TECHNICKÁ UNIVERZITA V LIBERCI
Ekonomická fakulta

# Nástroje pro analýzu Big Data se zaměřením na malé a střední podniky

## Diplomová práce

*Studijní program:* N6209 – Systémové inženýrství a informatika
*Studijní obor:* 6209T021 – Manažerská informatika

*Autor práce:* **Bc. Filip Le**
*Vedoucí práce:* Ing. Vladimíra Zádová, Ph.D.

Liberec 2018

Technická univerzita v Liberci
Ekonomická fakulta
Akademický rok: **2016/2017**

# ZADÁNÍ DIPLOMOVÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

| | |
|---|---|
| Jméno a příjmení: | **Bc. Filip Le** |
| Osobní číslo: | **E15000550** |
| Studijní program: | **N6209 Systémové inženýrství a informatika** |
| Studijní obor: | **Manažerská informatika** |
| Název tématu: | **Nástroje pro analýzu Big Data se zaměřením na malé a střední podniky** |
| Zadávající katedra: | **Katedra informatiky** |

Zásady pro vypracování:

1. Big Data - charakteristika, principy, způsoby zpracování
2. Specifika a informační potřeby MSP
3. Nástroje pro analýzu Big Data - charakteristika, porovnání
4. Zpracování dat vybranými nástroji
5. Doporučení pro MSP

Rozsah grafických prací:

Rozsah pracovní zprávy: **65 normostran**

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

Seznam odborné literatury:

**HOLUBOVÁ, Irena, Jiří KOSEK, Karel MINAŘÍK a David NOVÁK. Big Data a NoSQL databáze. Praha: Grada, 2015. Profesionál. ISBN 978-80-247-5466-6.**
**MAYER-SCHÖNBERGER, Viktor a Kenneth CUKIER. Big Data. Brno: Computer Press, 2014. ISBN 978-80-251-4119-9.**
**LOSHIN, David. Big data analytics: from strategic planning to enterprise integration with tools, techniques, NoSQL, and graph. Waltham, Mass.: Academic Press, 2013. ISBN 978-0-12-417319-4.**
**Elektronická databáze článků ProQuest (knihovna.tul.cz).**

Vedoucí diplomové práce: **Ing. Vladimíra Zádová, Ph.D.**
Katedra informatiky

Konzultant diplomové práce: **Bc. Martin Skalický**
The Information Factory, BI Architect

Datum zadání diplomové práce: **31. října 2016**

Termín odevzdání diplomové práce: **31. května 2018**

prof. Ing. Miroslav Žižka, Ph.D.
děkan

Technická univerzita v Liberci
Ekonomická fakulta
L.S.

doc. Ing. Jan Skrbek, Dr.
vedoucí katedry

V Liberci dne 31. října 2016

# Prohlášení

Byl jsem seznámen s tím, že na mou diplomovou práci se plně vztahuje zákon č. 121/2000 Sb., o právu autorském, zejména § 60 – školní dílo.

Beru na vědomí, že Technická univerzita v Liberci (TUL) nezasahuje do mých autorských práv užitím mé diplomové práce pro vnitřní potřebu TUL.

Užiji-li diplomovou práci nebo poskytnu-li licenci k jejímu využití, jsem si vědom povinnosti informovat o této skutečnosti TUL; v tomto případě má TUL právo ode mne požadovat úhradu nákladů, které vynaložila na vytvoření díla, až do jejich skutečné výše.

Diplomovou práci jsem vypracoval samostatně s použitím uvedené literatury a na základě konzultací s vedoucím mé diplomové práce a konzultantem.

Současně čestně prohlašuji, že tištěná verze práce se shoduje s elektronickou verzí, vloženou do IS STAG.

Datum:

Podpis:

## Poděkování

Touto cestou bych rád poděkoval vedoucí mé diplomové práce paní Ing. Vladimíře Zádové, Ph.D. za její ochotu, čas, věcné připomínky a cenné rady, které mi během zpracování práce věnovala.

## Anotace

Diplomová práce je zaměřena na Big Data, především pak na data ze sociálních sítí a možnosti jejich zpracování malými a středními podniky. Cílem práce je představit nástroje pro zpracování Big Data a následně je podle vydefinovaných kritérií porovnat a zvolit ten nejvhodnější pro malé a střední podniky. Dalším cílem je pak provést analýzu dat ze sociálních sítí vybraným nástrojem. V rámci analýzy také představit způsoby jak data ze sociálních sítí získat.

Součástí této práce je také obecná charakteristika Big Data a popsání způsobu zpracování těchto dat. Další důležitou částí je také představení jednotlivých technologií pro zpracování Big Data.

## Klíčová slova

analýza dat, Big Data, Facebook Graph API, sociální sítě, streamovaná data

## Annotation

Tools for Analyzing Big Data with Focus on Small and Medium Businesses

This thesis focuses on Big Data, especially data from social networks and how can small and medium enterprises process them. The main goal of this thesis is to introduce available Big Data tools, make a comparison based on predefined criteria and then to choose the most suitable tool for SME. Another goal is to analyze data from social networks on the chosen platform and as a part of the analysis to showcase ways how to get data from social networks.

This thesis also defines the term Big Data and how are they processed. Then it introduces some technologies that are used to process Big Data.

## Keywords

Big Data, Data analysis, data stream, Facebook Graph API, social network

# Obsah

# Seznam zkratek

| | |
|---|---|
| ACID | Atomicity, Consistency, Isolation, Durability |
| API | Application Programming Interface |
| AWS | Amazon Web Services |
| CSV | Comma-separated Values |
| DBMS | Database Management System |
| GFS | Google File System |
| HDFS | Hadoop Distributed File System |
| HTTP | Hypertext Transfer Protocol |
| JSON | JavaScript Object Notation |
| MSP | Malé a střední podniky |
| RDD | Resilient Distributed Dataset |
| SDK | Software Development Kit |
| SQL | Structured Query Language |
| TUL | Technická univerzita v Liberci |
| ICT | Informační a komunikační technologie |
| PC | Personal Computer (osobní počítač) |
| SPL | Search Processing Langugage |

# Seznam tabulek

# Seznam obrázků

# Úvod

Termín Big Data se v poslední době neustále skloňuje a zmiňuje se o něm čím dál více lidí pohybujících se v oboru informačních technologií. V minulosti se data shromažďovala těžko a byla generována především v korporátních společnostech a to jen v případě, že to příslušný pracovník zadal do systému. Bylo velmi málo systémů, které by generovaly data. Pracovat s velkými daty byla především výsada obrovských korporací, které na to měly jak dostatek financí, tak také příslušné zdroje dat, ze kterých by mohli čerpat. Tato doba již dávno pominula. Nyní generuje obrovské objemy dat každý člověk, ať už nákupem zboží a služeb, či přidáváním různých zážitků, pocitů a celkově informací do sociálních sítí. V dnešní době již tedy podniky nemusí spoléhat na data, která si sami pracně nasbírají, ale mají možnost data dolovat z nespočetně mnoha zdrojů.

Aby podnik mohl data využít, potřebuje nejdříve najít nástroj, který mu pomůže data z daného zdroje extrahovat a také dále transformovat pro další použití. Pak také potřebuje nástroj, kterým tato data zanalyzuje a dodá jim business význam.

Cílem této práce je představit nástroje pro analýzu Big Data, poté je dle vydefinovaných kritérií porovnat a zvolit nejvhodnější pro malé a střední podniky. Dále je cílem ukázat příklad postupu jak příslušná data extrahovat a jak je zanalyzovat. Analyzovaná data budou extrahována především ze sociálních sítí, jelikož se jedná o snadno přístupný zdroj dat, který mohou využít i příslušné podniky.

V práci je také uvedena obecná charakteristika Big Data, popsání způsobu jejich zpracování a představení jednotlivých technologií, pomocí kterých lze Big Data zpracovávat.

# Zhodnocení současného stavu

Big Data se stala fenoménem poslední doby, tudíž můžeme najít spousty odborných článků a několik publikací, které se tímto tématem zabývají. S daty pracuje každé odvětví, což způsobuje jejich rozdílné vnímání. Někdo je může vnímat jako něco naprosto nezbytného pro svůj podnik a naopak někdo jako nezbytně nutnou zátěž.

Publikovaná kniha, která byla přeložena do češtiny a zabývá se obecnou problematikou Big Data, nese název *Big Data: Revoluce, která mění způsob, jak žijeme a myslíme* od autorů V. Mayer-Schönbergera a K. Cukiera. V této publikaci autoři seznamují s termínem Big Data pomocí různých událostí, které byly relevantní k dané problematice. Na daných příkladech je poukázán přístup ke zpracování dat, který již nepracuje pouze se vzorkem, ale s celou množinou dat, díky čemuž lze získat přesnější a přínosnější informace.

Další českou publikací, tentokrát ale i od českých autorů je kniha *Big Data a NoSQL databáze* od autorů I. Holubové, J. Koska, K. Minaříka a D. Nováka. Velké objemy dat a jejich neustálý nárůst přináší řadu problémů se zpracováním, kdy běžné relační databáze již nestačí a vznikají NoSQL systémy, které nabízejí řešení v podobě efektivního uložení dat a dotazování. A právě o problematice NoSQL databází kniha pojednává. V publikaci jsou zmíněny jak různé formáty uložení dat, tak také základní principy, na kterých uložení dat v databázi stojí.

Co se týče zahraničních publikací tak těch je podstatně víc.

Kniha *Big Data Analysis* od D. Loshina popisuje hodnotu Big Data primárně z business pohledu. Autor zde popisuje, kdy by měl podnik začít uvažovat o technologiích, které dokáží zpracovávat velké objemy dat a jaký to bude mít efekt na business uživatele v podniku. V publikaci lze nalézt také popsané jednotlivé techniky a nástroje pro zpracování Big Data.

Další zahraniční publikací je kniha *Big Data Now: 2015* od společnosti O'Reilly Media, Inc. kde je popsána problematika bezpečnosti, dále pak také aplikace Big Data a popsána souvislost s pojmem Internet of things a jaké to má možné problémy.

Dalším klíčovým zdrojem jsou záznamy z konference DATAKON, konkrétně ročník 2014, který byl zaměřen na toto téma. Příspěvek *Big Data: jejich ukládání, zpracování a použití* od J. Pokorného je zaměřen na popis jednotlivých architektur a způsoby ukládání dat. Autor zde popisuje jednotlivé možnosti jak Big Data zpracovávat a jaká je vhodná architektura k ukládání velkého objemu dat. Dále je zde rozebírán problém škálovatelnosti jednotlivých systémů uložení dat. Nachází se tu také část věnovaná speciálně NoSQL databázím.

Ohledně NoSQL databází existuje také spousta článků v databázi ProQuest. Jedním z takových je článek *„NoSQL database technologies“* od autorů Madison, M., Barnhill, M., Napier, C., a Godin, J., který popisuje databázové technologie NoSQL. Autoři zde uvádějí podrobnou charakteristiku těchto technologií a dále uvádějí důvody, proč společnosti NoSQL technologie přebírají.

Akademických prací, které se zabývají problematikou, existuje několik. Například práce od autora M. Miloše s názvem „*Nástroje pro Big Data Analytics*“. Tato práce se zabývá oblastí Big Data a je konkrétně zaměřena na nástroje, které se v této oblasti využívají. V praktické části pak autor analyzuje data ze sociální sítě pomocí nástroje Hortonworks.

Další akademickou prací na toto téma, jejíž autorem je O. Linhart se nazývá „*Využití dat ze sociálních sítí pro BI*“. Tato práce je konkrétně zaměřena na využití dat ze sociálních sítí Twitter a Facebook za účelem získání konkurenční výhody. V práci je popsáno, jak by měl podnik postupovat, pokud chce získat data ze sociálních sítí.

# 1. Big Data

Objem dat pro každý podnik roste exponenciálně, neboť data relevantní pro podniky již nejsou jen z jejich zdrojových systémů. To, kdy se z dat stávají Big Data, není tak úplně přesně definované, ale existuje obecný předpoklad, že pokud na zpracování a ukládání dat nestačí běžné systémy a je třeba použít jiný přístup, aby k datovým transakcím docházelo v rozumném čase, pak se tato data považují za Big Data. Pojem Big Data nesouvisí jen s fyzickou velikostí dat (*volume*) a s tím kolik místa zabírají, nýbrž také s rychlostí nárůstu (*velocity*), různorodostí (*variety*) a dalšími aspekty. O těchto dimenzích se v souvislosti s termínem Big Data hovoří jako o ‚V's". Těchto aspektů neboli „V's" je několik, ale základní jsou tři. Tyto tři dimenze poprvé představil analytik společnosti Gartner, Doug Laney v roce 2001.[1]

Postupem času se začaly přidávat další dimenze, jelikož tři základní dimenze pro vymezení Big Data přestaly stačit. V článku „*How Many "V's" in Big Data? The Characteristics that Define Big Data",* jehož autorem je William Vorhies[2], je popsána autorova spolupráce s americkým oddělením NIST na projektu „Big Data Roadmap". Jedním z hlavních cílů tohoto projektu bylo definovat pojem Big Data. Na základě této analýzy byly specifikovány další dimenze jako je důvěryhodnost dat (*veracity*), hodnota dat (*value*), doba popisující prodlevu mezi realitou a uložením v databázi (**viscosity)** nebo také tempo, jakým se data dokážou šířit (**virality**).[3]

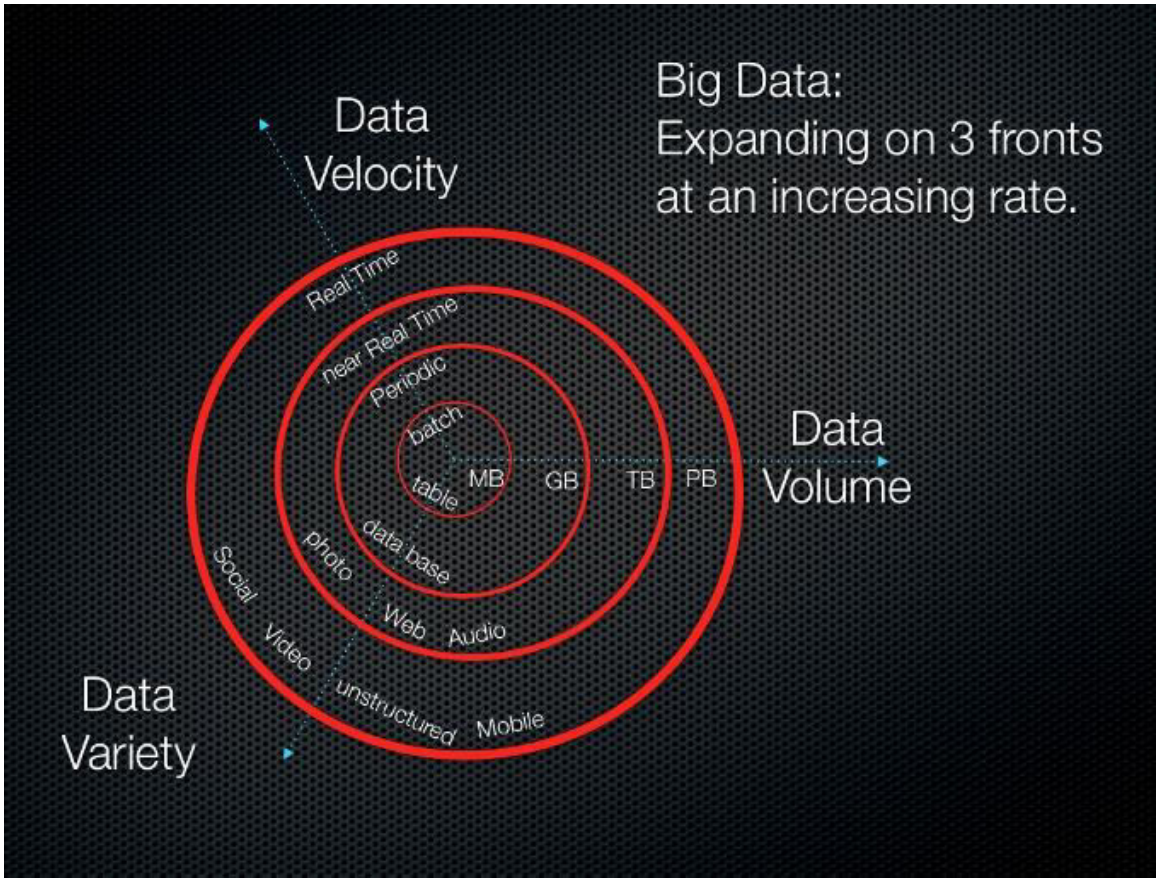
Na následujícím obrázku je znázorněna expanze a rozdělení dat do tří základních dimenzí vymezené Dougem Laney.

---

[1] LANEY, Doug. 3D Data management: Controlling Data Volume, Velocity, and Variety. In: *Application delivery strategies* [online]. 2001 [cit. 2016-12-10]. Dostupné z: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

[2] Prezident a hlavní datový analytik spolčnosti Data-Magnum, která se zabývá konzultacemi a technickými integracemi Big Data.

[3] VORHIES, William. *How Many "V's" in Big Data? The Characteristics that Define Big Data* [online]. In: 2014 [cit. 2016-10-12]. Dostupné z: http://www.datasciencecentral.com/profiles/blogs/how-many-v-s-in-big-data-the-characteristics-that-define-big-data




*Obrázek 1: Expanze a rozdělení dat do jednotlivých dimenzí*
Zdroj: Big Data [online]. In: [cit. 2017-03-29]. Dostupné z: http://itknowledgeexchange.techtarget.com/writing-for business/files/2013/02/BigData.001.jpg

## 1.1 Volume - objem

Zajímavou statistiku[4] ohledně nárůstu dat připravil Ben Walker, Marketing Executive ze společnosti Vouchercloud. Uvádí, že každý den se vyprodukuje přes 2,5 quintilionů[5] bajtů dat. Takovýto objem dat by se vešel zhruba na 100 milionů blue-ray disků (pokud by se tyto disky poskládaly na sebe, jejich výška by byla stejná jako výška čtyř Eiffelových věží poskládaných na sebe). To je neuvěřitelné množství dat, které se samozřejmě musí někde

[4] WALKER, Ben. *EVERY DAY BIG DATA STATISTICS – 2.5 QUINTILLION BYTES OF DATA CREATED DAILY* [online]. In: 2015 [cit. 2016-10-12]. Dostupné z: http://www.vcloudnews.com/every-day-big-data-statistics-2-5-quintillion-bytes-of-data-created-daily/

[5] 1 quintilion = $10^{18}$

ukládat a nějak zpracovávat. Dále statistika uvádí, že 90% těchto dat je nestrukturovaných, protože tato data celkově zahrnují například příspěvky na sociálních sítích, fotografie, ale také například historii nákupů zákazníků, telefonní logy a spoustu dalších dat.

To při jakém objemu se z obyčejných dat stanou Big Data, není definované, nicméně existují příklady firem na světě, o kterých lze s naprostou jistotou říci, že musí s Big Data pracovat. Jednou z takových firem je Facebook. Podle statistiky vydané samotnou společností, bylo na Facebook nahráno více než 250 biliónů fotografií, což je samo o sobě neskutečné množství dat, které musí být uloženo a zpracováno.

Velikost dat je problém, který se samozřejmě dá řešit určitým škálováním databáze, viz podkapitola 1.4.

## 1.2 Velocity – rychlost nárůstu

Rychlost nárůstu dat je v době sociálních sítí obrovská a největší výzva spočívá ve zpracování rychle narůstajících dat v co nejkratším čase, nejlépe pak zpracovávat data real-time. Komplikovanost uložení a transakčního zpracování dat je zřejmá. Ale co tato rychlost nárůstu znamená? Není to tak dávno, co bylo důležité umět z určitého vzorku dat předpovědět, jak se bude chovat celá množina. Nyní s takovouto rychlostí nárůstu dat, je k dispozici skoro celá množina, tudíž se musejí využít jiné metody, jak s tím pracovat.

Tuto podstatu vysvětlují na příkladu V. Mayer-Schönberger a K. Cukier. Uvádějí, příklad od Petera Norviga, experta na umělou inteligenci v Googlu, který říká, že Pablo Picasso viděl ve francouzské jeskyni Lascaux obraz koně, který vznikl v paleolitu (zhruba před 17000 lety), a prohlásil, že jsme od té doby nic nového nevymysleli.[6] Načež autoři uvádějí:

*„Picasso měl sice pravdu, ale jen v jistém smyslu. Vraťme se nyní k té fotografii koně. Nakreslení koně kdysi zabralo hodně času, ale dnes můžeme pomocí fotografie vytvořit reprezentaci koně mnohem rychleji. Je to sice změna, ale nejspíše nikoli klíčová, protože*

[6] MAYER-SCHÖNBERGER, Viktor a Kenneth CUKIER. *Big Data*. Brno: Computer Press, 2014, s. 18. ISBN 978-80-251-4119-9.

*v zásadě máme pořád totéž: obraz koně. Nyní nás však Norvig požádá, abychom myšlenkově zaznamenávali obraz koně a sérii obrazů začali přehrávat rychlostí 24 snímků za sekundu. Kvantitativní změna nyní přešla do stádia kvalitativní změny. Princip filmu a statické fotografie se zásadně liší. S veledaty je to obdobné: když změníme množství, mění se samotná podstata.“* [7]

Z předchozího textu je tedy patrné, že pokud bude docházet k exponenciálnímu nárůstu dat, celková podstata dat se bude měnit.

## 1.3 Variety – různorodost

Různorodost dat se vztahuje k heterogenitě zkoumaného datového setu. K lepšímu pochopení různorodosti dat je níže uvedena základní kategorizace dat.

Obecně dělíme data do tří základních kategorií:

- Strukturovaná data
- Nestrukturovaná
- Semistrukturovaná

**Strukturovaná data**

Tento typ dat se vyznačuje jasně definovanou strukturou. Běžně jsou tato data ukládána do relačních databází. Mají jasně definované schéma a přistupuje se k nim pomocí dotazovacích jazyků. Jako příklad může sloužit databáze studentů. Každý záznam v databázi bude tvořit id_studenta, jméno, adresa atd. Vzhledem k tomu, že mají jasně definovaný model, tak není složité využívat určitých transakcí, transformací a není problém se složitější analýzou.

**Nestrukturovaná**

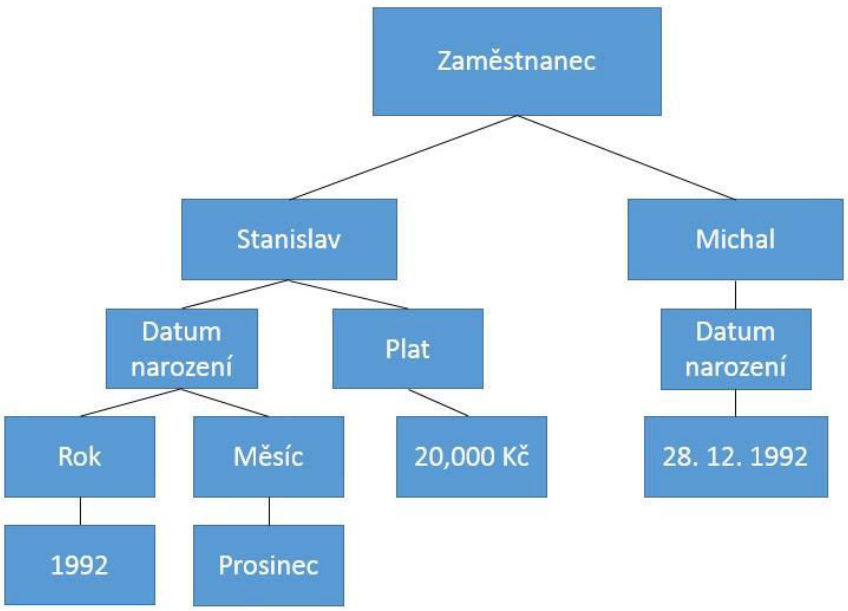
[7] MAYER-SCHÖNBERGER, Viktor a Kenneth CUKIER. *Big Data*. Brno: Computer Press, 2014, s. 18. ISBN 978-80-251-4119-9.

Již v roce 2008 prohlásil přední datový analytik Seth Grimes, že 80% relevantních dat pro business jsou data nestrukturovaná. Souvisí to také s obrovským nárůstem dat na sociálních sítích. Nestrukturovaná data nemají jasně definovanou strukturu, tedy nenajdeme pro ně jasně definované schéma v databázi a není tak jednoduché je analyzovat.

Jedná se o data ze sociálních sítí, ale i obrázky, audio, video. Důležitý aspekt také tvoří pojem Internet věcí. Přichází doba chytrých domácností, kde pomalu každý spotřebič bude mít své aplikační rozhraní (dále jen API), přes které bude komunikovat s určitým systémem, čili bude generovat data.

**Semistrukturovaná**

Často je tento typ dat popisován jako samo-popisující se či „data bez jasně definovaného schématu“. Tedy semistrukturovaná data můžou mít předem definované schéma, ale toto schéma není pevné, může se měnit, ale data tam stále zůstanou. Schéma lze definovat i ad-hoc podle jednotlivých objektů. Jednou z předností semistrukturovaných dat, je možnost ukládání dat, která nejsou úplná, jsou duplicitní, nebo se nedrží předem definované struktury. Na následujícím obrázku lze vidět příklad semistrukturovaných dat.




*Obrázek 2: Příklad semistrukturovaných dat*
Zdroj: vlastní

Typickým datovým formátem pro semistrukturovaná data je XML nebo JSON, viz podkapitola 1.7.1.

Podle A. Gupta může být heterogenita dat popsána také dimenzemi různorodosti.[8]

1. Strukturální dimenze různorodosti
2. Dimenze různorodosti médií
3. Sémantická dimenze
4. Dimenze přístupnosti

[8] GUPTA, Amarnath. Characteristics of Big Data - Variety. In: Coursera [online]. [cit. 2017-02-28]. Dostupné z: https://www.coursera.org/learn/big-data-introduction/lecture/oVg4p/characteristics-of-big-data-variety

**Strukturální dimenze různorodosti**

Strukturální dimenze popisuje rozdílnost v reprezentaci dat. Data mají různé formáty a modely. Například signál z EKG je jiná reprezentace dat než například příspěvek na sociální síti nebo než satelitní fotografie pořízené společností NASA.

**Dimenze různorodosti médií**

Tato dimenze se týká různorodosti médií, na kterých jsou data zaznamenána a přenášena. Audio záznam projevu a jeho následná transkripce, obsahují v zásadě velmi podobná data, ale média, na kterých jsou data zaznamenána, jsou velmi odlišná.

**Sémantická dimenze**

Sémantická dimenze je hlavně o významu dat, ve smyslu upřesňování kontextu. Například pokud děláme výzkum příjmu a zkoumáme dvě skupiny lidí, nemůžeme následně data z obou skupin spojit, protože kontext výzkumu u každé skupiny bude jiný.

**Dimenze přístupnosti**

Data mohou být přístupná v různých časových dimenzích. Mohou být přístupná v daný okamžik (real-time), například data ze sensorů. Nebo mohou být data přístupná kontinuálně, například z kamer.

Dimenze přístupnosti může následně určovat, jaké operace s daty budou vhodné a které již méně, obzvláště pokud se bude jednat o velké objemy dat.

## 1.4 Způsoby zpracování Big Data

Životní cyklus zpracování Big Data se podstatně liší od klasického (transakčního) životního cyklu. U klasického modelu se po úvodní datové analýze vytvoří specifika pro samotnou strukturu databáze. Poté přichází vytváření datového modelu. Následně je vytvářena databázová struktura na zpracování dat. Architektura celého řešení je pak optimalizovaná

jak z analytického hlediska, tak technického, jelikož struktura dat, jejich stav a forma jsou předem známy.

Při zpracování Big Data jsou data nejdříve shromážděna a nahrána do cílové platformy. Poté se na data typicky aplikuje metadatová vrstva, která zajistí vytvoření určité datové struktury. Jakmile se aplikuje metadatová vrstva, nastává transformační a analytická fáze. Vzhledem k dynamicky měnícím se datům a celkově flexibilnímu zpracování není typická databázově řízená architektura vhodná. Abychom dokázali zpracovat takové množství dat, která jsou velice komplexní a neustále přibývají, je vhodnější zvolit souborově řízenou architekturu s vhodným programovacím rozhraním.

K. Krishnan vyspecifikoval následující klíčové požadavky, co se týče infrastruktury a architektury zpracování.[9]

**Požadavky na architekturu zpracování**

- Data model-less
- Sběr dat by měl co nejblíže „real-time“
- Využívání mikrobatchů
- Minimální transformace dat
- Schopnost Multipartitionu
- Ukládání dat do file-systému či ne-relační databáze

**Požadavky na infrastrukturu**

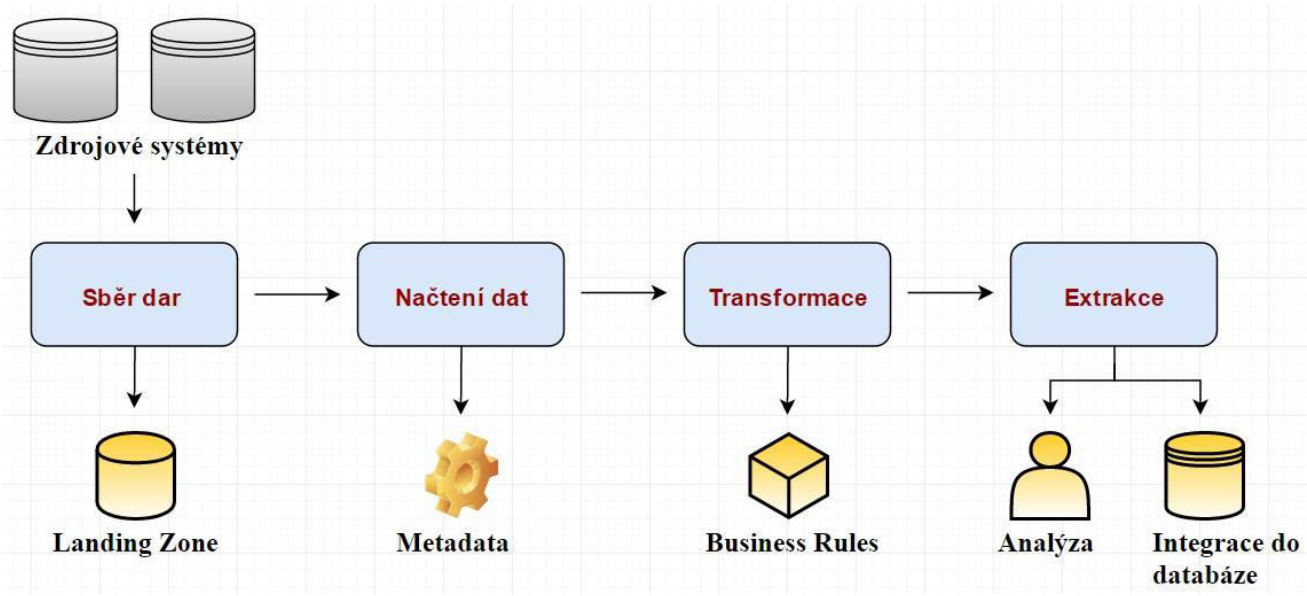
- Lineární škálování
- Vysoká propustnost
- Chybová tolerance
- Automatická obnova
- Vysoký stupeň paralelismu

[9] KRISHNAN, Krish. Data warehousing in the age of big data. Waltham: Elsevier, 2013, s. 37. ISBN 978-012-4058-910.

