Big Data – výskyt a využití ve firemní sféře

Ondřej Bouchal

Bakalářská práce
2017

Univerzita Tomáše Bati ve Zlíně
Fakulta aplikované informatiky

Univerzita Tomáše Bati ve Zlíně
Fakulta aplikované informatiky
akademický rok: 2016/2017

ZADÁNÍ BAKALÁŘSKÉ PRÁCE
(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Ondřej Bouchal**
Osobní číslo:	**A13211**
Studijní program:	**B3902 Inženýrská informatika**
Studijní obor:	**Informační technologie v administrativě**
Forma studia:	**prezenční**
Téma práce:	**Big Data – výskyt a využití ve firemní sféře**
Téma anglicky:	**Big Data – Its Presence and Use in the Corporate Sector**

Zásady pro vypracování:

1. **Vypracujte obecnou literární rešerši na dané téma.**
2. **Popište a identifikujte oblasti, kde se velká data prakticky využívají.**
3. **Proveďte analýzu a popis aktuálního stavu problematiky technologií zpracování a analýzy velkých dat.**
4. **Analyzujte použití dostupných technologií od Google, Apache, Microsoft, atd.**

Rozsah bakalářské práce:

Rozsah příloh:

Forma zpracování bakalářské práce: **tištěná/elektronická**

Seznam odborné literatury:

1. **Data science & big data analytics: discovering, analyzing, visualizing and presenting data. Indianapolis: Wiley, 2015, xviii, 410 stran. ISBN 978-1-118-87613-8.**
2. **HOLUBOVÁ, Irena, Jiří KOSEK, Karel MINAŘÍK a David NOVÁK. Big Data a NoSQL databáze. První vydání. Praha: Grada, 2015, 281 stran. ISBN 978-80-247-5466-6.**
3. **MARZ, Nathan a James WARREN. Big data: principles and best practices of scalable real-time data systems. Shelter Island: Manning, 2015, xx, 308 stran. ISBN 978-1-617290-34-3.**
4. **LABERGE, Robert. Datové sklady: agilní metody a business intelligence. 1. vyd. Brno: Computer Press, 2012, 350 s. ISBN 978-80-251-3729-1.**
5. **MAYER-SCHÖNBERGER, Viktor a Kenneth CUKIER. Big Data: revoluce, která změní způsob, jak žijeme, pracujeme a myslíme. 1. vyd. Brno: Computer Press, 2014, 256 s. ISBN 978-80-251-4119-9.**
6. **EAGLE, Nathan a Kate GREENE. Reality mining: using big data to engineer a better world. Cambridge, Massachusetts: The MIT Press, 2014, 1 online zdroj (vi, 199 pages). ISBN 9780262324564.**

Vedoucí bakalářské práce: **doc. Ing. Roman Šenkeřík, Ph.D.**
Ústav informatiky a umělé inteligence

Datum zadání bakalářské práce: **3. února 2017**

Termín odevzdání bakalářské práce: **30. května 2017**

Ve Zlíně dne 3. února 2017

UNIVERZITA TOMÁŠE BATI VE ZLÍNĚ
Fakulta aplikované informatiky
L.S.

doc. Mgr. Milan Adámek, Ph.D.
děkan

Ing. Miroslav Matýsek, Ph.D.
ředitel ústavu

Jméno, příjmení: Ondřej Bouchal

Název bakalářské/~~diplomové práce~~: Big Data – výskyt a využití ve firemní sféře

Prohlašuji, že

- beru na vědomí, že odevzdáním diplomové/bakalářské práce souhlasím se zveřejněním své práce podle zákona č. 111/1998 Sb. o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších právních předpisů, bez ohledu na výsledek obhajoby;
- beru na vědomí, že diplomová/bakalářská práce bude uložena v elektronické podobě v univerzitním informačním systému dostupná k prezenčnímu nahlédnutí, že jeden výtisk diplomové/bakalářské práce bude uložen v příruční knihovně Fakulty aplikované informatiky Univerzity Tomáše Bati ve Zlíně a jeden výtisk bude uložen u vedoucího práce;
- byl/a jsem seznámen/a s tím, že na moji diplomovou/bakalářskou práci se plně vztahuje zákon č. 121/2000 Sb. o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon) ve znění pozdějších právních předpisů, zejm. § 35 odst. 3;
- beru na vědomí, že podle § 60 odst. 1 autorského zákona má UTB ve Zlíně právo na uzavření licenční smlouvy o užití školního díla v rozsahu § 12 odst. 4 autorského zákona;
- beru na vědomí, že podle § 60 odst. 2 a 3 autorského zákona mohu užít své dílo – diplomovou/bakalářskou práci nebo poskytnout licenci k jejímu využití jen připouští-li tak licenční smlouva uzavřená mezi mnou a Univerzitou Tomáše Bati ve Zlíně s tím, že vyrovnání případného přiměřeného příspěvku na úhradu nákladů, které byly Univerzitou Tomáše Bati ve Zlíně na vytvoření díla vynaloženy (až do jejich skutečné výše) bude rovněž předmětem této licenční smlouvy;
- beru na vědomí, že pokud bylo k vypracování diplomové/bakalářské práce využito softwaru poskytnutého Univerzitou Tomáše Bati ve Zlíně nebo jinými subjekty pouze ke studijním a výzkumným účelům (tedy pouze k nekomerčnímu využití), nelze výsledky diplomové/bakalářské práce využít ke komerčním účelům;
- beru na vědomí, že pokud je výstupem diplomové/bakalářské práce jakýkoliv softwarový produkt, považují se za součást práce rovněž i zdrojové kódy, popř. soubory, ze kterých se projekt skládá. Neodevzdání této součásti může být důvodem k neobhájení práce.

Prohlašuji,

- že jsem na diplomové/bakalářské práci pracoval samostatně a použitou literaturu jsem citoval. V případě publikace výsledků budu uveden jako spoluautor.
- že odevzdaná verze diplomové práce a verze elektronická nahraná do IS/STAG jsou totožné.

Ve Zlíně, dne 30.5.2017

Bouchal

.....

podpis diplomanta

ABSTRAKT

Cílem této práce je zjistit, v jaké míře se vyskytují Big Data ve firemní sféře. Je zde vysvětlené, co to jsou Big Data, kde bylo možné se s nimi v minulosti setkat, zda jsou Big Data pro malé nebo velké podniky, způsoby jejich zpracování a krátce představený Big Data mining a Big Data Analytics. V této práci je také zpracovaný a vyhodnocený dotazník ohledně výskytu Big Dat ve firmách s různým zaměřením. Na konci práce je sestavený vzorový příklad analýzy z pohledu zákazníka, pro jakého poskytovatele platformy pro práci s Big Daty, by se mohl rozhodnout pomocí multikriteriální TOPSIS analýzy.

Klíčová slova: Big Data, Google, Hadoop, data, Microsoft, TOPSIS, AWS

ABSTRACT

The aim of this thesis is to find out frequency of occurrence of Big Data in businesses. The thesis explains the term Big Data, its history, usage in small or larger businesses and the ways of processing the data. The term Big Data Mining is also briefly mentioned. The questionnaire included in the thesis deals with occurrence of the Big Data in various types of companies. At the end of the thesis an example of an analysis from customer point of view is included. The analysis works as a tool for customers while choosing a provider of a platform for the Big Data using the TOPSIS analysis.

Keywords: Big Data, Google, Hadoop, data, Microsoft, TOPSIS, AWS

Děkuji svému vedoucímu bakalářské práce, doc. Ing. Romanu Šenkeříkovi Ph.D., za poskytnuté cenné rady, připomínky a odborné vedení.

Také bych chtěl poděkovat své přítelkyni, rodině a mým přátelům za jejich neutuchající víru v mé schopnosti a podporu, kterou mi poskytují.

Prohlašuji, že odevzdaná verze bakalářské práce a verze elektronická nahraná do IS/STAG jsou totožné.

OBSAH

ÚVOD

Tématem bakalářské práce jsou Big Data, jejich výskyt a využití ve firemní sféře. Cílem je podat informace o tom, zda firmy vědí o těchto datech a zda mají možnosti pro využití a zpracování Big Dat ve firmách.

Jak jednou řekl americký vědec Dan Ariely:

„*Big data is like teenage sex:*

everyone talks about it,

nobody really knows how to do it,

everyone thinks everyone else is doing it,

so everyone claims they are doing it." [1]

Tento výrok platí i dnes, spousta firem se ohání termínem Big Data, ale ne všechny přesně ví, jak velký přínos pro ně tato data mohou mít. Existuje mnoho případů, kdy firmy nedokáží zpracovat toto velké množství dat nebo ho zpracují, ale už neví, jak data dále využít a získat z nich to nejdůležitější a podstatné množství. Tudíž nedokáží využít pravý potenciál Big Dat.

Big data byla určitě i jeden z důvodů, proč začaly vznikat první počítače. Bylo totiž potřeba zpracovávat a ukládat více a více informací. Díky novým způsobům zpracování začala data dávat jiný smysl, dá se díky nim předvídat budoucí události nebo zaměřovat na cílové potřeby zákazníků. V dnešní době máme velice výkonné a vyspělé počítače než dříve, ale stále se přichází na nové způsoby, jak rychleji a efektivněji zpracovávat Big Data.

Celá bakalářská práce je rozdělena do několika kapitol. Jako první byl objasněn pojem Big Data, co je to, jak je lze charakterizovat a jaká je jejich velikost. Dále jsou představeny určité případy z minulosti, které se dají považovat za začátky nebo také průkopníky Big Dat. Další kapitolou je zapojení Big Dat ve firemní sféře. Není totiž nikde psané, zda jsou tato data pro velké nebo malé podniky, vše záleží jen na nich, zda se rozhodnou využít této výhody. Jsou zde dále představeny způsoby, jakými se Big Data zpracovávají nebo způsoby, kterými se data zpracovávaly dříve a také jednotlivé softwary.

Základem této bakalářské práce je praktická část, kde pomocí analýzy dotazníku lze ukázat, jaký vztah mají jednotlivé firmy k těmto datům, v jaké míře je využívají nebo jaká data přesně shromažďují. Poslední kapitola obsahuje zpracování analýzy pomocí multikriteriální TOPSIS metody. Zpracování probíhalo z pohledu budoucího zákazníka, který hledal vhodného poskytovatele platformy pro práci s Big Daty a pro finální rozhodnutí použil již zmiňovanou analýzu.

I. TEORETICKÁ ČÁST

1 CO JSOU TO BIG DATA?

Big Data, přímým překladem do češtiny veledata, jsou data, která jsou opravdu velká. Jelikož se odehrává 21. století, lze předpokládat, že jde o data digitální. Big Data nemají žádnou přesně ucelenou definici. Například se za Big Data dají označovat soubory dat, které mají takovou velikost, že se nedají zachycovat, zpracovávat či upravovat pomocí běžně používaných tradičních softwarových technologií v krátkých časových úsecích. Big Data jsou tedy soubory technologií, které se snaží spravovat a analyzovat velké množství nestrukturovaných dat pro získání výsledků, se kterými se dá dále pracovat nebo vyhodnocovat určité analýzy. Je možno mluvit o takzvaných stálých datech, protože zde teoreticky nedochází k žádnému mazání dat a jen se dále nabalují novější a novější data, tím se zpřesňují výsledky.

