu.

Ako pracovné prostredie jazyka R používame R Studio server, ktorý je nainštalovaný priamo na serveri, kde beží Cloudera Quickstart. Pre použitia tejto metódy sme sa rozhodli z dôvodu uľahčenia komunikácie Hadoopu a R. Prepojenie lokálneho R Studia so vzdialeným Hadoop klastrom nemá praktický význam a vyžadovali by si zložité nastavenie komunikácie služieb – použitím R Studio server túto potrebu eliminujeme. Výhodou je aj jeho dostupnosť zo vzdialených počítačov cez URL prakticky z ľubovoľného počítača, bez nutnosti inštalácie jazyka R na lokálnom počítači používateľa. Tento server sa vizuálne nijako nelíši od prostredia R Studio.

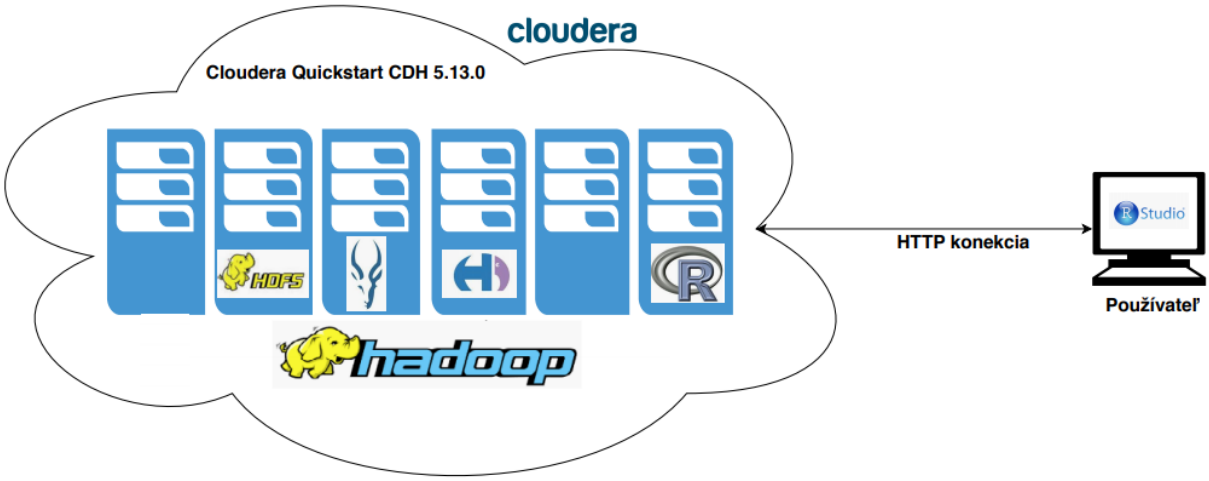
Na realizáciu konekcie medzi R a Hadoopom používame kolekciu balíčkov zvanú *RHadoop*. Skladá sa z viacerých balíčkov, ktoré umožňujú manipuláciu a analýzu dát uložených v Hadoop. V tejto práci opíšeme inštalačný proces balíčkov *rmr, rhdfs, plyrmr*.

Tabuľka 3.1**: Knižnice kolekcie *RHadoop***

Balíček	**Popis**
rhdfs	Poskytuje základnú konektivitu medzi R a HFDS. Používatelia môžu čítať, zapisovať a modifikovať súbory z HDFS priamo z R. Tento balíček stačí nainštalovať na nóde, na ktorej beží samotný R server.
rhbase	Poskytuje základnú konektivitu medzi R a databázou HBASE s použitím Apache Thrift servera. Používatelia môžu čítať, zapisovať a modifikovať tabuľky v databáze priamo z R. Tento balíček stačí nainštalovať na nóde, na ktorej beží samotný R server.
plyrmr	Umožňuje používateľovi vykonávať bežné dátové operácie na dátach uložených priamo v Hadoope. Vychádza z populárneho balíčku *plyr*. Tento balíček je nutné nainštalovať na každú nódu v klastri
rmr2	Umožňuje vykonávanie analýz s použitím Hadoop MapReduce funkcionality. Tento balíček je nutné nainštalovať na každú nódu v klastri
ravro	Poskytuje možnosť čítania a zápisu súborov typu *avro* pomocou R z HDFS. Tento balíček stačí nainštalovať na nóde, na ktorej beží samotný R server.

Zdroj: https://github.com/RevolutionAnalytics/RHadoop/wiki

Návrh architektúry riešenia sa nachádza na obrázku 3.1.



Obrázok 3.1: Architektúra použitého riešenia
Zdroj: vlastné spracovanie

Dáta používané na názorné príklady v tejto práci sú generované. Dátový set *poistne_data*, používaný v tejto práci obsahuje 12 premenných a 10 miliónov riadkov. Dáta sa týkajú poistných udalostí na automobiloch v členských krajinách Európskej únie, pričom reprezentujú dátový formát používaný v poisťovniach. Tieto dáta sme generovali priamo v R použitím funkcií *sample, rexp* a knižnice *stringi*. Dáta obsahujú nasledovné premenné:

- **cislo_zmluvy** – 9 miestne unikátne číslo zmluvy, prvé dve písmená sú vždy ID a potom náhodná kombinácia 7 čísiel, napr. ID8249478,
- **hodnota_vozidla –** náhodné číslo z intervalu (13000,100000), vyjadruje hodnotu vozidla v eurách,
- **pocet_skod –** vyjadruje, či nastala škoda na vozidle. Nadobúda hodnoty 0-3. Približne polovica hodnôt má nulovú škodu,
- **vyska_skody –** výška škody na vozidle v eurách, hodnoty exponenciálneho rozdelenia so strednou hodnotou 1500,
- **vek_vozidla –** počet rokov automobilu, náhodné číslo z intervalu (0,2),
- **pohlavie –** pohlavie zákazníka poisťovne,nadobúda hodnoty „zena" alebo „muz",
- **region –** 28 členských štátov EU, používajú sa oficiálne skratky delenie NUTS – spoločnej nomenklatúry územných jednotiek pre štatistické účely, napr. "CZ", "SK",
- **vek_zmluvy** – vek poistnej zmluvy, náhodné číslo z intervalu (0,5),
- **roky_praxe –** počet rokov od udelenia vodičského oprávnenia, náhodné číslo z intervalu (5,35).

V tejto práci používame nasledovné formy označenia:

- `font Courier New` označuje príkazy na úrovni operačného systému, v konzole R a Hadoop User Experience,
- `modrou farbou fontu Courier new` sú označené výstupy z konzoly,
- *kurzívou* sú označené kľúčové pojmy, príkazy, funkcie, názvy premenných a dátových setov v texte.

Použité verzie v tejto práci sú nasledovné:

- R 3.5.2 "Eggshell Igloo",
- R studio 1.2.1335,
- Cloudera Quickstart CDH 5.13.0,
- CentOS 6.7,
- Hadoop User Experience 4.0.

Autor tejto práce nezodpovedá za nekompatibilitu spôsobenú použitím odlišných verzií ako sú uvedené vyššie.

4. Výsledky práce a diskusia

4.1 Nastavenie a konfigurácia prostredí R a Hadoop

Pred naviazaním spojenia medzi prostrediami R a Hadoop, je východisková ich počiatočná konfigurácia. V tejto podkapitole opisujeme potrebné kroky, ktoré sú nutné na vytvorenie konekcie na strane Hadoopu, jazyka R a R servera. Najskôr sa budeme venovať nastaveniam Hadoopu. Na účely tejto práce je použité "demo" *Cloudera Quickstart* od spoločnosti Cloudera, pri inštalácií na niekoľko nódovom, ozajstnom klastri odporúčame nasledovné kroky vykonávať na tzv. *edge node* – na krajnej „nóde", tam, kde bude nainštalovaný aj samotný R server. Nasledujúce kroky by mal vykonať systémový administrátor. Začneme s vytvorením používateľa. Ide o používateľa na samotnom operačnom systéme, ktorého budeme používať na prácu s R Serverom, v prípade potreby je možné sa s ním prihlásiť na operačný systém klastra. Nasledujúci príkaz *adduser* spustíme ako *root* user:

```
[root@quickstart ~]# adduser jmasar
[root@quickstart ~]#
```

Vytvorili sme používateľa jmasar. Ďalej je nutné mu nastaviť heslo – toto heslo musí byť v unix resp. linux operačných systémoch špecifikované explicitne, po vytvorení. Nastavíme pomocou príkazu *passwd*.

```
[root@quickstart home]# passwd jmasar
Changing password for user jmasar.
New password:
Retype new password:
passwd: all authentication tokens updated successfully.
[root@quickstart ~]#
```

Akonáhle je heslo nastavené, používateľ sa dokáže prihlásiť do systému pomocou hesla. Toto isté heslo sa používa aj na prihlásenie na R server. Vytvorenie používateľa na úrovni operačného systému vytvorilo aj domovský adresár, avšak adresár na samotnom Hadoope, tj. na HDFS je potrebné vytvoriť manuálne. Na prácu s HDFS sa používa príkaz *hdfs dfs* a nasleduje požadovaná voľba – tieto príkazy sú podobné linuxovým príkazom. Kompletná dokumentácia príkazov je dostupná na stránke spoločnosti Apache. Na prechádzanie koreňového adresára HDFS použijeme nasledovný príkaz s možnosťou *ls*, čo znamená zobrazenie súborov v danom priečinku:

```
[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 6 items
drwxrwxrwx   - hdfs  supergroup          0 2017-10-23 10:29 /benchmarks
drwxr-xr-x   - hbase supergroup          0 2019-04-12 06:09 /hbase
drwxr-xr-x   - solr  solr                0 2017-10-23 10:32 /solr
drwxrwxrwt   - hdfs  supergroup          0 2019-04-16 11:53 /tmp
drwxr-xr-x   - hdfs  supergroup          0 2019-04-16 11:56 /user
drwxr-xr-x   - hdfs  supergroup          0 2017-10-23 10:31 /var
```

Na výstupe máme možnosť vidieť štruktúru HDFS. V priečinku *user* sú jednotlivé domovské adresáre používateľov, v tejto ceste je nutné vytvoriť domovský priečinok pre používateľa pomocou voľby *mkdir*. Tento príkaz musí byť realizovaný pomocou užívateľa, ktorý ma oprávnenia vytvoriť priečinok na HDFS.

```
-bash-4.1$ hdfs dfs -mkdir /user/jmasar
-bash-4.1$
```

Skontrolujeme vytvorenie priečinka pomocou už spomenutého príkazu *hdfs dfs -ls* a špecifikujeme cestu, ktorú chceme zobraziť. V našom prípade ide o priečinok /user.

```
-bash-4.1$ hdfs dfs -ls /user/
Found 10 items
drwxr-xr-x   - cloudera cloudera    0 2019-04-16 11:53 /user/cloudera
drwxr-xr-x   - mapred   hadoop      0 2017-10-23 10:29 /user/history
drwxrwxrwx   - hive     supergroup  0 2017-10-23 10:31 /user/hive
drwxrwxrwx   - hue      supergroup  0 2017-10-23 10:30 /user/hue
drwxrwxrwx   - jenkins  supergroup  0 2017-10-23 10:30 /user/jenkins
drwxr-xr-x   - hdfs     supergroup  0 2019-04-17 02:41 /user/jmasar
drwxrwxrwx   - oozie    supergroup  0 2017-10-23 10:30 /user/oozie
drwxrwxrwx   - root     supergroup  0 2017-10-23 10:30 /user/root
drwxr-xr-x   - hdfs     supergroup  0 2017-10-23 10:31 /user/spark
drwxr-xr-x   - hdfs     supergroup  0 2019-04-16 11:56 /user/test
```

Priečinok bol úspešne vytvorený, avšak môžeme vidieť v druhom resp. treťom stĺpci, že tento adresár nepatrí používateľovi jmasar, ale super používateľovi HDFS. To znamená, že by jmasar nemohol to toho priečinka zapisovať, čo predstavuje problém – v tomto priečinku budú totiž uložené jeho súbory.

Práva používateľovi nastavíme pomocou *hdfs dfs* s možnosťou *chown*, čo predstavuje zmenu majiteľa priečinka. Ako argument použijeme meno používateľa, ktorý bude novým vlastníkom a skupinu, do ktorej patrí.

```
-bash-4.1$ hdfs dfs -chown jmasar:users  /user/jmasar
```

Teraz už používateľ jmasar má práva na zapisovanie a čítanie súborov z adresára /user/jmasar na HDFS. Otestujeme pomocou možnosti *touchz*, čo predstavuje vytvorenie testovacieho súboru s nulovou veľkosťou. Ide o alternatívu linuxového príkazu *touch*. Príkaz musí byť spustený používateľom jmasar.

```
[jmasar@quickstart ~]$ hdfs dfs -touchz /user/jmasar/test_file
[jmasar@quickstart ~]$ hdfs dfs -ls /user/jmasar/
Found 1 items
-rw-r--r--   1 jmasar users   0 2019-04-17 02:44 /user/jmasar/test_file
[jmasar@quickstart ~]$
```

Máme možnosť vidieť, že testovací súbor bol vytvorený, čo znamená, že práva sú korektne nastavené. Do tohto priečinku budeme ďalej nahrávať, mazať, či ukladať súbory. Keďže ide o priečinok na distribuovanom systéme, bude dostupný z každého počítača zapojeného do klastra. V zásade nie je nutné vytvárať pre každého používateľa vlastný priečinok – najmä ak uvažujeme o veľkom počte užívateľov. Alternatívnou metódou je použitie spoločného, zdieľaného priečinka do ktorého bude mať prístup viacero používateľov. Ide o jednoduchšie riešenie, no treba zohľadniť viacero faktorov, napríklad:

- **manipulácia súborov navzájom** – používatelia môžu v rovnakom čase modifikovať súbory, prípadne zmazať či prepísať súbor niekomu inému,
- **bezpečnostné riziko –** nie všetci používatelia by mali vidieť všetky dáta ostatných,
- **housekeeping** – treba pravidelne vykonávať údržbu, používatelia totiž zvyknú ukladať veľké množstvo dočasných súborov.

Aký spôsob prístupu bude zvolený, závisí už od predpisov či procesov v danej organizácií. Dôležité je, aby používateľ – aktuár mal prístup do toho priečinka, kde sú uložené dáta, s ktorými môže pracovať. Všetky potrebné kroky na strane Hadoopu sú zosumarizované v prílohe 1.