- Programovací rozhraní

### Data-flow zpracování Big Data

Model architektury zpracování dat by měl vždy vycházet z určitého data-flow modelu. Na následujícím obrázku je znázorněn data-flow model pro zpracování Big Data.




*Obrázek 3: Data-flow Big Data*
Zdroj: vlastní

Z obrázku je patrné, že datový tok se skládá ze 4 částí:

### Sběr dat

V této fázi jsou data sbírána z mnoha různých zdrojů (sociální sítě, zdrojové systémy …) a následně ukládána do file-systému nebo ne-relační databáze, jinak řečeno do Landing Zone.

### Načtení dat

Ve fázi načítání dat je aplikována metadatová vrstva, která udává strukturu dat. Soubory dat se obvykle v této fázi rozkládají na menší soubory, které jsou dále řízeny metadaty. V této části se také dá využít partitioning (horizontální nebo vertikální).

**Transformace**

V této fázi dochází k samotné transformaci, která zahrnuje většinou aplikování nějakých business pravidel. Výstupem této fáze jsou typicky klíče metadat, které jsou dále napárovány na hodnoty (key-value pair).

**Extrakce**

V extrakční fázi se data mohou již analyzovat nebo může docházet k další integraci do databázového systému.

## 1.5 Distribuované zpracování

Z požadavků, které jsou uvedené v kapitole 1.4 a s ohledem na data-flow Big Data se jeví jako nejvhodnější zpracování dat tzv. distribuované. Tento typ zpracování dat se vyznačuje tím, že úlohy spojené se zpracováním dat distribuuje více jak jednomu místu, na rozdíl od centrálního zpracování, kdy se celé zpracovávání odehrává na jednom centrálním místě.[10]

Architektur, které využívají distribuované zpracování, je několik:

**Dvou vrstvá architektura (klient-server)**

V této architektuře klient obstarává prezentační vrstvu a případné vstupy. Server zpracovává data a reaguje na dotazy klienta.

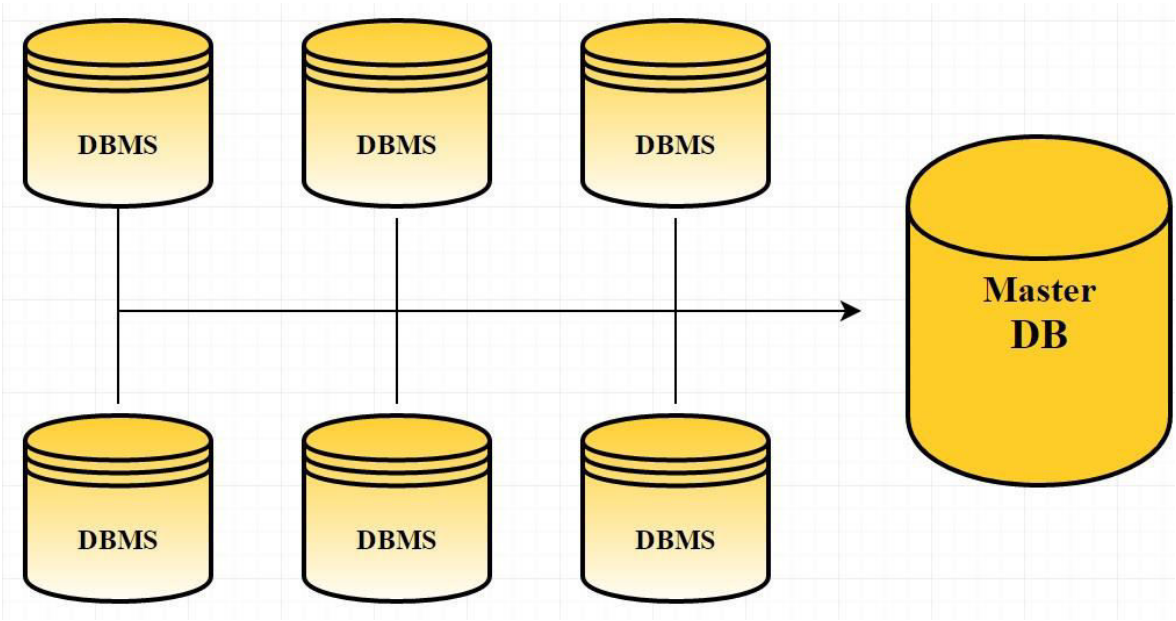
[10] KRISHNAN, Krish. Data warehousing in the age of big data. Waltham: Elsevier, 2013. ISBN 978-012-4058-910.

### Tří vrstvá architektura

Jedná se o nadstavbu dvou-vrstvé architektury (klient-server), kdy za účelem zvýšení efektivity zpracování byla přidána prostřední vrstva, která zahrnuje logiku zpracování dat, která probíhala na straně klienta. Jinými slovy klient už není přímo napojený na server, což vede k efektivnějšímu zpracování.

### Cluster architektura

Cluster architektura se vyznačuje zapojením více serverů či lokálních stanic a distribuováním jednotlivých úloh mezi veškeré články za účelem zvýšení efektivity. Každá stanice zapojená do clusteru má tedy svoje úlohy zpracování dat a následně jsou výsledky odesílány na master server, kde dochází k případné konsolidaci, či další distribuci dat. High-level koncept této architektury je zobrazen na následujícím obrázku.




*Obrázek 4: Cluster architektura*
Zdroj: vlastní

Mezi hlavní výhody distribuovaného zpracování patří horizontální škálování systému, které je efektivní a méně finančně náročné než škálování vertikální. Jednou z nevýhod může být výskyt redundantních dat.

## 1.6 Technologie pro zpracování Big Data

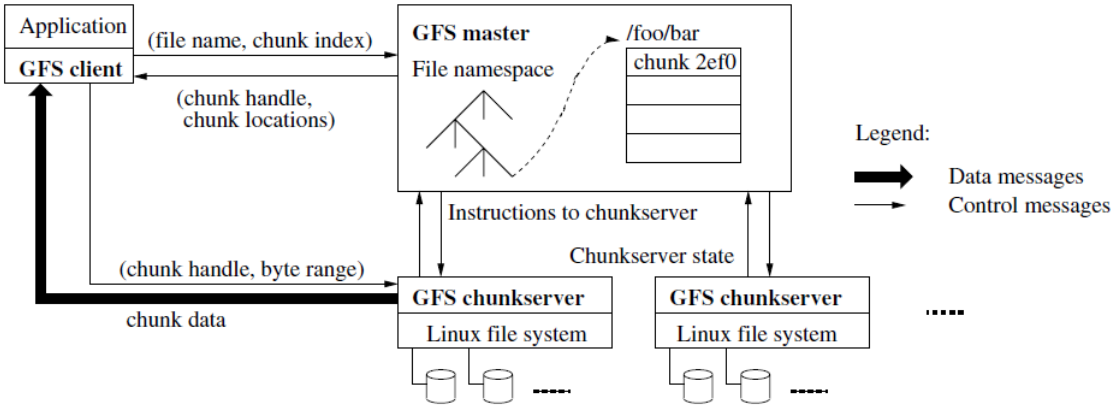
Aby bylo možné rozumně zpracovávat tak velké objemy dat, které mají navíc obrovský nárůst, je třeba extrémně výkonné prostředí, které je snadno ovladatelné a využívá horizontální škálování. Společnosti jako Google nebo Facebook musely řešit problém jak zvyšovat výkonnost dotazování a jak zařídit nekonečné škálování systému. Sám Google se rozhodl tento problém řešit interně a vyvinul si Google File Systém, dále jen GFS. A publikoval programovací model zvaný MapReduce.

### 1.6.1 GFS

Google začal vyvíjet tento file-systém pro své účely v 90. letech, jelikož v té době nebyly žádné systémy, které by pokryly jeho specifické potřeby. Tento file-systém stojí na principech, které v té době byly revoluční. Mezi tyto principy patří:

- Selhání komponent je více normou než výjimkou
- Standardem jsou velké soubory
- Většina dat je častěji měněna připojením dalších dat, nežli jejich přepsáním

Na následujícím obrázku je vyobrazena architektura GFS.




*Obrázek 5: Architektura GFS*
Zdroj: Architektura GFS [online]. In: . [cit. 2017-03-29]. Dostupné z: https://ofirm.files.wordpress.com/2013/01/gfs_architecture.png

GFS se skládá z jednoho Masteru a několika dalších chunkserverů. Veškerá data jsou uložena právě na těchto chunkserverech. Soubory jsou rozdělovány do tzv. chunků, které mají standardní velikost 64 MB. Každému chunku je GFS Masterem přiřazeno 64 bitové číslo při jeho vytvoření. Aby bylo zajištěno, že selhání komponent systému je běžné, je každý chunk minimálně jednou replikován na další jiný server. Standardní počet replik chunků jsou tři. Jako úzké hrdlo se v této architektuře jeví GFS Master, ale není tomu tak. Přes GFS Mastera nejdou žádné datové transakce, pouze sděluje klientovi, na jakém chunkserveru jsou uloženy jaké chunky a další metadatové informace. Klient pak často nekomunikuje vůbec s GFS Masterem, jelikož si informace od něj je schopen uložit do cache. [11]

Pokud dojde k selhání GFS Masteru, tak jediné co se ztratí, budou odkazy na chunky a metadata. Toto je ošetřené tím, že je master navržen tak, aby držel data v paměti a logoval na lokální disk, který je replikován do dalších uzlů.

GFS Master pravidelně kontroluje každý chuckserver a pomocí periodických kontrolních výpočtů hlídá poškození dat. Pokud GFS Master detekuje nějaké poškození, okamžitě spustí obnovu z replik. To znamená, že jsou data pouze po určitý čas nedostupná, nikoli poškozená.

[11] GHEMAWAT, Sanjay, Howard GOBIOFF a Shun-Tak LEUNG. The Google File System [online]. [cit. 2016-12-30]. Dostupné z: https://static.googleusercontent.com/media/research.google.com/cs//archive/gfs-sosp2003.pdf

### 1.6.2 MapReduce

Mezi základní principy pro zpracování velkých objemů dat patří programovací model MapReduce, který publikovali zaměstnanci společnosti Google, J. Dean a S. Ghemawat v článku „*MapReduce: Simplified Data Processing on Large Clusters“*.[12]

Jedná se o programovací model, který umožňuje paralelní zpracování a generování obrovských datových setů. Vhodná implementace tohoto modelu umožňuje dosahovat vysokých výkonů díky velkým clusterům komoditních počítačů. Implementace MapReduce modelu se nazývají MapReduce frameworky a patří k základním stavebním kamenům zpracování Big Data.

Princip programovacího modelu MapReduce je postaven na dvou základních funkcích. Jedna z funkcí se nazývá Map a druhá Reduce. Funkce Map zpracuje každý objekt vstupní množiny a převede data na strukturu klíč-hodnota. Po převedení dat na danou strukturu se ke stejnému klíči napárují všechny hodnoty přes danou množinu a tento výsledek se zašle funkci Reduce. Tato funkce pak k danému klíči sloučí všechny hodnoty, což znamená, že z velké množiny hodnot typicky zbyde jen hodnota jedna (tedy výsledek bude jeden klíč a k němu jedna hodnota). Hlavní výhodou funkce MapReduce je možnost zpracovávat velké objemy dat v distribuovaném prostředí. Na následujícím příkladu je jasná ukázka funkcionality. Problém, který bude řešen v příkladu, je zjišťování nejvyšší teploty v daných městech. Vstupní množina bude přes 1000 souborů, které budou mít následující strukturu. Každý soubor má v sobě informaci o městě a o teplotě, která v daném městě byla naměřena. Hodnoty se v souboru mohou opakovat a nic z uvedených informací neindikuje časovou složku, data jsou čistě náhodná, viz obrázek č. 6.

[12] DEAN, Jeffrey a Sanjay GHEMAWAT. MapReduce: Simplified Data Processing on Large Clusters. In: *Sixth Symposium on Operating System Design and Implementation* [online]. San Francisco, 2004 [cit. 2017-02-28]. Dostupné z: https://static.googleusercontent.com/media/research.google.com/cs//archive/mapreduce-osdi04.pdf

**Soubor 1**

| Město | Teplota |
|---|---|
| Liberec | 25 |
| Praha | 26 |
| Náchod | 28 |
| Frýdlant | 22 |
| Praha | 21 |
| Liberec | 24 |

**Soubor 2**

| Město | Teplota |
|---|---|
| Praha | 20 |
| Náchod | 31 |
| Náchod | 28 |
| Frýdlant | 15 |
| Praha | 18 |
| Frýdlant | 29 |

**Soubor 3**

| Město | Teplota |
|---|---|
| Náchod | 24 |
| Frýdlant | 26 |
| Liberec | 26 |
| Praha | 32 |
| Praha | 21 |
| Liberec | 24 |

*Obrázek 6: MapReduce - ukázka vstupních souborů*
Zdroj: vlastní

Funkce Map nejdříve zajistí namapování hodnot ze souboru tak, že klíč je název města a teplota bude příslušná hodnota. Poté funkce Map začne hledat v jednotlivých souborech nejvyšší teploty za daná města a výstup za každý soubor bude vždy pouze název města (klíč) a nejvyšší teplota (hodnota) v daném souboru. Implementovaná funkce Map pro tento konkrétní příklad by mohla vypadat takto.

Execute | Share Code main.rb ×

```ruby
def FunkceMap (soubor)

    soubor.each { |key, value|
    next if soubor[key] > value
    soubor_nejvyssi_teplota[key] = value
    }

end
```

*Obrázek 7: Funkce Map – příklad*
Zdroj: vlastní

Z 1000 souborů udělá funkce Map 1000 výstupů, které ale budou jednak namapovány jako klíč-hodnota a dále bude každý výstup obsahovat pouze jednou název města a k tomu nejvyšší teplotu, která byla v daném souboru nalezena.

*Tabulka 1: výstup funkce Map*

| Výstup 1 | | Výstup 2 | | Výstup 3 | |
|---|---|---|---|---|---|
| **Město** | **Teplota** | **Město** | **Teplota** | **Město** | **Teplota** |
| Liberec | 25 | Praha | 20 | Liberec | 26 |
| Praha | 26 | Náchod | 31 | Praha | 32 |
| Náchod | 28 | Frýdlant | 29 | Náchod | 24 |
| Frýdlant | 22 | | | Frýdlant | 26 |

Zdroj: vlastní

Z tabulky č. 1 jasně vidíme, že funkce Map odstranila redundantní hodnoty vzhledem k úloze (potřeba jen výskytů nejvyšších teplot v daném městě). Výstupy z funkce Map následně začne zpracovávat funkce Reduce. Tato funkce zpracuje všechny výstupy a pomocí daného algoritmu vrátí názvy měst pouze s nejvyšší teplotou, která byla nalezena napříč všemi dokumenty.

Celý tento myšlenkový pochod se využívá v mnoha úlohách, které mají za úkol zpracovat velké množství dat. Framework, který je na této funkčnosti postaven se využívá hlavně k efektivnímu rozdělování úloh mezi jednotlivé uzly v clusteru a tedy umožňuje paralelně zpracovávat danou úlohu.

### 1.6.3 Hadoop

Další systém pro zpracování velkých objemů dat, který začal být velmi populární je Hadoop. Jedná se o open-source Framework poskytující architektonické řešení na zpracování Big Data na levnější platformě, která má rychlou škálovatelnost a umožňuje paralelní zpracování dat.

Hadoop vznikl v rámci projektu Nutch, který spoluzaložili M. Cafarella a D. Coutting v roce 2002. Účelem projektu bylo vytvoření vyhledávacího systému, který by měl špičkový crawler systém[13]. To se nakonec povedlo, ale tým D. Cuttinga narazil na omezené škálování vyhledávácího systému. Problém se škálovatelností se vyřešil vyvinutím vlastního file-

[13] Crawler je speciální bot, který prochází celý web, za účelem vytvoření obrovské databáze – indexuje weby a tyto indexy se pak využívají přes webové vyhledávače.

systému s názvem Nutch Distributed File System (NDFS), který byl inspirován GFS. Dále pak přišla na řadu implementace programovacího modelu Map Reduce (také inspirováno Googlem) a open-source Hadoop byl na světě.[14]

Hadoop se skládá z pěti základních komponent.

- HDFS (Hadoop Distrubuted File System)
- MapReduce
- YARN
- HBase
- ZooKeeper

## HDFS

Hadoop Distributed File System je škálovatelný file-systém, který je určený pro komoditní hardware. Tento file-systém stojí na podobných principech jako GFS. Systém se sám dokáže efektivně obnovovat z výpadků, se kterými předem počítá a jsou spíše pravidlem než výjimkou.

HDFS stojí na 3 základních principech.

1. Zpracování souborů/dat ve velikosti petabytů
2. Zpracování streamovaných dat za účelem čtení dat s vysokou mírou propustnosti
3. Možnost implementace na komoditní hardware

### Architektura HDFS

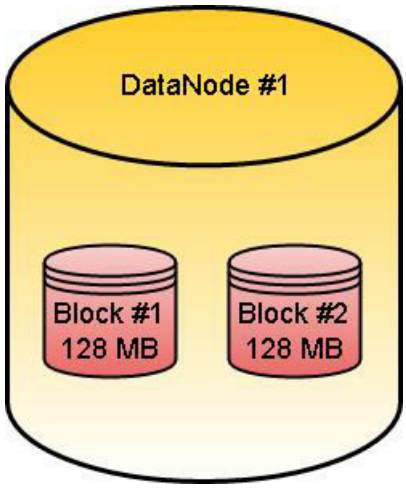
Architektura se opět odvíjí od GFS a obsahuje tyto hlavní stavební bloky.

- NameNode
- DataNode
- Image

[14] KRISHNAN, Krish. Data warehousing in the age of big data. Waltham: Elsevier, 2013. ISBN 978-012-4058-910.

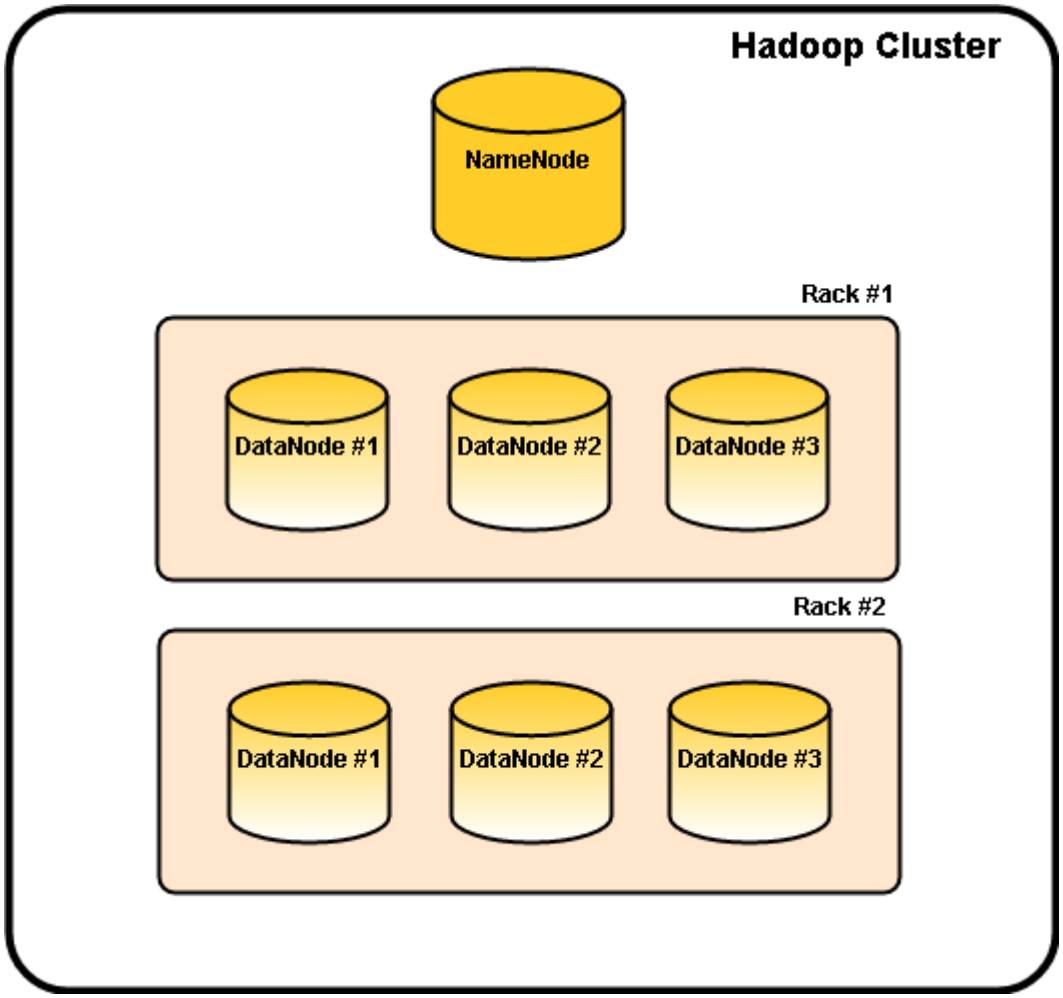
- Journal
- Checkpoint

HDFS je založený na architektuře master/slave, kde uzel nazvaný NameNode figuruje jako master a uzel nazvaný DataNode figuruje jako slave. NameNode spravuje jmenný prostor, metadata a řídí přístupy klientů k jednotlivým souborům a adresářům. DataNody slouží jako uložiště pro jednotlivé soubory, nebo spíše pro bloky souborů, viz obrázek 6. Bloky mají velikost většinou 64 MB nebo 128 MB. Tyto bloky jsou také samozřejmě replikovány do dalších uzlů, aby byla zajištěna odolnost a rychlé zotavení z výpadku. Důležitou vlastností těchto bloků je fakt, že velikost dat uložených na bloku se rovná následné velikosti bloku. To znamená, že pokud jsou v bloku pouze soubory s velikostí 64 MB, pak tento blok bude fyzicky zabírat 64 MB, nikoli 128 MB, což je jeho implicitní velikost. Tímto je optimalizováno ukládání souborů, ale také výkonnost daného řešení.




*Obrázek 8: Složení DataNode*
Zdroj: vlastní

Jednotlivé DataNode jsou seskupeny do tzv. racků, což jsou logické jednotky, jejichž seskupením vzniká celý Hadoop cluster. Data by měla být replikována jak v rámci jednoho racku, tak ale i v racku jiném, aby se zajistila co největší nezávislost a redukovala se tak pravděpodobnost ztráty dat. Níže uvedený obrázek č. 7 hrubě reflektuje architekturu jednoho clusteru.



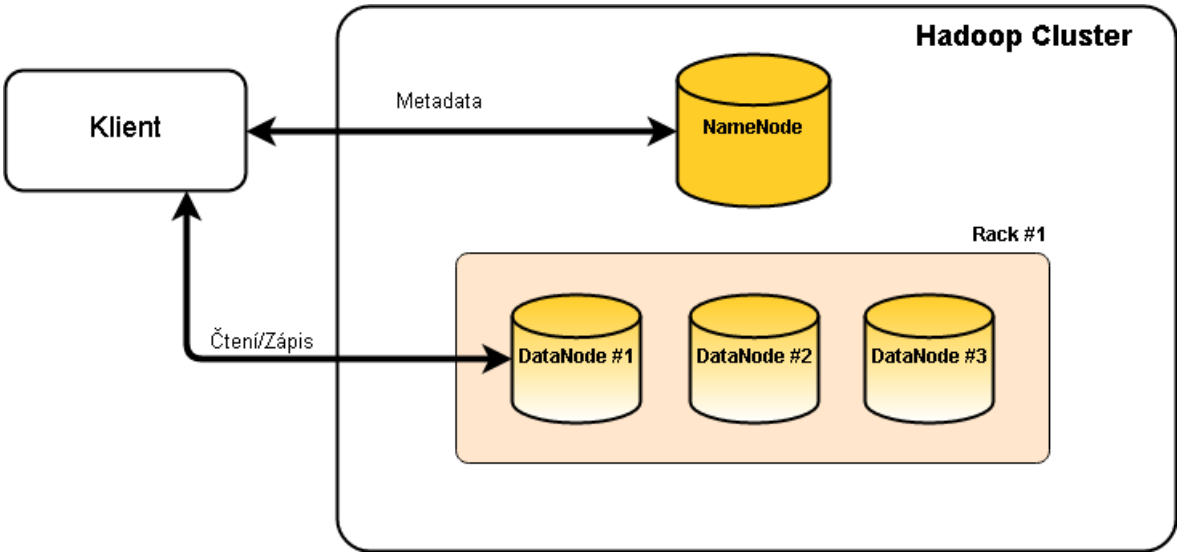

*Obrázek 9: Složení Hadoop Clusteru*
Zdroj: vlastní

K zajištění rychlého obnovení z výpadku NameNode v pravidelných intervalech ukládá na lokální disk tzv. checkpointy. Tento soubor je v podstatě Image uložený na lokálním file-systému. Image slouží jako snímek nastavení metadat celého jmenného prostoru. Journal je log transakcí, které proběhly od posledního uložení checkpointu. Obnovení systému pak vypadá tak, že Hadoop z checkpointu načte celou konfiguraci jmenného prostoru a poté provede všechny operace, které jsou v Journalu zalogovány.

**Operace v HDFS**

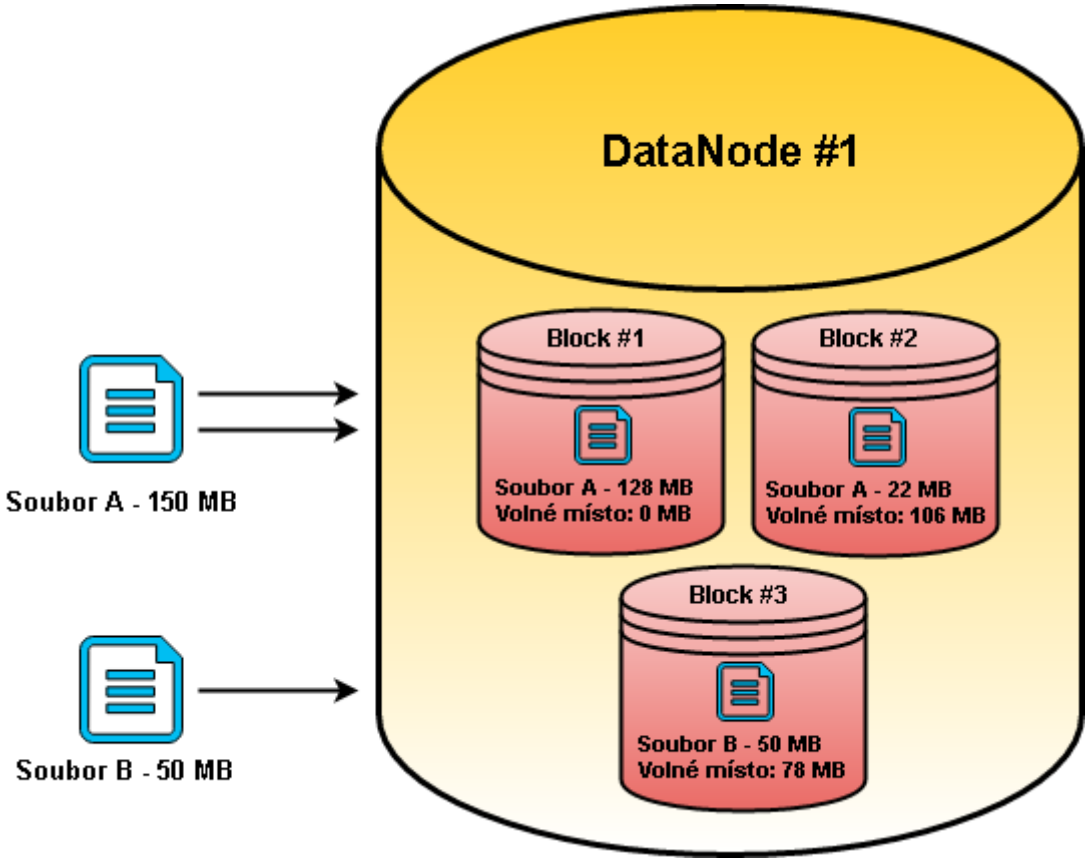
HDFS využívá pro komunikaci uživatele/softwaru a samotného file-systému rozhraní zvané Klient. Před jakoukoli operací musí Klient nejprve kontaktovat NameNode a vyžádat si metadata, která v sobě nesou například informace o umístění souborů a o tom, v jakých

blocích jsou jaké soubory uloženy. Poté již může Klient komunikovat přímo s jednotlivými DataNode a provádět čtení nebo zápis.