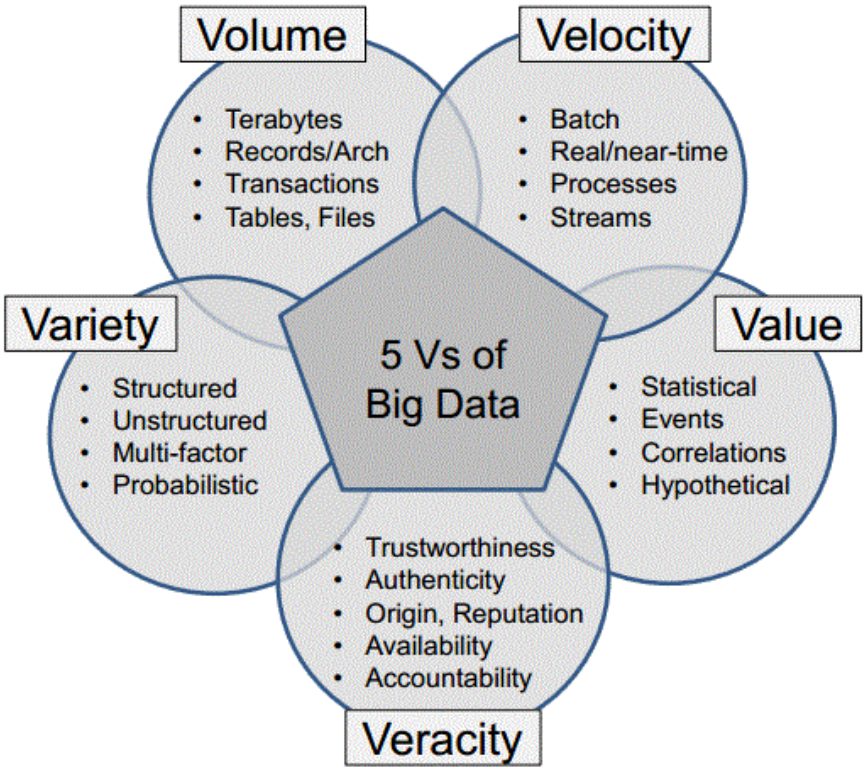
Big Data se z ničeho nic objevila při příchodu nových technologií, služeb a jejich kombinací navzájem. Příkladem může být zkoumání přírodních jevů pomocí různých vědeckých přístrojů nebo zkoumání sociálních sítí a souvisejících mobilních technologií a jejich rozšiřující aplikace. Tyto technologie a aplikace, díky svému velkému počtu uživatelů, generují každou chvilku četné množství dat, která musí být účinně uložena a chytře zpracována.

Díky tomuto uchovávání dat mohou firmy informace využívat a přizpůsobovat je své potřebě, což může být v podobě úpravy svých služeb a výrobků pro cílené zákazníky nebo mohou najít úplně jiné zdroje příjmů.

1.1 Vlastnosti Big Dat, 3V nebo také 5V

Neexistuje žádná přesná definice Big Dat, je jich spousta, ale jsou zde určité charakteristické vlastnosti, na kterých se všichni shodují. A to především tři základní vlastnosti spadající k definicím Big Dat: volume (objem), velocity (rychlost nárůstu) a variety (různorodost). Obecně se nazývají „3V“ vlastnosti. Dále k nim mohou patřit i další „V“ (viz Obr. 1). Jako jsou například veracity (nejistá věrohodnost), value (vysoká hodnota - pro firmu, která tyto hodnoty vlastní), validity (doba platnosti - omezené užívání) nebo také volatility (doba jejich nutného ukládání). [1]

Další vlastnosti budou zcela jistě časem přibývat. Je spousta zdrojů, které poukazují na základní vlastnosti Big Dat neboli „V“ a každý jim přikládá jiné hodnoty. Zmiňované 3 základní jsou ale vždy stejné (volume, velocity, variety). Na obrázku lze vidět, pod jaké části spadají určité odvětví. Například do variety (rozmanitost) patří data strukturovaná (structured), nestrukturovaná (unstructured) a pravděpodobnosti. Dále zde máme veracity (věrohodnost), zde patří důvěryhodnost (trustworthiness), pravost (authenticity), dostupnost (availability) atd. [1]



Obrázek 1 Big Data 5V model, zdroj: [2]

1.1.1 Volume (Objem)

Objem představuje celkovou velikost nebo množství aktuálně dostupných dat. Data, která nejsou stejná, narůstají obrovským tempem, tedy exponenciálně. Volume může být tedy složen z různých záznamů uživatelů, dat ze sociálních sítí, webů cílených na určité potřebné informace neboli témata, různých transakcí nebo ze všeho dohromady. Na sociálních sítích jsou k nalezení data ve formátech videí, hudebních souborů a objemných obrázků. Také je velmi běžné, že podniky mají terabytové a petabytové úložné systémy.

Vzhledem k tomu, že databáze exponenciálně rostou, aplikace a architekturu je potřeba často obměňovat. Někdy jsou stejné údaje přehodnocovány různě, z více úhlů pohledu. Přestože původní data jsou stejná, nově vytvořené systémy vytváří jiné vzorce pro vyhodnocování dat. Ovšem z takového množství dat může být vysoké procento jen nežádoucích informací, které mohou zkreslit finální úspěch.

Výskyt Big Dat se bude neustále vyvíjet, lze říci, že to, co je dnes považováno za Big Data, může být za pár let bezproblémově zvládnutelný objem dat. [3], [4]

1.1.2 Velocity (Rychlost)

Hlavní myšlenkou je zachytit celkovou dobu informace. Tedy dobu od vytvoření nové informace, přes její získání až po její finální zpracování. V minulosti bylo dávkované zpracování běžným krokem, informace se aktualizovaly z databáze každou noc nebo pouze jednou za týden, protože počítače vyžadovaly značný čas pro zpracování dat a aktualizaci databází. Teprve v poslední době se začal přikládat velký význam na rychlost zpracování dat a s vývojem nových technologií se na ni bude klást stále větší důraz. Už dnes je velmi důležité okamžité zpracování informací v reálném čase a rozhodnutí o budoucím naložení s nimi. Rychlost je měřítkem toho, jak rychle přicházejí data.

V éře Big Dat jsou data vytvářena v reálném čase nebo téměř reálném čase. S mnoha možnostmi připojení k internetu, jako například bezdrátové nebo kabelové zařízení a přístroje, lze předávat data v okamžiku jejich vytvoření. Jako příklad lze uvést Facebook, který musí zvládat návaly fotografií každý den. Tyto fotografie musí uložit, zpracovat, nahrát a musí být schopen je obnovovat. Také server Youtube, na který je nahráváno každou minutu přes 100 hodin videí, které musí být okamžitě dostupné k přehrání. Nebo je odesláno přes 200 miliónu e-mailů za minutu, které musí být hned doručeny. Výzvou každé organizace je tedy vypořádat se s obrovskou rychlostí, jakou jsou data vytvořena a používána v reálném čase. [3], [5]

1.1.3 Variety (Rozmanitost)

Data se nacházejí v odlišných formátech, například strukturovaná, nestrukturovaná, textová, obrazová a zvuková. V minulosti byla všechna data, která byla vytvořena, strukturována. A tím se odlišují Big Dat, zde protékají a filtrují se všechna tato data

dohromady. Tedy úhledně dány do sloupců a řádků, ale tyto dny jsou již pryč. V současné době je okolo 90% dat nestrukturovaných údajů, které generují různé firmy nebo organizace. Data dnes opravdu přicházejí v mnoha různých formátech. Široká škála dat vyžaduje odlišný přístup, stejně tak jako různé techniky pro ukládání nezpracovaných dat. Každý z nich požaduje různé typy analýz nebo různé nástroje k použití. Úlohou Big Dat je tedy všechny tyto informace získat, dát je dohromady, do jednotné podoby pro budoucí zpracování a nakonec data zpracovat a vybrat z nich podstatné a důležité informace. [3]

1.1.4 Veracity (Věrohodnost)

Mít velké, různě uspořádané množství dat přicházejících velkou rychlostí je bezcenné, pokud jsou tato data nesprávná. Zde se v Big Datech odkazuje na předsudky, abnormality nebo jiné zvláštní údaje v datech. Existují totiž data, která jsou vytvořena a uložena smysluplně k analyzovanému problému a data, která moc s problematikou nesouvisí nebo jen okrajově. Tedy při analýze a porovnávání je největší výzvou zjistit, která data jsou věrohodná a která ne, poté se až bere v potaz jejich objem dat a rychlost zpracování. Proto je důležité, aby se data uchovávala v čistotě a nedocházelo tak k hromadění tzv. „špinavých“ dat v systémech. Pokud je cílem tato data nashromáždit a analyzovat, je nutností být schopen důvěřovat. [5], [4]

1.1.5 Value (Hodnota)

Zde se přenáší schopnost přetvořit data na určitou hodnotu. Je tedy důležité, aby se podniky pokusily shromažďovat a využívat Big Data. Ale je snadné se dostat do pasti a začít se topit v Big Datech, pokud již dříve nebylo započato s inicializací určité obchodní hodnoty, které nám tato data přinesla. Big Data můžou přinést hodnotu téměř v jakékoliv oblasti podnikání nebo společnosti. [5]

- **Pomáhají firmám optimalizovat své procesy:** dokáží předpovídat poptávku, navyšovat nebo snižovat ceny výrobků.
- **Umožňují firmám lépe naslouchat zákazníkům:** nabízet jim doporučení, např. Amazon nebo Netflix.
- **Zlepšují zdravotní péči:** dokáží předpovídat výskyt chřipky.
- **Posouvají sportovní výkony:** GPS trackery.

Hodnoty v Big Datech je vhodné si určit. Protože brzy každá část podniku a společnosti změní své systémy kvůli tomu, že nyní mají mnohem více dat a také nové možnosti analyzování.

1.1.6 Validity (Doba platnosti)

Dobou platnosti se rozumí to, jak dlouho by měla být data platná, tedy od kdy do kdy budou uložena. V této době, kdy je potřeba dostávat data v reálném čase, je důležité určit, v jakém okamžiku nejsou data pro aktuální analýzu relevantní. [5]

1.2 Jaká je velikost Big Dat

Aby bylo možno velikost vůbec nějak objektivně měřit v přesných číslech, je nutno získat nějaké měřítko pro velikost dat. Lze vycházet z toho, že v dnešní době se velikost pevných disků pohybuje v hodnotách několika terabajtů (TB), což je 10^{12} bajtů. Ale Big Data jsou velice kapacitně obsáhlá, takže lze mluvit o objemu dat v petabytech. Přičemž je známo, že jeden petabyte je 1 000 000 000 000 000 bytů neboli 10^{15} bytů.

Společnost IBM (International Business Machines Corporation) pomocí Big Dat uvádí, že v roce 2020 bude podle jejich odhadů 6 miliard lidí na světě vlastnit mobilní telefon. Každý den pak vznikne 2,8 kvintiliónů, tedy $2{,}8 \times 10^{18}$ bajtů, dat a celkově bude uloženo na discích 40 zettabajtů dat (1 zettabajt je 10^{21}bajtů), což je ekvivalent miliardy pevných disků o velikosti 1TB. Nejčastějším zdrojem Big Dat jsou sociální sítě. Na serveru Youtube je denně shlédnuto přes miliardu hodin videí. Síť Twitter má přes 300 miliónů aktivních uživatelů měsíčně a přes 1 bilión tweetů (krátkých textových zpráv) každý měsíc. Dále je zde známý Facebook. Společnost Zephoria zveřejnila, že Facebook má více než 1,86 miliardy uživatelů aktivních každý měsíc, 300 miliónů nahraných fotek za den. Každých 60 sekund je na Facebook přidáno 510 000 komentářů, 293 000 aktualizovaných stavů a 136 000 nahraných fotografií a 42% obchodníků uvádí, že Facebook je pro jejich podnikání důležitý.

Dále lze uvést příklad u letadla Boeing. Jeden jeho motor vygeneruje každých 30 minut provozu 10 TB dat. Je-li vzato v potaz, že jeden zaoceánský let čtyřmotorového letounu vygeneruje 640 TB dat, která budou vynásobena asi 25 tisíci lety, které se uskuteční každý den, vyjde velké množství dat. [6], [1]

2 POČÁTKY BIG DAT

Nejvíce se o pojmu Big Data začalo mluvit na přechodu let 2012/2013, kdy dosáhl opravdového vzestupu. Pro mnoho lidí to byla velká záhada, nový termín a nulové množství informací, s čím se vlastně setkávají. Vědělo se jen, že jde o velká kvanta údajů a jsou důležité pro spoustu různých společností. Dalo by se říct, že všichni o nich mluvili nebo psali, ale jen malé množství je skutečně zpracovávalo a používalo. Už v minulosti byla Big Data velkým obchodním nástrojem. Nejednalo se tedy o pouhé shromažďování dat ve velkých podnicích, kde dat už bylo opravdu spousta, ale šlo i o nějaké uspořádání a filtrování těchto dat, dle zadaných požadavků, nebo pro vyvození informací, dle požadavků firem.