Po nastavení privilégií na strane Hadoopu, je potrebné vykonať viacero nastavení aj v samotnom jazyku R. Ako už bolo uvedené, nebudeme v práci používať lokálne R Studio, ale vzdialený R Studio server. Používateľ teda nemusí mať na svojom počítači nič nainštalované, ale pripojí na sa server pomocou ľubovoľného webového prehliadača. Viacero používateľov môže naraz zdieľať tento server a pracovať naraz, takisto inštalácia potrebných balíčkov prebehne iba raz – nemusí si každý inštalovať zvlášť. Tento server má takmer identické pracovné prostredie ako lokálne R Studio, takže pokiaľ je koncový používateľ zvyknutý na prácu s R Studio, nebude použitie tejto platformy predstavovať žiaden problém. Tento server odporúčame inštalovať na *egde* nóde hadoop klastra. Podobne ako pri inštalácií lokálneho R studia, je nutné nainštalovať najprv samotný jazyk R. Inštaláciu zabezpečí nasledovný príkaz:

```
[root@quickstart home]sudo yum install R
```

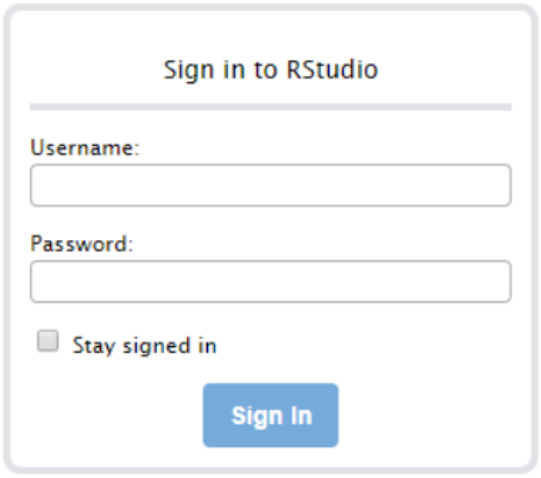
Tento krok je nutné realizovať na tom počítači, kde bude bežať samotný R server. V závislosti od požiadaviek organizácie, je možné R nainštalovať aj na ostatné počítače resp. servery, ktoré tvoria hadoop klaster, nie je to však nutnosťou. Ďalej inštalujeme samotný R Studio Server. Autor práce odporúča vždy postupovať podľa inštrukcií uvedenej na stránke R Studia[22]. Inštalujeme pomocou nasledujúcich príkazov:

```
[root@quickstarthome] wget https://download2.rstudio.org/server/cen-
tos6/x86_64/rstudio-server-rhel-1.2.1335-x86_64.rpm
[root@quickstart home] sudo yum install server/centos6/x86_64/rstudio-
server-rhel-1.2.1335-x86_64.rpm
```

Po úspešnej inštalácií je R Studio server automaticky spustený. Takisto sa automaticky spúšťa pri reštartoch servera. Pripojenie je možné pomocou URL adresy v tvare *http://hostname:8787*. Namiesto hostname je možné použiť aj IP adresu, ako v našom prípade, kde R Studio server beží na adrese *http://10.11.12.30:8787*. Po otvorení adresy sa objaví stránka, ktorá žiada prihlasovacie údaje. Prihlásenie funguje pomocou používateľa, ktorý už existuje na operačnom systéme – v našom prípade jmasar. Dokiaľ nie sú zmenené nastavenia R Studio server, každý existujúci používateľ sa dokáže prihlásiť. Administrátor dokáže obmedziť a zabezpečiť tieto prihlásenia pomocou konfiguračných súborov servera. V tejto práci sa tejto

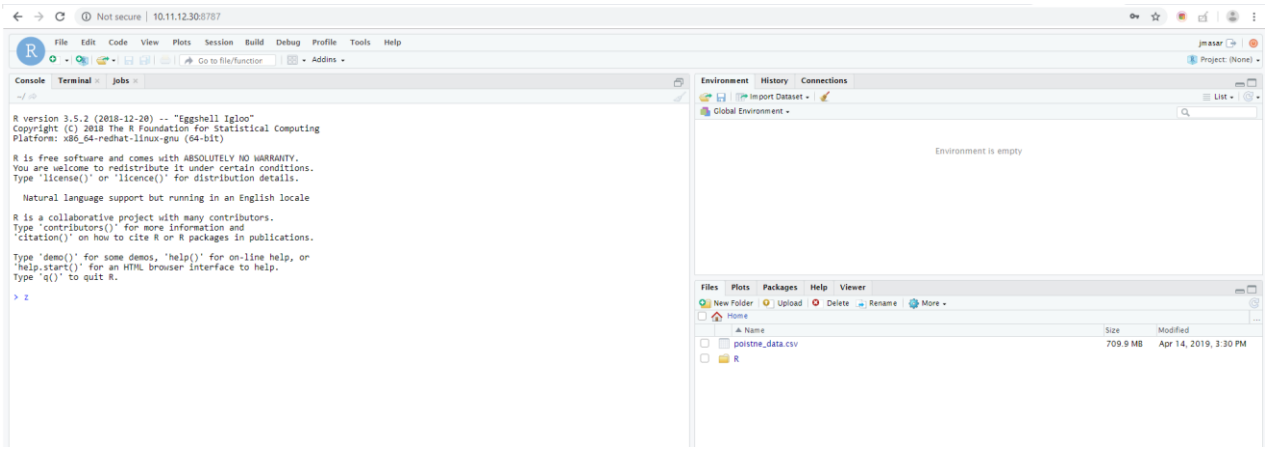
[22]Dostupné na: https://www.rstudio.com/products/rstudio/download-server/

konfigurácií nebudeme venovať, nakoľko každá organizácia má vlastné bezpečnostné predpisy a procesy.



Obrázok 4.1 : Prihlásenie do R Studio Server
Zdroj: vlastné spracovanie

Po autentifikácií sa nám zobrazí pracovné prostredie, ktoré bude používateľom R Studio už dobre známe a nelíši sa od lokálnej inštalácie, ako ukazuje obrázok nižšie. Automaticky sa vytvorí priečinok R v domovskom adresári používateľa.



Obrázok 4.2: Pracovné prostredie R Studio server
Zdroj: vlastné spracovanie

Po prihlásení do R Studio Server, môžeme pokračovať s konfiguráciou a inštaláciou jednotlivých balíčkov (packages), ktoré zabezpečia konektivitu a komunikáciu Hadoopu a R. Tieto balíčky môže inštalovať už koncový používateľ – v našom prípade akutár. Výhodou použitia R Studio server je, že po ich inštalácií tieto balíčky budú dostupné všetkým používateľom.

V našej práci budeme používať vybrané balíčky *RHadoop*, ktorý predstavuje súbor balíčkov dostupných pre integráciu R a Hadoopu. Tieto balíčky však nie sú dostupné v centrálnom repozitári CRAN, je nutné ich stiahnuť a nainštalovať manuálne. Balíčky je možné stiahnuť zo stránok vývojara[23]. Po stiahnutí a nahratí na server do domovského priečinku používateľa jmasar (/home/jmasar) môžeme inštalovať jednotlivé balíčky v poradí, ako opisujeme. Odporúčame postupovať podľa pokynov na stránke vývoja *RHadoop*. Inštalačný postup sa pre rôzne verzie a platformy operačných systémov môže líšiť. Nutnou podmienkou pre inštaláciou vybraných balíčkov je balíček *rJava*.

```
install.packages("rJava")
```

Po nainštalovaní začneme inštaláciou balíčka *rmr2*. Ešte predtým je však nutné nainštalovať jeho závislé balíčky, inak dostaneme chybovú hlášku. Až potom môžeme spustiť jeho inštaláciu.

```
install.packages(c("Rcpp","RJSONIO","digest","functional","reshape2",
"stringr","plyr","caTools"))
install.packages("~/rmr2_3.3.1.tar.gz", repos=NULL,type="source")
```

Samotný balíček *rmr2* inštalujeme pomocou definovania cesty k nemu s označením, že ide o lokálny súbor a jeho typom. Tento spôsob použijeme aj pri ostatných balíčkoch. Následne inštalujeme balíček *plyrmr*. Ešte pred jeho inštaláciou je nutné nastaviť premenné, ktoré budú určovať cestu k programom, ktoré používa Hadoop. Toto nastavíme v R pomocou príkazu *Sys.setenv*. Potrebujeme nastaviť premenné HADOOP_CMD a HADOOP_STREAMING. Prvá menovaná zabezpečuje, že dokážeme spustiť príkazy ako *hdfs dfs* priamo z prostredia R, druhá spustenie operácií pomocou integrovaných nástrojov Hadoop(napr. MapReduce). Premenné je možné nastaviť aj v profile používateľa priamo na úrovni operačného systému – potom nie je nutné ich explicitne v R definovať. Pre kompletnosť však uvádzame postup, kde ich definujeme.

```
Sys.setenv(HADOOP_CMD="/usr/bin/hadoop")
Sys.setenv(HADOOP_STREAMING="/usr/lib/hadoop-mapreduce/hadoop-
streaming-2.6.0-cdh5.13.0.jar")
```

Odporúčame tieto premenné v budúcnosti nastaviť priamo na operačnom systéme, prípadne ich definovať do R skriptu hneď na začiatok, pred spustením týchto knižníc. Po nastavení je potrebné nainštalovať závislé balíčky, potom inštalujeme balíček *plyrmr*.

```
install.packages(c("dplyr","R.methodsS3","Hmisc","memoise","lazyeval"
,"rjson","data,table")
```

[23]Dostupné na: https://github.com/RevolutionAnalytics/RHadoop/wiki/Downloads

```
install.packages("~/plyrmr_0.6.0", repos=NULL,type="source")
```

Tento balíček musí byť nainštalovaný až po inštalácií *rmr2*. Ako posledný inštalujeme *rhdfs*. Pred jeho inštaláciou a používaním musí byť definovaná premenná HADOOP_CMD. Tento balíček nemá žiadne ďalšie závislosti.

```
Sys.setenv(HADOOP_CMD="/usr/bin/hadoop")
install.packages("~/rhdfs_1.0.8.tar.gz", repos=NULL,type="source")
```

Po dokončení inštalácie máme k dispozícií balíčky, ktoré nám umožňujú integráciu s Hadoop prostredím. Ako test konektivity môžeme spustiť nasledujúce príkazy v R Studio Server. Pokúsime sa vylistovať súbory v koreňovom adresári HDFS (/).

```
library(rhdfs)
hdfs.init()
hdfs.ls("/")
```

```
  permission owner      group size          modtime        file
1 drwxrwxrwx  hdfs supergroup    0 2017-10-23 10:29 /benchmarks
2 drwxr-xr-x hbase supergroup    0 2019-04-12 06:09      /hbase
3 drwxr-xr-x  solr       solr    0 2017-10-23 10:32       /solr
4 drwxrwxrwt  hdfs supergroup    0 2019-04-18 06:34        /tmp
5 drwxr-xr-x  hdfs supergroup    0 2019-04-17 05:35       /user
6 drwxr-xr-x  hdfs supergroup    0 2017-10-23 10:31        /var
```

Najskôr je nutné načítať si knižnicu *rhdfs* pomocou príkazu *library*. Ďalej je potrebné použiť príkaz *hdfs.init()*, ktorým sa inicializuje použitie Hadoop príkazov priamo z konzoly R. Príkazy na operácie vždy začínajú *hdfs.* a po tom požadovaná operácia – napr. *ls* znamená zobrazenie obsahu, je možné použiť aj *mkdir* (tvorba priečinku), *rm* (odstránenie súboru) a mnohé ďalšie. Dokumentácia funkcií je k dispozícií na stránkach vývojára[24]. Do pozornosti pre aktuárov dávame najmä už spomínané *ls, mkdir, rm,* a ďalej *write, read, close, seek.* Na základe výstupu vidíme, že naša konfigurácia je správna a funguje – pri dodržaní všetkých opísaných krokov si čitateľ bude môcť takisto nastaviť korektné pracovné prostredie. Všetky potrebné kroky na strane R sú zosumarizované v prílohe 2.

[24]Dostupné na https://github.com/RevolutionAnalytics/RHadoop/wiki/user%3Erhdfs%3EHome

4.2 Načítanie a uloženie dát z Hadoop

V prostrediach organizácií, kde používajú Hadoop ako nástroj spracovania Big Data, sú dáta uložené v HDFS. Každá spoločnosť pristupuje k ukladaniu a členeniu súborov rozdielnym spôsobom. Prechádzajúca kapitola sa zaoberala nastavením práv používateľov na HDFS, preto je nutné, aby aktuár disponoval prístupom k priečinku, kde sú relevantné dáta uložené. V našom prípade používame ako úložisko dát priečinok /user/jmasar. Takisto predpokladáme, že dáta v HDFS už sú – v praxi dáta smerujú do HDFS automaticky, na základe procedúr, stratégií a predpisov danej spoločnosti. Pre účely tejto práce budeme používať už spomínaný dátový set *poistne_data.csv,* ktorý obsahuje 10 miliónov záznamov.