*Obrázek 10: Operace v HDFS – čtení a zápis*
Zdroj: vlastní

Ukládání souborů se od běžných file-systémů liší tím způsobem, že v každém bloku se nachází pouze jeden soubor. To znamená, že pokud má systém uložit dva soubory, jeden o velikosti 150 MB a druhý o velikosti 50 MB, tak z prvního souboru vezme 128 MB a uloží je na jeden blok, jelikož takový maximální objem dat se vejde do jednoho bloku. Zbylých 22 MB následně uloží do druhého bloku. Soubor o velikost 50 MB pak neuloží do druhého bloku, kde je zbývající kapacita 106 MB, ale uloží je do třetího bloku, aby bylo zajištěno pravidlo, že každý blok má v sobě právě jeden soubor. Tento konkrétní příklad je uvedený na obrázku č. 9.




*Obrázek 11: Operace v HDFS - ukládání souborů*
Zdroj: vlastní

## 1.7 NoSQL databáze

S rozmachem trendu Big Data a s rychle rostoucími objemy dat, není možné jednoduše vše zpracovávat a ukládat do pravidelných, strukturovaných a pevně daných tabulek. Relační databáze stojí na předpokladu, že je předem známá struktura dat a že uživatel nebo aplikace bude data do databáze často přidávat, často je aktualizovat a že nad daty probíhají často neznámé dotazy. To jsou důvody, proč jsou tato data rozdělena na menší kompaktní celky, které jsou uloženy v tabulce. Relační databáze také dále zajišťují konzistenci transakcí tím,

že využívají vlastnosti ACID, což zajišťuje pouze dva stavy transakce, buď proběhla kompletně, nebo nikoli.[15]

Vzhledem k tomu, že data od dob vzniku internetu začala růst neuvěřitelnou rychlostí a vzhledem k tomu, že byla potřeba tato data ukládat, začal nastávat problém s relačními databázemi. Vertikální škálování má totiž svá omezení a při horizontálním škálování naráží relační databáze na bariéry, jelikož při horizontálním škálování nejsou tak efektivní a dlouhodobě udržitelné. Proto v roce 2009 představil J. Oskarsson pojem „NoSQL“. Termín NoSQL znamená Not Only SQL a vyjadřuje odlišný přístup k datům. NoSQL databáze zažívají v posledních letech velký boom, jelikož je začíná využívat více systémů či aplikací. Existuje více typů NoSQL databází, které se navzájem liší a pracují na odlišných principech.

Primárním účelem NoSQL databází je data ukládat. Jejich hlavní předností je vysoká efektivita při vyhledávání, což je kompenzováno omezenou funkčností (databáze může sloužit jen pro čisté ukládání). Jednou z hlavních výhod je pak především možnost horizontálního škálování, které zajišťuje vyšší výkonnost.

### 1.7.1 Charakteristika NoSQL databází

I přes fakt, že se NoSQL databáze navzájem liší, lze nalézt některé společné charakteristiky, které je spojují. Každá NoSQL databáze navíc může být specifická konkrétní implementací, jelikož se může lišit konzistence, implementace dotazovacího jazyka nebo také řízení transakcí. Většinou se NoSQL databáze navzájem liší v těchto aspektech z důvodu výkonnosti a různých potřeb konkrétních podniků.

**Konzistence**

[15] HOLUBOVÁ, Irena, Jiří KOSEK, Karel MINAŘÍK a David NOVÁK. *Big Data a NoSQL databáze*. 1. vyd., Praha: Grada, 2015. Profesionál. ISBN 978-80-247-5466-6.

NoSQL databáze nemohou zajistit úplnou konzistenci transakcí, jelikož ve většině případů využívají plně vlastností ACID, ale využívá se tzv. „Eventual consistency“ – občasná konzistence. Díky způsobu občasné konzistence získávají databáze lepší dostupnost a možnost většího škálování, ale obětují přitom zaručení konzistence, které je pro některé aplikace či podniky klíčový. NoSQL databáze tedy nemají vlastnosti ACID, ale vlastnosti BASE. Tento akronym je definován jako Basically Available, Soft state, Eventual consistency a značí právě obětování konzistence za vyšší dostupnost a potřebu obnovovací periody.

**Dotazování**

Dotazování nad NoSQL databází je poněkud komplikované, jelikož není vyznačený jednoznačný standard jak k datům přistupovat. To je způsobeno různorodostí jednotlivých NoSQL databází. Většina podniků, které NoSQL databáze nabízí, si vyvinuly vlastní dotazovací jazyk, např. Cassandra CQL, Splunk SPL, CouchDB unQL atd. Dalším ze způsobů dotazování jsou programovací jazyky, které pomoci unikátního API přistupují k datům. V některých případech je pak využíváno MapReduce dotazování.[16]

**Flexibilita**

NoSQL databáze poskytují značnou míru flexibility díky tomu, že databáze nemá jasně danou detailní strukturu. Například aplikace, kde si uživatel může libovolně měnit následnou strukturu dat, jako jsou anketní aplikace. Jeden uživatel si může zvolit více zanořených otázek, jiný zase více polí atd. Technologicky je to samozřejmě možné provést s relační databází, nicméně není to optimální. NoSQL databáze to dokáží implementovat velmi lehce, kdy se jednotlivá pole budou vytvářet dynamicky a anketa se následně uloží jako celek s unikátním klíčem.

[16] HOLUBOVÁ, Irena, Jiří KOSEK, Karel MINAŘÍK a David NOVÁK. *Big Data a NoSQL databáze*. 1. vyd., Praha: Grada, 2015. Profesionál. ISBN 978-80-247-5466-6.

**Porovnání s relačními databázemi**

Co je charakteristické pro NoSQL databáze z pravidla pak není charakteristické pro SQL databáze a pokud by tyto dva druhy databází byly množiny, označily bychom je za disjunktní. V následující tabulce lze vidět porovnání těchto dvou typů databází.

| | **SQL databáze** | **NoSQL databáze** |
|---|---|---|
| **Typ databáze** | Relační | Nerelační |
| **Typ ukládaných dat** | Strukturovaná | Nestrukturovaná semistrukturovaná |
| **Škálování** | Vertikální | Horizontální |
| **Vlastnosti transakcí** | ACID | BASE |
| **Konzistence** | Silná | Občasná (liší se podle jednotlivých řešení) |
| **Schéma** | Statické schéma s jasně danou strukturou | Dynamické schéma/bez schéma |

*Tabulka 2: SQL vs NoSQL databáze*
Zdroj: vlastní

### 1.7.2 Typologie NoSQL databází

V rámci NoSQL databází existuje mnohem více odlišností mezi jednotlivými řešeními než u relačních. U NoSQL existují 4 základní typy databází:

- Databáze typu klíč-hodnota
- Grafové databáze
- Dokumentové databáze
- Sloupcové databáze

**Databáze typu klíč-hodnota**

Tento typ NoSQL databáze je postavený na velmi jednoduchém principu. Každý ukládaný objekt, ať už je to obrázek, video, dokument či prostý text bude mít svůj unikátní klíč. Tento klíč pak funguje jako reference na daný objekt (hodnotu), nicméně již tento vztah nefunguje zpětně. Jinými slovy velmi to ztěžuje vyhledávání podle hodnoty. Nicméně čím dál tím více databází typu klíč-hodnota začalo tento nedostatek obcházet pomocí sekundárních indexů. Tyto indexy jsou vytvořeny nad určitými atributy jednotlivých hodnot a následně lze podle nich vyhledávat. Databáze tohoto typu jsou jednoduché jak svou strukturou, tak operacemi, které zde probíhají. Standardní API pro páci s daty obsahuje tři druhy operací. Vložení hodnoty (PUT), získání hodnoty (GET) a mazání (DELETE). Příkladem databáze typu klíč-hodnota je například Oracle NoSQL Database nebo Redis.

**Dokumentové databáze**

Z názvu již vyplývá, že tento typ NoSQL databáze bude ukládat data v podobě dokumentů. Tyto dokumenty v sobě neobsahují jen typická data, ale jsou zde uložena i metadata. Databáze využívají formáty jako je JSON, XML nebo BSON. Dokumenty jsou dále typické svou stromovitou strukturou a tím, že dokáží efektivně komunikovat se systémy či aplikacemi. Při použití relačních databází vzniká často potřeba konverze dat, jelikož ke komunikaci s webovými aplikacemi jsou vyžadovány právě formáty jako JSON nebo XML, viz dále. Dokumentové databáze mohou uložená data přímo využít pro komunikaci s aplikacemi a nemusí docházet tak ke konverzi, která může být náročná. Další výhodou je fakt, že dokumenty uložené v databázi mohou mít vnitřní strukturu uzpůsobenou ke svým konkrétním účelům a nemusí se řídit předem definovaným schématem.

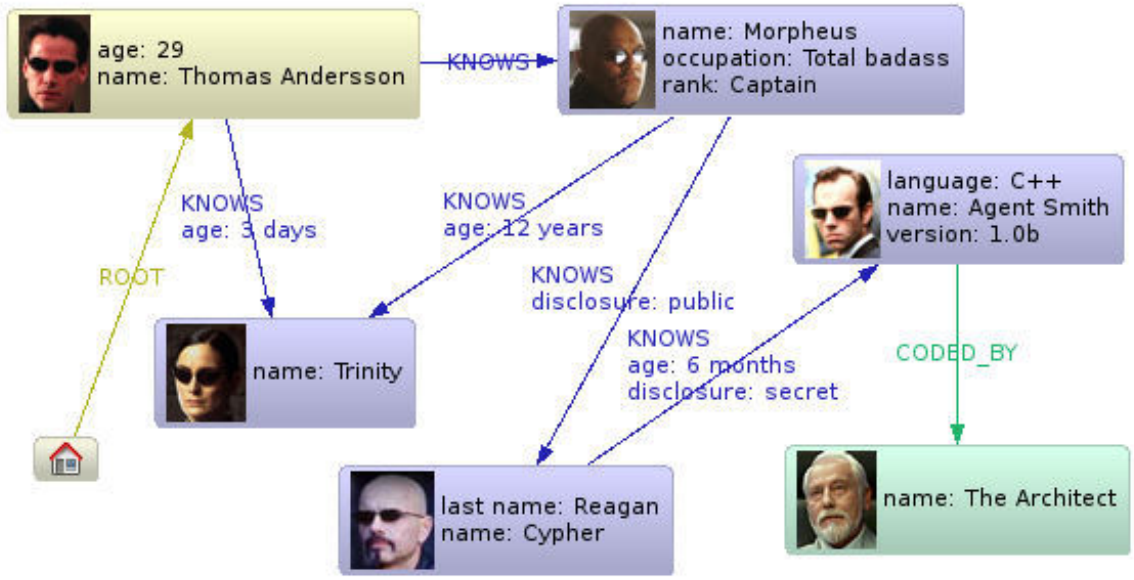
**Sloupcové databáze**

Tento typ databází vychází z principu, který představil Google jako BigTable databáze. Princip je podobný jako u databází typu klíč-hodnota. K datům se přistupuje prostřednictvím unikátního klíče, jenže v tomto případě tvoří tento klíč identifikace řádku a sloupce. Sloupcovou databázi si lze představit jako obrovskou excel tabulku, kde jsou v jednotlivých buňkách uloženy různá data v různých formátech. Tyto tabulky bývají obrovské, až miliardy řádků a statisíce sloupců. K vloženým datům se pak přidává ještě časové razítko za účelem verzování. Jako typický příklad aplikace, která tento typ databáze využívá je Google Earth.

Řádky označují zeměpisnou šířku a sloupce pak zeměpisnou délku. Uživatel pak přistupuje k datům zadáním těchto dvou atributů a aplikace zobrazí příslušný snímek uložený na daném místě. Na rozdíl od relačních databází umí sloupcové databáze efektivně pracovat s velkým množstvím prázdných hodnot. Hlavní nevýhodou je pak nemožnost indexace a nutnost znalosti celého klíče.

### Grafové databáze

Grafové databáze se od ostatních typů NoSQL databází podstatně liší. Data jsou totiž uložena v grafové struktuře, čili je zde množina uzlů, které jsou vzájemně propojeny hranami. Jednotlivé uzly představují objekty a hrany pak reprezentují vztahy mezi nimi. Na následujícím obrázku je příklad jednoho grafu v databázi.




*Obrázek 12: Příklad grafové databáze*
Zdroj: Graf DB [online]. In: . [cit. 2017-03-29]. Dostupné z: https://msdnshared.blob.core.windows.net/media/MSDNBlogsFS/prod.evol.blogs.msdn.com/CommunityServer.Blogs.Components.WeblogFiles/00/00/00/68/67/metablogapi/2311.image_42DC40D5.png

Na příkladu je zobrazen kultovní film Matrix, respektive postavy vyskytující se v tomto filmu. Jsou zde zobrazeny také jejich vztahy a vlastnosti. Lze si povšimnout toho, že i samotný vztah může mít nějaký atribut (na příkladu je označeno, jak dlouho se jednotlivé postavy znají). Hlavní výhodou grafových databází je tedy možnost zachycení vztahu mezi

jednotlivými objekty. To znamená, že lze provádět grafové analýzy, analýzy vazeb a zobrazovat vztahy mezi dokumenty a využívat tak výhody dokumentové databáze. Do grafových databází lze uložit jakýkoli datový formát. Hlavní nevýhodou, podobně jako u relačních databází, je neefektivní horizontální škálování.

### 1.7.3 Datové formáty v NoSQL databázích

Na rozdíl od relačních databází, které mají standardní datový model s komplikovanou strukturou, nemají NoSQL databáze typicky datový model vůbec, nebo je naprosto triviální (například NoSQL databáze typu klíč-hodnota, kde klíčem může být umělé ID objektu a hodnotou pak bude daný soubor). Tento soubor musí mít nějaký datový formát, záleží na využití NoSQL databáze.

#### CSV

CSV neboli Comma-Separeted Values je jedním z datových formátů se kterým se lze setkat v NoSQL databázích. Jedná se o formát, který byl podporován již v roce 1972 (ještě před prvním osobním počítačem). CSV ukládá data jako prostý text, kde je pokaždé jeden záznam na jednom řádku a pole jsou oddělena čárkou. Díky své jednoduché struktuře, kompatibilitě se všemi systémy a možnosti uložit tak velké množství dat, je formát CSV stále hojně využíván. CSV si lze představit jako tabulku v databázi s tím, že atributy jsou odděleny tzv. „oddělovačem“ (typicky čárkou) a jednotlivé záznamy jsou pod sebou v řádcích, viz obrázek.

```
 1 ItemID,Sentiment,SentimentSource,SentimentText
 2 1,0,Sentiment140,is so sad for my APL friend.............
 3 2,0,Sentiment140,I missed the New Moon trailer...
 4 3,1,Sentiment140,omg its already 7:30 :O
 5 4,0,Sentiment140,.. Omgaga. Im sooo  im gunna CRy. I've been at this dentist
 6 5,0,Sentiment140,i think mi bf is cheating on me!!!       T_T
 7 6,0,Sentiment140,or i just worry too much?
 8 7,1,Sentiment140,Juuuuuuuuuuuuuuuussssst Chillin!!
 9 8,0,Sentiment140,Sunny Again Twitter Work Tomorrow  :-|       TV Tonight
10 9,1,Sentiment140,handed in my uniform today . i miss you already
```

*Obrázek 13: Ukázka CSV*
Zdroj: Data ze sociální sítě Twitter [online]. In: . [cit. 2017-03-29]. Dostupné z: www.twitter.com

V CSV nemusí být oddělovačem pouze čárka, ale lze ho nastavit. Implicitně je jako oddělovač nastavena čárka, nicméně pro případy, kdy by se v datech čárka mohla vyskytovat, je lepší například zvolit středník, či tabulátor. Pokud by se čárka vyskytovala v datech, tak kromě zvolení jiného oddělovače lze také znak odescapovat. Toto řešení je poněkud kostrbaté, jelikož by se musela odescapovat každá čárka v datech, což by při velkém objemu dat nebylo nejvhodnější řešení.

Hlavním účelem tohoto formátu je snadný přenos dat mezi databázemi/aplikacemi. Tuto úlohu splňuje CSV skvěle při jednoduché struktuře dat. Jedním z hlavních problémů tohoto formátu je totiž nemožnost reprezentovat hierarchii dat, což je velký problém pokud by se měla v CSV formátu držet nějaká komplikovanější struktura dat.

**JSON**

JSON neboli JavaScript Object Notation je jedním z vůbec nejrozšířenějších datových formátů, využívající se především k výměně dat mezi webovými aplikacemi. Jedná se o jednoduchý formát, který umožňuje uchovávat komplexní datové struktury v pragmatickém zápisu dat. Formát JSON vychází z JavaScriptu a navrhl jej D. Crocford v roce 2000. Struktura JSONu vypadá zhruba takto


```
"comments": {
  "data": [
    {
      "created_time": "2017-06-21T17:01:05+0000",
      "from": {
        "name": "XYZ",
        "id": "4665964744203"
      },
      "message": "Tohle je boží fotka □",
      "id": "880912458741229_880913365407805"
    },
    {
      "created_time": "2017-06-21T17:02:12+0000",
      "from": {
        "name": "XZY",
        "id": "10203947419581975"
      },
      "message": "Vyraz... Jezis co jsem tam zapomel dat....",
      "id": "880912458741229_880914742074334"
    },
```


*Obrázek 14: Ukázka JSON*
Zdroj: Data ze sociální sítě Facebook [online]. In: . [cit. 2017-03-29]. Dostupné z: www.facebook.com

JSON dokáže uchovávat šest datových typů, z toho čtyři jednoduché a dva strukturované. Mezi jednoduché se řadí číslo, null, boolean a řetězec a mezi strukturované pak objekty a pole. Objekt v JSONu není tak úplně plnohodnotným objektem jako např. objekt v JavaScriptu. Jedná se spíše o asociativní pole. A právě pomocí tohoto asociativního pole neboli hashe se dají v JSONu uchovávat komplexnější datové struktury. JSON je jazykově nezávislý formát, tudíž se vyskytuje ve více programovacích jazycích. Kromě v JavaScriptu se s ním lze setkat např. v Ruby, PHP, Pythonu a dalších. Jednou z velkých nevýhod tohoto formátu je nemožnost jakéhokoli komentáře. Což podstatně komplikuje porozumění některých složitějších struktur.

Vzhledem k popularitě JSONu a stále většímu využívání tohoto formátu, bylo potřeba ho nějak standardizovat. Kvůli transparentnosti a konzistenci dat, kterou bylo potřeba zajistit jasnou definici zasílané struktury dat. Proto vzniklo JSON Schema, což je v podstatě standard struktury JSONu, který bude databáze či webová služba přijímat. Webová služba co využívá JSON Schema může obdržený JSON následně porovnat s JSON Schematem a podniknout určité kroky, pokud obdržený JSON není validní. Na obrázku lze vidět příklad JSON Schema.

# 2. Specifikace a informační potřeby MSP

V době bezplatné komunikace a cíleného marketingu, který nabízejí sociální sítě, si může jakýkoli malý a střední podnik obstarat data, díky kterým dokáže efektivněji řídit, komunikovat, nabízet, poptávat, ale i lépe poznat vlastní odvětví.

## 2.1 Specifikace MSP

Základním kritériem pro posouzení velikosti podniku je počet zaměstnanců, velikost ročního obratu a bilanční suma roční rozvahy. Evropská komise vydala v roce 2005 novou definici, čímž poukázala na důležitost malých a středních podniků, dále jen MSP, v podnikatelské sféře. G. Verhaugen, člen Evropské komise k nové definici uvádí:

*„Mikropodniky, malé a střední podniky jsou motorem evropského hospodářství. Jsou základním zdrojem pracovních příležitostí, vytvářejí podnikatelského ducha a inovace v EU, a jsou tedy rozhodující pro posílení konkurenceschopnosti a zaměstnanosti. Nová definice malých a středních podniků, která vstoupila v platnost dne 1. ledna 2005, představuje významný krok směrem k lepšímu podnikatelskému prostředí pro malé a střední podniky a zaměřuje se na podporu podnikání, investic a růstu. Tato definice byla vypracována po širokých konzultacích s dotčenými osobami, které prokázaly, že naslouchání malým a středním podnikům je klíčem k úspěšnému provedení lisabonských cílů.“*[17]

Definování MSP je tedy z ekonomického, ale i technického hlediska velmi klíčová záležitost. MSP tvoří páteř Evropské ekonomiky, což dokazují statistiky vytvořené Evropskou komisí, kde se uvádí, že v EU spadá do kategorie MSP přes 23 miliónů podniků (což je v Evropském průměru přes 99.81% všech podniků v zemi). Tyto podniky dokázaly vygenerovat přes 3.9 trilionů euro a zaměstnat 90 milionů lidí. Co se týče České republiky, tak poměr MSP vůči všem podnikům v zemi překračuje evropský průměr a dosahuje 99,85%

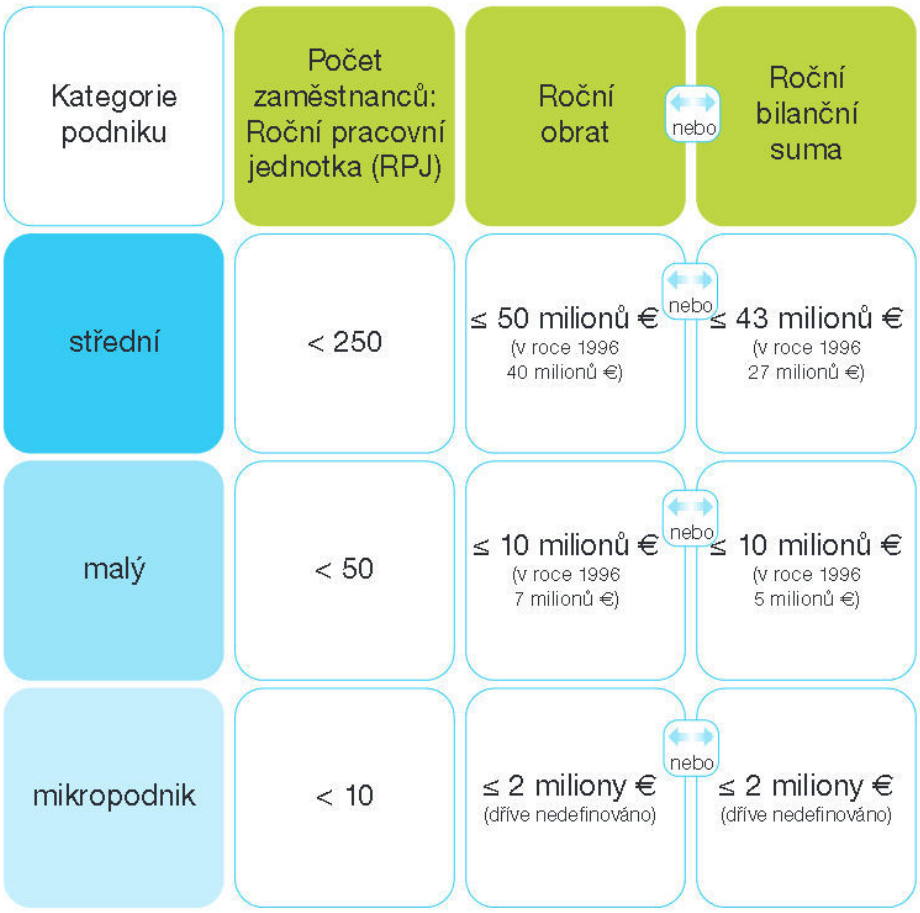
[17] Uživatelská příručka: k definici malých a středních podniků [online]. 1. Lucemburk: Úřad pro publikace Evropské unie, 2015 [cit. 2017-5-14]. ISBN 978-92-79-45316-8. Dostupné z: https://ec.europa.eu/docsroom/documents/15582/attachments/1/translations/cs/renditions/native.

(což je 992 616 podniků). MSP v České republice zaměstnávají 2 416 661 lidí (tj. 68,2%) a vygenerovaly 48.8 milionů Euro.[18]

Evropská komise stanovila jasnou a jednotnou definici MSP, za účelem větší soudržnosti a omezení narušování zdravé hospodářské soutěže. Stanovení jasné definice bylo nutné z důvodu jednotného trhu bez vnitřních hranic a finanční podpoře MSP.

Malý podnik je definován jako podnik s méně než 50 zaměstnanci, ročním obratem menším nebo rovným 10 milionům eur a roční bilanční sumou taktéž menší nebo rovnou 10 milionům eur. Střední podnik je podnik s méně než 250 zaměstnanci, ročním obratem menším nebo rovným 50 milionům eur a roční bilanční sumou menší nebo rovnou 43 milionům eur. Do kategorie MSP spadá také mikropodnik, který je definován jako podnik s méně než 10 zaměstnanci, ročním obratem menším nebo rovným 2 milionům eur a roční bilanční sumou taktéž menší nebo rovnou 2 milionům eur. Všechny tyto definice jsou zobrazeny na obrázku č. 14.

[18] *2016 SBA Fact Sheet - Czech Republic* [online]. 1. Europe: Europe Comission, 2016 [cit. 2017-12-14]. Dostupné z: https://ec.europa.eu/growth/smes/business-friendly-environment/performance-review_en#sba-fact-sheets



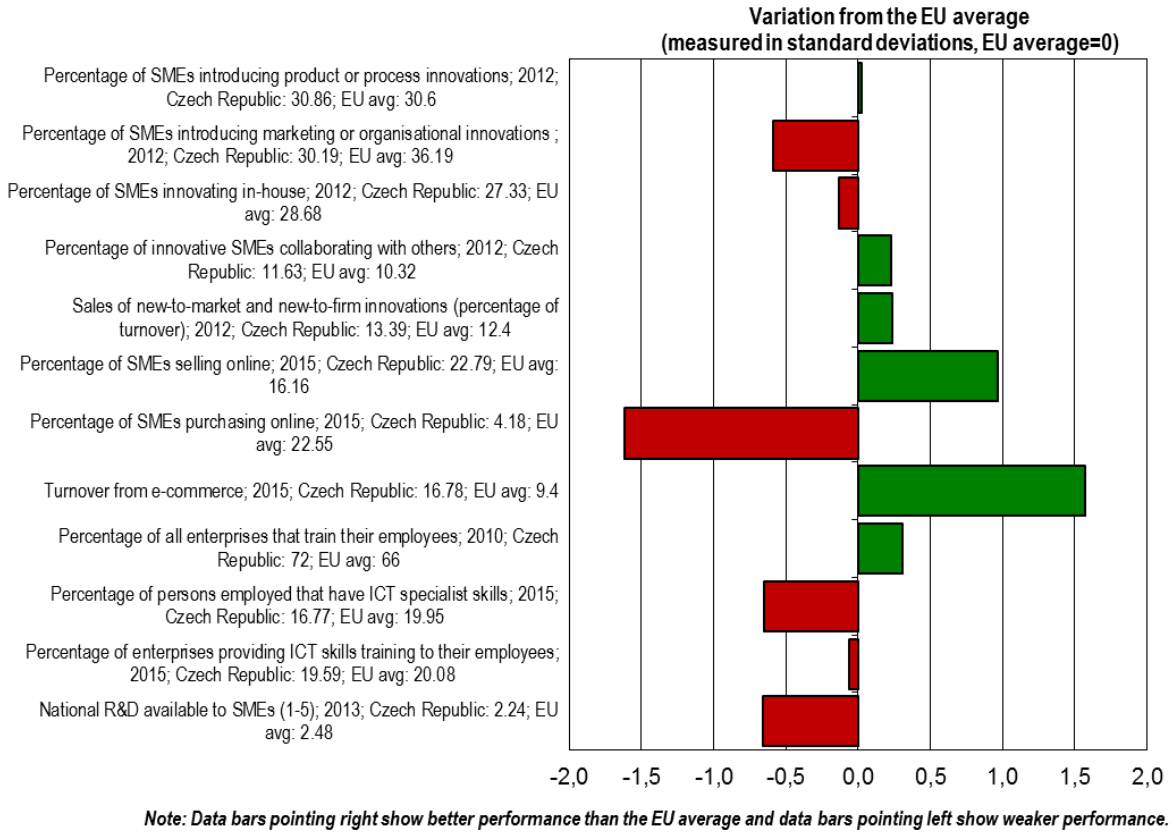

*Obrázek 15: Definice MSP*
Zdroj:http://www.czechinvest.org/data/files/definice-msp-uzivatelska-prirucka-4128cz.pdf

S rozmachem sociálních sítí získal každý podnik další distribuční kanál, kterým může komunikovat širší veřejnosti jak své produkty, tak prezentovat sám sebe. Využívání sociálních sítí také úzce souvisí s tím, jak moc je podnik vyspělý, co se týče ICT. Potenciál dnešní doby ve využívání služeb ICT je velký a proto by podnik měl být také vyspělý v určitých ICT kompetencích. To, jak moc jsou podniky kompetentní v ICT, uvádí statistika vypracovaná Evropskou komisí. Tato statistiky byla vypracována právě s ohledem na kompetence v oblasti ICT, což úzce souvisí s firemním vystupováním online. Evropská komise stanovila určité indikátory, dle kterých lze usoudit jak moc je podnik kompetentní v ICT. Mezi tyto indikátory ICT kompetencí se podle Evropské komise řadí například:

- Prodávání svých produktů online
- Nakupování od dodavatelů online

- Obraty z e-commerce
- Počet ICT specialistů pracujících v podniku
- Vzdělávání zaměstnanců v oblasti IT

Statistika, která je uvedená na obrázku č. 13, porovnává pomocí směrodatné odchylky údaje České republiky a EU průměru. Je patrné, že Česká republika v některých oblastech zaostává za evropským standardem a naopak v některých oblastech ho daleko převyšuje.