2.1 Počátek 19. století

Na počátcích devatenáctého století přišel námořník Matthew Fountaine Maury (americký důstojník námořnictva) k úrazu, díky kterému už se nadále nemohl plavit po moři. Proto dostal nabídku od námořnictva dělat v kanceláři vedoucího na oddělení Skladů námořnických map a přístrojů. Tohle místo bylo pro něj ideální. Jako mladý navigátor totiž nikdy nedovedl pochopit, proč lodě po moři křižují a neplují přímočarými trasami. Od kapitánů však dostával odpovědi v tom smyslu, že je bezpečnější se plavit po osvědčených trasách, protože v neznámých vodách číhají skrytá nebezpečí. Ovšem Maury věděl ze svých zkušeností, že to není tak úplně pravda. Pozoroval, jak se větry na moři střídají v přesném rozvrhu, silně vanoucí vítr náhle přestával při západu slunce. Ve všech přístavech, kde se Maury zastavil, hledal staré mořské vlky a shromažďoval jejich znalosti a zkušenosti. Učil se pravidelnost vln, větrů a mořských proudů, protože v námořnických mapách o tomto nebylo ani zmínky. Shromažďoval i staré lodní zápisky a mapy, které byly považovány za veteš. Maury s tuctem dalších pracovníků díky těmto informacím rozdělil celý Atlantik na bloky po pěti stupních zeměpisné šířky a délky. Zde zaznačil teploty, rychlost, směr větrů a vln v určitém ročním období. Zavedl lodní formuláře, které museli vyplňovat všichni námořníci a dle nich získával další data o trasách a podmínkách plavby. Mauryho námořní mapy zkrátily dlouhé plavby až o třetinu, což bylo výhodné pro obchodníky. Tyto mapy jsou využívány dodnes.

Lze tedy říct, že se se svou prací zařadil mezi průkopníky datafikace (Big Dat), neboli získávání informací z dat, kde ostatní žádnou cenu neviděli. [7]

2.2 Google Flu Trends

V roce 2009 byl objeven nový virus, chřipka H1N1. Byly to prvky virů způsobující prasečí a ptačí chřipku. Hygienici se proto začali obávat příchodu nebezpečné pandemie, která může zasáhnout celý svět. Proti tomuto novému viru nebyla dostupná žádná vakcína, doktoři jen doufali, že se jim postup nákazy podaří zpomalit. Ovšem k tomu potřebovali vědět, kde se chřipka vyskytuje.

Agentura CDC (Centers for Disease Control and Prevention), která byla součástí amerického ministerstva zdravotnictví, požádala lékaře, aby poskytovali údaje o nových výskytech chřipky. Ovšem tyto údaje byly vždy o týden nebo dva zpožděné. Lidé se dostavovali k lékařům vždy až po 2-3 dnech od projevení příznaků nákazy. Agentura tak vyhodnocovala výsledky o výskytu chřipky jen jednou týdně. V tu dobu přišel Google s Google Flu Trends, který shromáždil 50 milionů termínů, které Američani nejčastěji hledali a porovnal to s daty agentury CDC. Tento nástroj byl nastaven tak, aby monitoroval chřipkové případy na celém světě a v reálném čase, netrvalo to tedy týden ani dva. Podmínka monitorování byla založena na vyhledávání přes Google, ovšem zadávané výrazy musely mít něco společného s chřipkou. Google tento program představil takto: „*Hledali jsme úzký vztah mezi tím, kolik lidí hledá témata související s chřipkou a kolik lidí ve skutečnosti chřipku má. Bylo jasné, že ne každý, kdo vyhledává slovo chřipka, musí být opravdu nakažený. Když se shromáždily všechny hledané výrazy související s chřipkou, objevili jsme určitý vzorec. Porovnávali se velké počty dotazů s tradičními systémy sledování chřipky a zjistilo se, že určité vyhledávací dotazy bývají častěji zadávané v době, kdy se chřipka na určitém území nachází. Podle vyhledávaných dotazů se dá tedy spočítat, kde a kdy se chřipka nachází a hlavně v jakých zemích a regionech.*“ Dá se tedy říct, že díky velké databázi Googlu a 450 miliónů různých matematických modelů, kterými disponoval Google Flu Trends, je důvod proč se GFT stal symbolem Big Dat. [8], [9]

2.3 Kanály a New York

V New Yorku docházelo každoročně k několika set rozžhavením nebo výbuchům kanálů, protože pod nimi vypukl požár. Litinové kryty kanalizací, vážící přes 120 kilogramů, někdy vylétly do více než 15 metrů a poté dopadaly zpět na zem, což nebylo zrovna nejbezpečnější. Pravidelně se prováděla inspekce a údržba krytů, kterou vedla společnost Con Edison, která také zajišťovala elektrickou energii. Vždy se spoléhali na to, že kryty, které se chystají zkontrolovat, by mohly být ty, které se chystají vybuchnout. Tento přístup byl jen o náhodě. Ovšem v roce 2007 si podali žádost o statistiky z Kolumbijské univerzity. Doufali, že pomocí historických údajů o kanálech a jejich předchozích údržbách zjistí, kde by v budoucnu mohly problémy nastat. Předem by tedy věděli, kde investovat své zdroje a předejít tak problémům. V New Yorku je kolem 150 000 kilometrů podzemních kabelů a na ostrově Manhattan přes 51 000 krytů kanálu, z nichž více jak polovina pocházela z doby Thomase Edisona. Bylo jasné, že půjde o velký problém s veledaty. Záznamy se vedly již od půlky 19. století, ale měly velké množství formátů bez myšlenky o tom, že budou sloužit k datové analýze. Pouze termín „servise box“ se označoval minimálně 38 způsoby (SB, S, S BOX, S.B., S/B, S/BX, atd.). Data byla opravdu špatně zpracována, ale muselo se z nich vytáhnout jen užitečné jádro pro získání kvalitního prediktivního modelu. Nebylo třeba použít jen vzorek dat, musela se zpracovat opravdu všechna data. Hlavní otázkou tedy nebylo, proč kanály vybuchují, ale který kanál vybuchne. Nakonec se při dolování dat opravdu objevily užitečné informace. Jakmile se chaotická data naformátovala, aby je počítač byl schopen zpracovat, začalo se testovat. Vše fungovalo opravdu skvěle. Ve výsledku mezi 10% kanalizačních krytů na začátku seznamu patřilo 44% krytů, u kterých poté došlo k nehodám. Finálním výsledkem však bylo, že k nejvíce nehodám dochází tam, kde byly nejstarší kabely a také to, jestli u příslušného krytu došlo už v minulosti k potížím. [7]

3 BIG DATA VE FIREMNÍ SFÉŘE

Big Data ve firmách představují velké potenciální zisky. Jak už bylo uvedeno, velké množství dat může v mnoha firmách přiblížit jak potřeby zákazníků, tak jim více porozumět. Dá se zjistit, co a kdy zákazníci nejvíce nakupují, i kde se zrovna nacházejí. Z čehož vyplývá, že díky nabídkám přizpůsobeným zákazníkům, firmy mohou očekávat vyšší příjmy ze služeb nebo prodeje zboží a dělají si tak dobrou pověst. Existuje však mnoho firem, které s Big Daty neumí pracovat. Podle průzkumů se uvádí, že jde skoro o 80% podniků. Na obrázku č. 2 můžeme vidět pojmy spojující Big Data s firmami (business - podnik, logs - protokoly/záznamy/přihlášení, sharing - sdílení, statistics - statistiky, decision - rozhodnutí, transfer - převody, applications - aplikace, explore - posudky, optimization - optimalizace, atd.).

Obrázek 2 Big Data zdroj: [9]

Množství Big Dat roste stále více, tím jak zákazníci komunikuji, nakupují nebo využívají více elektronických služeb nebo jiné digitální technologie. Objevují se nové případy v podobě pokročilých analýz Big Dat, které lze nyní efektivně zpracovávat. K tomu jsou ale zapotřebí rychlé počítače, levné a velké úložné prostory a nové softwarové nástroje a postupy. Všechny tyto věci se nám momentálně dostávají na výborné úrovni, a to firmám dovoluje vytvářet nebo poskytovat služby či produkty na přání zákazníků, při tom také dochází ke snížení nákladů a zlepšení efektivity. [10]

Dá se tedy říci, že Big Data mohou přinášet lepší a efektivnější návrhy pro poskytování produktů a služeb a dále mohou být poskytovány třetím stranám za další cenné informace pro firmu. Takto si mohou firmy předávat svá vnitřní data, informace o zákaznících nebo vývoji na trhu v minulosti a mohou se zaměřit na budoucí marketingové cíle nebo předpovídat chování zákazníků. Díky tomu dokáží perfektně zacílit své obchodní aktivity. Podle studie od společnosti Accenture Analytics si 59% firemních manažerů nedokáže představit firmu bez analýzy Big Dat a 34% manažerů je považuje za velmi důležitá pro jejich působení. [10]

3.1 Big Data pro velké nebo malé firmy?

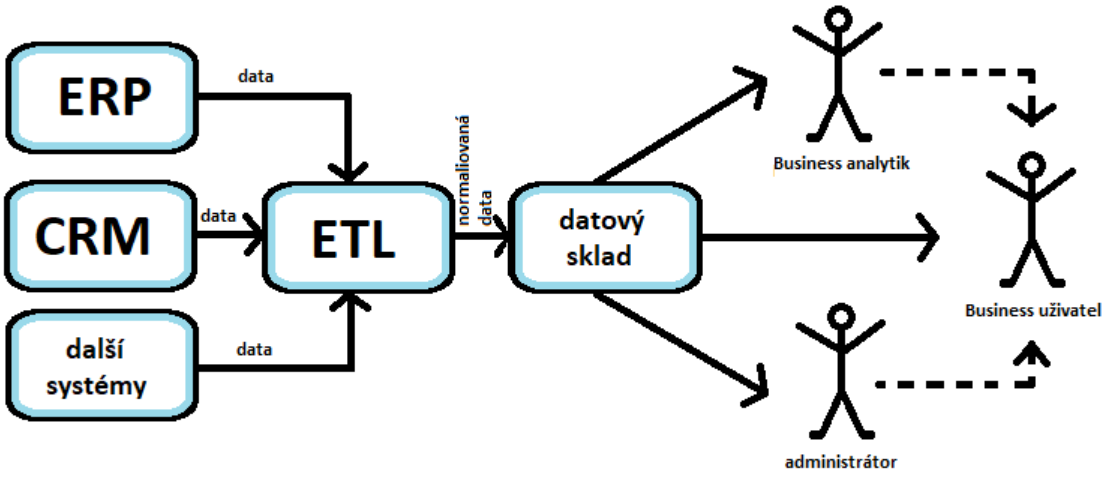
Nikde není přesně dáno, zda zvládne tyto údaje zpracovat malá nebo velká firma. Ale už více průzkumů ukázalo, že s Big Daty pracují spíše velké podniky oproti menším firmám. Právě tyto velké podniky těží z analýz Big Dat nejvíce. Určitě jedním z důvodů bude to, že velké firmy budou lépe umět odhadnout potencionální přínosy a více se zaměří na praktické využití těchto informací pro své reálné výhody. U velkých podniků se dá také vycházet z toho, že mají větší počty lidských nebo finančních zdrojů. U menších podniků je velký problém, protože ne vždy umějí dokonale zpracovat Big Data pro svoji potřebu nebo je dokonce ani nezpracují. Může to být i tím, že na tyto analýzy a problémy chybí specialisté, zdroje a rozpočet. Velice málo firem se výhradně spoléhá na své vlastní odborníky, proto s těmito daty nemohou počítat. Ty firmy, které ve velkém zpracovávají a vyhodnocují data, věří v obrovské revoluční změny, které Big data budou do byznysu vnášet. Těmto firmám je jasné, že musí měnit způsoby, jak do teď fungovaly a nesmějí si nechat tento pomyslný vlak ujet. Big Data nabízejí průnik do nových odvětví podnikání, dá se na nich postavit pilíř celofiremní strategie, tedy spousty nových příležitostí. Nejdůležitějším krokem je však změna způsobu chápání, myšlení a vnímání dat jako jedno z nejzákladnějších věcí pro firmu. [10]

4 ZPRACOVÁNÍ BIG DAT VE FIRMÁCH

Big Data mají odlišné vlastnosti a strukturu, to je odděluje od klasických firemních dat. Normální datové doky a nástroje nejsou schopny zpracovat a analyzovat velké objemy dat ve velmi krátkém čase (real-time), proto je potřeba hledat nové způsoby zpracování a analýzy Big Dat.