Jazyk R nedokáže priamo spracovávať dáta, ktoré sú uložené na HDFS, bez dodatočnej konfigurácie. Tieto dáta môžu byť uložené v rôznych formátoch (napr. csv, json či parquet), preto je potrebné ich najskôr do R načítať, prípadne upraviť dátový formát aby sme ich dokázali korektne čítať. Uvádzame dve možnosti načítania súborov, pomocou balíčka *rhdfs* a pomocou funkcie *fread*. Chceme načítať súbor *poistne_data.csv*, ktorý je umiestnený na HDFS v priečinku /user/jmasar/. Ako prvé je potrebné nastaviť globálnu premennú HADOOP_CMD. V prípade, že tak neurobíme, načítanie knižnice vypíše v R konzole chybovú hlášku.

```
Sys.setenv(HADOOP_CMD="/usr/bin/hadoop")
library(rhdfs)
HADOOP_CMD=/usr/bin/hadoop
Be sure to run hdfs.init()
hdfs.init()
```

Načítanie knižnice automaticky načíta aj požadovanú knižnicu *rJava* a pripomenie nám inicializovať pomocou funkcie *hdfs.init()*. Ďalej si skontrolujeme, či naozaj je súbor dostupný v HDFS pomocou *hdfs.ls()*.

```
hdfs.ls("/user/jmasar")
  permission  owner group      size       modtime                 file
1 drwx------ jmasar users 0 2019-04-19 16:19       /user/jmasar/.Trash
2 drwx------ jmasar users 0 2019-04-19 16:47     /user/jmasar/.staging
3-rw-r--r—jmasar  users  744395941  2019-04-18  06:28/user/jmasar/pois-
tne_data.csv
4 -rw-r--r-- jmasar user0 2019-04-17 02:44      /user/jmasar/test_file
```

Ako argument vo funkcii *hdfs.ls()* sme použili celú cestu a úvodzovky. Proces načítania sa skladá z nasledovných krokov. Najskôr je nutné si do premennej uložiť, že sa jedná o HDFS súbor. Následne túto premennú načítame, a potom konvertujeme z raw formátu (v ktorom bol súbor uložený v HDFS) do formátu, ktorý vieme ďalej používať v R. Potom načítame náš dataset pomocou funkcie *read.table* a *textConnection*. V našom prípade sme použili aj *header=TRUE,* čo naznačuje, že náš dátový set má hlavičku.

```
f = hdfs.file("/user/jmasar/poistne_data.csv"),"r")
m = hdfs.read(f, start = 0, n = hdfs.ls("/user/jmasar/poistne_da-
ta.csv")$size)
c = rawToChar(m)
poistne_data = read.table(textConnection(c), sep = ",", header=TRUE)
```

Konverziu z raw formátu vykonáme pomocou funkcie *rawToChar*. Ak by sme túto konverziu nevykonali, neboli by sme schopný načítať dáta a výstup z konzoly by vyzeral nasledovne(po vypísaní *m*):

```
> f = hdfs.file("/user/jmasar/poistne_data.csv","r")
> m = hdfs.read(f, start = 0, n = hdfs.ls("/user/jmasar/poistne_data.csv")$size)
> m
   [1] 22 22 2c 22 63 69 73 6c 6f 5f 7a 6d 6c 75 76 79 22 2c 22 68 6f 64 6e 6f 74 61 5f 76 6f 7a 69 64 6c 61 22 2c 22 70
  [39] 6f 63 65 74 5f 6b 6d 22 2c 22 70 6f 63 65 74 5f 73 6b 6f 64 22 2c 22 76 79 73 6b 61 5f 73 6b 6f 64 79 22 2c 22 74
  [77] 79 70 5f 76 6f 7a 69 64 6c 61 22 2c 22 76 65 6b 5f 76 6f 7a 69 64 6c 61 22 2c 22 70 6f 68 6c 61 76 69 65 22 2c 22
 [115] 72 65 67 69 6f 6e 22 2c 22 76 65 6b 5f 7a 6d 6c 75 76 79 22 2c 22 72 6f 6b 79 5f 70 72 61 78 65 22 0d 0a 22 31 22
 [153] 2c 22 49 44 32 39 38 31 34 31 34 22 2c 36 35 38 37 38 2c 38 30 39 39 33 2c 31 2c 36 38 32 2e 34 38 37 36 34 33 37
 [191] 38 39 35 2c 22 63 61 62 72 69 6f 22 2c 31 36 2c 22 6d 75 7a 22 2c 22 44 45 22 2c 30 2c 33 31 0d 0a 22 32 22 2c 22
 [229] 49 44 36 36 34 32 33 30 35 22 2c 33 36 34 31 39 2c 33 33 30 35 39 36 2c 31 2c 33 36 30 2e 39 34 33 37 36 35 31 32
 [267] 30 37 38 31 2c 22 73 70 6f 72 74 22 2c 34 2c 22 7a 65 6e 61 22 2c 22 55 4b 22 2c 33 2c 33 34 0d 0a 22 33 22 2c 22
 [305] 49 44 32 31 39 39 38 30 34 22 2c 36 31 31 36 32 2c 33 33 34 30 35 35 2c 31 2c 34 33 31 34 2e 35 39 35 36 36 35 35
 [343] 33 30 36 34 2c 22 63 6f 6d 62 69 22 2c 31 37 2c 22 7a 65 6e 61 22 2c 22 43 59 22 2c 34 2c 31 34 0d 0a 22 34 22 2c
 [381] 22 49 44 36 39 39 34 30 30 34 22 2c 35 36 36 39 30 2c 34 37 31 35 34 2c 31 2c 34 36 34 39 2e 37 38 31 36 33 36 36
 [419] 34 31 36 39 2c 22 63 6f 6d 62 69 22 2c 32 30 2c 22 7a 65 6e 61 22 2c 22 53 4b 22 2c 34 2c 32 30 0d 0a 22 35 22 2c
```

Obrázok 4.3: Dáta v raw formáte po načítaní z HDFS
Zdroj: vlastné spracovanie

Po konverzii a načítaní majú dáta požadovaný formát a štruktúru, ako je vidieť na nasledujúcom obrázku.

```
> c = rawToChar(m)
> poistne_data = read.table(textConnection(c), sep = ",", header=TRUE)
> head(poistne_data)
  X cislo_zmluvy hodnota_vozidla pocet_km pocet_skod vyska_skody typ_vozidla vek_vozidla pohlavie region vek_zmluvy
1 1    ID2981414           65878    80993          1    682.4876      cabrio          16      muz     DE          0
2 2    ID6642305           36419   330596          1    360.9438       sport           4     zena     UK          3
3 3    ID2199804           61162   334055          1   4314.5957       combi          17     zena     CY          4
4 4    ID6994004           56690    47154          1   4649.7816       combi          20     zena     SK          4
5 5    ID0890997           96519    93760          1   6445.4849       sedan           3      muz     PL          1
6 6    ID8583179           66555   217609          1    835.0665       coupe           2      muz     PL          3
  roky_praxe
1         31
2         34
3         14
4         20
5          6
6         16
```

Obrázok 4.4: Dáta v správnom formáte
Zdroj: vlastné spracovanie

Dáta boli načítané korektne, avšak môžeme si všimnúť, že bol pridaný stĺpec *X*, ktorý plní funkciu ID. Rýchlosť načítania 10 miliónov záznamov bolo 5 minút.

V premennej *n* definujeme, že súbor sa má čítať od začiatku do konca. V prípade, že by sme to nedefinovali, načítali by sme iba časť záznamov. Pri čítaní veľkých záznamov sa môžeme stretnúť s nasledujúcou chybou:

```
m = hdfs.read(f, start = 0, n = hdfs.ls("/user/jmasar/poistne_da-
ta.csv")$size)
Error in .jarray(mu) : java.lang.OutOfMemoryError: Java heap space
```

Ide o problém s pamäťou rezervovanou pre java proces, ktorý má počiatočnú hodnotu 512MB. Je potrebné reštartovať R konzolu a následne, ešte pred načítaním knižníc, problém odstránime nasledujúcim príkazom:

```
options(java.paramaters = "-Xmx8000m")
```

Uvedený spôsob čítania dát z HDFS funguje, avšak považujeme ho za ťažkopádny. Jednak ide o nutnosť použitia presného poradia príkazov, ale môžu nastať aj problémy s pamäťou. Preto odporúčame na čítanie dát z HDFS použiť funkciu *fread*. Táto funkcia je súčasťou balíčka *data.table*, ktorý sme už nainštalovali, kvôli závislostiam balíčkov *RHadoop*.

```
library("data.table")

poistne_data <- fread("/usr/bin/hadoop fs -text /user/jmasar/poi-
stne_data.csv")
```

V príkaze sa definuje binárny súbor, ktorý sa spustí, jeho parameter a cesta k súboru na HDFS. Týmto spôsobom čítanie súbory zaberie približne 40 sekúnd. Po jeho spustení sa priamo v konzole zobrazuje progres čítania, ako je možné vidieť na obrázku 4.5.

Podobným spôsobom ako čítanie dát funguje aj zápis dát z R Studio server do HDFS. Na zápis použijeme takisto sekvenciu príkazov. Najskôr inicializujeme zápis, potom prebehne samotný zápis a na záver ukončíme zápis ukončovacím príkazom. V tomto prípade zapisujeme na HDFS dátový set *vysoka_skoda* z R Studio, ktorý obsahuje iba pozorovania, kde je výška škody vyššia ako 10 tisíc eur.

```
> poistne_data <- fread("/usr/bin/hadoop fs -text /user/jmasar/poistne_data.csv")
|--------------------------------------------------|
|==================================================|
> head(poistne_data)
   V1 cislo_zmluvy hodnota_vozidla pocet_km pocet_skod vyska_skody typ_vozidla vek_vozidla pohlavie
1:  1    ID2981414           65878    80993          1    682.4876      cabrio          16      muz
2:  2    ID6642305           36419   330596          1    360.9438       sport           4     zena
3:  3    ID2199804           61162   334055          1   4314.5957       combi          17     zena
4:  4    ID6994004           56690    47154          1   4649.7816       combi          20     zena
5:  5    ID0890997           96519    93760          1   6445.4849       sedan           3      muz
6:  6    ID8583179           66555   217609          1    835.0665       coupe           2      muz
   region vek_zmluvy roky_praxe
1:     DE          0         31
2:     UK          3         34
3:     CY          4         14
4:     SK          4         20
5:     PL          1          6
6:     PL          3         16
>
```

Obrázok 4.5: Načítanie dát pomocou funkcie fread
Zdroj: vlastné spracovanie

```
zapis <- hdfs.file("/user/jmasar/vysoka_skoda.csv", "w")
hdfs.write(vysoka_skoda, zapis)

[1] TRUE
hdfs.close(zapis)

[1] TRUE
```

Prvým príkazom určujeme meno nového súboru na HDFS. Následne pomocou *write* príkazu vykonáme zápis, použité argumenty sú názov dátového setu v R Studio (v našom prípade vysoka_skoda) a premenná vytvorená v predchádzajúcom príkaze. Po vykonaní uzavrieme zápis pomocou príkazu *close*. Zapísanie skontrolujeme výpisom súborov v našom priečinku.

```
hdfs.ls("/user/jmasar")
  permission  owner group      size          modtime
file
1 drwx------ jmasar users 0 2019-04-20 09:00    /user/jmasar/.Trash
2 drwx------ jmasar users 0 2019-04-19 16:47    /user/jmasar/.staging
3 -rw-r--r-- jmasar users 744395941 2019-04-18 06:28 /user/jmasar/po-
istne_data.csv
4 -rw-r--r-- jmasar users 0 2019-04-17 05:02      /user/jmasar/test2
5 -rw-r--r-- jmasar users 0 2019-04-17 02:44   /user/jmasar/test_file
6 -rw-r--r-- jmasar users 116761511 2019-04-20 13:43 /user/jmasar/vy-
soka_skoda.csv
```

Súbor bol zapísaný úspešne. Ako aj v predchádzajúcom príklade, existuje alternatíva zápisu pomocou kombinácie funkcií *write.csv* a *pipe*. Zápis prebehne v rámci jedného príkazu, čo značne zjednodušuje proces.

```
write.csv(vysoka_skoda, file=pipe("hdfs dfs -put - /user/jmasar/vy-
soka_skoda.csv"))
```

write.csv nám zapisuje dáta do formátu .csv, funkcia *pipe* zase umožňuje beh príkazov (či už na operačnom systéme, alebo v rámci R). Rýchlosť zápisu je podobná pri oboch metódach, približne 20 sekúnd.

4.3 Operácie s dátami

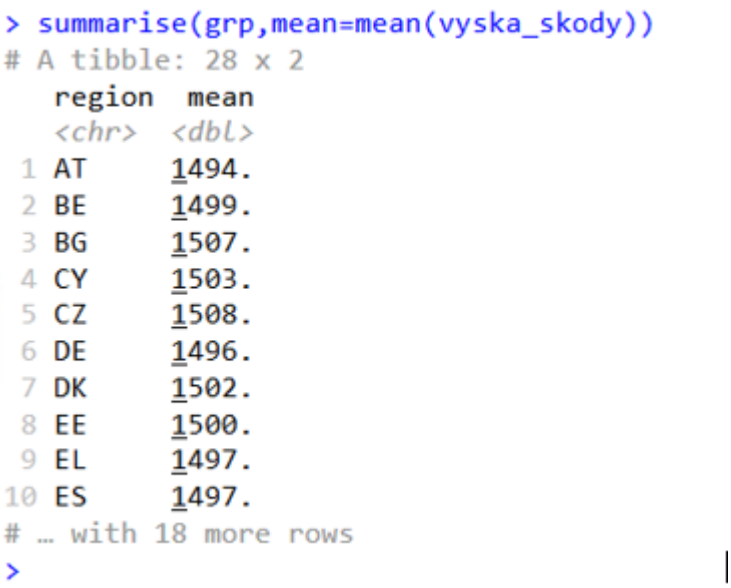
Po načítaní dát do prostredia R Studio server môže používateľ vykonávať operácie nad dátovými setmi podľa potreby. Napríklad aktuár si môže po načítaní dát o poistných udalostiach vypočítať priemernú škodu, vyjadriť počet škôd v jednotlivých krajinách, typoch auta, pri pohlaví a podobne. R vo všeobecnosti ponúka viacero nástrojov na prácu s dátami a ich úpravami. Jednou z najpopulárnejších knižníc na manipuláciami s dátami v R predstavuje *dplyr*. Umožňuje filtráciu dát, výber hodnôt, zhlukovanie hodnôt na základe kritérií a mnoho iných. Podporuje aj príkazy ako *select* či *group by*, čím pripomína databázový jazyk SQL. Práca s *dplyr* je vo všeobecnosti veľmi intuitívna a predstavuje jednu z najjednoduchších a najpopulárnejších spôsob spracovania dát. Balíček *dplyr* sme nainštalovali už v predchádzajúcej kapitole, preto pre jeho použitie stačí iba načítať knižnice pomocou známeho príkazu *library*.

```
library(dplyr)
```

Predpokladajme situáciu, že aktuára zaujíma priemerná výška škody v jednotlivých krajinách EÚ28. Ako prvé budeme musieť náš dátový set usporiadať podľa krajín a následne vypočítať ich priemer. Knižnica *dplyr* nám umožní dáta usporiadať a uložiť do nového dátového setu, a následne vypočítať priemer pomocou funkcie *mean* a *summarise*.

```
grp <- group_by(poistne_data,region)
summarise(grp,mean=mean(vyska_skody))
```

Výstup príkazov je zobrazený na obrázku nižšie. Vidíme, že priemer škody v jednotlivých štátoch osciluje okolo hodnoty 1500 – čo je očakávané, nakoľko stredná hodnota škody v celkom súbore je práve táto hodnota.



Obrázok 4.6: agregované hodnoty priemeru krajín
Zdroj: vlastné spracovanie

Knižnica *dplyr* so sebou prináša aj možnosť použitia znaku pipe (*%>%*), čo v praxi znamená nadväznosť jednotlivých príkazov. Nie je potrebné jednotlivé príkazy rozdeľovať, napríklad výpočet priemeru z predchádzajúcej strany by mal nasledovnú podobu:

```
poistne_data %>%
group_by(poistne_data,region) %>%
summarise(mean=mean(vyska_skody))
```

Použitie pipe zjednodušuje prácu s dátovými setmi, preto odporúčame tento znak používať, ako budeme aj my v tejto práci.