*Obrázek 16:ICT kompetence*
Zdroj: ICT competencies [online]. In: . [cit. 2017-03-29]. Dostupné z: https://ec.europa.eu/docsroom/documents/22382.jpg

V prodávání produktů online MSP v České republice převyšují Evropský standard o 6,63%, kdežto v nákupu online naopak zaostávají o více než 18%. To indikuje větší sklon k využívání internetových služeb jako komunikačního kanálu se zákazníky, než jako prostředek k nákupu. Dalším indikátorem, kterým MSP v České republice daleko převyšují evropský průměr je obrat z e-commerce. Tato forma obchodování se stala rychle rostoucím

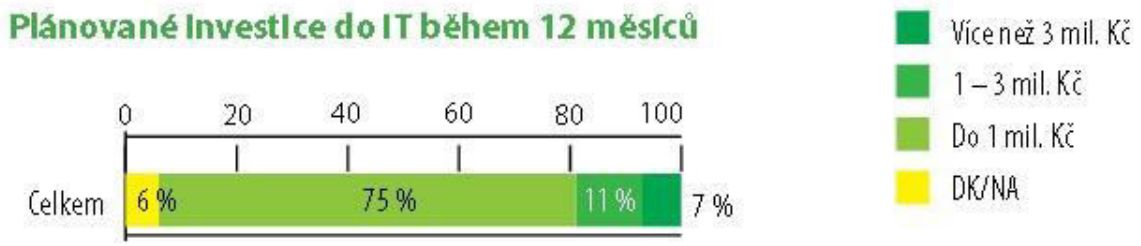
trendem, který přebírá mnoho podniků. Indikátor e-commerce úzce souvisí s nákupem a prodejem online a s využíváním moderních elektronických komunikačních prostředků, pod které spadají mimo jiné také sociální sítě. Počet pracujících ICT specialistů je v České republice naopak nižší o 3,18% než je evropský průměr. Poslední indikátor, kterým je vzdělávání zaměstnanců v oblasti IT, se skoro shoduje s evropským průměrem.

Z této statistiky je patrné, že MSP v České republice nijak nezaostávají za MSP v Evropě. Což může být i důsledek snahy Evropské komise zřídit pro malé a střední podniky v Evropě stejné podmínky.

## 2.2 Informační potřeby MSP

Informační potřeby podniků se v časovém horizontu mění a také jsou odvozovány z konkrétního zaměření činností daného podniku, nicméně jejich základ zůstává stejný. Podniky potřebují informace, které jim pomohou dosáhnout jejich dlouhodobých cílů. Většina MSP již své informační potřeby uspokojuje z informačního zdroje, kterým je internet. Čím dál tím více podniků chce investovat do IT a čím dál tím více podniků je na IT závislá. Dokazuje to průzkum z roku 2014 Asociace malých a středních podniků a živnostníků v ČR, dále jen AMSP, s názvem „*Investice do IT a práce s daty ve firmách"*. Cílem tohoto průzkumu bylo zjistit plánované investice do IT během 12 měsíců a odhalit priority pro budoucí rozvoj. Průzkum ukázal, že přes 36% podniků, je na IT zcela závislé, což znamená, že pokud by došlo k výpadku IT, nebyl by podnik schopen poskytovat základní služby. Dále z průzkumu vyplývá, že 18% podniků plánuje do IT investovat více než 1 milión Kč. Nabízí se tu srovnání s o rok starším průzkumem, který byl totožně strukturovaný. V průzkumu z roku 2013 plánovalo investovat přes 1 milión Kč do IT jen 10% podniků, čili skutečně chtějí podniky do toho odvětví čím dál tím více investovat.[19]

[19] Asociace malých a středních podniků a živnostníků České republiky (AMSP ČR). Investice do IT a práce s daty ve firmách [online]. 1. Praha: AMSP, 2014 [cit. 2017-12-14]. Dostupné z: http://amsp.cz/uploads/Pruzkumy/Vysledky_pruzkumu_Investice_do_IT_a_prace_s_daty_ve_firmach.pdf




*Obrázek 17: Investice do IT - AMSP 2014*
Zdroj:http://www.amsp.cz/uploads/Pruzkumy/Vysledky_pruzkumu_Investice_do_IT_a_prace_s_daty_ve_firmach.pdf

Data ze sociálních sítí se stávají pro podniky čím dál tím cennějším aktivem, které dokáží chytře využívat. Informace ze sociálních sítí stále nevyužívá většina MSP. Jelikož jsou informační potřeby stále vázány na činnosti konkrétních podniků, nemůže dojít ke stavu, že by všechny podniky potřebovali data z těchto zdrojů.

Český statistický úřad vypracoval za rok 2017 statistické šetření s názvem *„Šetření o využívání informačních a komunikačních technologií v podnikatelském sektoru“*. Předmětem tohoto šetření bylo rozšíření a využívání informačních a komunikačních technologií v podnikatelském sektoru. Jako základní soubor byly vybrány právnické a fyzické osoby s deseti a více zaměstnanými osobami ve vybraných odvětvích ekonomické činnosti. Vybranou technikou šetření byly dotazníky rozeslané zpravodajským jednotkám, kde byla možnost vyplnění jak klasického tištěného dotazníku, ale i elektronické verze. Výběrový soubor čítal 7 977 zpravodajských jednotek, které byly vybrány kombinací plošného, záměrného a stratifikovaného náhodného výběru. V tomto šetření se nachází segment sociálních médií, který poukazuje na využívanost tohoto komunikačního kanálu v podnicích.

**Podniky aktivně využívající sociální média**

Podle tohoto šetření, využívalo aktivně sociální sítě 32,7% malých podniků a 46,8% středních podniků. Nicméně je nutné také vzít v úvahu strukturu MSP v ČR, což naštěstí statistika zohledňuje. Nejvíce využívají sociální sítě MSP v odvětví *Informační a komunikační činnosti* (67,2% pro malé podniky a 83,2% pro střední podniky), nejméně pak odvětví *Stavebnictví* (pouze 20,7% pro malé podniky a 26,2% pro střední podniky). Ale ze

statistik je patrné, že každé odvětví nějak sociálně sítě využívá ať už méně aktivně či více. Kompletní statistiku lze nalézt v Příloze A.[20]

**Podniky využívající sociální média ke zlepšování obrazu firmy či uvádění produktů na trh**

Působení podniku na sociálních sítích může značně ovlivnit jeho image. Proto podniky investují nemalé peníze do Public relations, které dokáže s tímto komunikačním kanálem výborně pracovat. Český statistický úřad v rámci šetření vypracoval také statistiku, která se týkala opět podniků v ČR a sociálních sítí, tentokrát se zaměřil na to, kolik podniků v jakém odvětví, využívá sociální sítě pro zlepšování image podniku či uvádění produktů na trh. U malých podniků se největší podíl firem, které využívají sociální sítě ke zlepšení image či uvádění produktů na trh, nachází v odvětví *Informační a komunikační činnosti* (57,7%), na druhém místě je pak odvětví *Ubytování, stravování a pohostinství* (51,4%). U středních podniků je situace podobná. Největší podíl firem využívající sociální sítě ke zlepšení image či uvádění produktů na trh má také odvětví *Informační a komunikační činnosti* (75,7%). Je nutno podotknout, že u velkých podniků má 100% zastoupení, ve využívání sociálních médií ke zlepšování obrazu firmy či uvádění produktů na trh, odvětví *Ubytování,* viz Příloha B[20].

**Podniky využívající sociální média k získávání názorů/otázek od zákazníků**

Další vypracovanou statistikou v rámci šetření Českého statistického úřadu bylo, zda podniky využívají sociální sítě k získávání názoru/otázek od zákazníků. U malých podniků je největší podíl v odvětví *Ubytování, stravování a pohostinství* (43,5%). U středních podniků je to také v odvětví *Ubytování, stravování a pohostinství* (68,3%), viz Příloha C[20].

Z tohoto šetření je patrné, že informační potřeby ze sociálních sítí jsou přímo závislé na odvětví, ve kterém podnik působí. Také lze vyvodit závěr, že čím větší je podnik, tím více

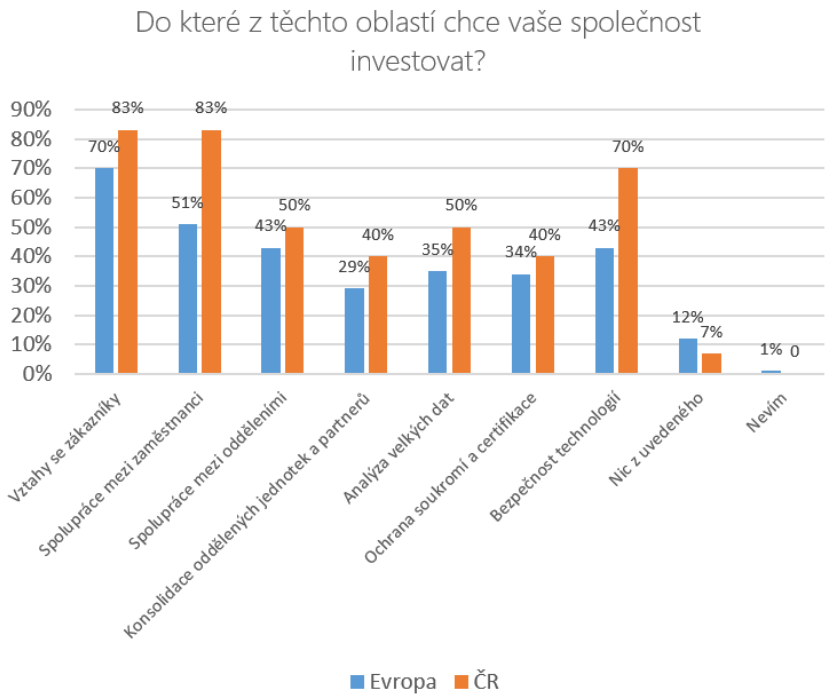
[20] ČESKÝ STATISTICKÝ ÚŘAD. *Využívání informačních a komunikačních technologií v podnikatelském sektoru - v roce 2017* [online]. 1. Praha: Odbor statistik rozvoje společnosti, 2017 [cit. 2017-12-14]. Dostupné z: https://www.czso.cz/csu/czso/vyuzivani-informacnich-a-komunikacnich-technologii-v-podnikatelskem-sektoru-2016-2017

sociální média využívá. MSP v ČR začínají postupně sociální média využívat, což potvrzuje i jeden ze závěrů šetření, že od roku 2013 se podíl firem, které mají profil na některé sociální síti, více než zdvojnásobil[21]. Opět to ale záleží na odvětví daného podniku, jelikož né všechny odvětví sociální sítě plně využíjí. Podniky v těch odvětvích, kde sociální sítě aktivně využívají, mají hlavně zájem získávat ze sociálních sítí informace o zákaznících.

Trend Big Data se nedotýká pouze odvětví Informačních technologií, ale dotýká se nepřímo všech odvětví, která jakýmkoli způsobem pracují s daty. O velké objemy dat, především pak o jejich zpracování, se zajímá čím dál tím více podniků a mají zájem do tohoto trendu investovat. Průzkum společnosti Microsoft s názvem *„Náskok díky technologiím"*, který byl realizován v roce 2016, měl za úkol zmapovat investice MSP v Evropě. Sběr dat probíhal prostřednictvím rozhovorů s respondenty ve 20 evropských zemích. Průzkumu se zúčastnili jen MSP. Jak je vidět na obrázku níže, tak do analýzy velkých dat by chtělo investovat 35% MSP z evropských zemí. V České republice jeví o tento trend značně větší zájem a chce do něj investovat 50% MSP. Průzkum mimo jiné poukazuje i na to, že většina podniků chce investovat do vztahu se zákazníkem, což se také týká sociálních sítí.[22]

[21] ČESKÝ STATISTICKÝ ÚŘAD. *Využívání informačních a komunikačních technologií v podnikatelském sektoru - v roce 2017* [online]. 1. Praha: Odbor statistik rozvoje společnosti, 2017 [cit. 2017-12-14]. Dostupné z: https://www.czso.cz/csu/czso/vyuzivani-informacnich-a-komunikacnich-technologii-v-podnikatelskem-sektoru-2016-2017

[22] MICROSOFT. *Náskok díky technologiím* [online]. 1. Praha: Ipsos Mori, 2016 [cit. 2017-12-14]. Dostupné z: https://news.microsoft.com/cs-cz/2016/09/02/ceske-male-a-stredni-podniky-chteji-investovat-vyrazne-vic-nez-firmy-v-ostatnich-evropskych-zemich/




*Obrázek 18: Náskok díky technologiím – investice*
Zdroj: https://ncmedia.azureedge.net/ncmedia/2016/09/SMB-investuji.png

Umění pracovat s velkými daty se stává klíčovou záležitostí už také pro MSP, protože díky tomu podniky získávají nové zákazníky a zvyšují svou produktivitu v rámci obchodu. Je tedy důležité, aby znaly možnosti analýzy velkých dat, potažmo nástroje, které toto umožnují, viz kapitola Nástroje pro analýzu Big Data - charakteristika, porovnání.

# 3. Nástroje pro analýzu Big Data

Analýza Big Data je trendem posledních let. Na trhu již existuje mnoho nástrojů, které umožňují analýzu velkých objemů dat. Většina z nich je postavená na výše zmiňovaném open-source frameworku Hadoop. Vzhledem k tomu, že nástroje mají být určeny pro analýzu Big Data v MSP, tak pro jejich výběr byla stanovena následující kritéria:

- Zpracování Big Data – platforma umožňuje zpracování velkých objemů dat
- Nástroj je k dispozici v nějaké podobě zdarma
- Možnost zprovoznění nástroje na PC
- Umožnění načtení vstupních dat ve formátu CSV

Na základě těchto kritérií byly vybrány nástroje Databricks, Splunk, Hortonworks Data Platform a Cloudera Data Hub. Všechny tyto nástroje nějakým způsobem splňují daná kritéria.

## 3.1 Databricks

Společnost Databricks byla založena v roce 2013 lidmi, kteří vytvořili Apache Spark. Jedním ze zakladatelů je Matei Zaharia, profesor na Standford University. Společnost vytvořila platformu, která využívá webové rozhraní spolupracující se Sparkem a poskytuje uživatelům automatizovaný cluster management.[23]

**Charakteristika nástroje**

Databricks je platforma, která je zaměřená čistě na Big Data. Je postavená nad frameworkem Apache Spark, který umožňuje rychlé a sofistikované analýzy dat. Databricks je čistě cloudová platforma, která využívá hosting od Amazonu. Databricks umožňuje svým zákazníkům pracovat s clustery v cloudu, přidávat a odebírat je přímo podle jejich potřeb. Každý zákazník má tedy svůj workspace, kde může provádět své analýzy. Dále pak

[23] Databricks Team. Databricks Platform. [online] 2015 [cit. 2017-03-21]. Dostupné z: https://databricks.com/product/unified-analytics-platform

poskytuje interaktivní vestavěný editor pro psaní příkazů v jazyce Python, Scala nebo SQL. Databricks umožňuje import dat z Hadoop, AWS, relačních databází a také z NoSQL databází. V rámci Databricks jsou tzv. Databricks Units (DBU), jedná se o jednotky, které určují výkonnost jeho řešení. Jedna DBU je typicky 30 GB operační paměť a 4 jádra. Zákazníci si pak určují kolik DBU ve svém řešení chtějí zahrnout.

**Cena nástroje**

Společnost Databricks nabízí 4 řady, které se liší úložným prostorem, počtem uživatelů a nabízenými službami. Jedná se o řady:

- Community
- Starter
- Professional
- Enterprise

Řada Community je zcela zdarma a nabízí 6GB cluster, vestavěný editor s podporou jazyků Python, Scala a SQL, možnost vytváření interaktivních vizualizací a také možnost vybrání si z několika verzí Apache Sparku.

V řadě Starter je vše co v řadě Community s tím rozdílem, že je neomezený počet clusterů (tedy i jejich velikostí), přidáné REST API a integrace IDE. Cena je 99$ měsíčně plus 0.40$ za každou DBU jednotku.

V řadách Professional a Enterprise je pak nabízena plná škála služeb, včetně podpory, GitHub integrace, neomezeného počtu uživatelů a s tím volitelná struktura rolí a práv a v neposlední řadě také automatické škálování clusterů. Tato řešení již nemají pevně danou cenu, ale vždy se to bude odvíjet od robustnosti a individuálních potřeb daného řešení.

## 3.2 Splunk

Společnost Splunk byla založena v roce 2002 R. Dasem a E. Swanem. První verze tohoto softwaru se objevila v roce 2004 a stala se velmi populární, tudíž si podniky začaly kupovat

Enterprise licence Splunku, což umožnilo společnosti růst. Splunk poskytuje svůj software více než 6000 společnostem ve více než 90 zemích. Mezi významné klienty patří Coca-cola, Adobe, Bosch, Ubisoft, Valve a mnoho dalších.

**Charakteristika nástroje**

Jedná se o softwarovou platformu, která umožňuje vyhledávání, analyzování, monitorování a vizualizaci Big Data. Splunk umožňuje import dat z různých zdrojů a je v tomto ohledu velmi flexibilní. Hlavní silou platformy je zpracování strojových dat, kam se zahrnují všechny aktivity aplikací. Silnou stránkou této platformy jsou customizované aplikace, které uživatelům pomáhají řešit specifické problémy. Dále je tato platforma schopna indexovat velkou škálu různorodých dat (logy, konfigurace, zprávy, skripty, různé statistiky…). Platforma Splunk využívá k práci s daty jazyk SPL (Search Processing Langugage), který obsahuje přes 140 commandů na prohledávání, korelaci, analýzu a vizualizaci dat.

**Cena nástroje**

Splunk nabízí dvě hlavní řady produktů, které se liší svou robustností a řešením. Prvním produktem je Splunk Light. Jedná se o produkt uzpůsobený pro menší společnosti, které nebudou denně potřebovat analyzovat obrovské množství dat. Splunk Light umožňuje snadnou analýzu, reportování a monitorování dat v reálném čase na jednom místě. Cena se odvíjí od denního zpracování dat společnosti. Pokud například společnost denně zpracuje (indexuje) 1 GB dat, tak ročně za produkt zaplatí 1035$. Druhým produktem je pak Splunk Enterprise, který je mnohem robustnější než Light a také v sobě nese více funkčností a služeb. Cena se opět odvíjí od objemu dat, které budou indexovány. V řadě Enterprise je to za 500MB denně, 6000$ ročně. Obě tato řešení Splunk nabízí také jako cloud, kde poskytuje 100% SLA. Splunk nabízí i verzi zdarma, která je omezená 500MB indexace.

## 3.3 Hortonworks

Jedná se o společnost, která nabízí platformu schopnou zpracovávat Big Data. Hortonworks byla založena v roce 2011 bývalými zaměstnanci společnosti Yahoo!. Nyní má společnost

přes 1000 zaměstnanců a poskytuje své služby ve více než 17 zemích světa. Mezi významné zákazníky pak patří například T-Mobile, Pandora, Samsung, E-bay, Spotify a další.

**Charakteristika nástroje**

Hortonworks poskytuje open-source platformu (Hortonworks Data Platform – HDP), která je postavena nad Hadoopem. Poskytuje dva primární produkty. Jedním z nich je Hortonworks Data Flow (HDF) a druhým výše zmiňovaná platforma HDP. HDF umožňuje sbírat, konsolidovat a dodávat data v reálném čase z Internetu věcí (IoT). To znamená, že streamovaná data nejsou pro tuto platformu cizí. Platforma HDP pak umožňuje společnostem vytvořit zabezpečený data lake, z kterého pak podnik může čerpat veškeré informace, provádět různé analýzy a vizualizace. Klíčovými komponentami HDP je Hadoop File System a YARN, viz kapitola Big Data.

**Cena nástroje**

Platforma HDP je zcela zdarma a není potřeba platit žádné poplatky. Podnik účtuje poplatky formou předplatného pouze za podporu, konzultace nebo školení.

## 3.4 Cloudera

Společnost Cloudera byla založena v roce 2008 C. Bisciglіem, A. Awadallahem, M. Olsonem a J. Hammerbacherem, kteří pracovali v Sillicon Valley pro velké společnosti jako Google, Yahoo!, Facebook nebo Oracle. Cloudtera je postavena nad Apache Hadoopem a poskytuje svým klientům velké know-how, technickou podporu, konzultace a školení. Cloudtera má přes 1600 zaměstnanců a své produkty poskytuje ve 28 zemích světa. Mezi hlavní zákazníky této společnosti patří Samsung, Cisco, MBank a další.

**Charakteristika nástroje**

Cloudera je moderní platforma vyznačující se zaměřením na Big Data. Tato platforma je opět postavena nad Apache Hadoopem. Cloudera se řadí mezi open-source software, nicméně nabízí komerční verzi Cloudera Enterprise Data Hub (CDH), která zahrnuje

management aplikace, integrace se systémy, knowledge base a podporu 24 hodin denně. CEDH umožňuje snadnější správu clusterů, které běží nad open-source Clouderou.

**Cena nástroje**

Cloudera je open-source nástroj a uživatel platí za její nadstavby, v podobě lepšího managementu nebo pak pokud má zájem o cloud. Cloudera poskytuje zdarma sandboxovou verzi – QuickStart VM, kde uživatel má k dispozici cluster s jedním nodem.

## 3.5 Porovnání nástrojů

Pro porovnání nástrojů byla zvolena metoda vícekriteriálního hodnocení. Hlavní výhodou této metody je zohlednění více kritérií při výběru vhodné varianty. Kriteria pro porovnání nástrojů byla stanovena tak, aby zohledňovala informační potřeby a možnosti MSP, obdobně jako kritéria výběru nástrojů.

### 3.5.1 Kritéria hodnocení platformy

S ohledem na potřeby MSP a na základě informací získaných z publikací jako je například *„Big Data Vendor Benchmark 2015“* od společnosti Experton Group[24] nebo publikovaných článků o srovnávání platforem jako je „*Comparing the leading big data analytics software options*“ od D. Loshina, byla stanovena níže uvedená kritéria.[25]

- Deployment model
- Dostupnost nástroje zdarma

[24] LANDROCK, Holm, Oliver SCHONSCHEK a Prof. Dr. Andreas GADATSCH. *Big Data Vendor Benchmark 2015* [online]. In: . Mnichov, Německo, D2015, s. 91 [cit. 2017-05-04]. Dostupné z: https://www.t-systems.com/solutions/big-data-vendor-download/1298900_1/blobBinary/Big+Data+Vendor_Download-ps.pdf

[25] LOSHIN, David. *Comparing the leading big data analytics software options* [online]. 2015 [cit. 2017-05-04]. Dostupné z: http://searchbusinessanalytics.techtarget.com/feature/Comparing-the-leading-big-data-analytics-software-options

- Funkcionalita
- Uživatelská přívětivost

**Deployment model**

Stanovené kritérium bude hodnotit jaký deployment model umožňuje využívat daná platforma, zda nabízí pouze on-premise variantu, nebo umožňuje uživateli využívat cloudové řešení.

**Dostupnost nástroje zdarma**

Jedním z hlavních kritérií bude dostupnost verze daného nástroje zdarma. Jelikož MSP mají oproti velkým korporacím značně omezené rozpočty, je potřeba hledat alternativy, které mohou poskytnout co nejlepší funkcionalitu vzhledem k vynaloženým nákladům. Hodnoceno tedy bude, zda nástroj lze využívat zdarma a za jakých podmínek.

**Funkcionalita**

Dalším kritériem je funkcionalita. Toto kritérium bylo zvoleno také jako jedno z klíčových. Jelikož každý nástroj může mít trošku odlišné funkčnosti, tak toto kritérium se skládá z dílčích subkritérií - kroků, kdy každý nástroj analyzoval stejný datový set a to pomocí následujících kroků.

1. Nahrání datového setu do platformy
2. Jednoduchá datová analýza – kolik tweetů je negativních a kolik pozitivních
3. Textová analýza – výskyt slova „Twitter“

Pro dané úlohy se bude měřit i doba trvání jednotlivých úloh, což by se dalo také označit za výkonnost. Datová analýza - kolik tweetů je negativních a kolik pozitivních probíhala vyhodnocením příslušného příznaku v datasetu. Textová analýza pak probíhala analýzou všech tweetů, zda obsahují slovo „Twitter“.

**Uživatelská přívětivost**

Uživatelská přívětivost je také jedním z kritérií, která jsou nezbytně nutná při hodnocení aplikací, potažmo softwaru obecně. Software se dělá za účelem zefektivnění nějakého procesu, což má za následek celkové zvýšení efektivity podniku. A jelikož je nezbytné, aby software obsluhoval uživatel, je klíčové, aby tomu rozuměl a bylo pro něj jednoduché se v prostředí softwaru pohybovat. V rámci tohoto kritéria je také přihlédnuto k dokumentaci daného řešení.

### 3.5.2 Stanovení vah kritérií

Jelikož se jedná o porovnání na základě vícekriteriálního rozhodování, je nezbytně nutné stanovit váhy jednotlivým kritériím, jelikož se důležitosti kritérií liší. Váhy byly stanoveny pomocí bodovací metody určování vah kritérií. Tato metoda umožňuje diferencovanější vyjádření subjektivních preferencí, než například metoda pořadí. Bylo tedy nutné kvantitativně ohodnotit jednotlivá kritéria a následně využít níže uvedený vzorec k výpočtu vah, kde $b_i$ označuje zvolenou stupnici, v tomto případě $b_i \in < 0{,}10 >$.

$$v_i = \frac{b_i}{\sum_{i=1}^{k} b_i}$$

*Tabulka 3: Váhy kritérií*

| Kritérium | Číslo kritéria | Body | Výsledná váha |
|---|---|---|---|
| **Deployment model** | 1 | 6 | 0,230 |
| **Dostupnost nástroje zdarma** | 2 | 8 | 0,308 |
| **Funkcionalita** | 3 | 8 | 0,308 |
| Nahrání datového setu | 3.1 | 2 | 0,250 |
| Jednoduchá datová analýza | 3.2 | 2 | 0,250 |
| Textová analýza | 3.3 | 4 | 0,500 |
| **Uživatelská přívětivost** | 4 | 4 | 0,154 |
| **Celkem** | | **26** | **1** |

Zdroj: vlastní

Největší váhu získala kritéria dostupnost nástroje zdarma a funkcionalita. Dostupnost nástroje zdarma je důležitým aspektem vzhledem k tomu, že se jedná o nástroj určený pro MSP, které často disponují s omezenými zdroji a větší finanční investice do nástroje pro analýzu dat, kde není jasná návratnost, je tak nemyslitelná. Funkcionalita je také klíčové kritérium, jelikož nástroj musí umět analyzovat data tak, aby to mělo pro podnik přidanou hodnotu. Nejmenší váhu získalo kritérium Uživatelská přívětivost. To však neznamená, že by nebylo důležité, ale v porovnáním s ostatními kritérii se jeví jako nejméně významné.

### 3.5.3 Hodnocení dle kritérií

#### 3.5.3.1 Deployment model

Prvním hodnoceným kritériem je technologické řešení, neboli „deployment model“. Technologické řešení, na kterém je platforma postavená. V rámci těchto čyř nástrojů, které byly vybrány, jsou aplikována řešení buď on-premise nebo cloud. Každý z těchto modelů má svá pro a proti, které by MSP měly při výběru zohlednit. Hlavní výhodou on-premise řešení je bezpečnost privátních dat. Různé podniky musejí kvůli určitým regulacím využívat jen toto řešení, protože není ze zákona možné mít uložená privátní data klientů. Nicméně bezpečnost je nyní mnohdy větší u cloudového řešení, ale některé pokročilejší a customizované bezpečnostní prvky v cloudových řešení nejsou. Další výhodou může být přístup k datům. Problémy u cloudového řešení mohou nastat například z důvodu dočasného nepřístupu služby, na druhou stranu se v rámci cloudu lze k datům dostat odkudkoli a kdykoli a odpadají starosti o infrastrukturu. Platforma Databricks nabízí čistě cloudové řešení. Splunk, Hortonworks a Cloudera mají ve své nabídce jak on-premise variantu, tak cloudové řešení.

#### 3.5.3.2 Dostupnost nástroje zdarma

Platforma Databricks nabízí verzi zdarma v prostředí cloudu s šesti gigabytovým mikro-clusterem s interaktivním prostředím. Splunk, Hortonworks a Cloudera nabízejí verze zdarma pouze v on-premise variantě. U platformy Splunk je omezena indexace dat na 500MB za den. Platformy Hortonworks a Cloudera mají jeden cluster s jedním pracovním nodem.

#### 3.5.3.3 Funkcionalita

Pro zjištění funkcionality byla jako datový set vybrána Analýza sentimentu ze sociální sítě Twitter, kterou vypracoval N. Sanders. Datový set je volně k dispozici na webové platformě Kaggle, kam různé společnosti a především datoví analytici dávají své analýzy, volné datové sety atd. Tento datový set obsahuje přesně 1 578 627 tweetů v anglickém jazyce. Každý řádek obsahuje ItemID, což je ID tweetu, Sentiment ID, daný tweet a indikátor zda je daný tweet pozitivní nebo negativní. U každého nástroje bylo postupně hodnoceno nahrání datového setu do platformy, jednoduchá datový analýza a textová analýza.