4.1 Klasický způsob využití dat z datových skladů

Není to tak dávno, kdy se data pro analýzy zpracovávala stálými úlohami. Každý podnik si produkoval spíše strukturovaná data ze stálých podnikových aplikací (CRM, ERP, ekonomika). Poté se pomocí ELT (Extrakce, Transformace, Load) nástrojů data kontrolovala, zpracovávala a poskytovala ostatním aplikacím podporu pro rozhodování a ukládali se do datového skladu. Proces se ve firmách opakoval v pravidelných denních nebo týdenních cyklech. Z těchto dat pak datoví analytici prováděli přes různé analytické nástroje výpočty nad daty ze skladů, viz Obr. 3. Objem dat z datových skladů málokdy přesáhl pár terabytů. [11]



Obrázek 3 Zpracování strukturovaných dat

4.2 Hadoop

Název Hadoop vznik v roce 2005, podle dětské, žluté, plyšové hračky slona od jednoho z tvůrců. Název je i se žlutým slonem ve znaku, viz Obr. 4. Hadoop je open source Framework, který zpracovává, analyzuje a ukládá velké množství nestrukturovaných dat. Bavíme se zde o velikosti několika petabajtů. Základem škálovatelnosti je určitý model, kde se výpočetní funkce přiřazují k datům, namísto přiřazení dat k výpočetním funkcím. Hadoop je tedy zaměřen na vyhledávání informací v obrovském množství dat, které je běžnými prostředky nemožné vyhledat. Byl dokonce navržen tak, aby se mohl rozšířit na tisíce počítačů z jednoho serveru. Také dokáže detekovat a řešit chyby na úrovni aplikace. Další výhodou je, že si poradí s většinou formátů nebo souborů, tudíž mu firmy mohou pokládat různé dotazy. Některé organizace mohou nabízet i Hadoop jako cloud službu, při tomto řešení není třeba instalovat servery, což jsou další investice navíc.

Důležitou myšlenkou je, že Hadoop je založen na distribuovaném uložení a zpracování dat. Proto se musí provozovat na několika vzájemně propojených clustrech (vzájemné seskupení volně vázaných počítačů nebo serverů, které spolu úzce spolupracují). Data, která jsou uložena v Hadoopu jsou taktéž uložena na různých serverech v clusteru. Pokud se některá kopie dat poškodí, zajistí se automatické přenesení na jiný dostupný server v clusteru. Taktéž Hadoop funguje pro výpočty. Jakmile jeden selže, celá operace se opakuje na jiném paralelně zapojeném serveru. [12], [13]

Obrázek 4 logo Hadoop, zdroj: [14]

Dá se tedy říct, že se Hadoop skládá ze dvou základních komponentů:

- HDFS – Hadoop Distributed File Systém (systém souborů)
- MapReduce – programovací paradigma

4.2.1 MapReduce (vývoj)

MapReduce je vysoce škálovatelná platforma pro ukládání dat, určená pro zpracování velkých datových sad ve stovkách až tisících výpočetních uzlů, které pracují paralelně. Hadoop je schopen spustit programy napsané v různých jazycích: Java, C++, Python. Díky tomu, že MapReduce pracuje paralelně, je velice užitečný pro provádění rozsáhlých datových analýz, díky více serverům v clusteru. MapReduce programy obsahují dva důležité úkoly: Map a Reduce. Map si bere určitou sadu dat a tu převádí zase na jinou, kde jednotlivé prvky jsou rozděleny na key/value (klíč a hodnotu). Druhý úkol Reduce zachycuje výstup z Map a kombinuje hodnoty key/value do menších sad. Obecně platí, že paradigma MapReduce je založeno na posílání počítače tam, kde jsou data uložena. Program pracuje ve třech etapách, kterými jsou: mapování, přehazování a redukování. Po dokončení zadaných úloh cluster sesbírá a zredukuje data, která potom odešle zpět na Hadoop server.

Lze uvést jednoduchý příklad počítání výskytu jednotlivých slov v textovém dokumentu. Map funkce si rozdělí celý dokument na samostatná slova a vytvoří key/value pár pro všechna jednotlivá slova. Následně jsou všechny páry seřazeny a předány funkci Reduce, která sečte všechny páry. [15]

4.2.2 HDFS (Hadoop Distributed File Systém, ukládání dat a správa)

HDFS představuje distribuovaný souborový systém, který je navržen na principu obyčejného hardwaru. Má své specifické vlastnosti, ale také je velice podobný jako jiné distribuované soubory. Je velice odolný vůči chybám, nabízí velkou propustnost k aplikačním datům a je ideální pro obsáhlé datové úložiště. Je klíčovým nástrojem pro správu Big Dat. Při přijímání dat informace rozdělí na jednotlivé části a rozvrhne je do různých uzlů v clusteru, což umožňuje paralelní zpracování. [15]

4.2.3 Pig

Apache Pig je ukázkou využití principu MapReduce pro analýzu Big Dat. Sestává se z jazyka vyšší úrovně nazývaného Pig Latin, který umožňuje definovat požadované analytické operace nad daty a infrastrukturami, které skripty v jazyce Pig Latin provádí. Výhodou je, že tyto programy je možné paralelizovat a aplikovat na Big Data. [1]

4.2.4 Cascading

Další zajímavou nadstavbou nad Hadoop MapReduce je Apache Cascading. Má podobný cíl jako nástroj Pig (umožnit pohodlnou a rychlou implementaci zpracování lokálních i distribuovaných dat pomocí Hadoop MapReduce). Cascading pro tyto účely používá vyšších programovacích jazyků jako např. Java. Práce s daty je zde založena na myšlence vytváření datových toků (streams) složených z rour (pipes) a datových filtrů. [1]

4.2.5 Hive

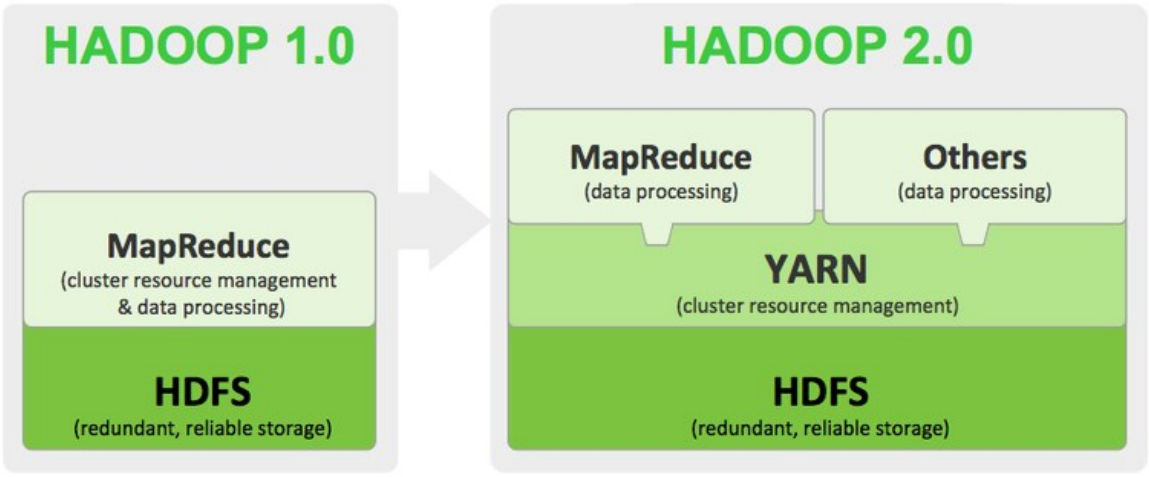
Hive na rozdíl od předchozích nástrojů neslouží ke zpříjemnění práce s Hadoop MapReduce frameworkem. Jedná se o infrastrukturu zajišťující funkcionalitu datového skladu. Apache Hive umožňuje definovat a přiřadit k datům, uloženým v distribuovaném uložišti, strukturu podobnou relačnímu modelu a pak se nad nimi dotazovat prostřednictvím jazyka HiveQL, vycházejícího z jazyka SQL. [1]

4.2.6 Zookeeper (řízení)

Jediná koordinační služba pro distribuované aplikace je Zookeeper. Slouží pro správu konfigurace, skupinové služby synchronizace a pojmenování. [15]

4.3 Hadoop 2.0 YARN

YARN je často nazýván operačním systémem Hadoopu, jelikož nese plnou zodpovědnost za správu, monitorování pracovních záležitostí, provádění bezpečnostních kontrol a řízení Hadoopu. Umožňuje spouštět Non-MapReduce aplikace. Přináší v některých aplikacích značné zvýšení výkonu, podporuje i další modely zpracování a zavádí pružnější spouštěcí mechanismus. YARN podporuje i jiné modely zpracování než MapReduce. Dále byla přidána podpora pro systém Microsoft Windows a všestranná kompatibilita s existujícími aplikacemi MapReduce pro verze Hadoop1.x. Hlavní rozdíl mezi Hadoop 1.0 a Hadoop 2.0 lze vidět na Obr. 5.



Obrázek 5 Hadoop architektura, zdroj: [16]

4.4 GFS (Google File System)

GFS je škálovatelný distribuovaný souborový systém od společnosti Google. Vznikl převážně pro potřeby Googlu. Poskytuje toleranci chyb, spolehlivost, škálovatelnost, dostupnost a výkonnost velkým sítím a připojeným uzlům. GFS se skládá z několika systémů pro ukládání dat, které jsou postaveny z nízko nákladových komoditních hardwarových komponent. Je také optimalizován tak, aby vyhovoval různým potřebám využití a ukládání dat společnosti Google, jako je třeba vyhledávač, který generuje velké množství dat, která se musí ukládat. [17]

GFS cluster má jeden master server s více blokovými chunk servery, které jsou nepřetržitě přístupné různým klientským systémům. Chunk servery ukládají data jako Linux soubory na lokální disky. Souborům je při vytvoření přidělen jedinečný a neměnný identifikátor. Chunk servery jsou minimálně 3x klonovány skrze servery, aby byla zajištěna vyšší spolehlivost proti chybám. Velikost chunku se stanovuje na 64 MB. Největší GFS clustery mají více než 1000 uzlů s kapacitou větší než 300TB, to umožňuje nepřetržitou přístupnost stovkám klientů. Hlavní funkce: [17]

- Odolnost proti chybám
- Replikace dat
- Automatické a efektivní obnovování dat
- Vysoká propustnost
- Vysoká dostupnost

4.4.1 HDFS vs GFS

V této kapitole si uvedeme srovnání dvou nejvýznamnějších souborových systémů pro práci s Big Daty. Oba dva systémy zahrnují hlavní uzel, který řídí celkový chod systému a komunikace mezi uzly. Google cluster obsahuje jeden master server a řadu chunk serverů, které jsou obvykle s Linux systémem a běží zde procesy uživatelské úrovně. HDFS je škálovatelný systém souborů napsaný v Javě, má jediný NameNode a řadu Datanodes, které slouží jako síť pomocí protokolů specifických pro HDFS. Další rozdíly v tabulce 1.

Tabulka 1 HDFS vs GFS

Parametry	Google File System	Hadoop Distributed File System
Vývojář	Google, Inc.	Apache Software Foundation
Programovací jazyk	C, C++	Java
Licence	Proprietary	Apache License 2.0
Rozdělení procesů	Master and chunk server	Name node and Data node
Základní rozdělení do bloků	Převážně je to 64 MB, dá se i přenastavit	Převážně je to 128 MB, dá se i přenastavit
Zabezpečení	Názvy souborů jsou náhodné. Všechna data zakódovaná. Měnící se algoritmy pro maskování	Implementuje model oprávnění pro soubory a adresáře. Založeno na povoleních z POSIX jako UserID a GroupID.
Komunikace	TCP protokol	Remote procedure call, TCP/IP
Implementace	Vytvořeno čistě pro Google	Pro spoustu společností: Facebook, IBM, Netflix

4.5 Microsoft Azure

Microsoft Azure je otevřená a flexibilní platforma, která poskytuje rychlé nasazení a správu cloud řešení. Nabízí opravdu širokou škálu služeb mezi hotovými aplikacemi, výpočetním výkonem, datovým uložištěm a síťovými službami. Aplikace se dají vytvářet v libovolných programovacích jazycích. Patří mezi globální službu, která se nachází na 38 regionech po celém světě. Azure provádí i georeplikaci, kdy se informace z různých datacenter replikují do dalších, čímž dochází i k zálohování. Data se mohou přesouvat nebo replikovat pouze v rámci jednoho GEO regionu (Brazílie, Evropa, Asie, Japonsko, Amerika), ale nikdy se nedostanou z jednoho regionu do druhého. Nabízí několik druhů služeb, například služby nabízející výpočetní výkon, služby pro data, podpůrné služby, služby pro síťování.