Chceme zistiť, koľko vysokých škôd nastalo v našom dátovom sete. Definujme podmienku, že škoda, ktorá bude presahovať 10 tisíc eur, bude považovaná za vysokú. Použijeme funkciu *filter*, do ktorej zadáme našu podmienku.

```
vysoka_skoda <- poistne_data %>%
+ filter(vyska_skody > 10000)
```

Týmto jednoduchým príkazom získame nový dátový set, kde bude škoda presahovať danú hranicu. Pomocou príkazu *head* sa môžeme presvedčiť o správnosti príkazu – na obrázku nižšie je vidieť len správne hodnoty. Celkovo bolo 179184 škôd vyšších ako 10 tisíc eur.

```
> vysoka_skoda <- poistne_data %>%
+ filter(vyska_skody > 10000)
> head(vysoka_skoda)
   V1 cislo_zmluvy hodnota_vozidla pocet_km pocet_skod vyska_skody typ_vozidla vek_vozidla pohlavie
1  23    ID7407935           20006   286129          2    10520.25      cabrio          11      muz
2  30    ID8067140           98608   353486          1    13112.59       sport          12      muz
3  37    ID0717084           35115    91908          3    14028.02       combi           3     zena
4  88    ID6296087           15000    73769          2    11150.16   limousine           6     zena
5 111    ID8266631           22301   140420          3    15306.10         SUV          15      muz
6 196    ID0157777           61282   225598          1    10253.29   limousine           2      muz
  region vek_zmluvy roky_praxe
1     DE          3         19
2     SE          5         30
3     CZ          1         19
4     EE          3         34
5     HR          0         24
6     LT          2          9
```

Obrázok 4.7: Filtrovanie datasetu pomocou dplyr
Zdroj: vlastné spracovanie

Aktuára ďalej môže zaujímať, aké boli počty udalostí s vysokou škodou v jednotlivých krajinách. Na zistenie opäť použijeme knižnicu *dplyr* a funkciu *count*, ktorá nám spočíta jednotlivé udalosti. Argumentami funkcie ďalej sú *sort=TRUE*, čo zabezpečí zoradenie od najvyššieho počtu po najmenší a názov stĺpca (*pocet_skod*). Tieto hodnoty si uložíme do novej premennej s názvom *pocet_vysokych_skod*.

```
pocet_vysokych_skod <- vysoka_skoda %>%
count(region,sort=TRUE, name= "pocet_skod")
```

Získali sme dátový set, ktorý obsahuje len dva stĺpce a to región a počet škôd. Ako máme možnosť vidieť na obrázku nižšie, dáta sú zoradené správnym spôsobom.

```
> pocet_vysokych_skod
# A tibble: 28 x 2
   region pocet_velkych_skod
   <chr>                <int>
 1 FR                    6564
 2 CZ                    6515
 3 LV                    6511
 4 SK                    6494
 5 PL                    6473
 6 DK                    6468
 7 NL                    6458
 8 HU                    6457
 9 BE                    6430
10 CY                    6425
# … with 18 more rows
```

Obrázok 4.8: Počet vysokých škôd v jednotlivých štátoch
Zdroj: vlastné spracovanie

Dáta v tomto tvare je možné už ľahko interpretovať, či graficky zobraziť – buď na grafe, alebo priamo na mape Európy. Kvôli vizualizácií dát chceme vytvoriť mapu Európy, ktorá bude zobrazovať počty vysokých škôd v jednotlivých krajinách. Na vytvorenie mapy budeme potrebovať dodatočné balíčky :

```
install.packages(c("maptools","rgeos"))
```

Okrem týchto balíčkov budeme potrebovať aj *shapefile* súbor, ktorý bude tvoriť samotnú mapu. Ide o dátový formát, ktorý sa používa na ukladanie a prenos vektorových priestorových dát pre geografické informačné systémy. Sú v nich popisované línie a plochy, ktoré môžu reprezentovať napríklad hranice štátov či iné geografické body. Zvyčajne majú tieto súbory príponu .shp. Okrem toho tieto súbory zvyčajne obsahujú aj ďalšie súbory, v ktorých sú popísané atribúty jednotlivých prvkov (napr. hranice krajín na základe geografických súradníc). Shapefile používaný v tomto príklade sme stiahli z Eurostatu. Tento súbor sme potom uploadli do domovského priečinka používateľa jmasar (/home/jmasar) a rozbalili. Ako prvé načítame naše potrebné knižnice:

```
library(rgeos)
library(maptools)
library(ggplot2)
```

Ďalej je potrebné načítať samotný shapefile súbor pomocou funkcie *readShapeSpatial.* Ako argument uvádzame cestu k .shp súboru.

```
mapa2 <- readShapeSpatial("~/map/NUTS_RG_60M_2016_3035_LEVL_0.shp")
```

Pomocou funkcie *fortify* konvertujeme dáta, ktoré obsahuje súbor mapa2 do použiteľného formátu pre prostredie R. Vyberáme si región NUTS_ID, čo nám zabezpečí kompletné dáta o členských krajinách a ich súradnice.

```
fortify_shape = fortify(mapa2, region = 'NUTS_ID')
```

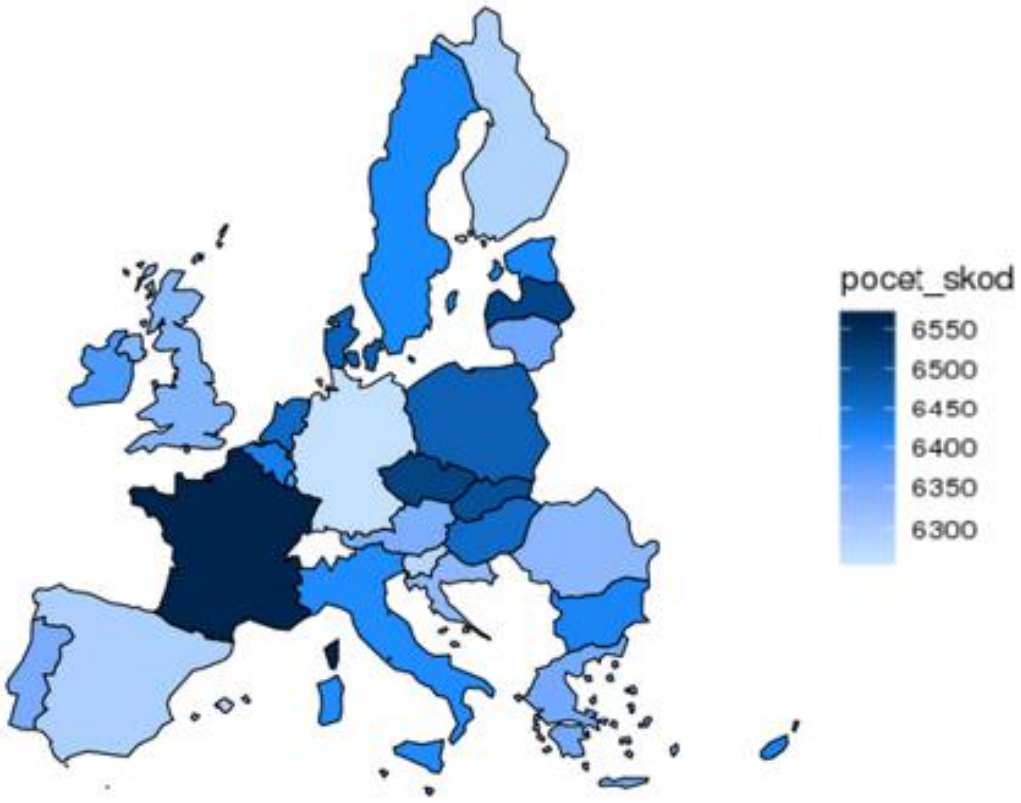
Potom je nutné spojiť náš dátový set, obsahujúci počet vysokých škôd a set *fortify_shape*. Spojíme ich pomocou funkcie *merge*, pričom definujeme prvky na základe ktorých sa spájajú. Je dôležité, aby *x* a *y* predstavovali rovnaké hodnoty – v našom prípade skratky krajín. Po spojení dát ešte tento dátový set zoradíme pomocou funkcie *order*, kvôli správnemu zobrazeniu hodnôt.

```
spojene_data = merge(fortify_shape, pocet_novych_skod, by.x="id",
by.y="region")
Map_plot = spojene_data[order(spojene_data$order), ]
```

Po týchto úpravách už môžeme vykresliť samotnú mapu pomocou knižnice *ggplot2* a jej funkcií.

```
ggplot(data=Map_plot) + geom_polygon(aes(x = long, y = lat, group =
group, fill = pocet_skod), color="black", size=0.5)
+ scale_fill_gradient2(low="#ffc4c4",mid="#ff5b5b" ,midpoint= 893,
high= "#350000")
+ theme_void()
+ coord_equal()
```

Knižnici *ggplot* a jej funkcionalite sme sa venovali v našej bakalárskej práci, preto nebudeme jej syntax podrobne opisovať. Po tomto príkaze je mapa už hotová, pričom rozdielne počty nehôd sú na mape viditeľne farebne rozlíšené.



Obrázok 4.9: Počet vysokých škôd v jednotlivých krajinách EÚ28
Zdroj: vlastné spracovanie

Ako ďalší príklad môžeme uviesť nasledovné – aktuárov budú zaujímať počty škôd v jednotlivých krajinách podľa nasledovných kritérií – budeme skúmať poistné udalosti žien a mužov, ktorí majú menej ako 15 rokov praxe, používajú SUV a celková škoda je vyššia ako 6 tisíc eur. Selektujeme si počty škôd podľa krajín nasledujúcim spôsobom a uložíme do premennej *pocet_skod_muzi*:

```
pocet_skod_muzi <- poistne_data %>%
  filter(pohlavie == "muz") %>%
  filter(roky_praxe < 15) %>%
  filter(vyska_skody > 6000)%>%
 filter(typ_vozidla == "SUV") %>%
  count(region,sort=TRUE, name= "pocet_skod_muzi")
```

Na korektné vykonanie príkazu vyššie je nutné mať načítanú knižnicu *dplyr*. Následne rovnakým spôsobom uložíme počet škôd žien (zmeníme argument prvého filtra na "zena") do premennej *pocet_skod_zeny*. Spojíme oba tieto dátové sety do jedného pomocou funkcie *merge*.

```
celkovy_pocet_skod = merge(pocet_skod_muzi, pocet_skod_zeny,
by.x="region", by.y="region")
```

Zobrazením tejto premennej uvidíme počet jednotlivých škôd podľa krajín a pohlaví. Ukážku dát môžeme vidieť na obrázku nižšie.

```
> celkovy_pocet_skod
   region pocet_skod_muzi pocet_skod_zeny
1      AT             510              416
2      BE             461              361
3      BG             486              374
4      CY             452              372
5      CZ             497              391
6      DE             477              387
7      DK             480              417
8      EE             493              384
9      EL             495              383
10     ES             505              378
11     FI             482              383
12     FR             484              390
13     HR             460              367
14     HU             471              410
```

Obrázok 4.10: Počet škôd mužov a žien v členských krajinách
Zdroj: vlastné spracovanie

Dáta v takomto formáte sú výsledkom našej požiadavky, avšak prostredie R ponúka silné vizualizačné možnosti, preto bude vhodné, ak výstup z obrázka vyššie aj graficky znázorníme. Použijeme na to funkciu *barplot*. Ešte predtým, je však nutné dáta upraviť – *barplot*

na korektné zobrazenie požaduje dáta v tvare matice. V našom prípade bude mať matica 28 stĺpcov a 2 riadky. Maticu vytvoríme pomocou funkcie *matrix*.

```
matica_skod <- matrix(ncol=28, nrow=2)
```

Vytvorenej matici priradíme názvy stĺpcov, pričom jednotlivé hodnoty sú z dátového setup *celkovy_pocet_skod*. Následne pridáme aj názvy riadkov, čo v našom prípade bude jednoduché, nakoľko máme iba dve pohlavia.

```
colnames(matica_skod)=as.character(celkovy_pocet_skod$region)
rownames(matica_skod)=c("muzi", "zeny")
```

```
> matica_skod <- matrix(ncol=28, nrow=2)
> colnames(matica_skod)=as.character(celkovy_pocet_skod$region)
> rownames(matica_skod)=c("muzi", "zeny")
> matica_skod
     AT BE BG CY CZ DE DK EE EL ES FI FR HR HU IE IT LT LU LV MT NL PL PT RO SE SI SK UK
muzi NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
zeny NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
>
```

Obrázok 4.11:Vytovrenie matice pre barplot
Zdroj: vlastné spracovanie

Obrázok na predchádzajúcej strane nám zobrazuje správne vytvorenie matice a priradenie názvov stĺpcov a riadkov. Matica je však prázdna, je potrebné do nej načítať dáta z dátového setup *celkovy_pocet_skod*.

```
matica_skod[1,] <- celkovy_pocet_skod$pocet_skod_muzi
matica_skod[2,] <- celkovy_pocet_skod$pocet_skod_zeny
```

Hranaté zátvorky naznačujú, že napĺňame celý riadok matice. Matica je hotová a vyzerá nasledovne:

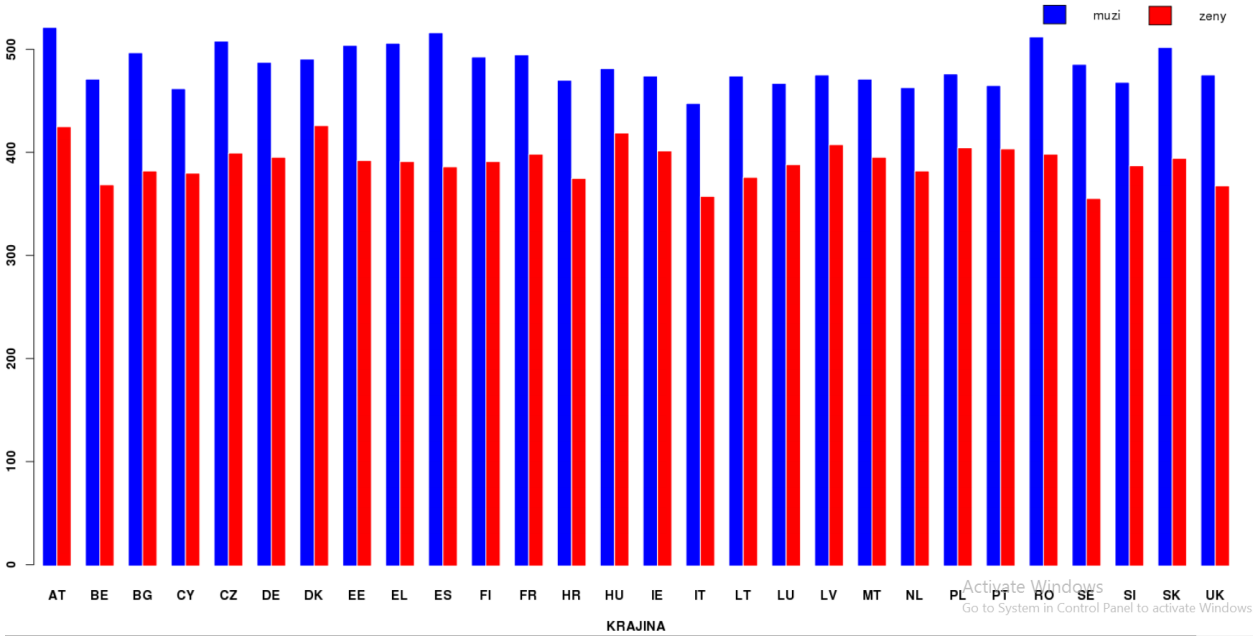
```
> matica_skod[1,] <- celkovy_pocet_skod$pocet_skod_muzi
> matica_skod[2,] <- celkovy_pocet_skod$pocet_skod_zeny
> matica_skod
      AT  BE  BG  CY  CZ  DE  DK  EE  EL  ES  FI  FR  HR  HU  IE  IT  LT  LU  LV  MT  NL  PL  PT  RO  SE  SI  SK
muzi 510 461 486 452 497 477 480 493 495 505 482 484 460 471 464 438 464 457 465 461 453 466 455 501 475 458 491
zeny 416 361 374 372 391 387 417 384 383 378 383 390 367 410 393 350 368 380 399 387 374 396 395 390 348 379 386
      UK
muzi 465
zeny 360
```

Obrázok 4:12: Finálna verzia matice
Zdroj: vlastné spracovanie

Graf vytvoríme pomocou už spomínanej funkcie *barplot*:

```
barplot(matica_skod, col=c("blue","red") , border="white", font.axis=2,
beside=T, legend=rownames(matica_skod), xlab="KRAJINA",font.lab=2)
```

V argumentoch funkcie definujeme farby, hranicu medzi jednotlivými stĺpcami, legendu, použitý font a to, že graf má vertikálny a nie horizontálny charakter (argument *beside=T*).



Obrázok 4.13:Počty škôd v jednotlivých krajinách podľa pohlaví
Zdroj: vlastné spracovanie

Na príkladoch vyššie sme demonštrovali funkcionalitu cieľa našej práce. Dáta sme načítali v HDFS do R Studio server a následne sme ich pomocou R upravili a vizualizovali do želanej podoby. Pri ilustrácií funkcionality sme vždy dáta z HDFS načítali priamo do internej pamäte R Studio. V reálnom svete však môže nastať, že dátové sety na HDFS sú tak veľké, že ich nahratie do pamäte je z kapacitných dôvodov nemožné. Preto je nutné tento dátový set spracovať priamo na HDFS. Aj táto možnosť je dostupná priamo z prostredia R – pomocou balíčka *plyrmr*. Ten umožňuje vykonávať vybrané operácie a ako vstupné dáta použiť priamo súbor na HDFS. Takisto ponúka aj možnosť dáta priamo zapísať naspäť na Hadoop, bez nutnosti načítania do pamäte. V takom prípade sa spustí proces priamo na strane Hadoopu (MapReduce job), ktorý sa postará o vykonanie úlohy. V závislosti od veľkosti dátového setu je nutné korektné nastavenie nielen na strane R, ale aj správna konfigurácia na strane Hadoopu. Ide o nastavenie pokročilých pamäťových parametrov a vlastností algoritmov klastra, vrátane replikačných faktorov, preto tento spôsob v tejto práci nebudeme ďalej opisovať. V prípade

záujmu čitateľa o túto problematiku odporúčame tutoriál na stránkach vývojára, kde sú popísané základné funkcie tohto balíčka.[25].

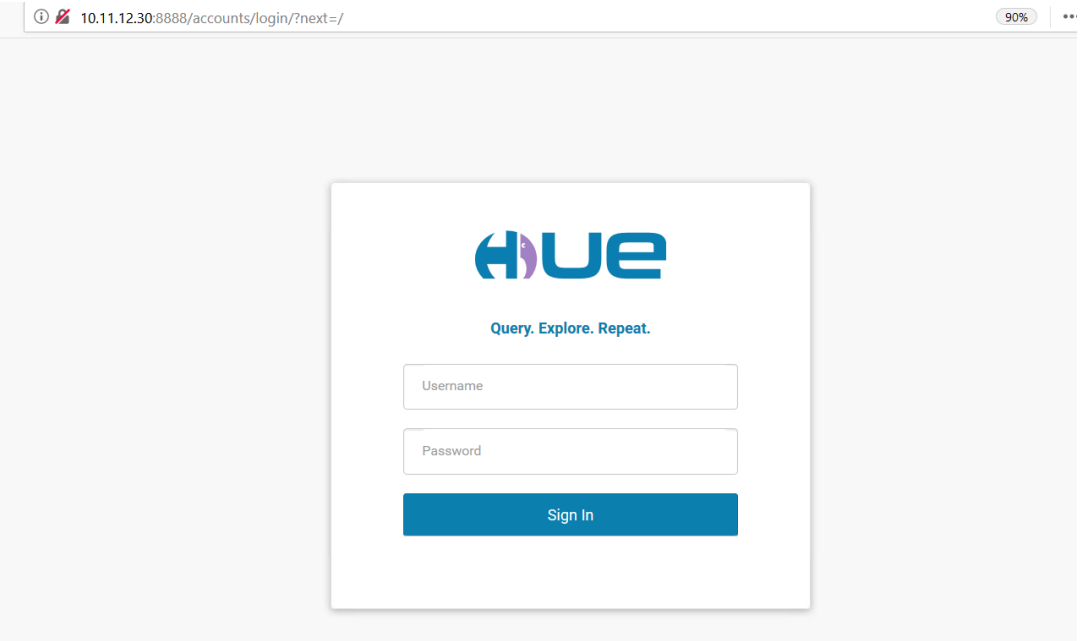
Ako máme možnosť vidieť na príkladoch vyššie, po načítaní dát z HDFS je možné s dátami jednoducho manipulovať pomocou štandardných funkcií jazyka R. Prepojenie HDFS ako zdroja súborov a R ako nástroj Data science umožňuje aktuárom nielen získavať dáta z nových zdrojov, ale aj prináša nové možnosti integrácie a využitia nástrojov a metód Big Data pri aktuárskej praxi.

4.4 Spracovanie údajov pomocou Hadoop User Experience

V súvislosti s Big Data a aktuárstvom, pochopiteľne, existuje viacero možností spracovania dát uložených v prostredí Hadoop. Okrem použitie externých nástrojov ako napríklad R a ich spojením s Big Data technológiou, Hadoop ponúka svoje vlastné, integrované riešenie na pristupovanie a prácu s jeho ekosystémom – Hadoop User Experience (ďalej len HUE). Ide o open-source webový interface pre Hadoop, ktorý podporuje prechádzanie, vizualizáciu či dopytovanie dát v klastri. Tento komponent je súčasťou všetkých hlavných distribúcií Hadoopu, prípadne je možné ho dodatočne na klaster nainštalovať. HUE je dostupné na adrese http://hostname:8888. Autentifikácia používateľa prebieha priamo na úrovni HUE, za prideľovanie prístupov je zodpovedný systémový administrátor.

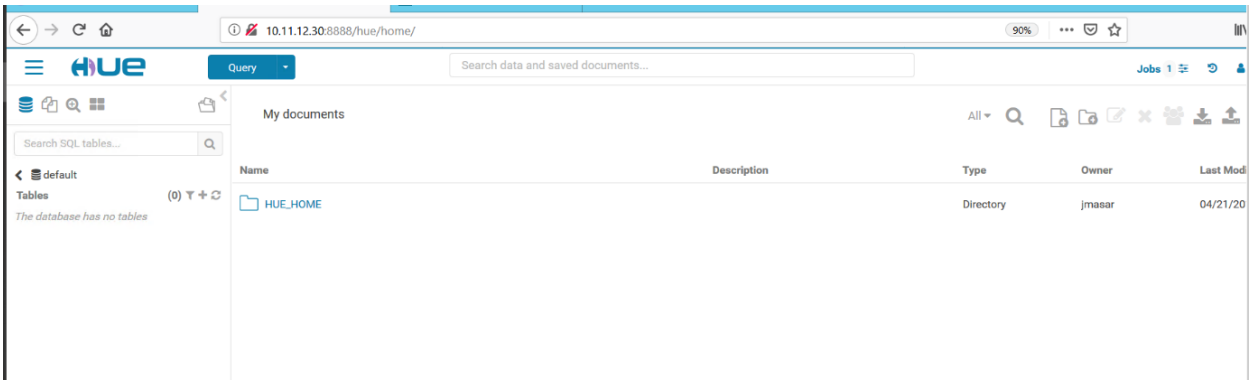
V prípade, že používateľ nemá vytvorený priečinok na HDFS, po registrácii v HUE sa tento priečinok automaticky vytvorí. Ak používateľ existuje, HUE sa zosynchronizuje s už existujúcim priečinkom a bude ho registrovať ako domovský. Po prihlásení do HUE sa spustí krátky tutoriál o jeho základných funkciách. Odporúčame tento krátky úvod pozorne sledovať a absolvovať, skúsenejší používatelia ho môžu preskočiť.

[25] Dostupné na https://github.com/RevolutionAnalytics/plyrmr/blob/master/docs/tutorial.md



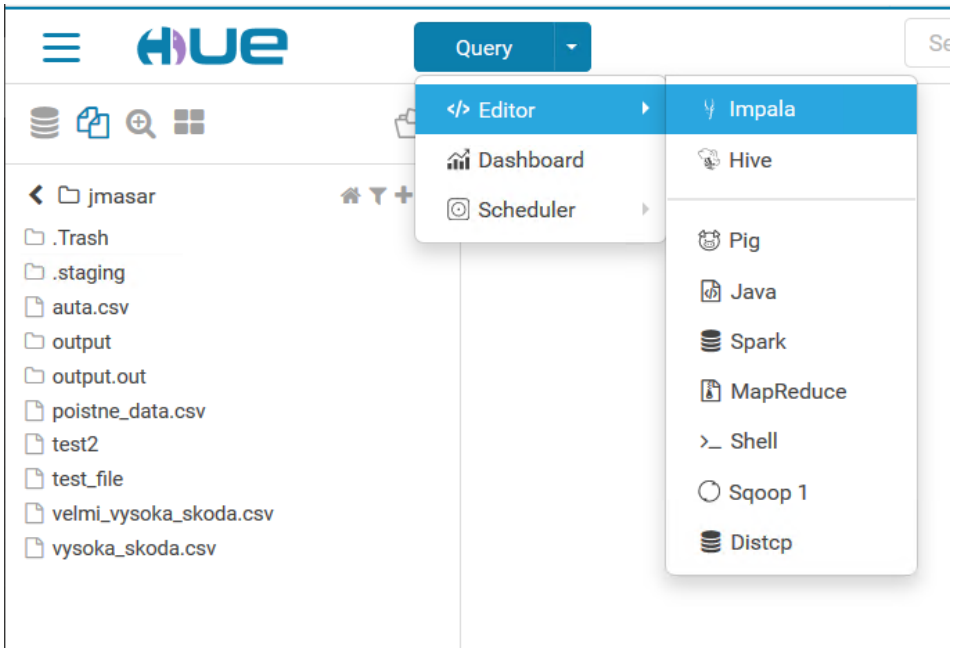
Obrázok 4.14: Webový interface HUE – prihlasovacie okno
Zdroj: vlastné spracovanie

Po prihlásení sa zobrazí domovská obrazovka používateľa, ktorú si používateľ môže personalizovať podľa preferencií.



Obrázok 4.15: Pracovné prostredie HUE
Zdroj: vlastné spracovanie,2019.

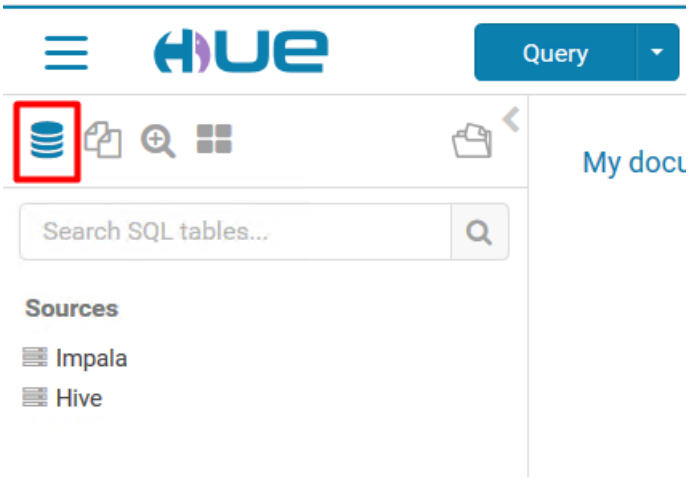
Na tejto obrazovke je možné ukladať si dokumenty, vytvárať priečinky, vytvárať skripty, nahrať či stiahnuť súbory a mnohé iné. HUE takisto umožňuje prehliadanie HDFS. Jednou z jeho hlavných úloh je však umožnenie práce používateľovi s rozličnými Hadoop službami, napríklad:



Obrázok 4.16: Query Editor HUE
Zdroj: vlastné spracovanie

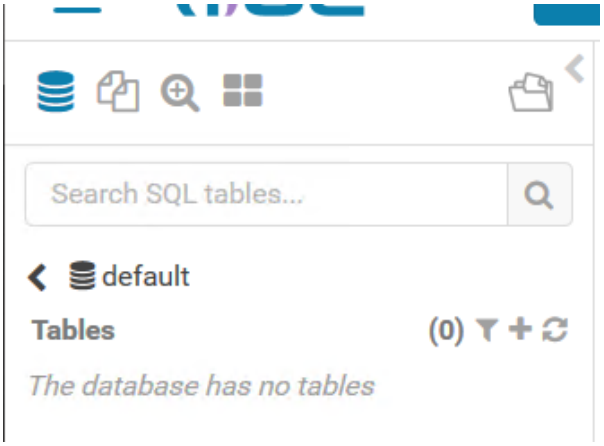
Máme možnosť pracovať nielen s Hadoop službami, ale aj operačným systémom či databázami. Do pozornosti pre aktuárov a vo všeobecnosti dávame najmä prvé dve služby – *Hive* a *Impala*. Vo svojej podstate ich môžeme označiť ako databázy – ide o služby, ktoré dokážu vykonávať čítanie, zapisovanie a manažovanie dátových setov v distribuovanom úložisku. To znamená, že umožňujú používateľom pracovať s dátami v Hadoope pomerne jednoducho. Hive používa svoj vlastný jazyk a to HiveQL – ide o modifikáciu SQL s jemne odlišnou syntaxou. Najväčším rozdielom je, že nepodporuje funkciu UPDATE. Impala používa jazyk SQL, pričom je kom-patibilná aj s jazykom HiveQL, funkcia UPDATE je však takisto nepodporovaná. Rozdiel je aj v rýchlosti spracovania jednotlivých dotazov. Vo všeobecnosti je Impala rýchlejšia, nakoľko ide o službu in-memory, čiže údaje má uložené v pamäti RAM. To v praxi znamená, že poskytuje veľmi rýchle spracovanie výsledkov, no s určitým obmedzením. Ak by sme totiž spustili veľmi náročnú operáciu, Impala si môže alokovať príliš veľa pamäte a následne spôsobiť spomalenie iných služieb. To však závisí od konfigurácie a výkonnosti klastra. V tejto práci budeme ilustrovať spracovanie dát pomocou služby Impala.