## Databricks

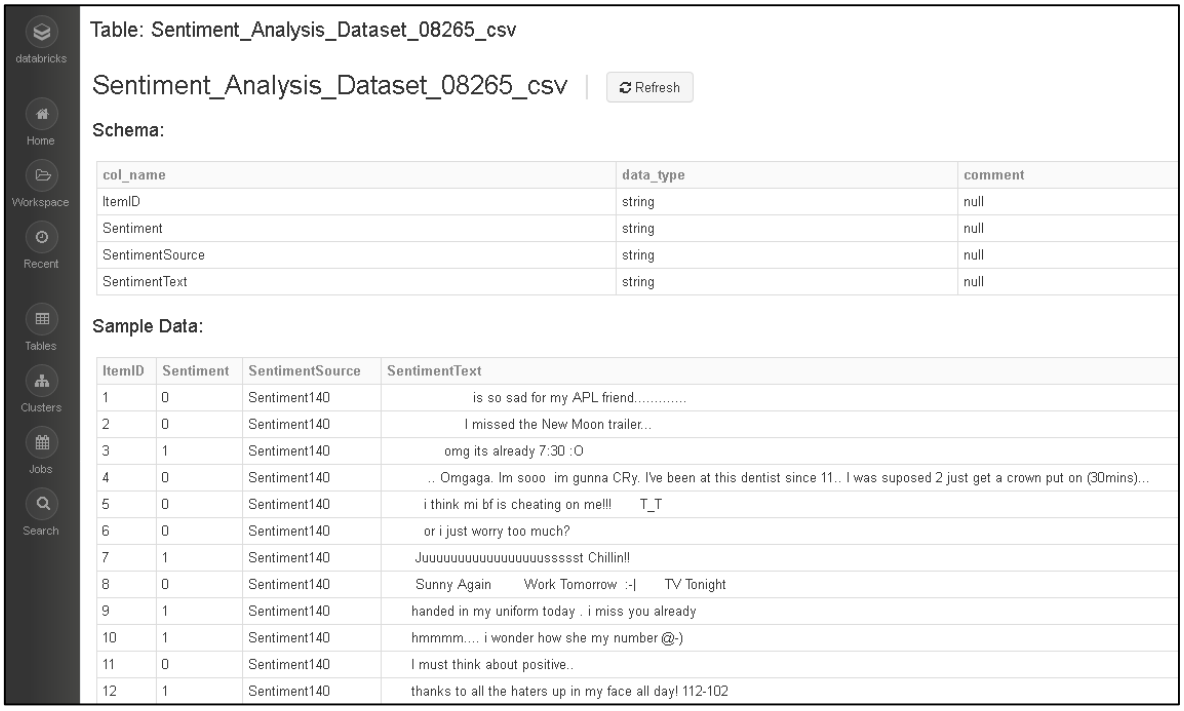
Platforma umožňuje uživateli vytvářet on-demand clustery na kterých je framework Apache Spark. Jde zde předinstalovaná Java, Python a Scala, což umožňuje vývoj skriptů nad vstupními daty. Uživateli je dále umožněno vytvářet klasické tabulky a pomocí dotazovacího jazyka SQL analyzovat data.

**Nahrání datového setu do platformy**

Nahrání datového setu do prostředí bylo velice jednoduché a intuitivní. Platforma podporuje funkčnost Drag&Drop, tudíž stačilo datový set jen přetáhnout a okamžitě začalo nahrávání. Soubor o velikosti 149 MB při rychlosti uploadu 60 Mbps trval 7 minut a 25 sekund.

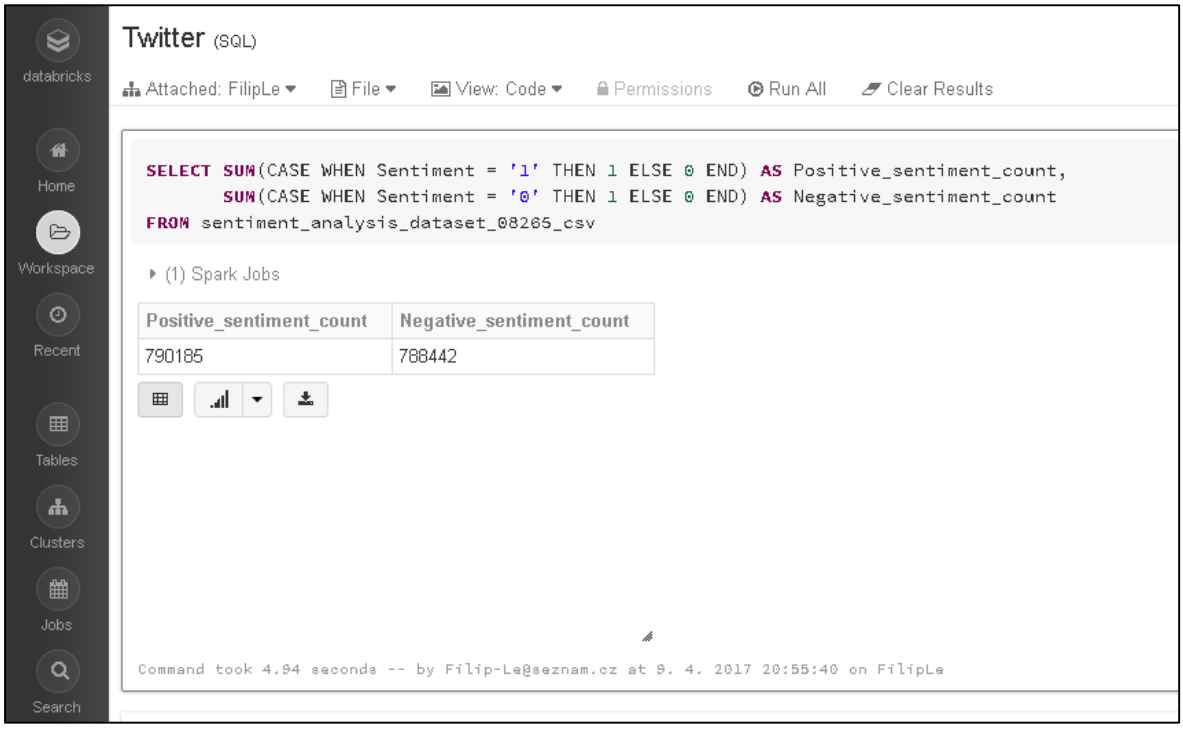
**Jednoduchá datová analýza – kolik tweetů je negativních a kolik pozitivních**

Platforma Databricks podporuje okamžitý převod dat do tabulky. Datový set ve formátu CSV tedy šlo okamžitě vložit do tabulky a platforma automaticky vytvořila schéma a naindexovala sloupce, které se při nahrání souboru zvolily jako hlavička. Na obrázku níže lze vidět strukturu vytvořeného schématu a vzorek dat v tabulce.




*Obrázek 19: Databricks - vytvoření tabulky*
Zdroj: vlastní

Pro zjištění kolik je pozitivních a negativních tweetů byl použít příznak ve sloupci Sentiment, pro pozitivní tweet je hodnota 1 a pro negativní 0. Databricks podporuje dotazovací jazyk SQL, tudíž byl napsán dotaz do tabulky, který sečetl všechny pozitivní a negativní tweety. SQL příkaz běžel 4.94 sekund. Celý SQL příkaz a výsledky jsou zobrazeny na následujícím obrázku.




*Obrázek 20: Databricks - analýza pozitivních a negativních tweetů*
Zdroj: vlastní

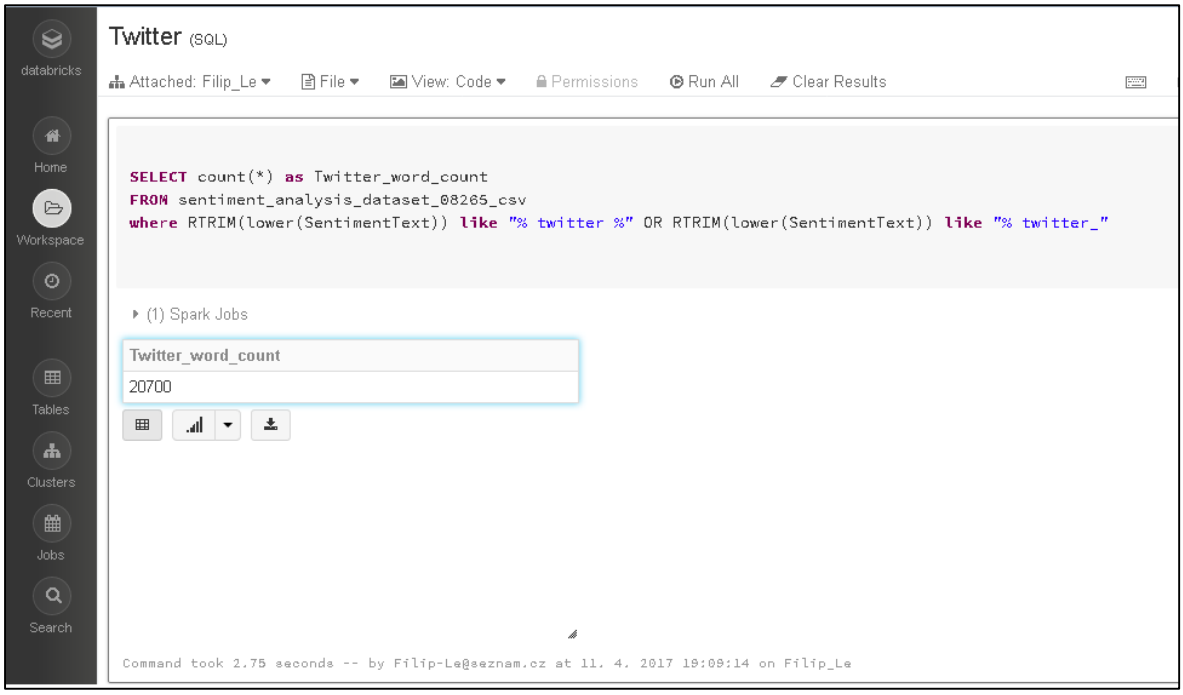
### Textová analýza – výskyt slova „Twitter“

Pro textovou analýzu v Databricks je možné zvolit více postupů. Prvním postupem je analýza s využitím programovacího jazyka Python. Napsaný algoritmus nejdříve načte dataset ve formátu CSV a uloží ho jako formát dataframe. Následně každý tweet uloží do pole. Poté dochází k iterování přes pole, kde se každé slovo ve větě analyzuje, zda to není slovo „Twitter“, pokud ano, tak se přičte jednička do vydefinované proměnné. Analýza běžela 5 minut a 51 sekund, zanalyzovalo se přes 1 500 000 tweetů a slovo „Twitter“ se vyskytovalo 20680 krát. Celý kód je vidět na následujícím obrázku.


*Obrázek 21: Databricks – Wordcount*
Zdroj: vlastní

Z obrázku je vidět, že algoritmus přeskočil dva tweety, jelikož nebyl validní počet sloupců. Což vzhledem k velkému objemu dat není relevantní. Tento způsob analýzy není nejvhodnější, jelikož zde probíhají operace navíc, které zatěžují cluster. Nicméně poukazuje to na určitou dynamičnost nástroje, kde složitější algoritmy budou probíhat právě s využitím jazyka Python, nikoli SQL.

Dalším způsobem je klasický SQL dotaz do tabulky. Tento způsob se zdá sofistikovanější a rychlejší. Dotaz proběhl za 2,75 sekundy a slovo „Twitter“ bylo nalezeno 20700krát.




*Obrázek 22: Databricks - WordCount SQL*
Zdroj: vlastní

Na platformě Databricks nebylo komplikované provést výše uvedené úlohy. Podpora více programovacích jazyků dává uživateli možnost zvolit nejvhodnější variantu pro analýzu Big Data. Databricks dále také umožňují vizualizaci dat přímo v prostředí.

# Splunk

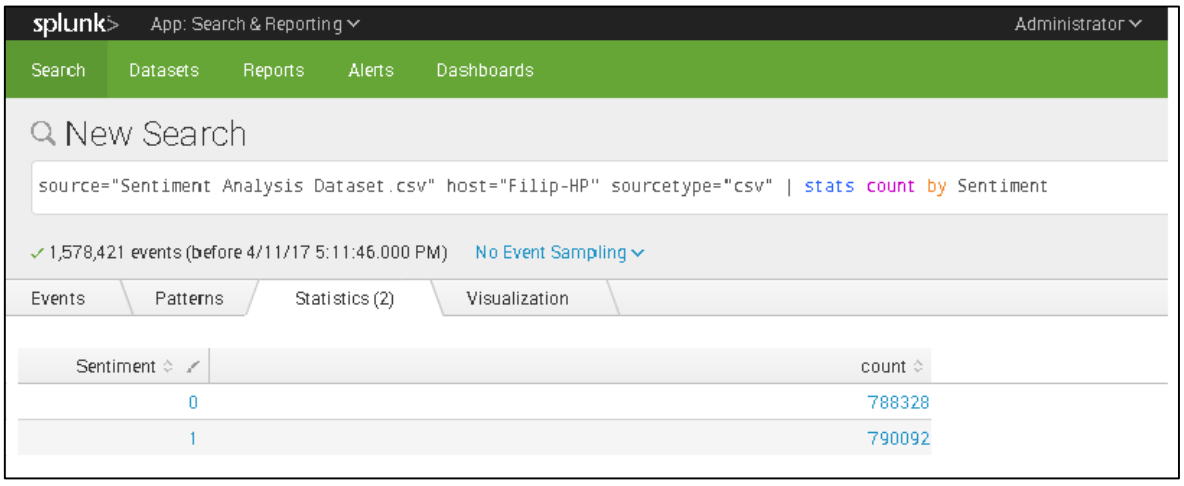
Platforma je primárně založena na indexaci dat, která následně umožňuje rychlou analýzu dat. K indexování využívá Splunk tzv. Peer nodes. Peer node přijme data a následně je nejen naindexuje, ale také je replikuje na ostatní nody. Pro vyhledávání dat se pak využívá komponenta Search head, která spouští vyhledávání indexů napříč nody. V rámci samotné indexace dochází také k normalizaci dat, což zajistí konzistenci při reportování či analýze. K dotazování využívá Splunk vlastní dotazovací jazyk postavený na syntaxi SQL s Splunk Processing Langugage, dále jen SPL. SPL umožňuje prohledávání dat, jejich modifikaci, update i mazání.

**Nahrání datového setu do platformy**

Prostředí Splunk podporuje nahrání různorodých dat, tudíž nebyl problém při nahrání datového setu ve formátu CSV. Vzhledem k tomu, že verze zdarma od Splunk (Splunk Enterprise Free) je on-premise software, tak se jednalo jen o přesun souboru, což trvalo 8 sekund. Platforma Splunk po nahrání externích dat vyžaduje specifikování typu dat (platforma sama přednastaví doporučovaný typ). Jsou zde možnosti jako strukturovaná, nestrukturovaná data nebo také přímo datový typ, např. CSV. Splunk díky specifikaci datového typu dokáže efektivně normalizovat a kategorizovat data při samotné indexaci. Po nahrání datového setu a následné specifikaci dat musí dojít k indexaci. Samotná indexace trvala 45 sekund, tudíž nahrání datového setu i s indexací trval celkově 50 sekund.

**Jednoduchá datová analýza – kolik tweetů je negativních a kolik pozitivních**

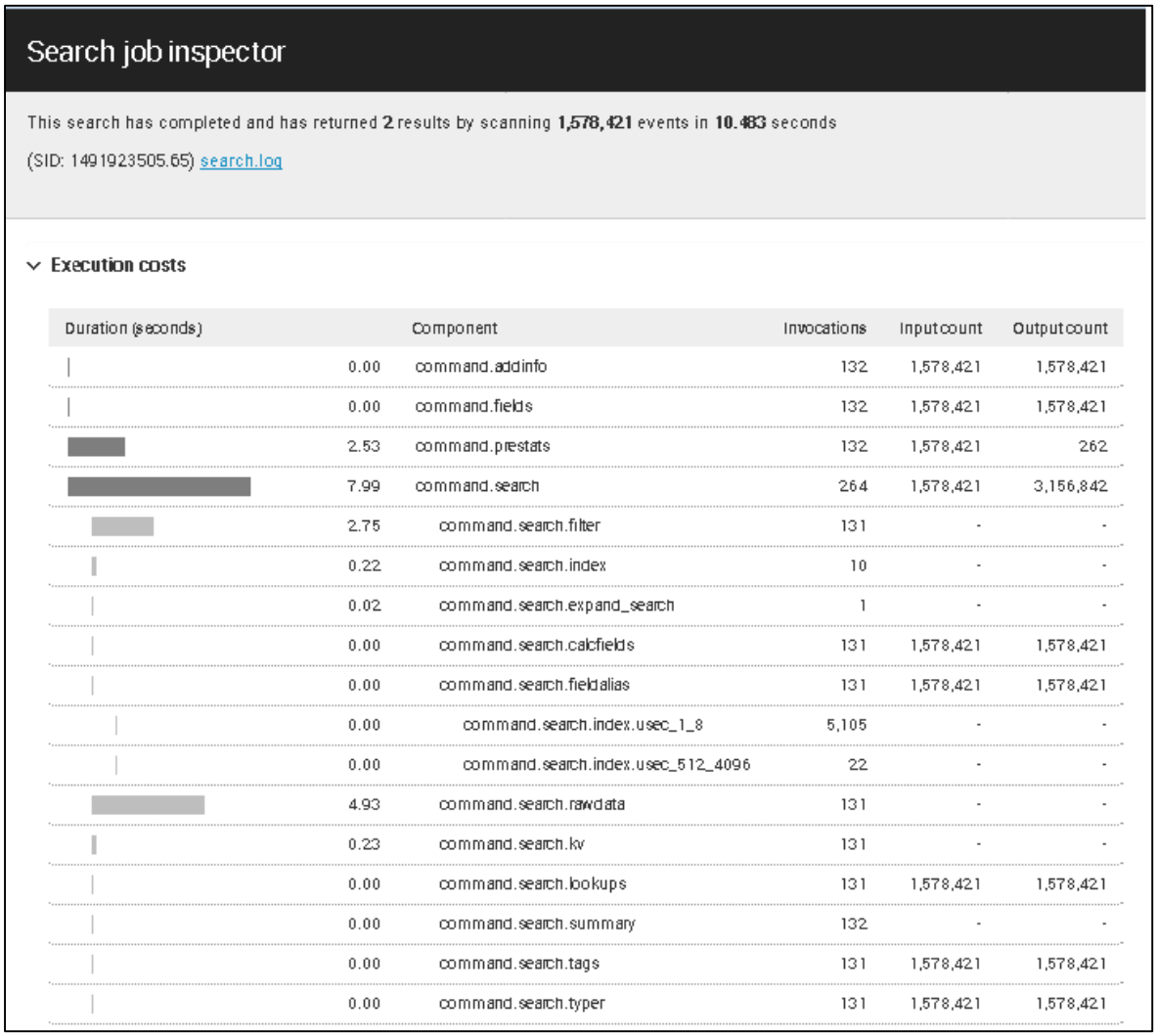
Samotná datová analýza kolik tweetů je negativních a kolik pozitivních je pak velmi jednoduchá, jelikož pomocí SPL specifikujeme hodnotu daného atributu a zabere to minimální čas díky předešlé indexaci dat. Vzhledem k tomu, že jsou již data naindexovaná, analýza kolik tweetů je negativních a kolik pozitivních zabrala jen pár sekund. Využit byl jazyk SPL, příkaz a výsledné počty jsou vidět na následujícím obrázku.




*Obrázek 23: Splunk - analýza pozitivních a negativních tweetů*
Zdroj: vlastní

Z obrázku lze vidět, že bylo zpracováno 1 578 421 eventů, což je o 205 záznamů méně, než obsahuje datový set. To je způsobené normalizací a kategorizací při indexování. Příkaz trval

přesně 10,483 sekund. Splunk poskytuje také funkčnost Job Inspector, díky které lze jasně dohledat, jak dlouho jednotlivé příkazy běžely. Na základě toho se dají dotazy optimalizovat, což přináší efektivnější a ve výsledku méně nákladné řešení.



Search job inspector

This search has completed and has returned **2** results by scanning **1,578,421** events in **10.483** seconds

(SID: 1491923505.65) search.log

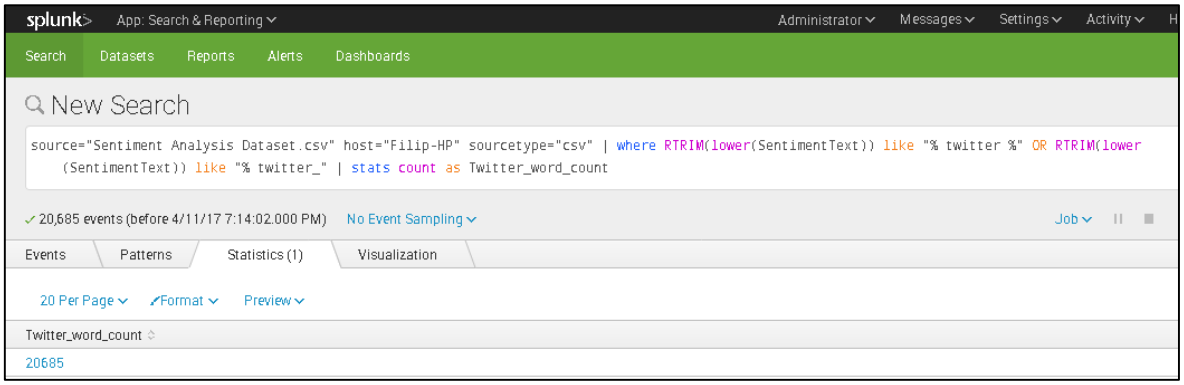
**Execution costs**

| Duration (seconds) | Component | Invocations | Input count | Output count |
|---|---|---|---|---|
| 0.00 | command.addinfo | 132 | 1,578,421 | 1,578,421 |
| 0.00 | command.fields | 132 | 1,578,421 | 1,578,421 |
| 2.53 | command.prestats | 132 | 1,578,421 | 262 |
| 7.99 | command.search | 264 | 1,578,421 | 3,156,842 |
| 2.75 | command.search.filter | 131 | - | - |
| 0.22 | command.search.index | 10 | - | - |
| 0.02 | command.search.expand_search | 1 | - | - |
| 0.00 | command.search.calcfields | 131 | 1,578,421 | 1,578,421 |
| 0.00 | command.search.fieldalias | 131 | 1,578,421 | 1,578,421 |
| 0.00 | command.search.index.usec_1_8 | 5,105 | - | - |
| 0.00 | command.search.index.usec_512_4096 | 22 | - | - |
| 4.93 | command.search.rawdata | 131 | - | - |
| 0.23 | command.search.kv | 131 | - | - |
| 0.00 | command.search.lookups | 131 | 1,578,421 | 1,578,421 |
| 0.00 | command.search.summary | 132 | - | - |
| 0.00 | command.search.tags | 131 | 1,578,421 | 1,578,421 |
| 0.00 | command.search.typer | 131 | 1,578,421 | 1,578,421 |

*Obrázek 24: Splunk - Job Inspector*
Zdroj: vlastní

### Textová analýza – výskyt slova „Twitter“

Textová analýza probíhala opět za pomocí jazyka SPL. Kód byl zvolený totožný jako u platformy Databricks, jelikož SPL je založen na jazyku SQL, dala se využít stejná where podmínka. Příkaz trval přesně 17,18 sekund a vyhodnotil, že slovo „Twitter“ se v datovém setu vyskytuje 20 685 krát. Rozdílnost je způsobená úpravou datového setu při indexaci.




*Obrázek 25: Splunk – WordCount*
Zdroj: vlastní

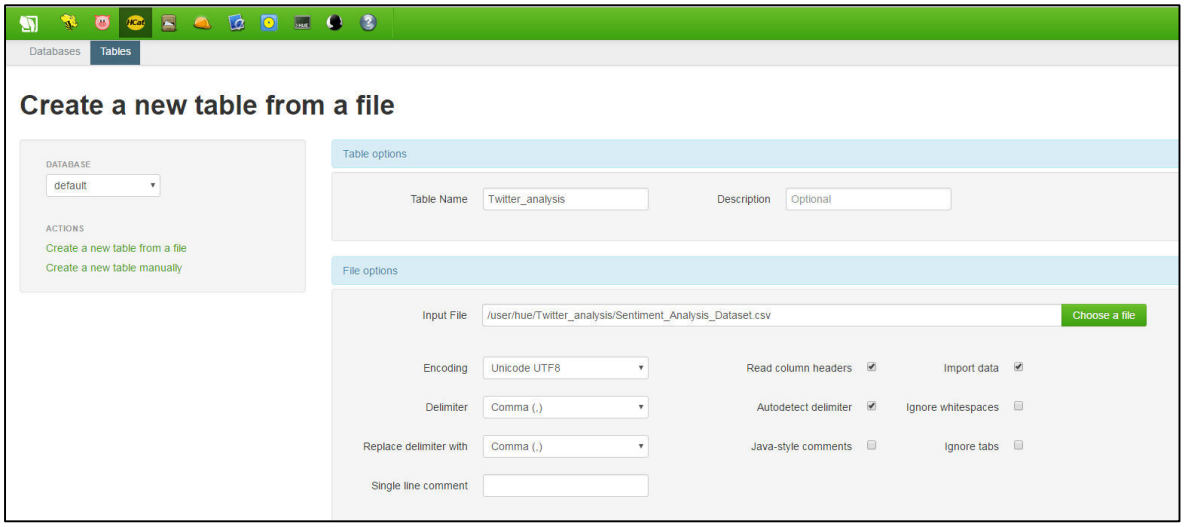
Platforma Splunk je díky indexaci dobrým nástrojem pro práci s velkými daty. Poskytuje uživatelům také dynamickou vizualizaci dat, díky které jsou analýzy přehlednější a srozumitelnější. Nicméně při práci s cloudovou verzí se vyskytly problémy ohledně indexace většího objemu dat, tudíž musela být zvolena on-premise varianta, která byla omezena výkonností pracovní stanice.

# Hortonworks

Hortonworks poskytuje verzi zdarma v podobě Sandboxu. Ke zprovoznění Hortonworks Sandboxu je nutné mít aplikaci, která vytváří virtuální prostředí. V tomto případě byla využita aplikace od Oraclu VirtualBox. Hortonworks sandbox obsahuje nástroje Apache, jako jsou Apache Pig, Hive, Hue a další. Veškerá interakce uživatele s Hadoop clusterem probíhá přes webové rozhraní Hue.

### Nahrání datového setu do platformy

Jelikož je využívána sandboxová verze, nahrání souboru do platformy probíhal v rámci stejného prostředí, čili šlo jen o přesun do souboru. To trvalo 15 sekund. Do části nahrání datového setu lze v tomto případě zahrnout i to, jak dlouho trvalo ze souboru udělat tabulku. To je zajištěno pomocí HCatalog, což je vrstva v Hadoopu, která řídí práci s tabulkami a úložným prostorem. Prostředí HCatalog lze vidět na následujícím obrázku.



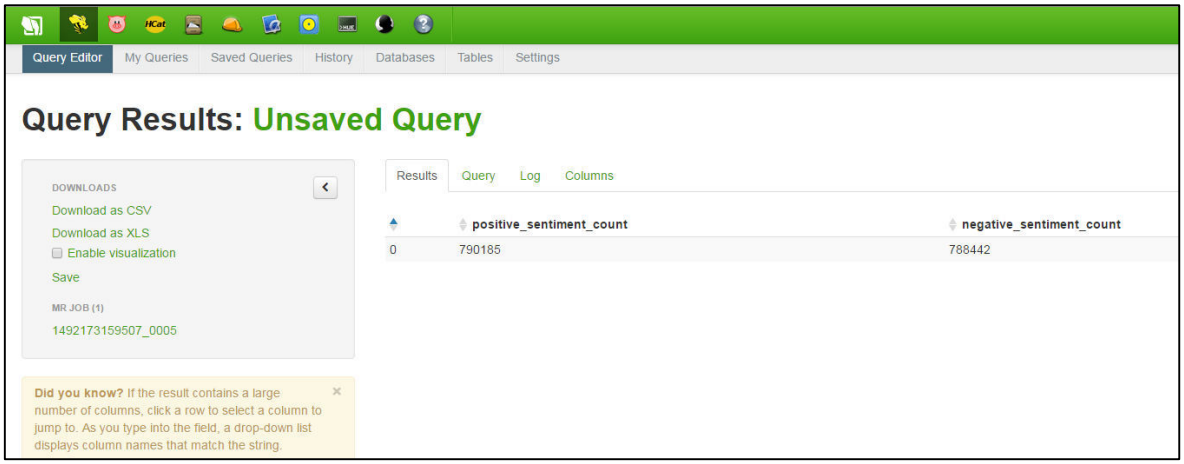

*Obrázek 26: Hortonworks – HCatalog*
Zdroj: vlastní

Tabulka byla vytvořena přímo z datového setu ve formátu CSV. HCatalog umožňuje uživateli určit názvy sloupců, datové typy a také delimiter v souboru CSV.

Vytvoření tabulky a import dat trval přesně 2 minuty a 19 sekund. Což znamená, že celkové nahrání dat trval 2 minuty a 34 sekund.

**Jednoduchá datová analýza – kolik tweetů je negativních a kolik pozitivních**

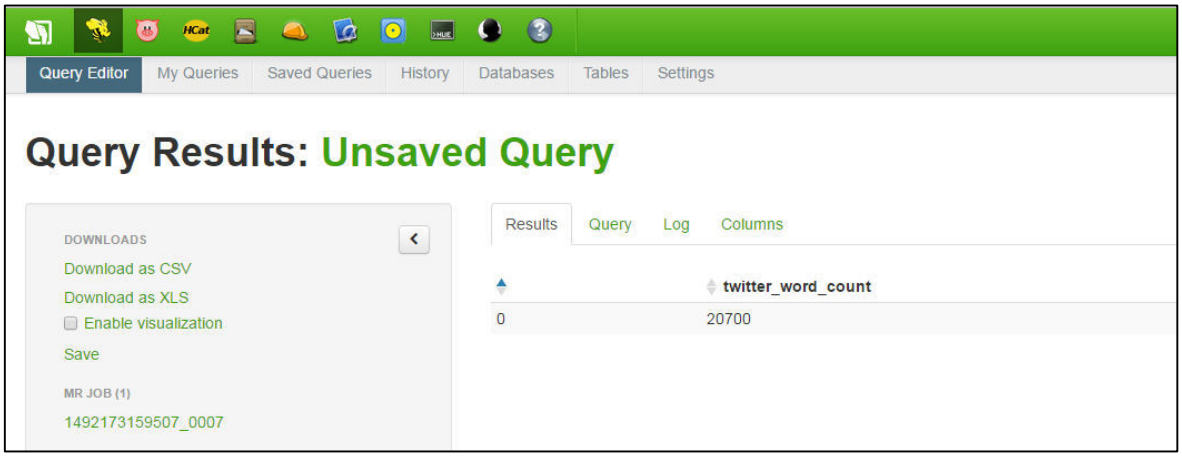
Pro jednoduchou datovou analýzu byl zvolen nástroj Beeswax, což je vestavěný Hive klient, který zajišťuje exekuci SQL dotazů nad tabulkami. SQL dotaz je totožný jako v předchozích případech. Job, který vykonával SQL příkaz, trval přesně 1 minutu a 5 sekund. Při zpracovávání dat byly využity funkce Map i Reduce. Na níže uvedeném obrázku, lze vidět výsledky, které se shodují s výsledky předchozích nástrojů.