4.5.1 HDInsight

Azure HDInsight je jediná plně spravovaná cloudová nabídka Apache Hadoopu, která poskytuje optimalizované opensource analytické clustery s 99,9% dostupností. Převážně tato služba slouží pro velké objemy dat, jako jsou spravované clustery se zabezpečením a monitorováním na podnikové úrovni. Tato služba je navržena pro plnou redudanci a vysokou dostupnost. Dochází zde k replikaci hlavního uzlu a geografické replikaci dat s NameNode. Díky tomuto nedochází k závažným chybám. Clustery prostředí HDInsight (Hadoop, HBase, Storm a Spark) podporují řadu programovacích jazyků (Java, Python) a mnoho další se dá nainstalovat pomocí skriptů. [18]

4.6 AWS (Amazon Web Servis)

Společnost Amazon není potřeba nijak představovat. Nejenže je to známý a pravděpodobně jeden z největších internetových obchodů, ale je také jeden z prvních poskytovatelů široce nabízených cloud služeb. AWS nabízí mnoho služeb, které nemusí pracovat vzájemně na sobě. Dále nabízí velkou škálu služeb, od výpočetních až po obchodní. [19]

4.6.1 Amazon EMR (Amazon Elastic MapReduce)

Amazon EMR poskytuje řízený Hadoop, který umožňuje snadné, rychlé a nákladově efektivní zpracování obrovských dat v rámci dynamických škálovatelných instancí EC2. Může také bezpečně a spolehlivě zpracovávat širokou škálu velkých případů využití dat, včetně log analýz, indexování webu, transformací dat, finanční analýzy a vědecké analýzy. Pomocí Amazon EMR lze poskytnout stovky nebo tisíce výpočetních instancí pro zpracování dat v libovolném měřítku. [19]

4.6.2 Amazon Simple Storage Servis (Amazon 3S)

Amazon 3S je uložiště objektů s jednoduchým webovým rozhraním pro libovolné množství dat z libovolného místa na internetu. Je navržen tak, aby poskytoval 99,99% dostupnost a měřítko skrze trilióny objektů celosvětově. [19]

5 BIG DATA MINING VE FIRMÁCH

Datamining neboli dolování dat je určitý proces, při kterém se získávají užitečné informace. Tyto informace jsou ovšem složitější a užitečnější než ty, které se získávaly doposud. Probíhá zde zkoumání vzorků nebo vzájemných vztahů v datech a výsledkem by měla být analýza, která dokáže předpovědět nebo určit trendy, pokud je dostatek dat. Data mining se dá použít skoro ve většině typů firemních aplikací, kde odpovídají na různé typy otázek. Především se používá tam, kde se shromažďuje velké množství dat. Datamining se dá definovat jako proces, který vyhledává potřebné vazby a trendy v Big Datech. Dá se říct, že datamining je fáze Online Analytical Processing, ale to není úplně pravda. Datamining přidává pokročilejší analýzy, než pouhé sumarizační analytické zpracování dat. Na trhu se dá objevit spoustu softwarů, které jsou označovány jako datamining systémy, ale v mnoha případech jsou nedostatečné základní požadavky. Do dataminingu zahrnujeme: statistiku, rozpoznávání podobností, databázové a výkonné výpočetní technologie, datové skladiště, získávání dat, bankovnictví (stav účtů, loginy), telefonní operátory, obchodní řetězce, průmysl, atd. Ovšem tento způsob získávání dat není jen pro velké firmy, ale má i své významné uplatnění v těch menších. [20]

5.1 CRoss-Industry Standard Proces for Data Mining (CRISP-DM)

Jedná se o standardizovaný postup pro veškeré obory, bez ohledu na to z jakého oboru data pocházejí. Vznikl za účelem sjednocení implementace dataminingu do strategie. Datamining lze popsat v šesti částech. [20]

- **Business/Research Understanding Phase:** je důležité porozumět potřebám zákazníka a stanovit cíle, tvoří se návrh a plán pro řešení.
- **Data Understanding Phase:** zde se vytváří hypotézy, které se dále v průběhu procesu snaží potvrdit, ale někdy se i vyvrátí nebo se najde jiné řešení.
- **Data Preparation Phase:** integrace více datových zdrojů. Špatná integrace dat může způsobit znehodnocení zdrojů dat, to se promítne na celkové kvalitě řešení.
- **Modeling Phase:** testuje vhodné metody a nastavení parametrů pro řešení daného problému. Zde se vybere několik řešení, které postupují do další fáze.

- **Evaluation Phase:** konečné hodnocení a selekce získaných modelů podle vlastností a správnosti řešení. Dle výsledku je možné zvážit implementaci celého systému.
- **Deployment Phase:** poslední krok. Zde proces nekončí, ale začíná se cyklicky opakovat. Je nezbytné udržovat modely a zdroje dat aktuální.

Datamining se dá také rozdělit do dvou skupin:

- **Predikce:** předpovídá budoucí vývoj podle získaných informací, předpověď počasí, ceny na burze, atd.
- **Deskripce:** popsání dané skutečnosti.

5.2 Elektronický obchod

Datamining je nejčastěji používán pro segmentaci zákazníků v marketingu, pro reklamní kampaně a sledování rizika obchodu. Analýzu lze provádět z účtenek kamenných obchodů nebo i z objednávek e-shopů. U e-shopu je výhoda, že lze také sledovat zákazníky prohlížené produkty. Zde se ukazují a porovnávají pravděpodobnosti nákupů jednoho zboží s jinými. A tyto výsledky se pak v kamenných obchodech používají třeba pro lepší umístění výrobku na viditelnější místo nebo na e-shopu jako doporučení nejprodávanějších věcí. [20]

6 BIG DATA ANALYTICS

Big data analýza je proces, který zkoumá rozsáhlé a rozmanité datové sady pro odhalení skrytých vzorců, neznámých korelací, trendů na trhu, preferencí zákazníka a dalších užitečných informací, které mohou pomoci organizacím.

V Big Datech se nejčastěji uplatňují 3 přístupy:

- **Důraz na korelace**
- **Integrace dat z různých zdrojů**
- **N = vše** - probíhá analýza celé datové sady

Analytické metody můžeme dělit podle analyzačního předmětu nebo hloubky analýzy.

6.1 Dělení podle předmětu analýzy

V této kapitole bude představeno nejzákladnější dělení podle předmětu analýzy.

6.1.1 Analýza strukturovaných dat

Základem této analýzy je zpracování podle jedno z typů, tím je OLAP (Online analytics processing). Tento typ operací efektivně využívá struktury datového skladu typu hvězda nebo sněhová vločka a umožňuje několik druhů operací s využitím jednotlivých dimenzí. Data se pak dále promítnou do multidimenzionální kostky. Podle jednotlivých dimenzí takové kostky pak lze realizovat zejména tyto operace: [1]

- **slice** – filtrování podle jedné zvolené hodnoty vybrané dimenze a tedy snížení kostky o jedna
- **dice** – výběr pouze několika hodnot z vybraných dimenzí
- **roll-up** – vybrané seskupení hodnot
- **dril-down** – inverzní operace k roll-up

6.1.2 Analýza textu

Analýza textu neboli Text mining je hledání informace v mnoha textových dokumentech a různých formulářích. Získanou informaci lze vyjádřit jasněji a rázněji, protože cílové sdělení je podstatně menší. Uplatňuje se zde především: [1]

- Extrakce informací – rozpoznání, řešení, detekce.
- Sumarizace – zde můžeme extrahovat klíčové informace o textu nebo vytvářet souhrnné parafrázování originálního textu.
- Analýzy sentimentu – analyzování textu lidských názorů (marketing).
- QA systém – systém vyhledávání informací, systém odpovídání dle znalostí a systém hybridní.

6.1.3 Analýza webu

Analýza webu nebo také Web mining je proces dolování dat a algoritmů k extrahování informací přímo z webu. Cílem analýzy webu je hledání vzorů, údajů a jejich shromažďování a celková analýza těchto dat pro získání přehledu o trendech, průmyslu a uživatelích obecně. Obsah získaných dat se nejčastěji skládá z textů a strukturovaných údajů (seznamy, tabulky, obrázky, videa, zvukové stopy). Web mining lze rozdělit do 3 kategorií: [21]

- **Web content mining** – dolování obsahu, proces získávání užitečných informací z webu.
- **Web structure mining** – dolování struktury, analyzování uzlů a struktury připojení webové stránky pro odhalení propojení webové sítě.
- **Web usage mining** – dolování využití, informace o záznamu serveru pro získání přehledu o aktivitě uživatelů, odkud jsou, kolik lidí kde kliklo atd.

6.2 Dělení podle hloubky analýzy

Zde jsou rozebrány tři typy analytických metod: deskriptivní, prediktivní a preskriptivní. Tyto tři typy analytických metod by měly koexistovat, ani jedna není lepší než ta druhá, každá je odlišná od té předešlé, ale všechny jsou nezbytně nutné pro získání úplného přehledu o organizaci a navzájem na sebe navazují. [22]

6.2.1 Deskriptivní analýza

Deskriptivní analýza pomáhá organizacím pochopit, co se stalo v minulosti. Jedná se o souvislost uplynulé minuty nebo také několika let zpět. Také pomáhá pochopit vztahy mezi zákazníky a produkty. Cílem je pochopit, jaký přístup je potřeba přijmout v budoucnosti. Tedy poučit se z minulosti a ovlivnit budoucnost. Běžnými příklady deskriptivní analytiky jsou reporty o řízení organizace (informace o prodeji, zákaznících, operacích, financích, atd.). [22]

6.2.2 Prediktivní analýza

Prediktivní analýza poskytuje jakýsi návod, který se provádí na základě údajů. Jedná se o odhad pravděpodobnosti budoucího výsledku. Jde o strojové učení, dolování dat, modelování a spoustu teorií. Díky této analýze lze identifikovat případná rizika nebo příležitosti v budoucnu. Jako příklad prediktivní analýzy je možno uvést předpověď chování zákazníků v oblasti prodeje a marketingu. Pro prediktivní analýzu je také dobré mít co nejvíce dat, protože zde platí: čím více dat, tím lepší předpovědi. [23]

6.2.3 Preskriptivní analýza

Preskriptivní analýza je závislá právě na výsledcích deskriptivní a prediktivní analýzy. Je konečnou fází porozumění podnikání. Neočekává pouze to, co se stane a kdy se to stane, ale také důvody proč se to stane a poskytne doporučení, jak dále pokračovat za pomocí předpovědí. Tuto analýzu však využívají pouze 3% firem a i tak je stále dost chybová. Příkladem je samoobslužný vůz společnosti Google, který přijímá rozhodnutí na základě různých předpovědí a budoucích výsledků. [22]

II. PRAKTICKÁ ČÁST

7 ÚVOD DO PRAKTICKÉ ČÁSTI

Praktická část je rozdělena do dvou bodů:

- V prvním bodě je provedena analýza dotazníku, která pojednává o znalostech firem ohledně Big Dat. Jednotlivé otázky jsou rozebrány a u každé je pomocí grafů znázorněno výsledné vyhodnocení. Pomocí těchto výsledných grafů je vyvozen závěr.
- V druhé části je uveden vzorový příklad analýzy z pohledu zákazníka, pro kterého poskytovatele platformy pro práci s Big Daty by se mohl rozhodnout na základě multikriteriální TOPSIS analýzy. Zde je provedeno celkové sestavení multikriteriální TOPSIS analýzy i s výsledným vyhodnocením.

8 DOTAZNÍKOVÉ ŠETŘENÍ

V rámci vypracování bakalářské práce byl vytvořen dotazník, který byl zaslán do firem s různým zaměřením (IT, zemědělství, zábava a média, telekomunikace, výzkum, finanční služby, doprava, atd.). V úvodu dotazníku je krátké představení o čem dotazník pojednává a k čemu budou sloužit jeho výsledky.