Po prihlásení máme na ľavom okraji dostupné dva základné dátové zdroje:



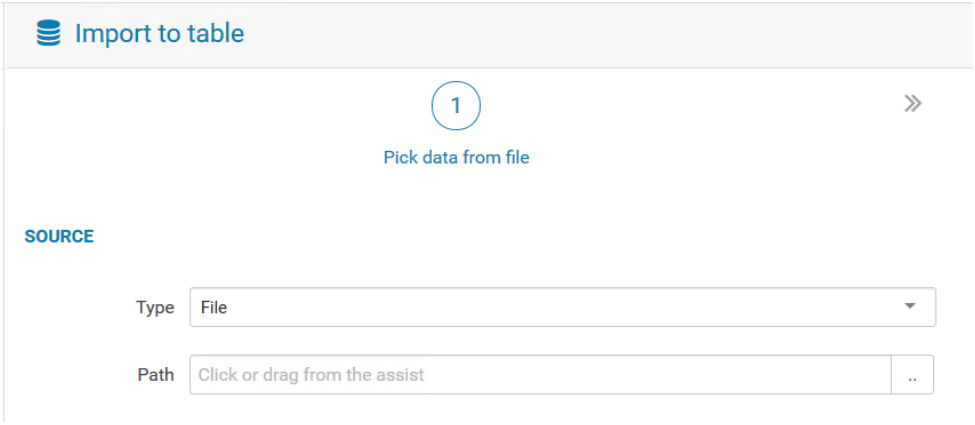
Obrázok 4.17: Dostupné databázy v HUE
Zdroj: vlastné spracovanie

Vyberieme si službu Impala, ktorá nám zobrazí dostupné databázy. Vidíme, že v našom prípade v základnej databáze nie sú zatiaľ dostupné žiadne tabuľky:



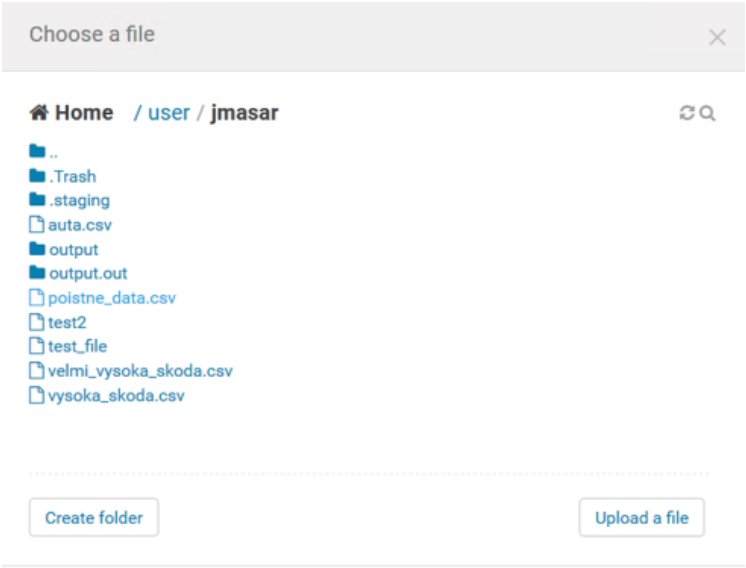
Obrázok 4.18: Tabuľky v databáze default
Zdroj: vlastné spracovanie

V takomto prípade máme možnosť si tabuľku do databázy načítať po kliknutí na ikonu +. Objaví sa nám intuitívny sprievodca, ktorý nám pomôže importovať súbor a konvertovať do tabuľky. Ako prvý krok definujeme cestu k súboru, z ktorého budeme vytvárať tabuľku:



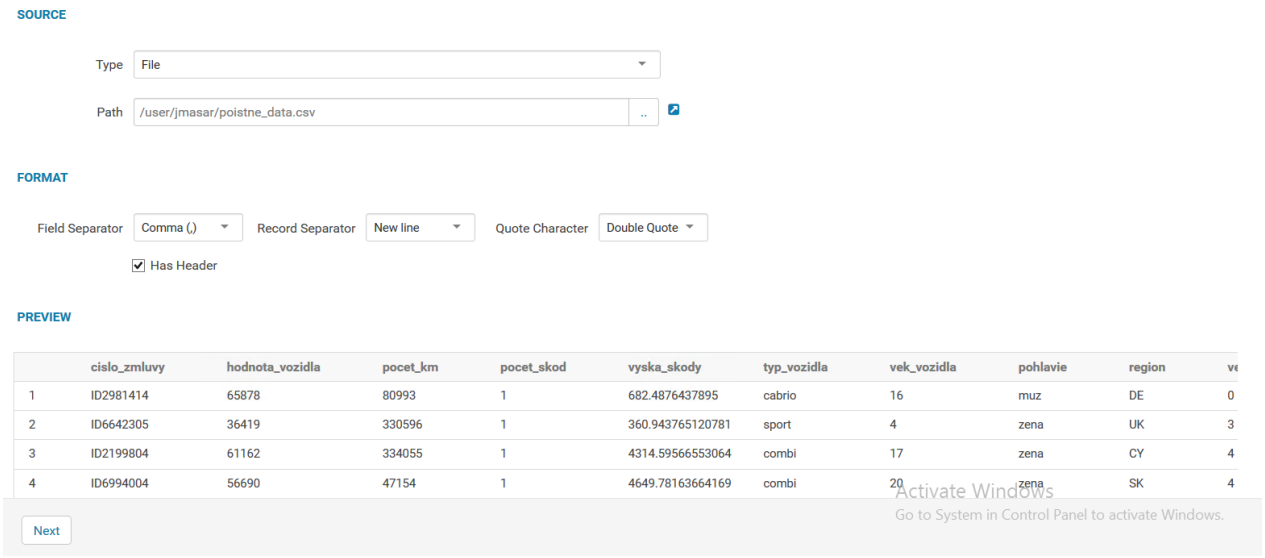
Obrázok 4.19: Import tabuľky v prostredí HUE
Zdroj: vlastné spracovanie

V našom prípade budeme importovať súbor *poistne_data.csv*, ktorý sme používali aj v prechádzajúcej kapitole. Pri výbere súboru sa automaticky zobrazí domovský priečinok používateľa, ako môžeme vidieť na nasledujúcom obrázku:



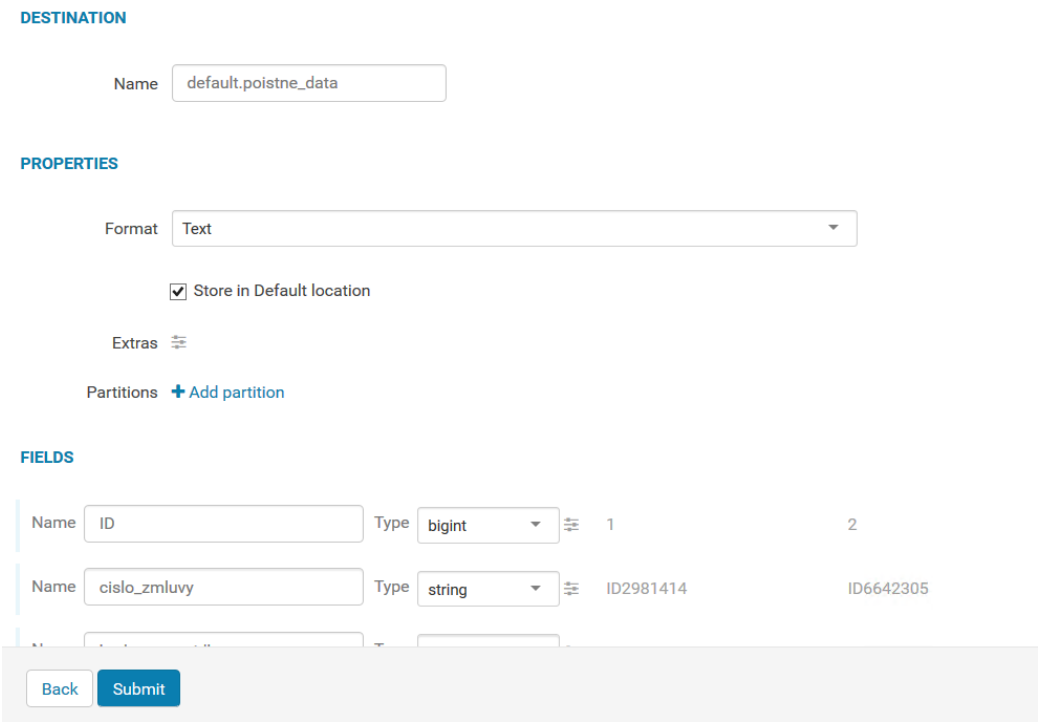
Obrázok 4.20: Domovský priečinok na HDFS v prostredí HUE
Zdroj: vlastné spracovanie

Po vybratí súboru sa nám automaticky v dolnej časti obrazovky vytvorí náhľad našich dát a možnosti ich importu. Tu môžeme manuálne definovať, akým znakom sú hodnoty oddelené, ako budú oddelené jednotlivé záznamy a textové polia. Vo väčšine prípadov nám HUE automaticky ponúkne správne nastavenia a nie je potrebná manuálna intervencia. Takisto nám je ponúknutý náhľad našich dát:



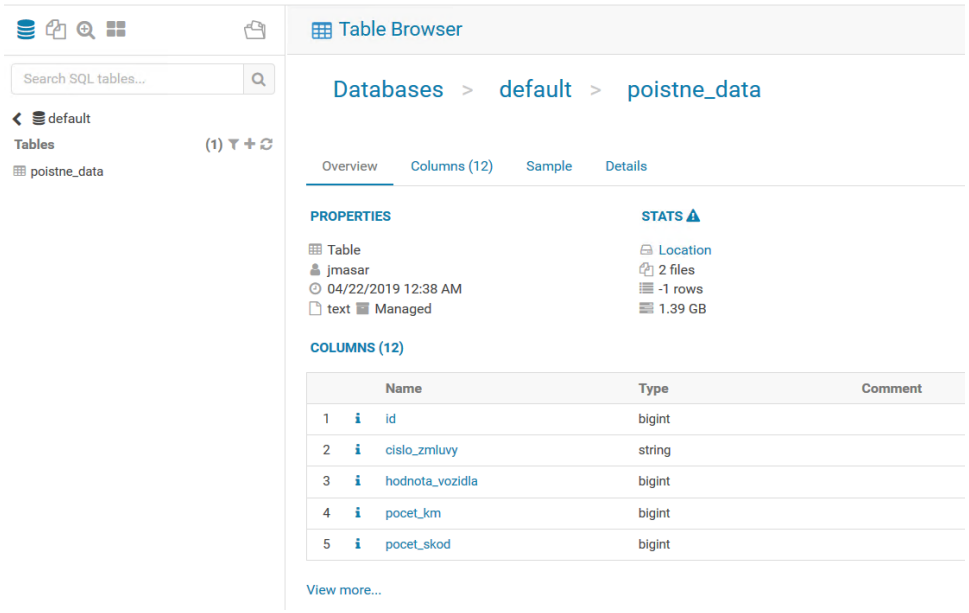
Obrázok 4.21: Náhľad dát pred importom
Zdroj: vlastné spracovanie

Po kontrole nastavení a stlačením *Next* sa dostaneme do finálneho kroku, kde môžeme upravovať resp. meniť dátové typy, zvoliť meno tabuľky a formát uloženia. Odporúčame používať preddefinovaný formát text. Po nastavení môžeme pokračovať s importom dát:



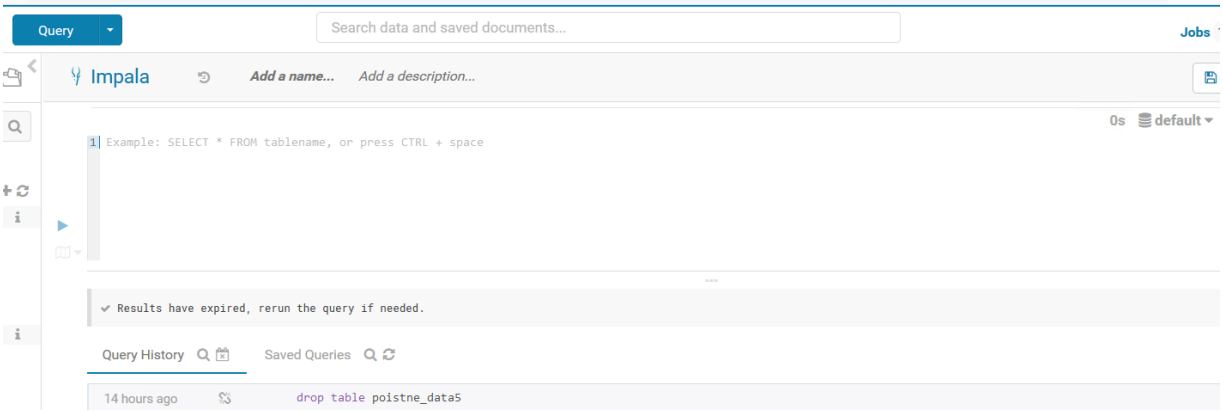
Obrázok 4.22: Nastavenie parametrov importu dát
Zdroj: vlastné spracovanie

Po dokončení nás systém informuje a novovytvorená tabuľka bude k dispozícií v databáze *default*. Vytvorenie tabuľky prebehlo veľmi rýchlo, trvalo približne 3 sekundy.