*Obrázek 27: Hortonworks - Jednoduchá datová analýza*
Zdroj: vlastní

### Textová analýza – výskyt slova „Twitter“

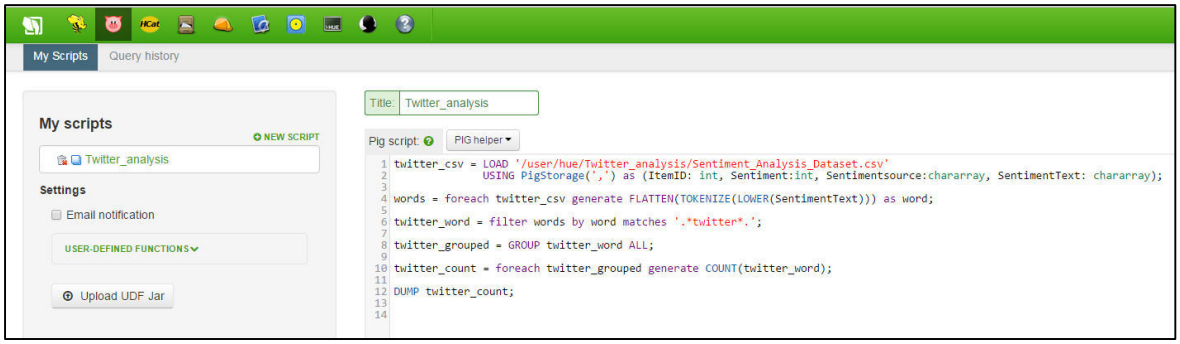
Hortonworks podobně jako Databricks poskytuje více řešení jak tuto úlohu vyřešit. Prvním řešením bude SQL dotaz totožný jako u řešení na platformě Databricks. Dotaz vrátil správný počet výsledků, což je 20 700 a trval přesně 1 minutu a 8 sekund.




*Obrázek 28: Hortonworks – SQL*
Zdroj: vlastní

Druhým řešením je využití integrovaného nástroje Apache Pig, který využívá programovací jazyk Pig. Tento programovací jazyk byl navržen přímo na práci s Hadoopem. Dokáže rychle analyzovat data a umožňuje paralelizaci. Využití tohoto postupu není optimální, ale poukazuje to opět na rozšířené funkčnosti platformy, kdy lze Apache Pig využít ke

složitějším úlohám. Algoritmus pro tuto úlohu nejprve načte datový set CSV do proměnné, určí se přitom oddělovač a také sloupce a jejich datové typy. Následně algoritmus postupuje po řádcích a převede obsah sloupce SentimentText na malá písmena pomocí funkce Lower, dále pak pomocí funkce Tokenize rozdělí text na jednotlivá slova a zanoří je do bagu a tzv. tuplů. Aby se dala jednotlivá slova procházet, tak se musí zanoření slov odstranit, k čemuž slouží funkce Flatten. Následně pomocí funkce Filter se hledá jen slovo „Twitter“ v definovaném bagu slov. Poté se jen využije funkce Group, která výsledky spojí a pomocí funkce Count se jednotlivé výskyty slova spočítají. Skript napočítal 20 700 krát slovo „Twitter“ a trval přesně 6 minut a 28 sekund.




*Obrázek 29: Hortonworks - Pig skript*
Zdroj: vlastní

Platforma Hortonworks má integrováno dost Apache nástrojů na to, aby šlo dynamicky a efektivně pracovat s velkými objemy dat. Verze zdarma je poskytována formou Sandboxu, tudíž uživatel je závislý na výkonu jeho pracovní stanice, což může analýzy dat dost komplikovat. Hortonworks poskytuje uživateli různé možnosti jak data analyzovat i vizualizovat.

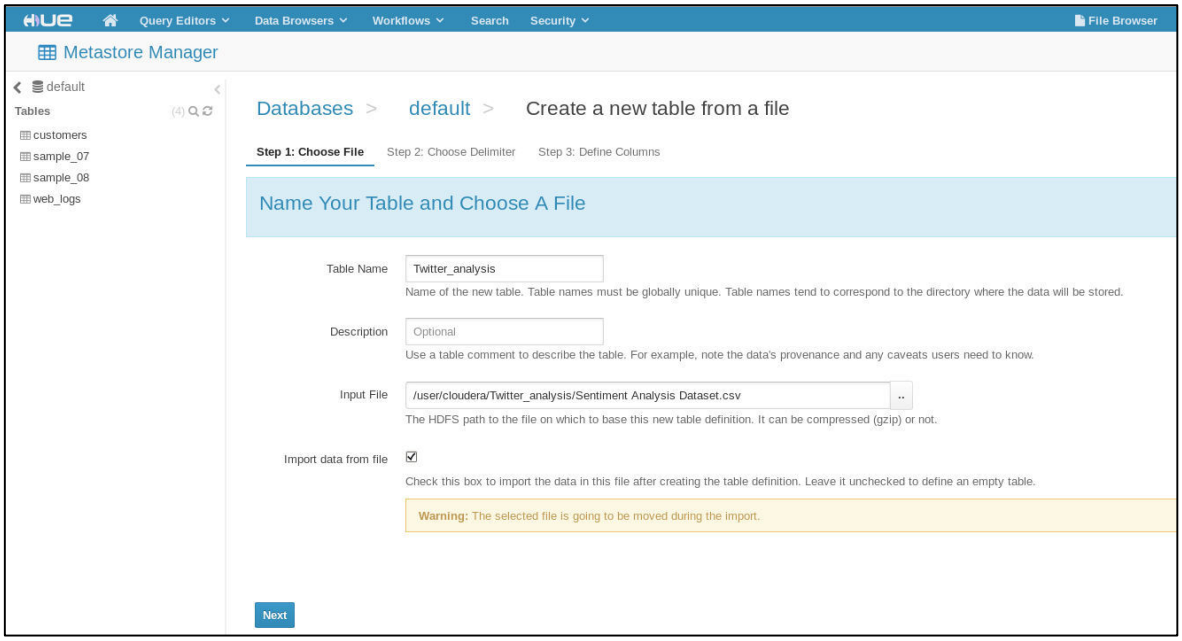
## Cloudera

Cloudera poskytuje zdarma CDH prostřednictvím virtuálního prostředí Cloudera QuickStart VM. Toto virtuální prostředí je třeba spustit v aplikaci, která virtuální prostředí umí vytvořit, v tomto případě byla použita aplikace od společnosti Oracle s názvem VirtualBox. Ve vytvořeném prostředí se již nachází platforma CDH, která uživateli poskytuje nástroje pro práci s daty, jako jsou Hadoop Impala, Apache Hive, HBase, Hadoop Hue, Hadoop YARN,

Spark atd. Data jsou zpracovávána na jednom clusteru s jedním nodem. Jelikož je to On-premise model, výkon je opět vázaný na konkrétní pracovní stanici.

### Nahrání datového setu do platformy

Nahrání datového setu do prostředí trval 3,12 sekund, jelikož to bylo jen přesunutí do příslušné složky. Nicméně do nahrávací části lze jako v případě Hortonworks platformy zahrnout i to, jak dlouho trvalo platformě z dat udělat tabulku, nad níž by se dalo data více analyzovat. Nahrání dat do tabulky probíhal přes webové rozhraní Hue, které slouží k práci s Apache Hadoopem. Vytvoření tabulky z datového setu ve formátu CSV, který má přes 1 500 000 záznamů trvalo přesně 32 sekund. Celkově tedy nahrání datového setu trval 35,12 sekund. Webové rozhraní Hue lze vidět na níže přiloženém obrázku.

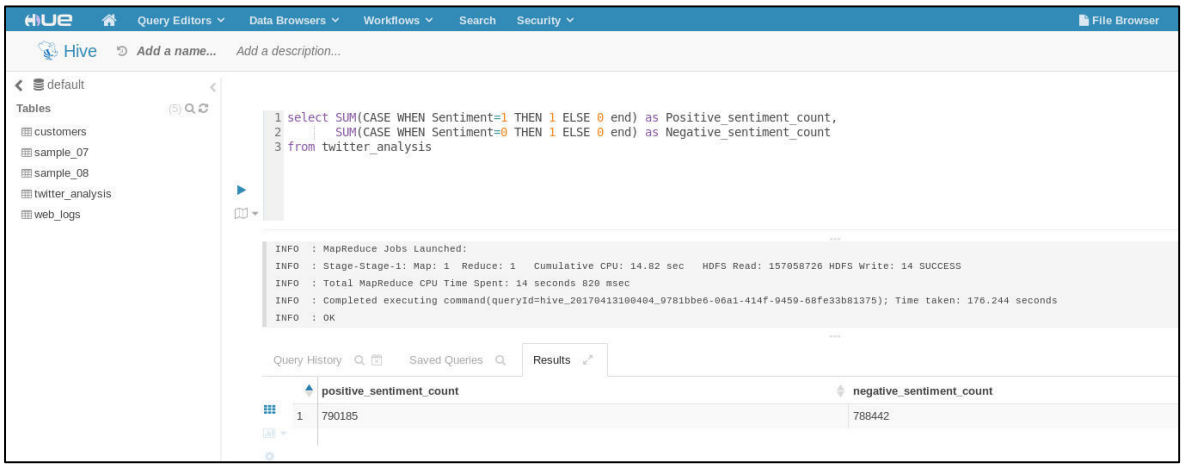



*Obrázek 30: Cloudera - webové rozhraní Hue*
Zdroj: vlastní

### Jednoduchá datová analýza – kolik tweetů je negativních a kolik pozitivních

V rámci platformy se dá pracovat s komponentou Apache Hive, která umožňuje uživateli v Hadoopu pracovat s daty prostřednictvím dotazovacího jazyka SQL. Datová analýza kolik tweetů je negativních a kolik pozitivních byla vykonána pomocí stejného SQL dotazu jako v případě platformy Databricks. Vykonání SQL dotazu trvalo 2 minuty a 56 sekund. Na níže

uvedeném obrázku lze vidět i to, že byla využita funkce MapReduce a že job trval skutečně 2 minuty a 56 sekund.

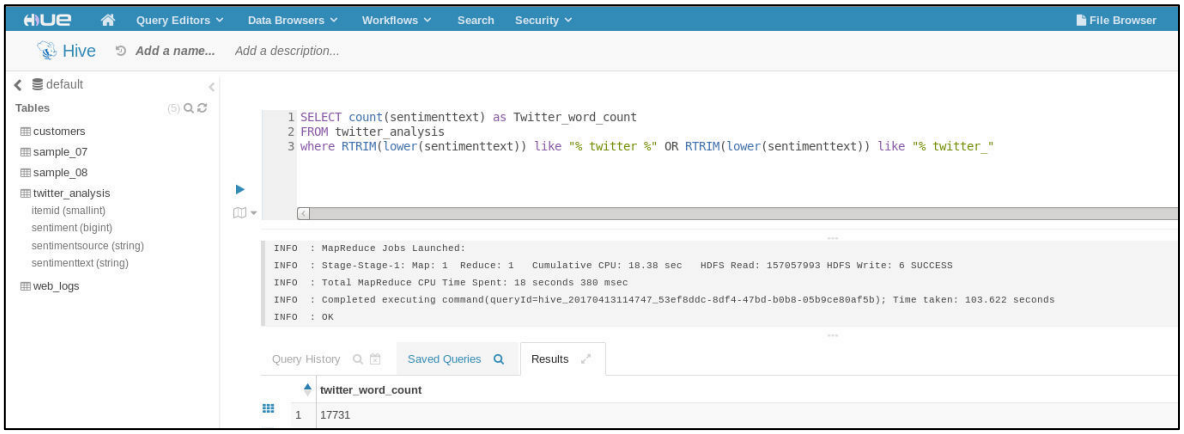



*Obrázek 31: Cloudera - Jednoduchá datová analýza*
Zdroj: vlastní

### Textová analýza – výskyt slova „Twitter“

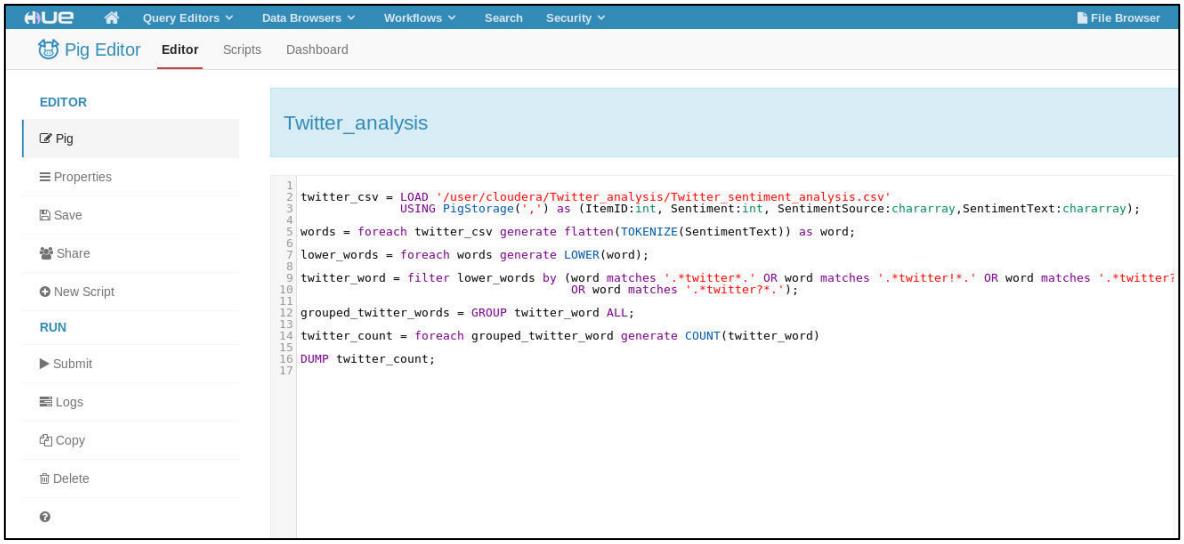
Tuto úlohu lze na platformě Cloudera vyřešit více způsoby. Jedním z nich je klasický SQL dotaz a druhým z nich je analýza s využitím integrovaného nástroje Apache Pig, stejně jako u platformy Hortonworks.

První řešení pomocí SQL dotazu proběhlo za 1 minutu a 43 sekund a slovo „Twitter“ v datovém setu dotaz našel 20700 krát.




*Obrázek 32: Cloudera – WordCount*
Zdroj: vlastní

Druhým způsobem řešení této úlohy je skript v programovacím jazyce Pig. Tento programovací jazyk je uzpůsoben pro práci s Hadoopem a využíváním funkce MapReduce. Syntaxe je proto specifická. Skript je založen na podobném algoritmu jako v platformě Databricks s tím rozdílem, že se zde využívá funkce tokenizace, která rozdělí text na jednotlivá slova. Exekuce skriptu trvala 5 minut a 12 sekund a slovo „Twitter“ se v datovém setu našlo 20 700 krát.




*Obrázek 33: Cloudera - Pig skript*
Zdroj: vlastní

Platforma Cloudera je díky předem připravené integrace všech Hadoop nástrojů vhodná pro práci s velkými objemy dat. Cloudera dává uživateli možnost zvolit si způsob zpracování dat a následnou vizualizaci pomocí Apache komponent.

#### 3.5.3.4 Uživatelská přívětivost

Hodnocení uživatelské přívětivosti je ze všech nejsubjektivnější. Všechny platformy mají přehledný design s intuitivním rozmístěním ovládacích prvků. Na první pohled uživatel ví co dělat a snadno nalezne jednotlivé funkce, které platformy nabízí.

Nicméně jako nejvíce uživatelsky přívětivou platformou bylo zvoleno Databricks. Kromě přehledného designu je hlavní výhodou přehledná a věcná dokumentace s konkrétními příklady datových analýz. Hlavním problémem je poněkud komplikované vytváření tabulek. Platforma Splunk je ze všech nejméně intuitivní, ale díky dobře napsané dokumentaci a návodům, to není až tak velký problém. Hlavní výhodou je pak přehledné vytváření tabulek. Hortonworks je přehledná platforma u které je však hlavním problémem poměrně složitá instalace kvůli sandboxu. Platforma Cloudera je intuitivní a má přehlednou dokumentaci, která popisuje jednotlivé funkčnosti na příkladech. Hlavním problémem je kromě složitější instalace také né příliš povedené grafické rozhraní.

### 3.5.4 Výsledné hodnocení

Každý z testovaných nástrojů se něčím liší, ať už funkčností, deployment modelem či uživatelskou přívětivostí, nicméně každý z nich je použitelný pro hlubší analýzu Big Data. V tabulce č. 4 je detailní vypsání faktických výsledků u všech kritérií, které byly hodnoceny.

*Tabulka 4: Detailní vypsání faktických výsledků*

| Platforms | databricks | splunk> | Hortonworks | cloudera |
|---|---|---|---|---|
| **Deployment model** | Cloud | On-premise/Cloud | On-premise/Cloud | On-premise/Cloud |
| **Dostupnost nástroje zdarma** | Platforma poskytuje zdarma 6GB cluster s interaktivním prostředí v cloudu | On-premise software - omezena indexace dat (upload dat) 500 MB/den | On-premise sandboxová verze Potřeba aplikace na vytvoření virtuálního prostředí | On-premise sandboxová verze Potřeba aplikace na vytvoření virtuálního prostředí |
| **Funkcionalita - Upload datového setu** | 7 minut a 25 sekund | 50 sekund | 2 minuty a 34 sekund | 35,12 sekund |
| **Funkcionalita - Jednoduchá datová analýza – kolik tweetů je negativních a kolik pozitivních** | 4,94 sekund | 10,483 sekund | 1 minuta a 5 sekund | 2 minuty a 56 sekund |
| **Funkcionalita - Textová analýza – výskyt slova „Twitter“ - SQL příkaz** | 2,75 sekund | 17,18 sekund | 1 minuta a 8 sekund | 1 minuta a 43 sekund |
| **Funkcionalita - Textová analýza – výskyt slova „Twitter“ - skript** | 5 minut a 51 sekund | Chybí | 6 minut a 28 sekund | 5 minut a 12 sekund |
| **Uživatelská přívětivost** | Přehledný design. Hlavní výhodou je přehledná a věcná dokumentace s konkrétními příklady datových analýz. Hlavním problémem je poněkud komplikované vytváření tabulek. | Platforma Splunk je ze všech nejméně intuitivní, ale díky dobře napsané dokumentaci a návodům, to není až tak velký problém. Hlavní výhodou je pak přehledné vytváření tabulek. | Hortonworks je přehledná platforma u které je však hlavním problémem poměrně složitá instalace kvůli sandboxu. | Platforma Cloudera je intuitivní a má přehlednou dokumentaci, která popisuje jednotlivé funkčnosti na příkladech. Hlavním problémem je kromě složitější instalace také né příliš povedené grafické rozhraní. |

Zdroj: vlastní

Na základě těchto výsledků pak byla vytvořena rozhodovací matice, kde bylo každé kritérium bodováno od 0 (nejhorší) do 10 (nejlepší) pro jednotlivé platformy. Po obodování došlo k pronásobení vahami, které byly stanoveny v kapitole 3.5.2. Výsledky lze vidět v tabulce č. 6.

*Tabulka 5: Bodování kritérií pro jednotlivé platformy*

| Platforms | Váha kritéria | databricks | splunk> | Hortonworks | cloudera |
|---|---|---|---|---|---|
| **1. Deployment model** | 0,23 | 6 | 8 | 8 | 8 |
| **2. Dostupnost nástroje zdarma** | 0,308 | 8 | 7 | 7 | 7 |
| **3. Funkcionalita** | 0,308 | 8,25 | 5,25 | 6,25 | 6,5 |
| 3.1 Funkcionalita - Upload datového setu | 0,25 | 3 | 7 | 5 | 8 |
| 3.2 Funkcionalita - Jednoduchá datová analýza | 0,25 | 10 | 8 | 6 | 4 |
| 3.3 Funkcionalita - Textová analýza | 0,5 | 10 | 3 | 7 | 7 |
| **4. Uživatelská přívětivost** | 0,154 | 8 | 5 | 6 | 7 |
| **Celkem** | | **30,25** | **25,25** | **27,25** | **28,5** |

Zdroj: vlastní

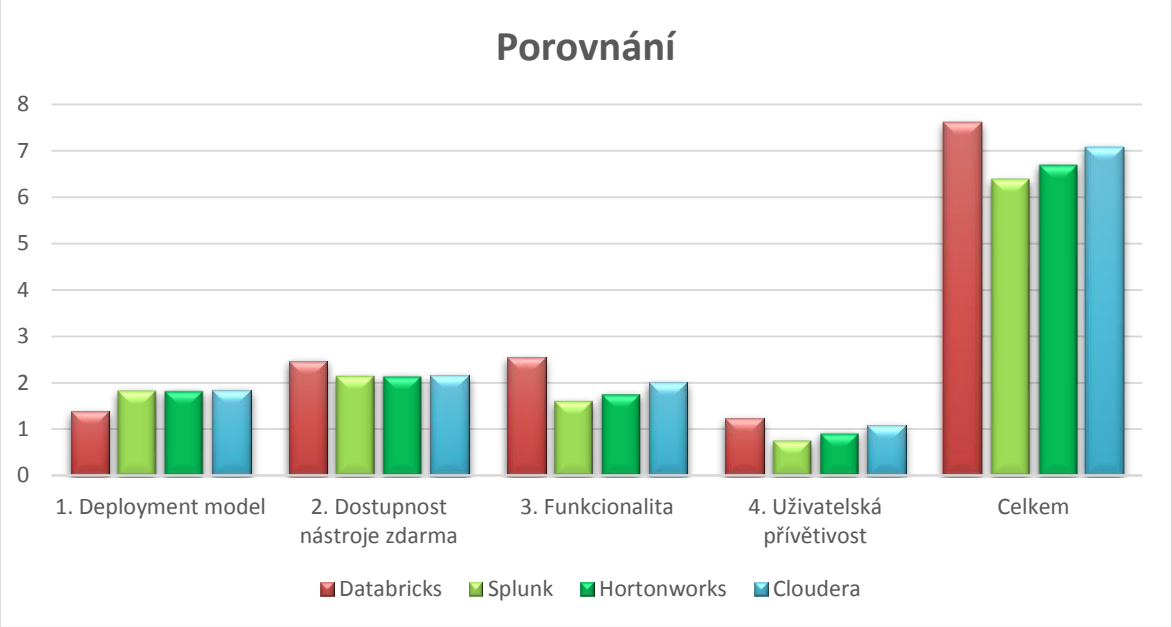
Nejvíce bodů získala platforma Databricks, nicméně je ještě potřeba pronásobit bodování jednotlivými váhami, aby tak došlo k validnímu stanovení nejvhodnější platformy, viz tabulka č. 6.

*Tabulka 6: Výsledné hodnocení*

| Platforms | Váha kritéria | databricks | splunk> | Hortonworks | cloudera |
|---|---|---|---|---|---|
| **1. Deployment model** | 0,23 | 1,38 | 1,84 | 1,84 | 1,84 |
| **2. Dostupnost nástroje zdarma** | 0,308 | 2,464 | 2,156 | 2,156 | 2,156 |
| **3. Funkcionalita** | 0,308 | 2,541 | 1,617 | 1,771 | 2,002 |
| 3.1 Funkcionalita - Upload datového setu | 0,25 | 0,75 | 1,75 | 1,25 | 2 |
| 3.2 Funkcionalita - Jednoduchá datová analýza | 0,25 | 2,5 | 2 | 1,5 | 1 |
| 3.3 Funkcionalita - Textová analýza | 0,5 | 5 | 1,5 | 3 | 3,5 |
| **4. Uživatelská přívětivost** | 0,154 | 1,232 | 0,77 | 0,924 | 1,078 |
| **Celkem** | | **7,617** | **6,383** | **6,691** | **7,076** |

Zdroj: vlastní

Po zohlednění vah kritérií je nejvhodnější variantou platforma Databricks. Kromě kritéria deployment model, získala platforma Databricks nejlepší hodnocení za každé kritérium, viz graf níže.




*Obrázek 34: Graf porovnání nástrojů*
Zdroj: vlastní

Databricks nabízí verzi zdarma formou cloudu. Ostatní poskytovatelé využívají on-premise model, což nutí uživatele využít vlastní hardware. Nejlepší výsledky, co se týče analýzy dat, byly také zaznamenány při práci s platformou Databricks. Nicméně to je ovlivněno výkonem pracovní stanice, na které probíhaly analýzy s využitím ostatních nástrojů.

# 4. Zpracování dat vybraným nástrojem

Pro zpracování byla zvolena data ze sociálních sítí, konkrétně z Twitteru a Facebooku. Důvodem vybrání těchto dvou sociálních sítí je fakt, že se jedná o dvě nepoužívanější sociální sítě na světě a také to, že umožňují poměrně jednoduché získání relevantních dat.

## 4.1 Zpracování dat ze sociální sítě Twitter

Twitter je jednou z nejpopulárnějších sociálních sítí poslední doby. K roku 2017 má přes 300 miliónů aktivních uživatelů. Twitter je založen na principu odesílání textových příspěvků dlouhých maximálně 140 znaků, označované jako tweety. Pokročilejší uživatelé nebo vývojáři mohou k datům z Twitteru přistupovat prostřednictvím aplikačního rozhraní nazvané Twitter API. Toto aplikační rozhraní se dělí na dvě hlavní API.

- Twitter REST API
- Twitter Streaming API

Twitter REST API poskytuje uživatelům přístup k čtení i zápisu dat sociální sítě Twitter. Pomocí REST API lze vytvářet nové tweety, ale také dohledávat staré, získat podrobnější informace o uživatelích, jako je lokalita, stáří účtu, počet tweetů atd. REST API vrací defaultně na dotaz odpověď ve formátu JSON.

Twitter Streaming API poskytuje uživatelům přístup přímo ke streamovaným datům sociální sítě Twitter. To znamená, že uživatel je schopen obdržet tweety podle zvolených kritérií, provést analýzu či vizualizaci a to vše v reálném čase. V příkladu zpracování dat na platformě Databricks bude využito právě Twitter Streaming API.

Pokud chce uživatel využívat aplikační rozhraní Twitter API, je nejprve nutné aby se daná aplikace autentizovala a autorizovala vůči Twitteru. To je zajištěno pomocí tzv. Tokenů, což jsou textové řetězce, pomocí kterých se uživatel či aplikace identifikuje. Pro získání těchto Tokenů je nutné vytvořit na Twitteru aplikaci. Po vytvoření aplikace stačí kliknout do záložky Keys and Access Tokens a zde lze nalézt Consumer Token/Consumer secret a Access Token/Access Token secret.

- Consumer Token

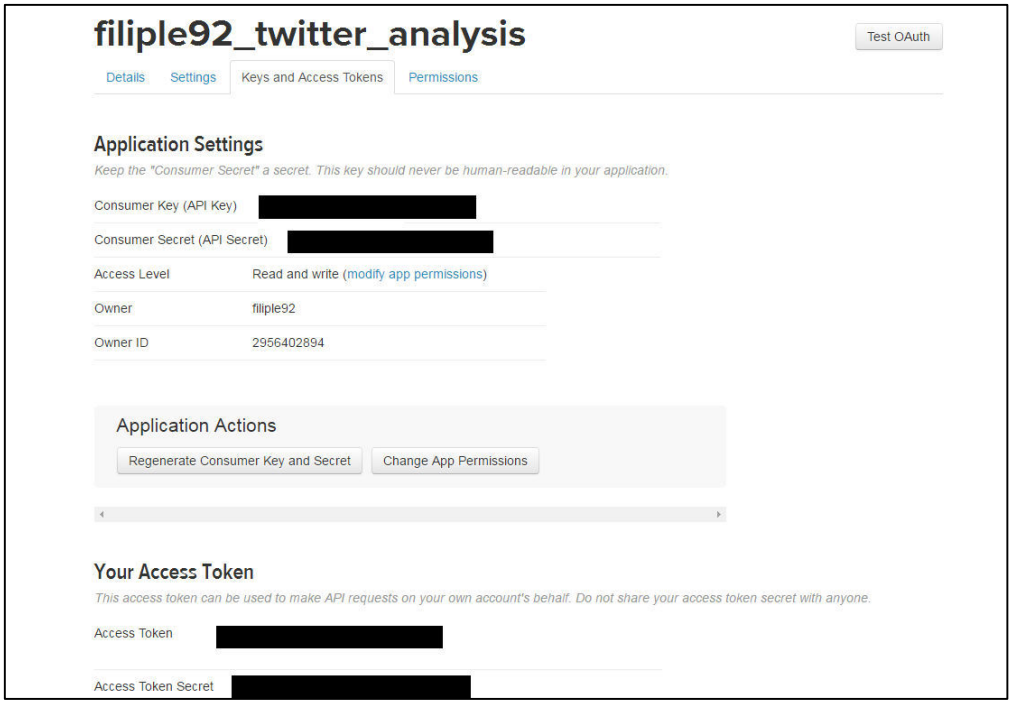
  Jedná se o API Key, který je svázán s danou aplikací a identifikuje jí. Využívá se při autentizaci.
- Consumer secret

  Funguje jako heslo k danému tokenu, pomocí něhož dochází k autentizaci vůči Twitteru.
- Access Token

  Umožňuje autorizaci vůči Twitteru.
- Access Token secret

  Funguje jako heslo k Access Tokenu.

Následující obrázek ukazuje příslušnou záložku na Twitteru s Tokeny.