Cílem dotazníku je zjistit, zda se firmy už potkaly s Big Daty, jaké shromažďují údaje neboli data a zda mají správné strategie nebo software pro jejich zpracování. Dotazník byl vytvořen přes Google Docs, což je sada kancelářských aplikací poskytovaných online od společnosti Google. Vše bylo zpracováno tak, aby byly otázky a odpovědi srozumitelné, a aby vyplnění dotazníku trvalo max. 2 minuty. Pro cílové respondenty byl zasílán přes e-mailovou adresu. Celý dotazník byl zpracován graficky a dále jsou rozebrány jednotlivé otázky.

8.1 Zpracování dotazníku

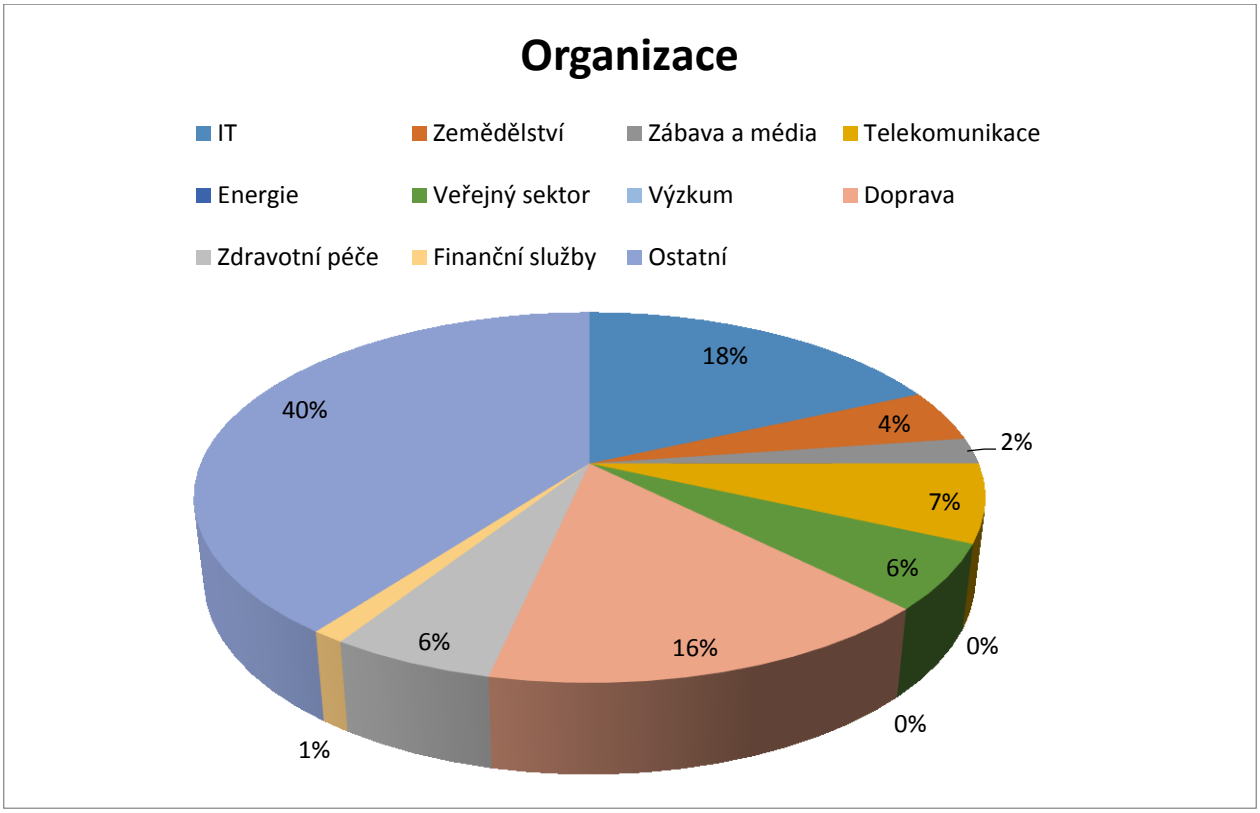
Zde budou představeny a rozebrány jednotlivé otázky z dotazníku a poskytnuty grafické výsledky, podle odpovědí respondentů.

- **Setkali jste se někdy s výrazem Big Data?**

Graf 1 První otázka dotazníku: Setkali jste se někdy s výrazem Big Data?

Jako první položená otázka v dotazníku byla, zda se už respondenti setkali s pojmem Big data. Z Grafu č. 1 nám vyplývá, že pouze 41% respondentů ví, co je to pojem Big Data nebo o něm alespoň někdy slyšeli a zbylých 59% se s tímto pojmem nikdy nesetkali. Zde bylo očekáváno daleko větší procento těch, kteří pojem Big Data znají.

- **Zaměření organizace**



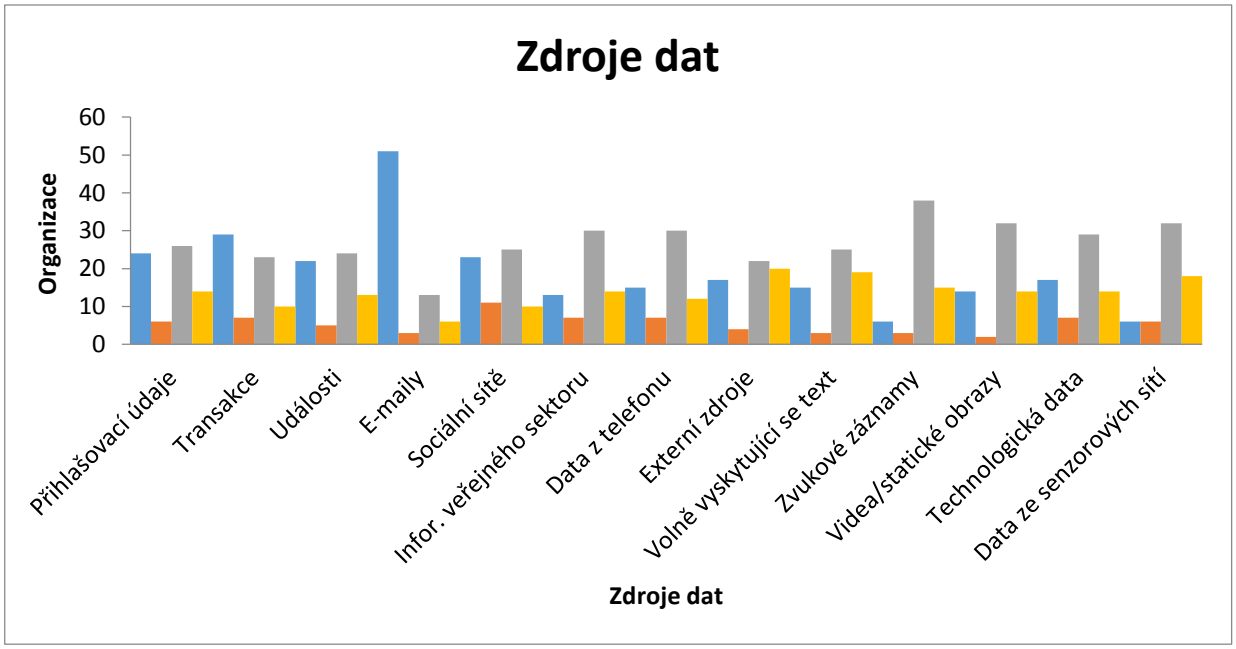
Graf 2 Druhá otázka dotazníku: Zaměření organizace

Další otázka měla za cíl zjistit, v jakém oboru firmy podnikají neboli výskyt jejich působnosti. Z tohoto výsledného Grafu a Tabulky č. 2 můžeme vyčíst, že se do vypracování dotazníku dostalo nejvíce firem zaměřených na IT (18%), na druhém místě se s 16% zapojily firmy zaměřené na dopravu. A ostatní zaměření firem vychází téměř srovnatelně. Ale mezi 40% dotazovaných, kteří odpověděli, že spadají pod jiné zaměření, než byly výše vypsané, se objevovaly firmy zaměřené na reklamy, ubytování, cestovní ruch, reality, služby na cestovní nebo turistický ruch a také pojišťovny.

Tabulka 2 Zaměření organizace

Zaměření organizace	%
Ostatní	40
IT	18
Doprava	16
Telekomunikace	7
Zdravotní péče	6
Veřejný sektor	6
Zemědělství	4
Zábava a média	2
Finanční služby	1
Energie	0
Výzkum	0

- **Zdroje dat**



Graf 3 Třetí otázka dotazníku: Zdroje dat

Z další otázky, z jakých zdrojů organizace získávají nebo očekávají shromažďování dat, vyšlo, že nejvíce dat se získává z e-mailů, viz Graf 3. E-mailové účty nebo data z e-mailů získává více než 62% dotazovaných firem. Mezi další nejvíce získávané data patří informace o transakcích (35%), sociální sítě (30%) a přihlašovací údaje (29%). Ze sociálních sítí plánuje do budoucna shromažďovat data dalších 13 % firem. Mezi nejméně

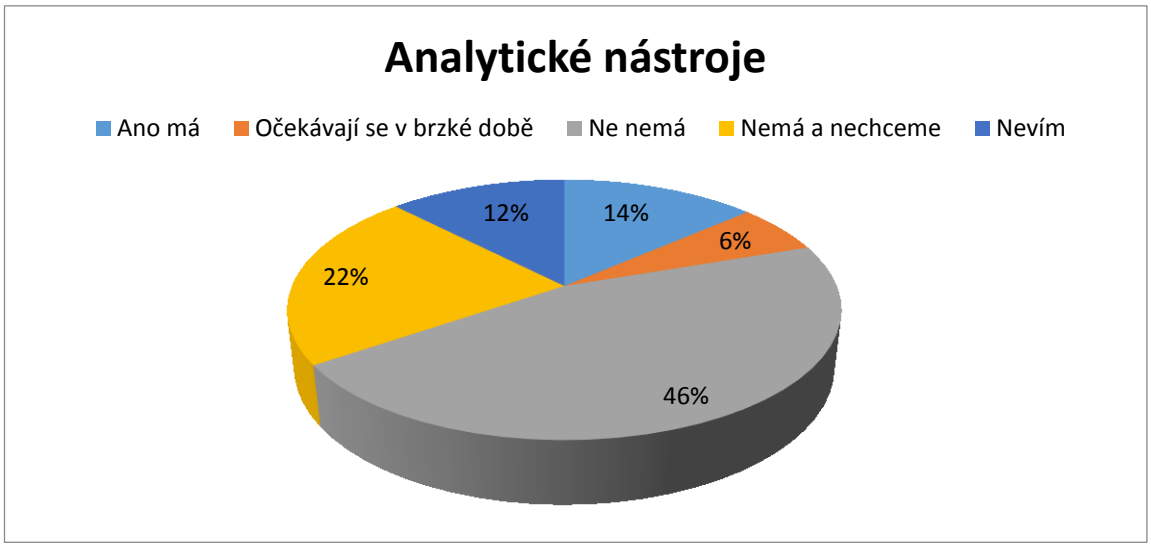
shromažďované nebo odmítané údaje patřila data ze zvukových záznamů, videí, statických obrazů nebo také data z telefonů a informací z veřejného sektoru.

Strategie

Graf 4 Čtvrtá otázka dotazníku: Strategie

Největším překvapením bylo, když 76% firem uvedlo, že nemají strategii k získávání nebo ke zpracování Big Dat viz Graf 4. Oproti tomu se našlo pouhých 15%, které tuto strategii ve firmě zavedlo. A 9% firem nevědělo, zda se ve firmě vůbec nějaká taková strategie vyskytuje.