Obrázok 4.23:Vytvorená tabuľka pomocou HUE
Zdroj: vlastné spracovanie

Po načítaní tabuľky už je všetko pripravené a môžeme začať vykonávať na tabuľke dopyty. Nový dopyt vykonáme po kliknutí na button *Query* v ľavej hornej časti obrazovky.



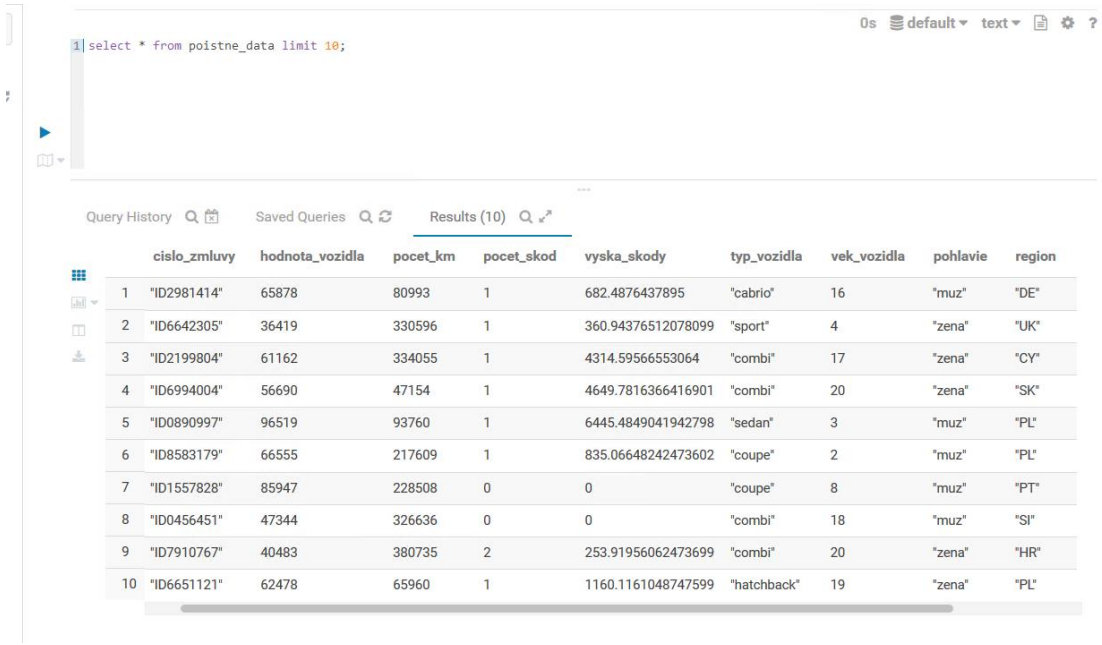
Obrázok 4.24: Impala query v prostredie HUE
Zdroj: vlastné spracovanie

Vidíme, že HUE nám priamo zobrazuje SQL konzolu, pričom nám ponúka aj príklad. Samotná práca s SQL je v prostredí HUE intuitívna – interaktívne SQL dopĺňa a navrhuje predpokladané operácie, či priamo validuje náš dopyt. Pod konzolou je história našich predchádzajúcich

dopytov, v prípade potreby je možné jednoducho zopakovať SQL príkaz. Zadajme jednoduchý príkaz na zobrazenie prvých 10 záznamov z tabuľky:

```
SELECT * FROM POISTNE_DATA LIMIT 10;
```

Po spracovaní výstupu je výsledok nám vypísaný priamo pod konzolou:



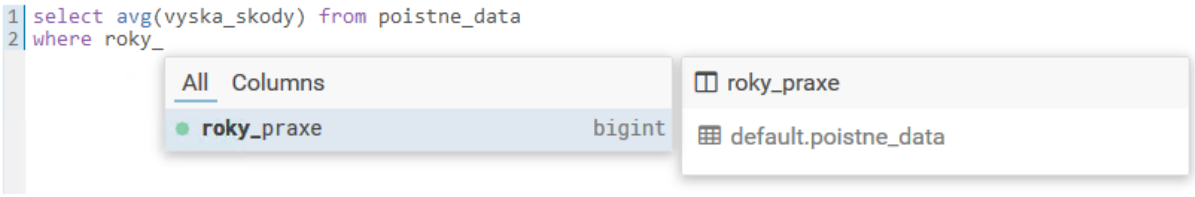
	cislo_zmluvy	hodnota_vozidla	pocet_km	pocet_skod	vyska_skody	typ_vozidla	vek_vozidla	pohlavie	region
1	"ID2981414"	65878	80993	1	682.4876437895	"cabrio"	16	"muz"	"DE"
2	"ID6642305"	36419	330596	1	360.94376512078099	"sport"	4	"zena"	"UK"
3	"ID2199804"	61162	334055	1	4314.59566553064	"combi"	17	"zena"	"CY"
4	"ID6994004"	56690	47154	1	4649.7816366416901	"combi"	20	"zena"	"SK"
5	"ID0890997"	96519	93760	1	6445.4849041942798	"sedan"	3	"muz"	"PL"
6	"ID8583179"	66555	217609	1	835.06648242473602	"coupe"	2	"muz"	"PL"
7	"ID1557828"	85947	228508	0	0	"coupe"	8	"muz"	"PT"
8	"ID0456451"	47344	326636	0	0	"combi"	18	"muz"	"SI"
9	"ID7910767"	40483	380735	2	253.91956062473699	"combi"	20	"zena"	"HR"
10	"ID6651121"	62478	65960	1	1160.1161048747599	"hatchback"	19	"zena"	"PL"

Obrázok 4.25: Ukážka dát v prostredí HUE
Zdroj: vlastné spracovanie

Vďaka in-memory technológií trvajú dopyty nad dátami veľmi krátko, rýchlosť spracovania je oveľa rýchlejšia ako klasické databázy. Vidíme, že dáta sú zobrazené v správnej štruktúre, a teda môžeme sa zamerať na dopyty, ktoré môžu aktuárov zaujímať.

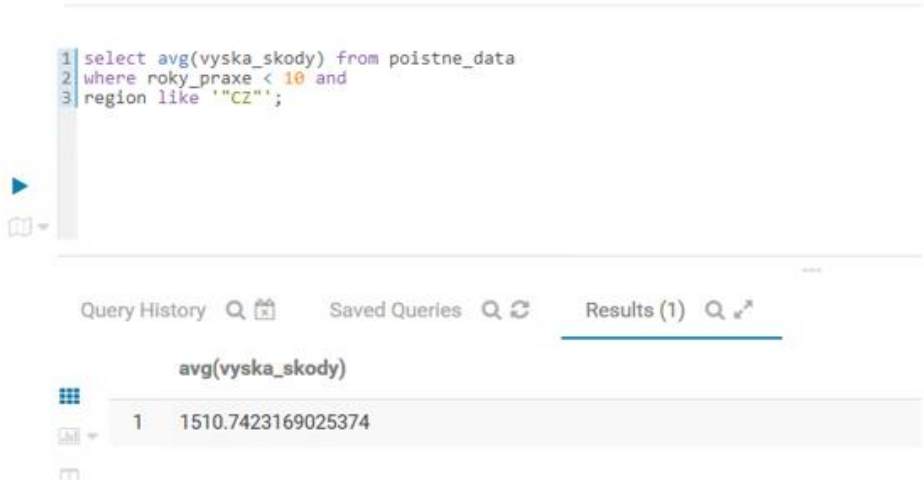
Snažme sa zistiť priemernú výšku škody v Česku pre vodičov s nízkou praxou, tj. do 10 rokov. SQL dopyt bude mať tvar:

```
select avg(vyska_skody) from poistne_data
where roky_praxe < 10 and
region like '"CZ"';
```



Obrázok 4.26: Interaktívne dopĺňanie dotazov v prostredí HUE
Zdroj: vlastné spracovanie

Výsledkom príkazu je hodnota 1510.74 eur. Celkové spracovanie príkazu trvalo približne 5 sekúnd, čo je dobrý čas vzhľadom na to, že ide o tabuľku, ktorá má 10 miliónov záznamov.

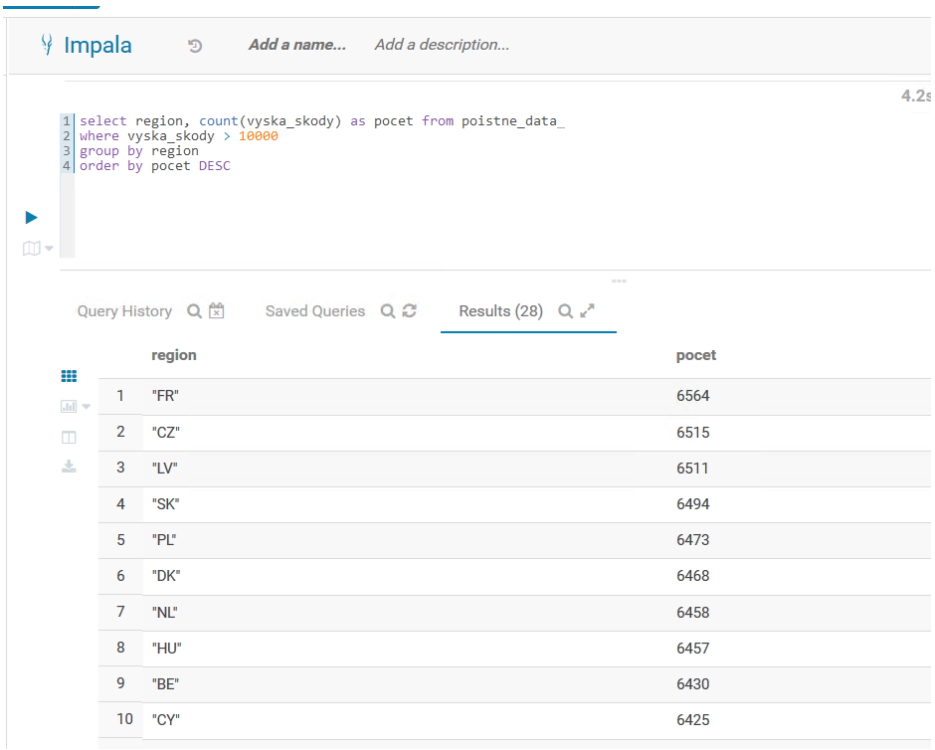


Obrázok 4.27: Priemerná škoda vodičov s nízkou praxou z ČR
Zdroj: vlastné spracovanie

Práca s SQL je v prostredí HUE veľmi jednoduchá, aj keď používateľ pozná len základy tohto jazyka. V kapitole 4.3 sme v prostredí R Studio server pomocou knižnice *dplyr* hľadali počty vysokých škôd v jednotlivých členských štátoch. S použitím SQL dostaneme rovnaký výsledok týmto dopytom:

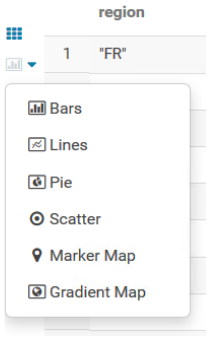
```
select region, count(vyska_skody) as pocet from poistne_data
where vyska_skody > 10000
group by region
order by pocet DESC;
```

Dopyt trvá približne 4 sekundy, následne je výsledok zobrazený pod konzolou. Vidíme, že hodnoty sú rovnaké ako aj pri spracovaní v prostredí R. Celkový čas spracovania bol pri oboch spôsoboch približne identický – R ako štatistický softvér dokáže veľmi rýchlo vykonávať operácie s dátami. Rýchlosť Impaly by však naplno vynikla v prípade ešte väčších dátových setov – služba dokáže bez problémov vykonávať dopyty za približne rovnaký čas ako v našom prípade aj na veľmi rozsiahlych dátach, ktoré môžu mať stovky miliónov záznamov.



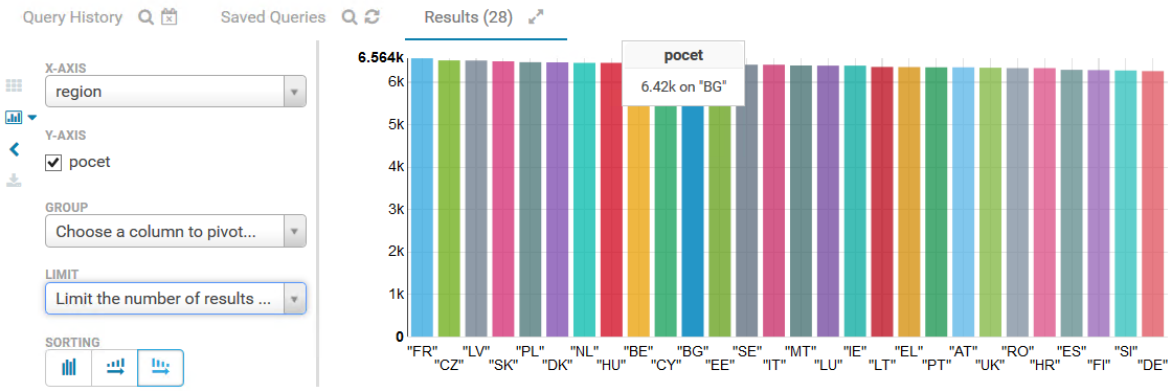
Obrázok 4.30: Počty vysokých škôd podľa krajín
Zdroj: vlastné spracovanie

Prostredie HUE prináša aj rôzne možnosti vizualizácie dát, ich používanie je veľmi jednoduché. Stačí si vybrať možnosť grafického zobrazenia pri výstupe:



Obrázok 4.31: Grafické možnosti v prostredí HUE
Zdroj: vlastné spracovanie

Máme možnosť výberu viacerých typov grafov – v našom prípade sme zvolili možnosť *Bars*, nakoľko sa najviac hodí vzhľadom na charakter dát. HUE vytvorí graf a následne ho v prípade potreby vieme dodatočne modifikovať. Názvy premenných a farby sú automaticky pridané už pri vytvorení grafu.