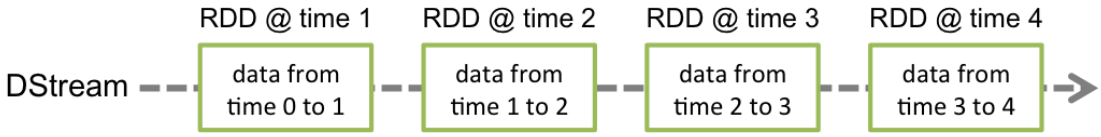

*Obrázek 35: Twitter tokens*
Zdroj: vlastní

Platforma Databricks je postavená nad frameworkem Apache Spark. Součástí tohoto frameworku je rozšíření s názvem Spark Streaming, které umožňuje škálovatelné, vysoko propustné zpracování streamovaných dat. Tato data mohou být přijímána z mnoha zdrojů, jako jsou například Apache Flume nebo TCP Sockety (Twitter) a další. Zpracování probíhá

tak, že Spark Streaming obdrží živý datový stream a tato data rozděluje do dávek, které jsou následně zpracovány Spark Enginem do finálního streamu dat, který se nazývá DStream.


*Obrázek 36: Spark Streaming data*
Zdroj: https://spark.apache.org/docs/latest/img/streaming-flow.png

DStream je tedy označení pro kontinuální stream dat, který je buď přímo ze zdroje (např. Twitter) nebo již zpracovaný stream dat generovaný transformací vstupních dat. DStream je složen z jednotlivých dávek dat ze zdroje ve formě neměnitelných spark datasetů, které se nazývají RDD (Resilient distributed datasets). Každý RDD obsahuje data z určitého intervalu ze zdroje.



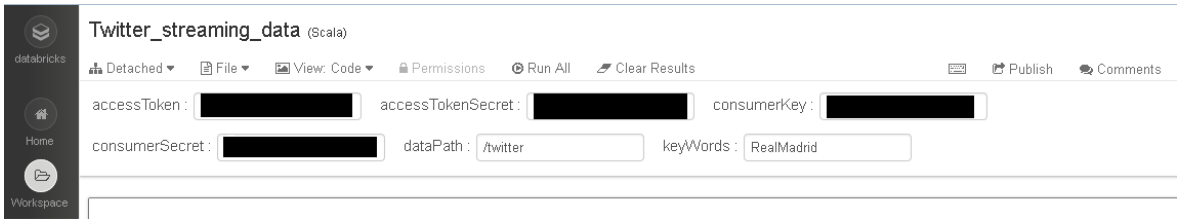

*Obrázek 37: DStream*
Zdroj: https://spark.apache.org/docs/latest/img/streaming-dstream.png

V následujícím příkladu bude využit způsob zpracování streamovaných dat, který představil na Spark Summitu 2016 M. Franklin[26]. S malými úpravami bude tento způsob aplikován na

[26] Na konferenci Spark Summit 2016, která se konala v New Yorku 16. 1. 2016, bylo představeno zpracování streamovaných dat na platformě Databricks. M. Franklin, hlavní vývojář na projektu Spark SQL v Databricks, na konferenci představil způsob jak na platformě Databricks zpracovávat streamovaná data.

jinou problematiku. Ke zpracování streamovaných dat je využit programovací jazyk Scala. V Databricks bylo potřeba tedy založit notebook kompatibilní s jazykem Scala. Notebook v Databricks je jakési rozhraní, které umožňuje komunikaci s touto platformou. Při vytváření notebooku si lze zvolit, jaký programovací jazyk bude využit ke komunikaci s platformou. V tomto případě to bude jazyk Scala.

Prvním krokem je získání příslušných Tokenů na autentizaci a autorizaci. Pro zadání tokenů jsou v notebooku zavedené widgety, které mají podobu prázdných polí, do kterých uživatel získané tokeny vloží, viz níže uvedený obrázek.




*Obrázek 38: Databricks – Tokens*
Zdroj: vlastní

Na obrázku jsou ještě vidět pole s názvem „Path“ a „keyWords“. Do pole Path se zadá cesta, kam se budou streamovaná data přenášet. V tomto případě to je složka s názvem twitter. Do pole keyWords zadá uživatel klíčová slova, podle kterých se mají streamovaná data filtrovat. Tento filtr se předá jako jeden z parametrů v komunikaci s TwitterAPI, čili streamovaná data budou chodit již odfiltrovaná. V tomto příkladu bylo jako klíčové slovo zvoleno „RealMadrid“, což je nejslavnější španělský fotbalový klub. Následně stačí jen spustit celý skript najednou, kdy dojde ke streamování dat do předem zvoleného uložiště.

Pro vytvoření streamu a následné stahování dat je napsaná ve skriptu třída TwitterStream, která přebírá některé principy z knihovny Sparku TwitterUtils, kde najdeme funkce ke komunikaci s Twitter API (REST i Stream).

Po spuštění skriptu se tedy začnou odchytávat streamovaná data sociální sítě Twitter ještě než jsou fyzicky uloženy do databáze. Tato data se ukládají do předem stanoveného uložiště jako RDD soubory. Ve skriptu se dá také stanovit interval, po kterém se má vytvořit nový RDD do kterého se budou streamovat data. V tomto případě jsou to 2 sekundy. Streamování

dat probíhalo něco přes 60 minut a bylo staženo 23915 tweetů s filtrem na slovo „RealMadrid“. Tyto tweety jsou uchovány v 1925 RDD objektech. Tyto RDD objekty lze pak díky modulu Spark SQL jednoduše všechny načíst do strukturovaného objektu. Tento objekt se nazývá DataFrame a je to tedy dataset s klasickou sloupcovou strukturou. Tento typ objektu může být vytvořen z různých zdrojů, jako jsou například tabulky, externí databáze nebo právě i RDD objekty. DataFrame API je možné využívat prostřednictvím jazyků jako je Scala, Java či Python. V tomto případě se všechny tweety do DataFramu načtou pomocí jednoduchého příkazu.

```
// Načtení všech RDD do jednoho DataFramu, který se v tomto případě jmenuje df_tweets
val df_tweets = spark.read.json(path)
```

*Obrázek 39: Databricks - DataFrame creation*
Zdroj: vlastní

Tento DataFrame df_tweets obsahuje tedy všech 23915 tweetů ve sloupcové struktuře, čili lze snadno provádět analýzu dat, ale uživatel musí znát DataFrame API. Součástí tohoto API, je také funkce, která z DataFrame vytvoří dočasné view, kde lze již standardně využívat jazyk SQL.

```
df_tweets.createOrReplaceTempView("tweets")
```

*Obrázek 40: Databricks - DataFrame – view*
Zdroj: vlastní

Není potřeba se omezovat jen na SQL. Tím, že lze volně přistupovat k datům pomocí vhodného API, lze psát různé algoritmy, pomocí nichž lze data analyzovat. Pro tuto ukázku analýzy jazyk SQL nicméně postačí.

Jako první se lze podívat na to, kdo vůbec nejvíce tweetoval za dobu ukládání streamovacích dat. Pro tuto úlohu lze napsat jednoduchý select, který seskupí záznamy dle jména a provede count. Bylo stanoveno ještě jedno omezení a to pouze na 10 uživatelů s nejvíce tweety, kvůli přehlednosti.

```
// Který uživatel odeslal nejvíce tweetů ?
display(sql("""
  SELECT user.screen_name as Nickname, count(*) as Tweets_count
  FROM tweets
  GROUP BY user.screen_name
  ORDER BY count(*) DESC limit 10
"""))
```

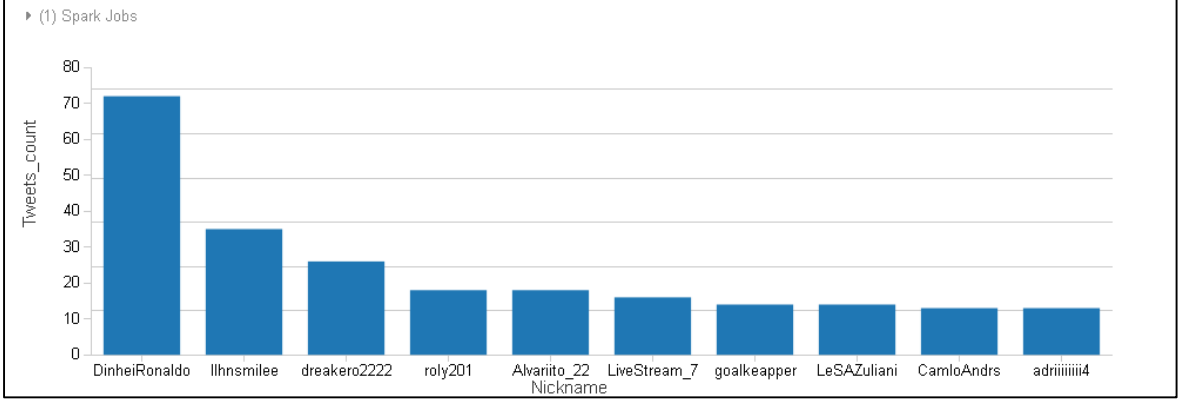
▸ (1) Spark Jobs

| Nickname | Tweets_count |
|---|---|
| DinheiRonaldo | 72 |
| Ilhnsmilee | 35 |
| dreakero2222 | 26 |
| roly201 | 18 |
| Alvariito_22 | 18 |
| LiveStream_7 | 16 |
| goalkeapper | 14 |
| LeSAZuliani | 14 |
| CamloAndrs | 13 |
| adriiiiiiii4 | 13 |

*Obrázek 41:Databricks - Kdo odeslal nejvíce tweetů ?*
Zdroj: vlastní

Z obrázku je patrné, že nejvíce tweetů za zhruba 64 minut stihl napsat uživatel s přezdívkou „DinheiRonaldo“. Databricks poskytuje uživatelům i možnost vizualizace dat. Lze vybírat z nejrůznějších grafů jako je například sloupcový graf, koláčový graf, bodový graf atd. Pro tento případ byl zvolen sloupcový graf.

```
// Který uživatel odeslal nejvíce tweetů ?
display(sql("""
  SELECT user.screen_name as Nickname, count(*) as Tweets_count
  FROM tweets
  GROUP BY user.screen_name
  ORDER BY count(*) DESC limit 10
"""))
```



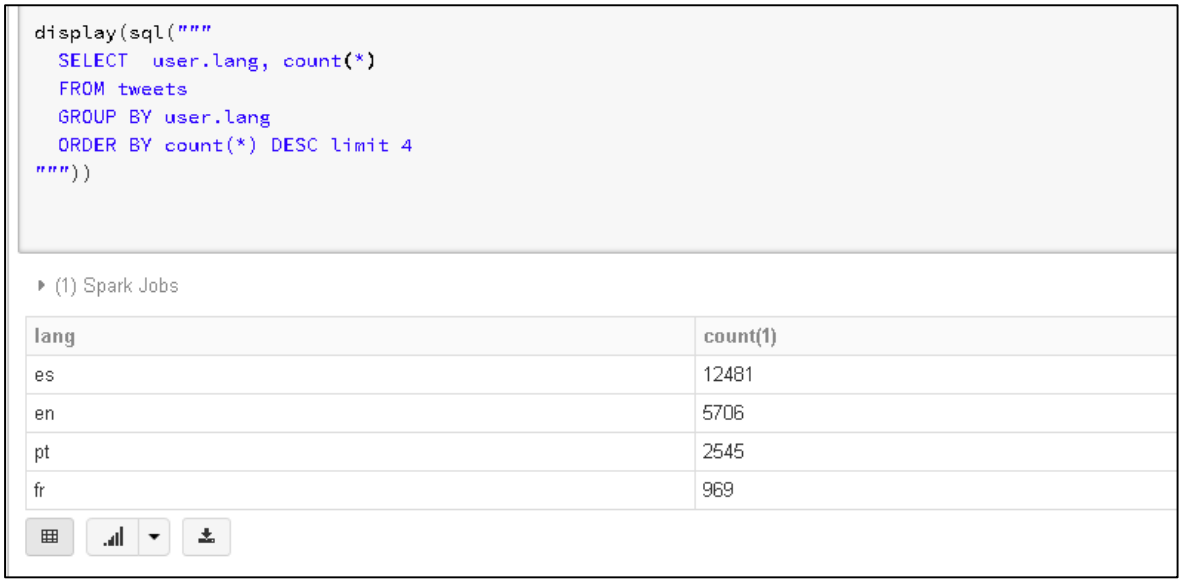

*Obrázek 42: Databricks – graf*
Zdroj: vlastní

Další analýza bude zaměřena na zjištění země, odkud pochází uživatelé z datového setu. K tomu bude využit atribut „location“ z TwitterAPI, který se zadává při registraci. V platformě Databricks je možné vizualizovat získaná data pomocí widgetu Map, který zobrazuje celosvětovou mapu. Tento widget podporuje pouze kódy států v ISO normě 3166 systém alpha 3. Jelikož uživatelé Twitteru při registraci většinou zadávají jako atribut location celý název země, bylo nutné provést namapování názvů zemí na ISO kódy. To bylo provedeno pomocí nahrání CSV souboru s mapováním zemí na ISO kódy a následnou klauzulí JOIN, která nám hodnoty „namapuje“. Nejvíce uživatelů, kteří tweetovali bylo pochopitelně ze Španělska, jelikož výběrová podmínka pro streamovaná data byla jen na výskyt španělského klubu Real Madrid. Následně pak hodně uživatelů bylo z Británie či Portugalska. Mapa odkud jsou uživatelé, kteří tweetovali vypadá následovně.


*Obrázek 43: Databricks - Mapa odkud pochází uživatelé*
Zdroj: vlastní

Využitý atribut location poskytuje informaci odkud je uživatel, nicméně reálné místo, odkud tweet přišel, to s největší pravděpodobností nezjistí. To, odkud přesně tweet přišel, by se dalo zjistit, jen pokud by daný záznam měl geolokační souřadnice (atribut coordinates). V tomto atributu se nachází zeměpisná šířka i délka. Podle Twitteru obsahuje tyto geolokační souřadnice v průměru pouze 4% tweetů. V tomto konkrétním případě obsahuje geolokační souřadnice pouze 55 záznamů, což je zhruba 0,2% z celkové množiny. Tato analýza by tedy postrádala smysl, jelikož vzorek by nedosahoval ani průměrné hodnoty, kterou uvádí Twitter.

Jako další lze provést analýzu nejčastěji používaného jazyku. To může podniku více prozradit o tom, jaký jazyk by měli více využívat pro komunikaci se zákazníky, případně na jaké trhy více cílit a podobně. TwitterAPI poskytuje atribut lang, pomocí kterého se dá lehce zjistit jazyk daného tweetu. Opět lze využít SQL dotaz, kde se zobrazí pouze čtyři nejvíce vyskytující se jazyky z toho důvodu, že ostatní jazyky jsou zastoupeny ve velmi malém měřítku.

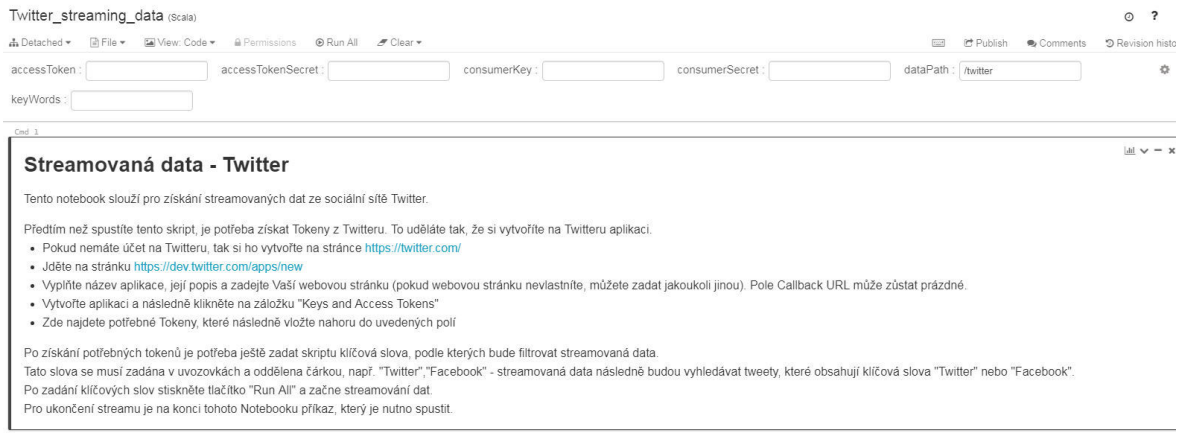



*Obrázek 44: Databricks – lang*
Zdroj: vlastní

Notebook vytvořený na platformě Databricks, pomocí kterého lze analyzovat streamovaná data ze sociální sítě Twitter je volně dostupný přes webové rozhraní. Jelikož platforma umožňuje volné sdílení notebooků mezi uživateli, tak se stačí pouze na Databricks zaregistrovat a následně notebook naimportovat do pracovního prostředí. Pak již lze notebook zaměřený na streamovaná data volne využívat[27].

Notebook je včetně postupu v češtině, jak streamovaná data získat, viz obrázek.

[27] Notebook je volně dostupný na adrese: https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/5647495502287019/4355490338349702/327186963973146/latest.html




*Obrázek 45: Databricks notebook - streamovaná data*
Zdroj: vlastní

## 4.2 Zpracování dat ze sociální sítě Facebook

Facebook je již dlouhodobě nejvýznamnější a nejrozsáhlejší sociální síť na světě. Dokazuje to fakt, že jako první přesáhl milník miliardy registrovaných uživatelů a k roku 2017 má dokonce 1,71 miliardy aktivních uživatelů[28]. To mimo jiné také znamená obrovské množství dat, které lze získat a využít je. K datům Facebooku, lze přistupovat přes aplikační rozhraní Graph API. Toto nízko úrovňové aplikační rozhraní postavené na http protokolu, je primární způsob jak dostat požadovaná data z této platformy.

Vzhledem k tomu, že se také webové technologie posouvají rychle dopředu, tak se přirozeně API Facebooku vyvíjí a vydávají se neustále nové verze. K datu 20. 7. 2017 je poslední verze 2. 10. Facebook zajišťuje verzování svého API, což pomáhá především ostatním vývojářům přizpůsobit se změnám okolí a dává jim to potřebný čas na implementaci novějších řešení. Každá verze API je totiž platná ještě 2 roky od vydání novější verze, takže vývojáři aplikací mají dva roky na přechod na nové API. Vývojáři Graph API nicméně počítají s rigidními aplikacemi a garantují, že pokud některá aplikace volá příkazy z již deprekované verze, tak budou příkazy automaticky přesměrovány na nejstarší dostupnou verzi, čímž se zajistí integrita aplikací.

[28] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Podobně jako u Twitteru, tak také Facebook využívá tokeny na identifikaci toho, kdo k omunikuje s Facebookem prostřednictvím jeho aplikačního rozhraní. Tyto tokeny slouží primárně jako autorizační prostředek. Facebook rozeznává čtyři typy těchto tokenů.

**User Access Token**

Jedná se o jeden z nejvíce využívaných access tokenů. Pokaždé když aplikace chce získat či nějakým způsobem modifikovat data konkrétního uživatele na Facebooku, tak potřebuje tento typ tokenu. Pro získání tohoto tokenu musí nejdříve aplikace odeslat konkrétnímu uživateli požadavek na udělení oprávnění.

**App Access Token**

Tento token slouží k přístupu do nastavení Facebook aplikace.

**Page Access Token**

Tento token je podobný jako User Access Token jen s tím rozdílem, že objektem není samotný uživatel, ale přímo facebooková stránka. Pro získání tohoto tokenu je nutné nejprve získat user access token uživatele, který má administrátorská oprávnění k dané stránce.

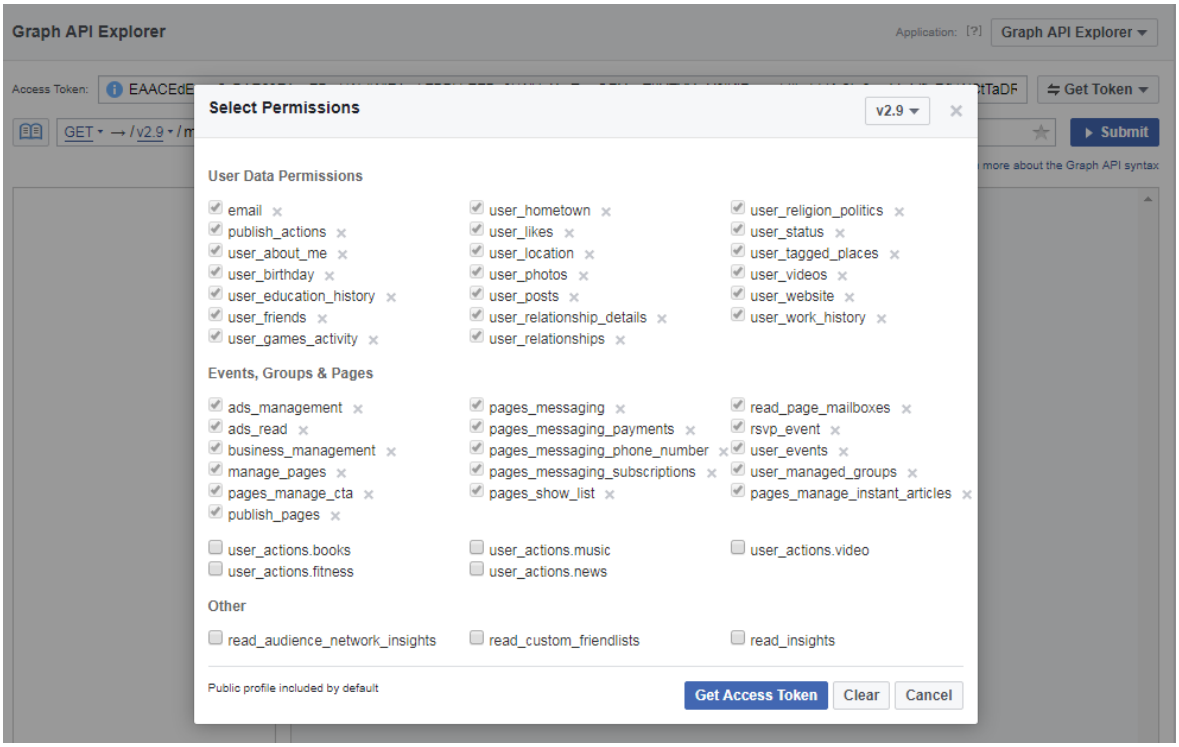
**Client Access Token**

Jedná se o token, který lze vložit do binárních souborů v mobilních zařízeních či do desktopových aplikací za účelem identifikace této aplikace vůči Facebooku. Tento typ tokenu se využívá velmi zřídka.

Pro účely této práce bude potřeba získat User Access Token. To lze provést hned několika způsoby a primárně záleží na platformě, ve které budou data využívána. Například JavaScript získává a uchovává token automaticky v cookies daného prohlížeče. Platformy jako iOS či Android si token automaticky generují skrz předem připravené třídy. Pro získání tokenu a také následné získání dat bude využit nástroj Graph API Explorer.

**Graph API Explorer**

Jedná se o nástroj, který byl vyvinut přímo společností Facebook Tento nástroj slouží pro psaní dotazů, přidávání či mazání dat přímo z Facebooku za využití Graph API. Tento nástroj také zvyšuje transparentnost daného API, jelikož po složení dotazu jsou vrácena data ve formátu JSON, což umožňuje okamžitou kontrolu dat vůči datům vráceným na určité platformě. Před samotným skládáním dotazu je potřeba získat Access Token. Graph API Explorer má přímo implementovanou funkci, která slouží k získání Access Tokenu. Kliknutím na tlačítko Get Acces Token je uživateli dále předloženo dialogové okno, kde může přímo vyspecifikovat práva získaného Access Tokenu. Nabídku s přístupovými právy lze vidět na následujícím obrázku.




*Obrázek 46: Dialogové okno s právy*
Zdroj: vlastní

V pravé části obrázku lze vidět tlačítko „Get Token“ pomocí něhož se přistupuje do dialogového okna s právy. Po získání tokenu už je potřeba jen specifikovat dotaz pomocí Graph API. Aplikační rozhraní Facebooku přistupuje k datům přes princip sociogramu, což je graf, který znázorňuje vztahy mezi členy určité skupiny. Informace jsou tedy reprezentovány jako:

- Nodes (uzly)
- Edges (hrany)
- Fields (pole)

Nodes neboli uzly jsou hlavními objekty, s kterými Graph API pracuje. Jedná se o objekty, jako jsou uživatel, stránka či událost. Každý uzel má unikátní ID, díky kterému k němu lze přistoupit a získat tak potřebná data k další analýze. Hrany (Edges) figurují pak jako vztahy mezi uzly, například komentář k fotografii nebo fotografie jednotlivých stránek. Každá hrana se tedy váže na určitý uzel či více uzlů a poskytuje především referenci na tyto uzly. Uzel příspěvek, má hranu s názvem Likes, která obsahuje seznam uživatelů (reference na uživatele, respektive uzly), kteří příspěvek označili jako „To se mi líbí“. Pole (Fields) jsou atributy příslušného uzlu. Například uzel, který představuje profil má pole jako ID, Name, Link atp. Kompletní seznam polí a hran lze nalézt v dokumentaci GraphAPI[29] či přímo přes nástroj Graph API Explorer.

**Dotazování pomocí Graph API Exploreru**

Generování dotazů, jelikož je Graph API postavené na protokolu HTTP, využívá jednu z dotazovacích metod právě protokolu HTTP a to sice metodu GET. Graph API Explorer má implementovaný vlastní GUI, přes který se dají dotazy skládat. Grafické rozhraní nástroje, lze vidět na následujícím obrázku.


*Obrázek 47: Graph API Explorer*
Zdroj: vlastní

V prostoru na dotaz (pod Access Tokenem) je potřeba vybrat metodu pro dotazování, tedy GET. Dále pak vybrat příslušnou verzi Graph API, která bude vycházet z potřeb daného

[29] https://developers.facebook.com/docs/graph-api/reference/user

dotazu, typicky však lze ponechat defaultní nastavení, což je nejaktuálnější verze rozhraní. Dalším krokem je do dotazovacího okna zadat unikátní ID objektu (hrany či uzlu) z kterého je potřeba získat data. Na následujícím obrázku je dotaz na uzel přihlášeného uživatele.


*Obrázek 48: Graph API Explorer - dotaz na uzel "me"*
Zdroj: vlastní

Unikátní ID objektu bývá typicky uměle vygenerované číslo ze sekvence. Nicméně v tomto případě byl použit uzel „me“. Tento uzel je speciálním koncovým bodem, který odkazuje na uživatele či stránku (respektive uživatelovo unikátní user_id, v případě stránky je to pak page_id), jejíž Access token je využíván k volání Graph API. V příkladu uvedeném na obrázku dotaz vrátil standardní atributy (jméno a unikátní ID uzlu), které vrací API pokaždé, pokud dotaz není specifikován. Data byla vrácena ve formátu JSON. Nástroj také nabízí možnost exportu dotazu do různých SDK (Android SDK, iOS SDK, PHP SDK, JavaScript SDK a cURL SDK), což podstatně usnadňuje implementaci do různých aplikací. Je tedy možné složit si dotaz v tomto nástroji a poté vyexportovat zdrojový kód do aplikace např. pro Android.

V nástroji Graph API Explorer existuje také rolovací menu se seznamem polí (viz obrázek č. 44 - vlevo) pro daný uzel, na který se lze dotázat, což podstatně ulehčuje práci s nástrojem. Odpadá tak totiž potřeba znát pole pro dané typy uzlů, nebo je hledat v dokumentaci. Stačí kliknout na tlačítko se znakem plus a rolovací menu zobrazí veškerá pole, na která se pro daný uzel lze dotázat.

Data uzlu, která API vrací, se tedy dají ovlivnit vyspecifikováním jednotlivých polí daného uzlu. Tato pole (nebo také atributy objektu) se vyspecifikují přímo v dotazu. V příkladu,

který je uvedený níže na obrázku, jsou dotazovány pole jméno (name), zařízení (devices) a lokalita (location) uzlu „me“.


*Obrázek 49: Graph API Explorer - dotazování na pole*
Zdroj: vlastní

Pole name určuje jméno uživatele, pole devices odkazuje zařízení, z kterých se uživatel přihlašoval na Facebook pomocí aplikace (pokud se uživatel přihlašoval například přes mobilní telefon s operačním systémem Android, ale přes webové rozhraní, tento atribut to nezachytí) a nakonec atribut location, který odkazuje na lokalitu, z které uživatel pochází. Atribut location poskytuje pouze přesnou lokalitu uživatele (kterou zadal při vyplňování informací), nikoli pak zemi, ve které se dané město nachází. Nicméně při dotazu na lokalitu je vrácen nejen název města, ale také ID daného místa. To znamená, že se jedná také o uzel, který má další atributy, z kterých se dá zjistit více informací (například země ve které město leží). Ke zjištění země, ze které uživatel pochází, lze využít zanořování dotazů.