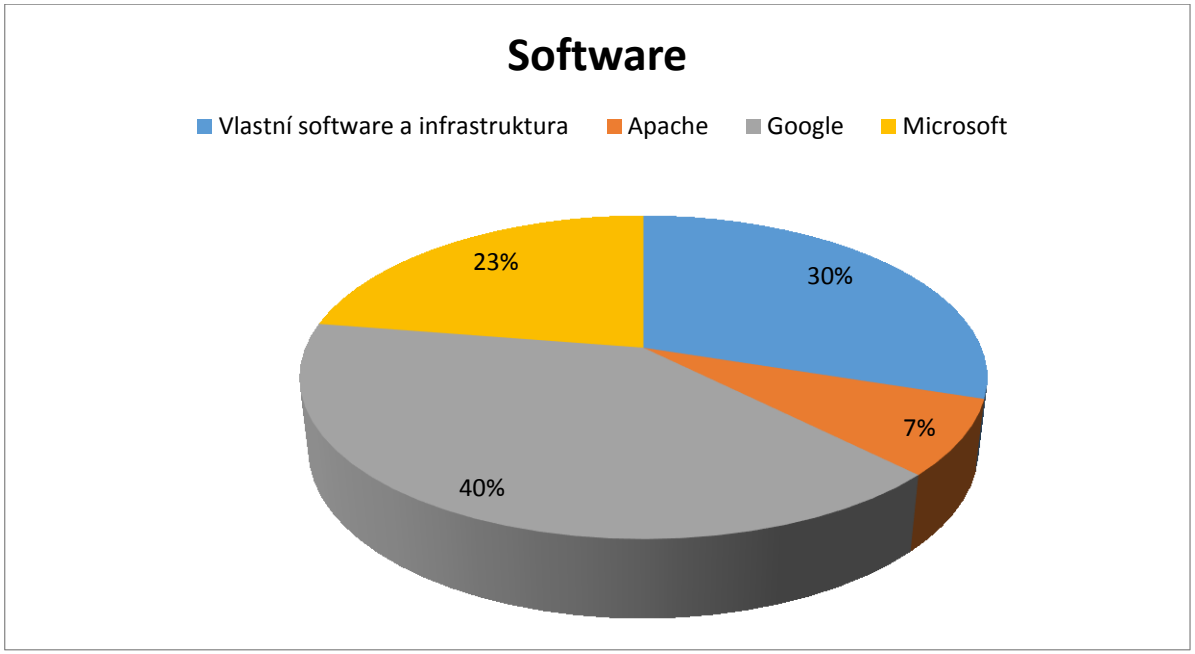
- **Analytické nástroje**



Graf 5 Pátá otázka dotazníku: Analytické nástroje

U otázky zda mají organizace správné analytické nástroje pro zpracování, odpovědělo 68,8% dotazovaných firem, že tyto analytické nástroje nemají, z toho 22,5% je ani neočekává nebo spíše nechce. Jak můžeme vidět v Grafu č. 5 pouhých 18,8% má správné analytické nástroje nebo je v brzké době očekává.

- **Software**



Graf 6 Šestá otázka dotazníku: Software

V poslední otázce, zda firmy používají nějaký software od uvedených poskytovatelů, nám z Grafu č. 6 vyšlo, že firmy nejvíce využívají softwary od společnosti Google a to v počtu 29%. Dále 21% firem uvedlo, že mají svůj vlastní software. Software od společnosti Microsoft používá 16% a 5% od společnosti Apache. Zbylých 29% uvedlo, že nemají žádný software pro zpracování dat.

8.2 Vyhodnocení dotazníku

Z dotazníku vyplývá, že se Big Data ještě nedostala úplně do všech firem nebo se o nich dostatečně neví. Ovšem firmy, které se snaží zpracovávat Big Data, někdy nemají strategie pro zpracování dat a ani ty správné analytické nástroje. Nedokáží tak využít pravý přínos těchto Big Dat. Ale mezi nejčastěji získávané informace patří data z přihlašovacích údajů, transakcí, událostí, e-mailů a sociálních sítí, což se dalo očekávat a obecně se ví, že tyto zdroje generují opravdu velké množství dat. Jedním z překvapení bylo, že 21% firem mělo vlastní software i přes to, že v dnešní době se jich nabízí velké množství. Nejde ani přesně určit, které odvětví firem s Big Daty pracuje nejvíce, je to individuální a záleží na firmách nebo společnostech, zda budou chtít do budoucna udržet krok s ostatními.

9 PŘÍKLAD VÝBĚRU POSKYTOVATELE POMOCÍ METODY TOPSIS

Multikriteriální metoda TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) je založena na výběru varianty, která je nejblíže ideální variantě, která je charakterizovaná vektorem nejlepších kriteriálních hodnot. A zároveň nejdále od bazální varianty, která je reprezentována vektorem nejhorších kriteriálních hodnot. Předpokládá se maximalizační charakter všech kritérií. Pokud nejsou všechna kritéria maximalizační, je nutné je na maximalizační převést.

9.1 Poskytovatelé

Zde budou krátce představeni poskytovatelé platforem pro zpracování Big Dat a v dalším kroku budou analyzováni multikriteriální TOPSIS metodou.

9.1.1 Microsoft Azure

Obrázek 6 logo Microsoft Azure, zdroj: [13]

Microsoft Azure je stále se rozšiřující kolekce integrovaných cloudových služeb, která je využívaná vývojáři a odborníky na IT. Poskytuje možnosti sestavovat, nasazovat a využívat nástroje, aplikace a architektury dle vlastního výběru. Nabízí i cloudové uložiště a ochranu osobních údajů. Microsoft Azure podporuje širokou škálu operačních systémů, programovacích jazyků, databází a zařízení. Poskytuje také velké množství ostatních služeb, jako je vytváření aplikací, trvalé a široce škálovatelné cloudové uložiště s vysokou dostupností, správu relačních databází SQL, zabezpečení, vývojářské nástroje a monitorování. Nabízí také 30 dní zkušebního provozu, kde přidělí do začátku 200 dolarů. Jinak uvádí, že se vždy platí pouze za to, co používáte. [13]

9.1.2 Google Cloud Platform

Obrázek 7 logo Google Cloud Platform, zdroj: [24]

Služba GCP osvobozuje od nákladů na správu za infrastruktury, poskytování serverů a konfiguraci sítí. Ohledně Big Dat nabízí ukládání a analyzování Big Dat v rámci jediné platformy GCP. Nabízí také další služby, stejné jako Microsoft Azure, ale ne tak obsáhlé. Je zde možnost vyzkoušení free verze, kdy po skončení nedojde ke smazání účtu, ale k nabídce na placený účet. Připisují 300 dolarů kreditu zdarma při registraci, které jsou k dispozici po dobu dvanácti měsíců. Přistup ke všem Cloud platformám je možný z různých zařízení.

9.1.3 Amazon Web Servis

Obrázek 8 logo AWS, zdroj: [19]

AWS nabízí velkou škálu služeb, které pomáhají rychle a snadno implementovat aplikace pro analýzu Big Dat. AWS poskytuje rychlý přístup k flexibilním a levným IT zdrojům, takže se může prakticky rychle měnit jakákoliv velká datová aplikace, datové skladování, bezpečnostní server atd. Má v nabídce přesně ty typy zdrojů, které jsou potřebné pro analytické aplikace. Služba Amazon Web Services nabízí širokou škálu globálních cloudových produktů včetně výpočetních, úložných, databázových, analytických, síťových, mobilních, vývojářských nástrojů, nástrojů pro správu, IO, bezpečnostních a podnikových aplikací. I zde najdeme free přístup po dobu jednoho měsíce s mnoha výhodami.

9.1.4 IBM SPSS (Statistical Package for the Social Sciences)

Obrázek 9 logo IBM SPSS, zdroj: [25]

IBM SPSS získává z dat hlubší smysluplnější poznatky a předpovědi o tom, co se pravděpodobně stane. Nabízí pokročilé metody, pomáhá nalézt nové příležitosti a zvyšuje efektivitu. Pod celý proces spadá plánování, shromažďování dat, analýza, tvorba sestav a implementace. Taktéž nabízí třiceti denní zkušební verzi.

9.2 Postup multikriteriální metody TOPSIS

- Prvním krokem je výběr 4 softwarů pro zpracování Big Dat:
  - Microsoft azure
  - GCP
  - AWS
  - IBM SPSS
- Dalším krokem je zvolení si základních kritérií a přiřazení váhy kritérií, které odpovídají tomu, jak je dané kritérium důležité. Celkový součet vah se musí rovnat 1. Kritéria a jim přiřazené váhy jsou uvedeny v následující tabulce.

Tabulka 3 Kritéria a Váhy

Kritéria	Váhy
Zkušební verze	0.1
Přehled/dojem	0.2
Cloud uložiště	0.2
Cena	0.3
Služby	0.2

- Poté je vytvořena tabulka, kde budou 4 zmíněné softwary a u nich 4 kritéria. Přidáme ke každému softwaru a kritériu hodnocení od 1 do 10, kdy 1 je nejhorší a 10 nejlepší hodnocení, viz tabulka 4.

Tabulka 4 Hodnoty

KRITÉRIUM	Microsoft Azure	GCP	AWS	IBM SPSS
Zkušební verze	8	8	9	7
Přehled/dojem	9	7	8	8
Cloud uložiště	8	9	8	6
Cena	5	7	7	9
Služby	9	8	10	5

- Dále je vytvořena normalizovaná kriteriální matice.

$$r_{ij} = \frac{y_{ij}}{\sqrt{\sum_{i=1}^{m} y_{ij}^2}}$$; R=r_{ij}, kde pro i=1,2,…, m; j=1,2,…, n;

- Poté jsou umocněny hodnoty v řádcích, sečteny hodnoty a tento součet odmocněn. Vznikne tedy 5 pomocných výpočtů:
  - Zkušební verze: $\sqrt{8^2 + 8^2 + 9^2 + 7^2} = 16{,}06$
  - Přehled/dojem: $\sqrt{9^2 + 7^2 + 8^2 + 8^2} = 16{,}06$
  - Cloud uložiště: $\sqrt{8^2 + 9^2 + 8^2 + 6^2} = 15{,}65$
  - Cena: $\sqrt{5^2 + 7^2 + 7^2 + 9^2} = 14{,}28$
  - Služby: $\sqrt{9^2 + 8^2 + 10^2 + 5^2} = 16{,}43$
- Čísla, která vyšla, se vydělí každým číslem v řádku (příklad: 8/16,06; 8/16,06; 9/16,06; 7/16,06). Tímto vyjdou čísla mezi 0 a 1, tyto výsledky určí tzv. přitažlivost, která se zaokrouhlí na 2 desetinná místa, výsledky viz tabulka 5.

Tabulka 5 Přitažlivost

KRITÉRIUM	Microsoft Azure	GCP	AWS	IBM SPSS
Zkušební verze	0,50	0,50	0,56	0,44
Přehled/dojem	0,56	0,44	0,50	0,50
Cloud uložiště	0,51	0,58	0,51	0,38
Cena	0,35	0,49	0,49	0,63
Služby	0,55	0,49	0,61	0,30

- Konkrétní hodnoty kritérií v matici jsou vynásobeny váhou kritéria $z_{ij} = wjg_{ij}$, viz tabulka 6 a 7.

Tabulka 6 Přitažlivost a váhy

KRITÉRIUM	Microsoft Azure	GCP	AWS	IBM SPSS	Váhy
Zkušební verze	0,50	0,50	0,56	0,44	0.1
Přehled/dojem	0,56	0,44	0,50	0,50	0.2
Cloud uložiště	0,51	0,58	0,51	0,38	0.2
Cena	0,35	0,49	0,49	0,63	0.3
Služby	0,55	0,49	0,61	0,30	0.2

Tabulka 7 Váhy kritérií

KRITÉRIUM	Microsoft Azure	GCP	AWS	IBM SPSS
Zkušební verze	0,050	0,050	0,056	0,044
Přehled/dojem	0,112	0,088	0,100	0,100
Cloud uložiště	0,102	0,116	0,102	0,076
Cena	0,105	0,147	0,147	0,189
Služby	0,110	0,098	0,122	0,060

- Nyní se vytvoří ideální (h_1, h_2,…) a bazální (d_1,d_2,…) varianty.
  - h_j= max zij
  - d_j= min zij
- Vypočítá se vzdálenost od ideální varianty (IH), viz rovnice 1.

$$d_i^+ = \sqrt{\sum_{j=1}^{n}(z_{ij} - h_j)^2; i = 1,2,..m} \quad (1)$$

- Dále se vypočítá vzdálenost bazální varianty (BH), viz rovnice 2

$$d_i^- = \sqrt{\sum_{j=1}^{n}(z_{ij} - d_j)^2; i = 1,2,..m} \quad (2)$$

- V následující tabulce jsou ideální hodnoty kritérií, jedná se o maximální hodnoty kritérií v řádku. Tyto hodnoty je vhodné zvýraznit.

Tabulka 8 Maxima

KRITÉRIUM	Microsoft Azure	GCP	AWS	IBM SPSS
Zkušební verze	0,050	0,050	0,056	0,044
Přehled/dojem	0,112	0,088	0,100	0,100
Cloud uložiště	0,102	0,116	0,102	0,076
Cena	0,105	0,147	0,147	0,189
Služby	0,110	0,098	0,122	0,060

- V každém řádku se odečtou hodnoty maxima od ostatních hodnot v řádku a jednotlivé výsledky se umocní na druhou. Poté se sečtou všechny hodnoty ve sloupcích a výsledky se odmocní, vyjde tato tabulka.