Obrázok 4.32: Grafické zobrazenie dát v prostredí HUE
Zdroj: vlastné spracovanie

Potom, čo spracujeme dáta máme možnosť ich v HUE aj uložiť – dopyty, aj grafy. Používateľ tak má často používané operácie, dopyty, grafy a súbory kedykoľvek k dispozícii. HUE predstavuje používateľsky príjemné prostredie na spracovanie veľkých dát, ktoré je veľmi intuitívne a rýchle. Aktuári často pri svojej práci používajú aj jazyk SQL. Práve pre týchto aktuárov prináša HUE najväčšiu pridanú hodnotu, a to v jeho jednoduchosti a množstve funkcií.

4.5 Porovnanie jednotlivých prístupov

Oba prístupy predstavené v prechádzajúcich kapitolách umožňujú spracovanie Big Data podľa potrieb používateľa. Každý z prístupov má svoje výhody aj nevýhody, pričom aj R aj HUE obsahujú veľké množstvo funkcií, ktoré aktuár pri svojej práci dokáže využiť. Základné hodnotenie oboch je uvedené v tabuľke č. 4.1. Pri hodnotení sme použili nasledovné hodnotenie: 1 je najlepšie, 5 najhoršie. Po vyhodnotení všetkých kritérií máme možnosť vidieť, že obe technológie majú približne rovnaké hodnotenie, každý z nich vyniká v niečom inom. Môžeme konštatovať, že obe prinášajú široké možnosti spracovania Big Data a ďalej je už na stratégií jednotlivých spoločností resp. rozhodnutí aktuárov, ktoré budú používať. Je logické, že aktuári so silnou znalosťou jazyka R budú preferovať spojenie jazyka R a Hadoopu pre svoje analýzy. Ako sme však mali možnosť vidieť, pre prácu s HDFS a inými službami, je nutná aspoň základná znalosť princípov technológií Big Data, preto aktuár pre lepšie pochopenie prepojenia technológií by mal absolvovať rôzne školenia ohľadom Hadoopu. Znalosť systému Linux, resp.

aspoň jeho základných príkazov bude takisto veľkou výhodou. Výhodou R je, že aktuár môže použiť všetky funkcie a balíčky, ktoré R ponúka a následne ich aplikovať na Big Data.

Tabuľka 4.1: **Komparácia prostredí R a HUE (subjektívne vyjadrenie autora práce)**

	R	**Hadoop User Experience**
Prívetivosť prostredia	2	1
Rýchlosť spracovania	2	2
Náročnosť nastavenia a prepojenia s Hadoop	3	1
Komplexnosť riešenia	2	3
Integrácia s inými technológiami	1	2
Variabilita možností spracovania	1	3
Riešenie pre Big Data	3	1
Možnosť rozšírenia funkcionality	1	3

Zdroj: vlastné spracovanie,2019.

HUE môžeme považovať za jednoduchšie prostredie z hľadiska nastavení potrebných na zabezpečenie konektivity. Ide o nástroj, ktorý prichádza s väčšinou Hadoop distribúcií ako integrovaný. HUE bolo vytvorené pre priame spracovanie Big Data používateľmi, čo znamená, že ponúka rozsiahlejšie a robustnejšie nástroje ako R. Z hľadiska výkonnosti má HUE takisto výhodu oproti R – využívaním spojenej výpočtovej sily počítačového klastra môžeme dosiahnuť vynikajúce výsledky v rýchlosti spracovania dát aj v reálnom čase. Prostredie HUE je veľmi intuitívne a za krátky čas si ho osvoja aj noví používatelia. Ťažiskom pre prácu s touto službou je znalosť SQL. Aktuár si následne môže vybrať, ktorú z Hadoop technológií použije pri svojich dopytoch. Používateľsky príjemný front-end služby prináša aj rozsiahle možnosti prechádzania súborov na HDFS a vizualizácie dát. Súbory môžeme jednoducho, na pár kliknutí nahrať resp. stiahnuť aj z lokálneho počítača. Práca s dátami je takisto jednoduchá a intuitívna – vrátane vytvárania nových tabuliek na databáze. HUE je nepochybne nástrojom, ktorý hrá hlavnú úlohu v spracovaní Big Data v spoločnostiach, aktuárstvo nie je výnimkou. Aktuári nájdu v HUE široké spektrum možností, ktoré môžu použiť pri svojej práci.

Ktorú technológiu bude poisťovňa resp. oddelenie aktuárstva používať, závisí od ich preferencie či internej firemnej stratégie. Obe ponúkajú komplexné možnosti práce s Big Data a závisí na rozhodnutí kompetentných osôb, ktorý zo spôsobov bude použitý pre každodennú aktuársku prax.

Záver

V úvodnej časti sme charakterizovali Big Data a vedný odbor priamo súvisiac s touto problematikou, Data Science. Uviedli sme analógiu medzi povolaním dátových vedcov a aktuárov. Môžeme konštatovať, že aktuári plnia úlohu dátových vedcov v poisťovniach. Na základe tohto predpokladu sme konštatovali potrebu použitia Big Data v tejto profesii. Uviedli sme konkrétne aplikačné príklady a spôsoby používania Big Data v poisťovníctve. Práca aktuára vyžadujem znalosť mnohých softvérových nástrojov, ktoré im pomáhajú dosahovať želaná výsledky. Populárnym sa najmä v posledných rokoch stáva jazyk R, ktorý predstavuje výkonnú platformu na spracovanie, analýzu a vizualizáciu dát. V závere úvodnej časti sme predstavili Apache Hadoop, jeden z najpopulárnejších open source nástrojov používaných na spracovanie Big Data. Načrtli sme jeho architektúru a princípy jeho fungovania, ako aj dôvody jeho popularity.

Druhá časť práce obsahuje formuláciu cieľa tejto diplomovej práce. Okrem hlavného cieľa sme si určili aj čiastkové ciele, ktorým naplnením sme dosiahneme lepšie pochopenie danej problematiky.

Po dôkladnej analýze sme vypracovali návrh riešenia integrácie dvoch populárnych platforiem, jazyka R a Hadoopu. V tejto časti sa nachádzajú aj údaje o použitých knižniciach, softvéroch verziách a dátach používaných v tejto práci.

Vo finálnej kapitole práce sme postupným opisom krokov potrebných na integráciu oboch riešení a postupným naplnení čiastkových cieľov dosiahli hlavný cieľ – a to tvorbu manuálnu, ktorý umožňuje ich spojenie. Tento manuál môže slúžiť systémovým administrátorom a aktuárom ako príručka a odporúčanie, ako postupovať v prípade potreby použitia jazyka R na spracovanie dát, ktoré sú uložené v HDFS. Pri testovaní spojenia a načítania dát pomocou knižnice *rhdfs* sme sa stretli problém spôsobeným konfigurácie jazyku Java, Išlo o nedostatočné množstvo pamäte, ktoré mohla používať pre jednotlivý proces. Chybu sme odstránili dodatočným nastavením parametra priamo v R Studio server. Následne proces ukladania a načítania dát z HDFS fungoval bez problémov. Okrem týchto dvoch možností sme popísali aj spôsob prechádzanie HDFS a nastavenia potrebných globálnych premenných, ktoré zabezpečujú konekciu. Po úspešnom načítaní dát sme názorne ilustrovali operácie dátovými setmi priamo v R Studio pomocou knižnice *dplyr*. Získané výsledky sme ďalej vizualizovali použitím knižníc *ggplot2* a *barplot*. Tieto vizualizácie predstavujú záverečnú a dôležitú fázu

Data Science a to vo forme grafickej interpretácie dát. Ďalej sme v práci ako alternatívu k navrhnutému riešenie uviedli možnosť spracovania veľkých dát s využitím integrovaného nástroj Apache Hadoop – Hadoop User Experience. Popísali sme možnosti, ktoré tento nástroj ponúka a pomocou *Apache Impala* sme demonštrovali proces vytvárania tabuliek zo súborov uložených na HDFS. Po vytvorení týchto tabuliek sme ukázali spracovanie dát, pričom naším cieľom bolo dosiahnutie podobných výsledkov v ako v prostredí R Studio server. Následne sme tieto výsledky aj graficky zobrazili v HUE. V záverečnej časti kapitoly sme uviedli porovnanie jazyka R a Hadoop User Experience a možností, ktoré pri spracovaní veľkých dát ponúkaju. Ohodnotili sme jednotlivé funkcie a formulovali odporúčania, ktorá platforma môže predstavovať ideálny nástroj pre prácu aktuárov.

Zoznam použitej literatúry

[1] ALLERIN. *Big Data for Insurance* [online]. 2019. [cit. 18.01.2019]. Dostupné na: https://www.allerin.com/services/big-data-analytics/insurance

[2] AMERICAN ACADEMY OF ACTUARIES. *Big Data and The Role of The Actuary* [online]. 2018. [cit. 24.01.2019]. Dostupné na: https://www.actuary.org/sites/default/files/files/publications/BigDataAndTheRoleOfTheActuary.pdf

[3] BERKLEY. *What is Data science* [online]. UC Berkley School of Information, 2018. [cit. 18.01.2019].Dostupné na: https://datascience.berkeley.edu/about/what-is-data-science/

[4] BROWN, Mike. *Are People OK With Insurance Companies Using Big Data.* [online]. 2018. [cit. 24.01.2019]. Dostupné na: https://lendedu.com/blog/insurance-companies-using-data-survey/

[5] CAO, Longbing. *Data Science: Challenges and Directions. Communications of the ACM* [online]. 2017, **60**(8), 59-68 [cit. 29.04.2019]. DOI: 10.1145/3015456. ISSN 00010782.

[6] COLUMBUS, Louis. *Big Data Analytics Adoption Soared In The Enterprise In 2018.* [online]. 2018. [cit. 24.01.2019]. Dostupné na: https://www.forbes.com/sites/louiscolumbus-/2018/12/23/big-data-analytics-adoption-soared-in-the-enterprise-in-2018/

[7] FINANCIAL WEB. *Pros and Cons of Pay as You Drive Insurance.* [online]. 2016. [cit. 02.02.2019]. Dostupné na: https://www.finweb.com/insurance/pros-and-cons-of-pay-as-you-drive-insurance.html

[8] GLASDOOR. *50 Best Jobs in America* [online].2018. [cit. 18.01.2019]. Dostupné na: https://www.glassdoor.com/List/Best-Jobs-in-America-2018-LST_KQ0,25.htm

[9] GSMA. *2025 Every Car Connected* [online]. 2012. [cit. 18.01.2019].Dostupné na: https://www.gsma.com/iot/wp-content/uploads/2012/03/gsma2025everycarconnected.pdf

[10] JAMES, Josh. *Data Never Sleeps 6.0* [online]. 2018. [cit. 23.01.2019]. Dostupné na: https://www.domo.com/blog/data-never-sleeps-6/

[11] LANEY, Douglas. *3D Data Management: Controlling Data Volume, Velocity and Variety* [online]. 2001. [cit. 20.01.2019]. Dostupné na https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

[12] MALLON, Sean. *Predictive Modelling And Big Data Are Insurance Industry Powerhouses* [online]. 2018. [cit. 27.01.2019]. Dostupné na: https://www.smartdatacollectiv-e.com/predictive-modeling-and-big-data-are-insurance-industry-powerhouses/

[13] MARR, Bernard. *Big Data in Practice.* [online]. 2018. [cit. 23.01.2019]. Dostupné na: https://www.bernardmarr.com/default.asp?contentID=1076

[14] MARR, Bernard. *Big Data: 20 Mind-Blowing Facts Everyone Must Read* [online]. 2015. [cit. 23.02.2019]. Dostupné na: https://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/

[15] MARR, Bernard. *How Big Data is Changing Insurance Forever* [online]. 2015. [cit. 27.01.2019]. Dostupné na: https://www.forbes.com/sites/bernardmarr/2015/12/16/how-big-data-is-changing-the-insurance-industry-forever/

[16] MARR, Bernard. *How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read* [online]. 2018. [cit. 20.01.2019]. Dostupné na:https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/

[17] MEARIAN, Lucas. *Insurance Company Now Offers Discounts* [online]. 2015. [cit. 28.01.2019]. Dostupné na: https://www.computerworld.com/article/2911594/insurance-company-now-offers-discounts-if-you-let-it-track-your-fitbit.html

[18] NEW GEN APPS. *6 Reasons Why Choose R Programming For Data Science Projects* [online]. 2017. [cit. 02.02.2019]. Dostupné na: https://www.newgenapps.com/blog/6-reasons-why-choose-r-programming-for-data-science-projects

[19] OXFORD ENGLISH DICTIONARIES. *Big Data Definition*, [online]. 2019. [cit. 20.01.2019]. Dostupné na: https://en.oxforddictionaries.com/definition/big_data https://en.oxforddictionaries.com/definition/big_data

[20] PÁLEŠ, Michal. Kvalita údajov a jej význam pre aktuárov. In *Slovenská štatistika a Demografia*. Bratislava: Štatistický úrad SR. 2019. 64 s. ISSN 1210-1095.

[21] PEARSON, Lillian. *Data Science for Dummies*. Hoboken : John Wiley & Sonc, Inc., 2017. 384 s. ISBN 978-1-119-32763-9

[22] PIATETSKY, Gregory. *19th Annual KDnuggets Sofware Poll* [online]. 2018. [cit. 02.02.2019].Dostupné na: https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html

[23] PROVOST, Foster – FAWCETT, Tom. *Data Science And Its Relationship to Big Data and Data-Driven Decision Making* [online]. 2013. [cit. 18.01.2019. Dostupné na: https://www.liebert-pub.com/doi/pdfplus/10.1089/big.2013.1508

[24] RESEARCH AND MARKETS. The *Big Data Market: 2018 - 2030 - Opportunities, Challenges, Strategies, Industry Verticals & Forecasts* [online]. 2018. [cit. 27.01.2019]. Dostupné na: https://www.researchandmarkets.com/reports/4564313/the-big-data-market-2018-2030-opportunities

[25] SCHNEIDER, Ján. *Hadoop Essentials*. Bratislava : Datavard s.r.o., 2018.

[26] SEAGATE. *Digitalization of the World.* [online]. 2018. [cit. 23.01.2019]. Dostupné na: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

[27] THE RADICATI GROUP, INC. *Email Statistics Report, 2018-2022* [online]. 2018. [cit. 23.01.2019].Dostupné na: https://www.radicati.com/wp/wp-content/uploads/2017/12/Email-Statistics-Report-2018-2022-Brochure.pdf

[28] ZEPHORIA. *Top 15 Valuable Facebook Statistic*.[online]. 2019. [cit. 23.01.2019]. Dostupné na: https://zephoria.com/top-15-valuable-facebook-statistics/