Tabulka 9 Ideální hodnota

KRITÉRIUM	Microsoft Azure	GCP	AWS	IBM SPSS
Zkušební verze	0,000036	0,000036	0	0,000144
Přehled/dojem	0	0,000576	0,000144	0,000144
Cloud uložiště	0,000196	0	0,000196	0,0016
Cena	0,007056	0,001764	0,001764	0
Služby	0,000144	0,000576	0	0,003844

Ideální hodnoty	0,0862090	0,0543323	0,0458694	0,0757100

- Bazální hodnoty kritérií (minimální hodnoty kritérií v řádku) budou zvýrazněny a zapsány do Tabulky 10.

Tabulka 10 Minima

KRITÉRIUM	Microsoft Azure	GCP	AWS	IBM SPSS
Zkušební verze	0,050	0,050	0,056	0,044
Přehled/dojem	0,112	0,088	0,100	0,100
Cloud uložiště	0,102	0,116	0,102	0,076
Cena	0,105	0,147	0,147	0,189
Služby	0,110	0,098	0,122	0,060

- V každém řádku se odečtou hodnoty minima od ostatních hodnot v řádku a jednotlivé výsledky se umocní na druhou. Poté jsou sečteny všechny hodnoty ve sloupcích a výsledky se odmocní, viz Tabulka 11.

Tabulka 11 Bazální hodnoty

KRITÉRIUM	Microsoft Azure	GCP	AWS	IBM SPSS
Zkušební verze	0,000036	0,000036	0,000144	0
Přehled/dojem	0,000576	0	0,000144	0,000144
Cloud uložiště	0,000676	0,0016	0,000676	0
Cena	0	0,001764	0,001764	0,007056
Služby	0,0025	0,001444	0,003844	0
Bazální hodnoty	0,0615467	0,0695989	0,0810679	0,0848528

- Dále bude užit finální vzorec pro výpočet relativního ukazatele vzdálenosti od bazální varianty, viz rovnice 3.

$$UV = \frac{BH}{(IH+BH)} \tag{3}$$

- Tyto hodnoty byly vypočítány v předchozích tabulkách, IH - ideální hodnota, BH - bazální hodnota, UV - ukazatel vzdálenosti.

Tabulka 12 Výsledek

KRITÉRIUM	Microsoft Azure	GCP	AWS	IBM SPSS
Ideální hodnoty-IH	0,0862090	0,0543323	0,0458694	0,0757100
Bazální hodnoty-BH	0,0615467	0,0695989	0,0810679	0,0848528
Ukazatel vzdálenosti-UV	0,416544	0,561593	**0,638645**	0,528471

9.3 Výsledek

Díky sestavení multikriteriální TOPSIS metody se nejvíce blíží ideálu software pro zpracování Big Dat od společnosti Amazon web servis. Na druhém místě je software od společnosti Google Cloud Platform, na třetím místě je software od IBM SPSS a na posledním místě je software od Microsoft Azure. Celkový výsledek můžeme vidět v Tabulce 12.

ZÁVĚR

Teoretická část v této bakalářské práci je pomyslná sonda do unikátní oblasti Big Dat a jejich rozšíření ve firemním sektoru v ČR. V práci byly poskytnuty základní a klíčové informace o této problematice. Začátek práce se týká vzniku Big Dat, základními pojmy, definicemi a jejich vlastnostmi. Také jsou zde zmíněné příklady využití Big Dat v minulosti, které jsou považovány za průkopníky těchto dat. Dále je zde ukázáno, jak zpracovávat tato data nebo pomocí jakých programů s nimi nakládat. K vypracování této části byly použity vědecké články a odborná literatura.

V praktické části je jako první zpracován dotazník. Z výsledků dotazníku vyšlo, že více jak polovina dotazovaných firem se nikdy nesetkala s pojmem Big Data, z toho vyplývá, že tato oblast není stále dostatečně rozšířená a ani se o ní pořádně neví. Ovšem firmy, které se věnují analýzám Big Dat, nejvíce sbírají data z e-mailů, transakcí, sociálních sítí a přihlašovacích údajů. I přes to, že necelá polovina dotazovaných firem pracuje s Big Daty, tak 76% uvádí, že nemají strategie týkající se zpracování těchto dat. Zbylých 46% odpovědělo, že nemají ani správné analytické nástroje. Opět se zde naráží na problém, že i když se firmy snaží pracovat s Big Daty, nedokáží je plně využít ve svůj prospěch. Posledním krokem v praktické části byl výběr vhodného poskytovatele platformy pro zpracování Big Dat pomocí multikriteriální TOPSIS analýzy. Zde podle zadaných kritérií se nejlépe umístil software od poskytovatele Amazon web servis, který nabízí řadu dalších sad a rozšíření pro zpracování Big Dat.

Jedním z přínosů této práce je poukázat na současný stav Big Dat ve firemní sféře. Firmy by se měly naučit pracovat s těmito daty. Dá se říci, že je to pomyslný vlak, který se žene stále vpřed a přináší sebou spoustu výhod a zisků.

SEZNAM POUŽITÉ LITERATURY

[1] HOLUBOVÁ, Irena, Jiří KOSEK, Karel MINAŘÍK a David NOVÁK. Big Data a NoSQL databáze. První vydání. Praha: Grada, 2015, 281 stran. ISBN 978-80-247-5466-6.

[2] Big Data: 5 v's of big data. In: *Pinterest* [online]. [cit. 2017-05-25]. Dostupné z: https://cz.pinterest.com/pin/550776229409314191/

[3] DAVE, Pinal. Big Data – What is Big Data – 3 Vs of Big Data: Volume, Velocity and Variety. In: SQLAuthority [online]. 2013 [cit. 2017-05-25]. Dostupné z: https://blog.sqlauthority.com/2013/10/02/big-data-what-is-big-data-3-vs-of-big-data-volume-velocity-and-variety-day-2-of-21/

[4] GEWIRTZ, David. Big Data Analytics: Volume, velocity, and variety: Understanding the three V's of big data. In: ZDNet [online]. 2016 [cit. 2017-05-25]. Dostupné z: http://www.zdnet.com/article/volume-velocity-and-variety-understanding-the-three-vs-of-big-data/

[5] VAN RIJMENAM, Mark. Why The 3V's Are Not Sufficient To Describe Big Data. In: Datafloq [online]. 2007 [cit. 2017-05-25]. Dostupné z: https://datafloq.com/read/3vs-sufficient-describe-big-data/166

[6] The Top 20 Valuable Facebook Statistics. In: Zephoria: Digital Marketing [online]. 2017 [cit. 2017-05-25]. Dostupné z: https://zephoria.com/top-15-valuable-facebook-statistics/

[7] MAYER-SCHÖNBERGER, Viktor a Kenneth CUKIER. Big Data: revoluce, která změní způsob, jak žijeme, pracujeme a myslíme. 1. vyd. Brno: Computer Press, 2014, 256 s. ISBN 978-80-251-4119-9.

[8] WALSH, Bryan. BIG DATA: Google's Flu Project Shows the Failings of Big Data. In: Time [online]. 2013 [cit. 2017-05-25]. Dostupné z: http://time.com/23782/google-flu-trends-big-data-problems/

[9] CAN, Andrew. MONETISING BIG DATA IN TELECOMS. In: Neural Technologies [online]. 2016 [cit. 2017-05-25]. Dostupné z: https://www.neuralt.com/34/545/monetising-big-data-in-telecoms

[10] DOSTÁL, Dalibor. Big data slibují zisky, firmy s nimi však neumějí zacházet. In: BusinessInfo: Oficiální portál pro podnikání a export [online]. 2015 [cit. 2017-05-25]. Dostupné z: https://www.businessinfo.cz/cs/clanky/big-data-slibuji-zisky-firmy-s-nimi-vsak-neumeji-zachazet--63177.html

[11] DOLÁK, Ondřej. Big Data: Nové způsoby zpracování a analýzy velkých objemů dat. In: SystemOnline: S přehledem ve světě informačních technologií [online]. [cit. 2017-05-25]. Dostupné z: https://www.systemonline.cz/clanky/big-data.htm

[12] AUGUSTÍN, Jakub. BIG DATA A MOŽNOSTI JEJICH VYUŽITÍ. In: ADASTRA [online]. 2014 [cit. 2017-05-25]. Dostupné z: http://www.adastra.cz/clanky/big-data-a-moznosti-jejich-vyuziti.

[13] Hadoop. In: Microsoft Azure [online]. 2017 [cit. 2017-05-25]. Dostupné z: https://azure.microsoft.com/cs-cz/solutions/hadoop/]

[14] How To Setup Apache Hadoop On CentOS. In: UnixMan: Linux/Unix news and reviews [online]. 2015 [cit. 2017-05-25]. Dostupné z: https://www.unixmen.com/setup-apache-hadoop-centos/

[15] DOLÁK, Ondřej. Když se řekne „Hadoop“. In: Linuxexpres [online]. 2013 [cit. 2017-05-25]. Dostupné z: https://www.linuxexpres.cz/software/kdyz-se-rekne-hadoop

[16] SULLIVAN, Dan. Getting Started with Hadoop 2.0. In: Tom´s IT PRO: REAL-WORLD BUSINESS TECHNOLOGY [online]. 2014 [cit. 2017-05-25]. Dostupné z: http://www.tomsitpro.com/articles/hadoop-2-vs-1,2-718.html

[17] Ghemawat, Sanjay, Gobioff, Howard a Leung, Shun-Tak. The Google File System. [Online] 2003.[Citace: 25. 5. 2017.] http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/c s//archive/gfs-sosp2003.pdf

[18] HDInsight: Cloudová služba Sparku a Hadoopu pro podniky. In: Microsoft Azure [online]. 2017 [cit. 2017-05-25]. Dostupné z: https://azure.microsoft.com/cs-cz/services/hdinsight/

[19] Start Building on AWS Today. In: Amazon Web Services [online]. 2017 [cit. 2017-05-25]. Dostupné z: https://aws.amazon.com/?nc2=h_lg

[20] PROCHÁZKA, Michal. Data mining: jiný pohled na problém. In: VTM [online]. 2017 [cit. 2017-05-25]. Dostupné z: http://vtm.e15.cz/aktuality/data-mining-jiny-pohled-na-problem

[21] Web Mining. In: Technopedia [online]. 2016 [cit. 2017-05-27]. Dostupné z: https://www.techopedia.com/definition/15634/web-mining

[22] VAN RIJMENAM, Mark. Understanding Your Business With Descriptive, Predictive And Prescriptive Analytics. In: *Datafloq* [online]. 2016 [cit. 2017-05-27]. Dostupné z: https://datafloq.com/read/descriptive-predictive-prescriptive-analytics/151

[23] Data science & big data analytics discovering, analyzing, visualizing and presenting data. Indianapolis: Wiley, 2015, xviii, 410 stran. ISBN 978-1-118-87613-8.

[24] Big data solutions. In: *Google Cloud Platform* [online]. 2017 [cit. 2017-05-25]. Dostupné z: https://cloud.google.com/products/big-data/

[25] IBM SPSS. In: *StatWorks: Computational Analytics* [online]. 2017 [cit. 2017-05-25]. Dostupné z: http://www.statwks.com/index.php/analytics/ibm-spss/

SEZNAM POUŽITÝCH SYMBOLŮ A ZKRATEK

AWS	Amazon Web Services
CDC	Centers for Disease Control and Prevention
CRISP-DM	CRoss-Industry Standard Proces for Data Mining
CRM	Customer relationship management
ELT	Extrakce, Transformace, Load
EMR	Elastic MapReduce
ERP	Enterprise Resource Planning
GCP	Google Cloud Platform
GFS	Google File Systém
HDFS	Hadoop Distributed File Systém
IBM	International Business Machines Corporation
IT	Information Technology
SPSS	Statistical Package for the Social Sciences
TB	TeraByte
TOPSIS	Technique for Order Preference by Similarity to Ideal Solution

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

SEZNAM GRAFŮ

SEZNAM PŘÍLOH

PŘÍLOHA P I: OBSAH DISKU CD

:/fulltext.pdf