Zanořování dotazů je další z funkcí, kterou nástroj podporuje. Prakticky to znamená, že pro informaci, která by jinak vyžadovala více dotazů do běžného API, stačí vytvořit jeden robustní dotaz, který jí poskytne. Obecný formát dotazů vypadá následovně.

```
GET graph.facebook.com
  /{node-id}?
    fields=<first-level>{<second-level>}
```

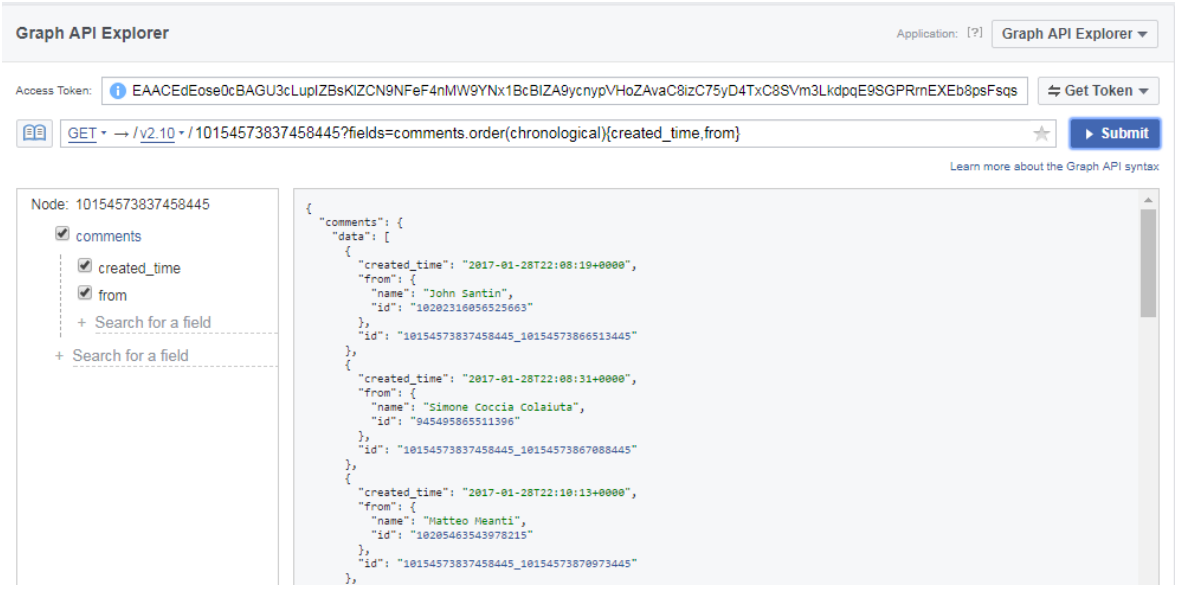
*Obrázek 50: Zanořování dotazů*
Zdroj: https://developers.facebook.com/docs/graph-api/using-graph-api/

První úrovní se v tomto případě myslí vybraná pole (uzly) nadřazeného uzlu. Druhá úroveň je daný pod dotaz a zadávají se tam pole uzlů, které jsou vyspecifikované v první úrovni. Jako příklad poslouží dotaz z obrázku č. 45. Země, ze které uživatel pochází, lze zjistit právě zanořením dotazu.


*Obrázek 51: Zanořování dotazů – příklad*
Zdroj: vlastní

První úrovní je v tomto případě atribut Location uzlu Me. Vzhledem k tomu, že atribut location vrací také ID, lze vytvořit zanořený dotaz a tedy druhou úrovní dotazu je atribut Location uzlu Location. Ten vrací konkrétní zemi, ve které město leží, dále pak zeměpisnou šířku, délku a poštovní směrovací číslo města.

Graph API podporuje také řazení vrácených dat. Lze si tak seřadit komentáře u fotografií od nejaktuálnějšího po ten nejstarší a naopak. Parametry, které této funkci lze přiřadit jsou chronological nebo reverse_chronological. Tato funkce se standardně využije právě při zkoumání komentářů k určitým uzlům. Na obrázku jsou komentáře k fotografii na stránce Nike a jsou seřazeny chronologicky, tedy od prvního komentáře po poslední. Řazení komentářů je podle atributu created_time, který nese časovou složku dle standardu ISO 8601.




*Obrázek 52: Řazení výsledných dat*
Zdroj: vlastní

Mimo standardní analýzu stránky zmíněnou výše, které poskytuje také přímo Facebook Insight, by bylo užitečné vědět, jak lidé reagují na jednotlivé příspěvky, případně jak se liší reakce dle skupin, určených podle vybraných metrik. Facebook již neposkytuje konkrétní analýzu každého příspěvku, nicméně pomocí Graph API lze vyhledat potřebná data, pokud budeme znát ID daného příspěvku.

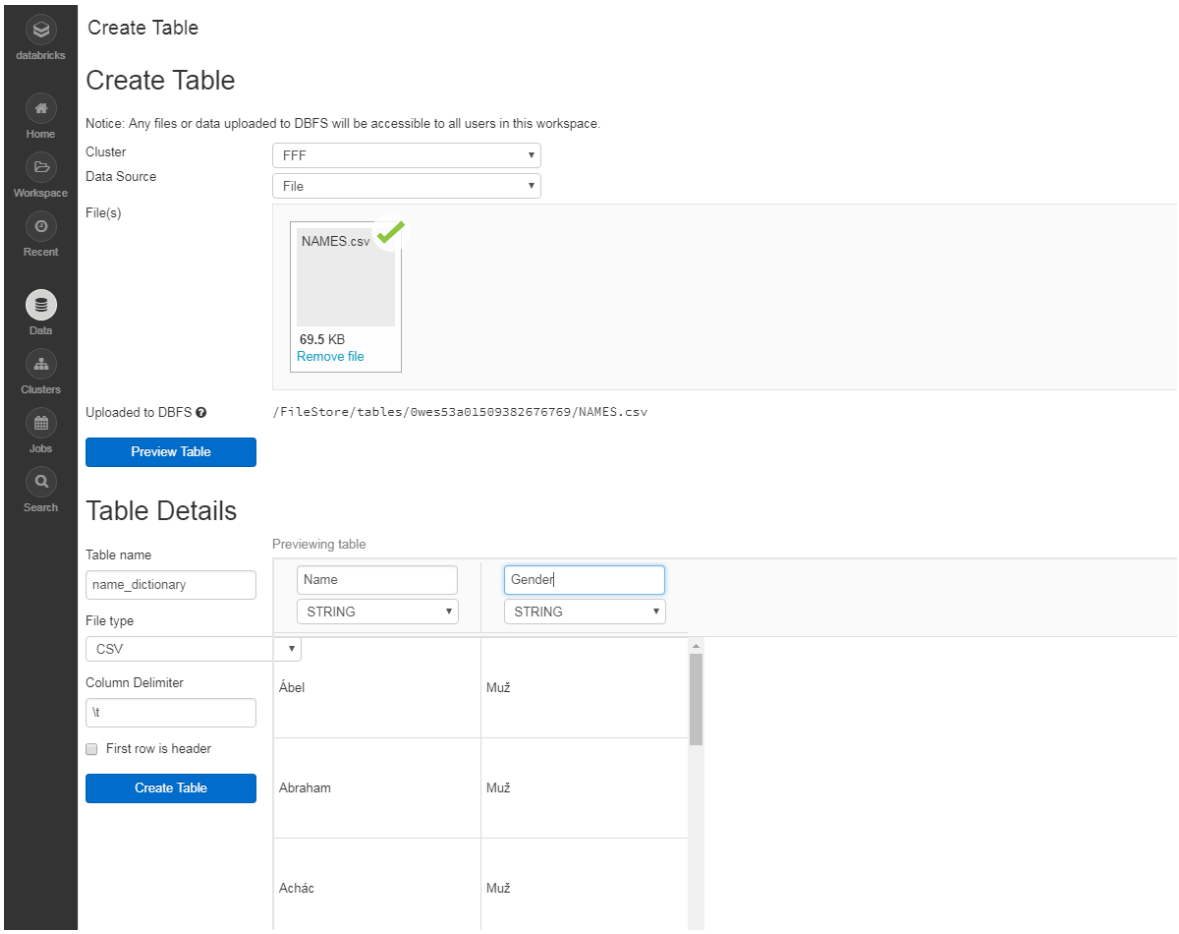
Následující analýza, bude zaměřena na odhad pohlaví uživatele, který klikl na tlačítko „To se mi líbí“ u libovolného příspěvku. Analýzou konkrétního příspěvku se dá následně zjistit, jak příspěvek zaujal jednotlivé pohlavní skupiny a přizpůsobit pak vkládání dalších příspěvků s ohledem na předešlé analýzy. To umožní nejen mnohem lépe zacílit jednotlivé příspěvky, ale také to poskytne cenná data, která se dají využít při volení marketingových strategií.

Určováním pohlaví na základě jména se zabývá spousta softwarů, stránek a vzniklo dokonce i pár aplikačních rozhraní, které se touto úlohou zaobírají. To, že se jedná o poměrně komplexní úlohu, dokazuje fakt, že se dodnes v registračních formulářích setkáváme s manuálním určením pohlaví (standardně je ve formuláři checkbox či jiná komponenta na určení pohlaví uživatele). Znát pohlaví uživatele je naprosto zásadní věc při marketingových

kampaních, kdy je žádoucí co největší personalizace. Právě například při personalizaci e-mailů se s touto úlohou lze setkat.

Pro určení pohlaví ze jména užívají softwary nejrůznější algoritmy, které dokážou pohlaví určit s pravděpodobností blížící se sto procentům. V rámci této analýzy bude využit jednoduchý algoritmus, který bude pouze porovnávat křestní jméno uživatele s číselníkem ženských nebo mužských jmen.

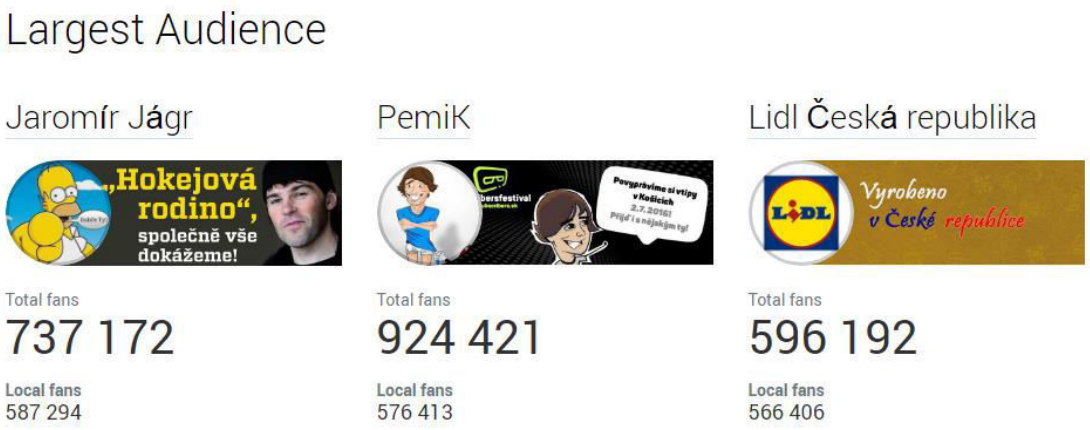
Prvním krokem bylo vydefinování mužských a ženských jmen, které se vyskytují na území České republiky. Data, která se používají v analýze, jsou k dispozici na webu krestni-jmena.cz. Tento web má vlastní databázi všech jmen, které jsou registrované v České republice. Po stažení všech jmen bylo potřeba konvertovat soubor se jmény do formátu CSV kvůli jednoduššímu nahrání do platformy Databricks, viz obrázek níže.




*Obrázek 53: Nahrání jmen do Databricks*
Zdroj: vlastní

Po nahrání seznamu jmen bylo potřeba zvolit vhodná testovací data. Při výběru vhodného testovacího vzorku bylo potřeba vybrat stránku na Facebooku tak, aby měla pokud možno, co největší podíl českých fanoušků. Další kritérium pro výběr testovacích dat byla pak aktivita fanoušků stránky, primárně pak suma všech „To se mi líbí“ (dále jen Likes), které stránka dokázala nasbírat. Kritéria výběru korespondují s cílem získat co největší počet testovacích dat, tedy vybrat český příspěvek s co nejvíce počty likes. K výběru testovacích dat byla použita platforma české společnosti Socialbakers, která se zabývá analýzou všech sociálních sítí a je největší a nejúspěšnější českou společností v tomto odvětví.

První metrikou byl tedy počet českých fanoušků Facebook stránky, potažmo pak jejich podíl. Na obrázku níže jsou vidět tři facebook stránky, které mají českých fanoušků nejvíce.




*Obrázek 54: Nejvíce českých fanoušků*
Zdroj: platforma Socialbakers

Jak je patrné z obrázku, prvenství v této metrice nese facebook stránka hokejové hvězdy Jaromíra Jágr, která čítá přesně 587 294 českých fanoušků. Na druhém místě je pak oficiální stránka československého baviče PemiKa. Třetí místo obsadila stránka Lidlu.

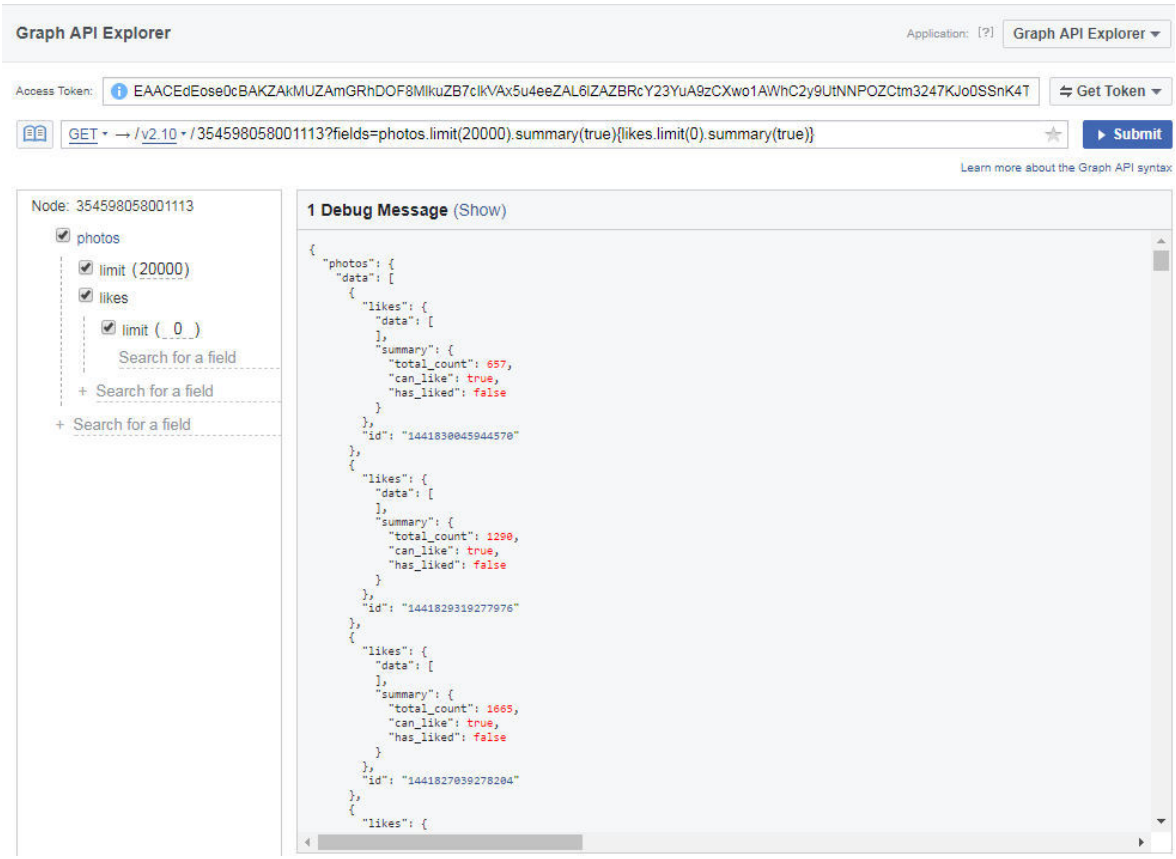
Po vyhodnocení této metriky, bylo potřeba zjistit, která z těchto tří stránek má nejaktivnější fanoušky v oblasti klikání na tlačítko „To se mi líbí“. To, že stránka má nejvíce českých fanoušků ještě neznamená, že nutně má také všechny fanoušky aktivní.

| Name | Sum of Like Reactions |
|---|---|
| PemiK / PemiKstranka | 2 237 036 |
| Jaromír Jágr / 68Jagr | 558 466 |
| Lidl Česká republika / lidlcesko | 46 527 |
| | N/A |

*Obrázek 55: Počet celkových "To se mi líbí"*
Zdroj: platforma Socialbakers

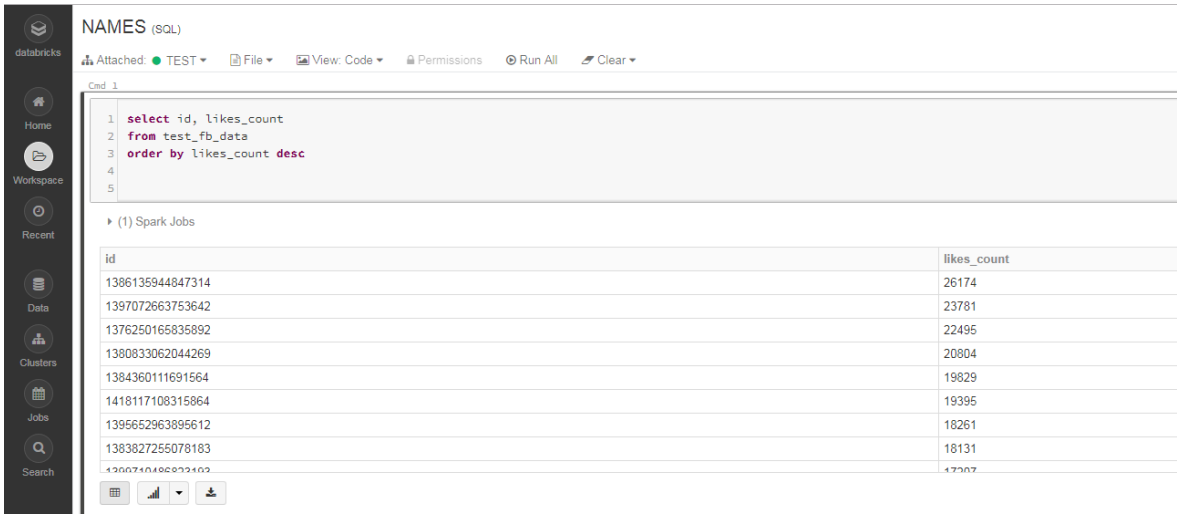
Z obrázku je vidět, že jasným vítězem co se týče sumy likes je stránka PemiK. Fanoušci s touto stránkou celkově mnohem více interagují než s těmi ostatními. Je to dáno také tím, jak často správce stránky přidává příspěvky a také jaký obsah je prezentován. Obsah stránky PemiK je vtipný a dost virální, tudíž je zřejmé, že také tato stránka bude mít nejvíce aktivních fanoušků. Jako testovací data poslouží tedy jeden z příspěvků ze stránky PemiK, jelikož právě tam bude největší testovací vzorek.

Po zvolení vhodné stránky bylo potřeba najít post s nejvíce likes, která kdy tato stránka vytvořila. Zdánlivě triviální úloha je však poměrně komplikovaná. Tuto úlohu značně ulehčuje fakt, že stránka PemiK je založena hlavně na vtipných fotografiích, tudíž stačí prohledat všechny fotografie/obrázky a zjistit, která má nejvíce likes. Tato úloha byla vyřešena pomocí Graph API Exploreru, kdy byl vytvořen tzv. zanořený dotaz, viz obrázek níže.




*Obrázek 56: Zanořený dotaz pro dohledání fotky s nejvíce likes*
Zdroj: vlastní

Tento výsledný JSON byl dále nahrán do platformy Databricks a pak pomocí SQL zjištěno ID fotografie/obrázku s nejvíce likes, viz obrázek č. 57.

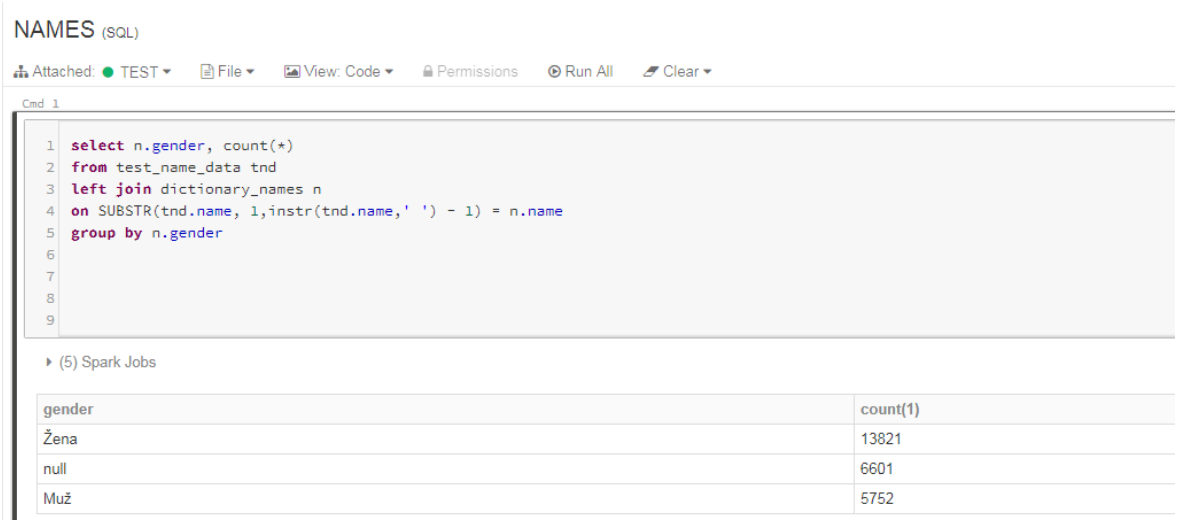



*Obrázek 57: ID fotky s nejvíce likes*
Zdroj: vlastní

Fotografie/obrázek, který získal nejvíce likes (přes 26000) byl následně zvolen k testování.

Potom co bylo nalezeno ID objektu s nejvíce likes, tak byl využit znovu nástroj Graph API Explorer. Přes API byl nalezen hledaný objekt a po spuštění složeného příkazu byla vyhledána jména všech lidí, co tomuto objektu dali „To se mi líbí“. Seznam všech lidí se vrátil standardně ve struktuře JSON, která byla převedena pomocí open-source nástroje konkolone.io do CSV formátu a následně nahrána do platformy Databricks, kde proběhla následná analýza.

Pro zjištění kolik žen a kolik mužů dalo „To se mi líbí“ danému objektu, bylo potřeba k množině jmen připojit vydefinovaný číselník se jmény a jejich pohlavími. Číselník se k množině připojil left joinem, jelikož bylo předpokládáno nedohledání některých křestních jmen v číselníku (na platformě facebook si často uživatelé volí smyšlená jména u kterých je samozřejmě až nemožné určit pohlaví, jelikož to nejsou křestní jména, ale spíše přezdívky). Číselník byl připojen k testovacím datům klauzulí ON na základě jména. Nicméně jelikož Facebook API vrací pouze celé jméno uživatele, který dal „To se mi líbí“, tak bylo potřeba zajistit výběr jen křestního jména. V rámci připojovací klauzule byla tedy použita funkce SUBSTR a v jejím rámci ještě funkce INSTR, které společně zajistily výběr jen prvního slova do mezery (to je předpokládané křestní jméno, nicméně funkce pro vybírání křestního jména by šla ještě zoptimalizovat




*Obrázek 58: Facebook analýza*
Zdroj: vlastní

Jak lze vidět z obrázku, objektu dalo „To se mi líbí“ 13821 Žen a 5752 Mužů. Hodnota null se vyskytla 6601krát, což znamená, že se jména nedohledala v číselníku jmen s pohlavím. To může být způsobeno několika důvody:

- Uživatel nezadal své pravé křestní jméno
- Uživatel zadal nejprve příjmení a až poté křestní jméno
- Uživatelovo křestní jméno se na území České republiky ještě nevyskytlo, tudíž není v databázi
- Uživatel zadal hypokoristickou podobu křestního jméno[30]

Z výsledků analýzy je patrné, že se daný objekt spíše líbil ženám. Z toho lze vyvodit několik závěrů, například že tématika daného objektu byla spíše ženského zaměření. Díky této analýze lze publikovat různé objekty a zkoumat reakci jednotlivých cílových skupin a zjišťovat tak na co skupiny reagují lépe, na co hůře a podle toho pak přizpůsobit svou reklamní strategii.

[30] Jméno užívané v neformálním nebo domáckém prostředí např. místo Vojtěch - Vojta

Tato analýza sloužila jen jako jednoduchý příklad toho, jak by se dala využít data ze sociální sítě Facebook na platformě Databricks.

# 5. Doporučení pro MSP

Každý MSP by měl mít určenou základní koncepci pro analýzu dat. Tato základní koncepce může být vydefinována pomocí následujících kroků:

- Vydefinování cíle datové analýzy
- Určení datových zdrojů v organizaci využitelných pro analýzu dat
- Zvolení nástroje pro analýzu dat
- Promyslet, jakým způsobem bude realizován sběr dat a následná analýza
- Prezentace dat

Jedním z nejdůležitějších kroků při analýze dat, je určit její cíl, jelikož to bude využito jako základní stavební kámen ke všem dalším krokům. Cíle datové analýzy by měly vycházet z informační strategie podniku. Například zjištění jakým způsobem zákazníci reagují na nový produkt, kterým jazykem primárně komunikovat se zákazníky přes sociální sítě nebo třeba jak vylepšit křížový prodej[31].

Zvolení datových zdrojů musí korespondovat se stanoveným cílem., tedy pokud je cílem zjistit jak zákazníci reagují na nový produkt, jedním z možných datových zdrojů jsou sociální sítě. Lze využívat jak interní datové zdroje (typicky jsou zdrojem interní systémy, např. skladový systém, software na účetnictví atd.), tak externí datové zdroje. Jako externí datové zdroje lze označit všechna volně dostupná data. Kromě dat ze sociálních sítí, ze kterých mohou podniky vytěžit nejvíce informací, existuje také spousta volně dostupných dat, která mohou mít pro MSP přínos. Program Fórum pro otevřená data poskytuje na webové stránce http://www.otevrenadata.cz/ přehled otevřených dat, která jsou volně dostupná. Lze zde nalézt jak data veřejné správy, tak volně dostupná data z jednotlivých krajů v České republice.

Na základě porovnání nástrojů pro analýzu Big Data podle kritérií, která zohledňují informační potřeby a možnosti MSP (podkapitola 3.5) je doporučeno využít nástroj

[31] Jedná se o obchodní techniku, která má za cíl prodat více produktů doporučením souvisejícího produktu. Např. když si zákazník koupí myš k počítači, tak mu nabídnout také podložku pod myš

Databricks. Tento nástroj umožňuje snadnou, ale zároveň komplexní analýzu dat díky své rozmanitosti. Hlavní nevýhoda tkví v tom, že je nástroj postaven čistě na cloudovém řešení, což pro některé podniky nemusí být zrovna pohodlné. Pokud nástroj z některých důvodů nebude MSP vyhovovat, lze využít stejný postup pro vybrání ideálního nástroje, který je v podkapitole 3.5 (tj. určení kritérií pro hodnocení, stanovení vah kritérií a následně vybrat nejvhodnější nástroj).

Existuje mnoho způsobů jak data sbírat a analyzovat, případně jaké datové analýzy využít. Opěrným bodem by měl být opět předem stanovený cíl analýzy a také nástroj, který byl pro analýzu dat zvolen. V kapitole 4 je uveden postup, jak lze nástroj Databricks využít pro sběr a analýzu dat ze sociálních sítí Twitter a Facebook. Analýza dat ze sociální sítě Twitter je zaměřená na streamovaná data, což znamená, že lze v reálném čase sledovat reakce uživatelů například z jaké země, je reakcí nejvíce. S tímto faktem lze následně pracovat v rámci plánování zásob a distribuce daného produktu. Největším přínosem této analýzy je fakt, že umožňuje okamžité reakce podniku. Kromě konkrétní analýzy, která je provedena v této práci si mohou MSP zpracovat např. analýzu sentimentu nad streamovanými daty. To by podnikům pomohlo porozumět co se v dané chvíli děje s jejich značkou a dynamicky tak reagovat na jakoukoli změnu názoru uživatelů na daný podnik. Analýza sentimentu v reálném čase může být velmi silný nástroj, co se týče sledování názoru na produkt a v některých případech může i značně snížit náklady[32].

Analýza dat ze sociální sítě Facebook není postavená nad streamovanými daty jako analýza dat z Twitteru, ale pouze nad statickými daty, která se extrahovala pomocí Facebookového Graph API Exploreru, postup extrahování dat je popsán v podkapitole 4.2. Tato analýza by se dala ještě rozšířit tak, aby dávala přesnější rozčlenění pohlaví dle jména. Nabízí se hned několik vylepšení a to například zohlednit hypokoristickou podobu jména. Nejjednodušším

[32] V práci „*Detecting Sentiment Change in Twitter Streaming Data*“ od autorů A. Biffeta, G. Holmese a B. Pfahringera je uveden případ krize automobilového podniku Toyota z roku 2010, kdy měly některé vozy této značky v USA problém s plynovým pedálem. Toyotě trvalo kolem tří měsíců, než si uvědomila, že se jedná o krizi většího rozsahu. V dané práci je dokázáno, že v tomto období se sentiment uživatelů Twitter na značku Toyota výrazně zhoršil (cca o 50%) a spolu s hashtagem #Toyota se začala v jednotlivých tweetech objevovat slova jako „gas“ či „pedals“. Změna sentimentu v reálném čase by Toyotě umožnila rozpoznat, že se jedná o mnohem větší problém a následně by podnik mohl lépe zareagovat

způsobem jak toto implementovat je přidat tvary jmen do číselníku. Poté by se klauzulí ON, která se využívá pro propojení číselníku a tabulky, připojila daná jména. Dalším rozšířením by pak mohlo být zahrnutí pořadí jmen. To znamená, že pokud uživatel zadal své jméno v pořadí: 1. Příjmení, 2. Křestní, tak se pohlaví neurčí. Implementace tohoto vylepšení by zpřesnila analýzu, nicméně by bylo potřeba vzít v potaz to, že uživatel může mít jako příjmení standardní křestní jméno např. Tomáš Lukáš.

Prezentace výsledných dat je velmi důležitá úloha, která napomáhá celkovému porozumění výstupu analýzy. V nástroji Databricks lze data prezentovat pomocí vizualizací, jako jsou grafy, tabulky, ale i mapy.

Největším předností nástroje Databricks je možnost psaní vlastních skriptů (v jazycích Python či Scala) a dělat složitější analýzy dat. Jako příklad je v kapitole 3.5.3 uvedeno hledání slova pomocí skriptu v jazyce Python. Nicméně lze dělat i mnohem složitější analýzu, jako např. analýza sentimentu. Naopak největším omezením platformy Databricks (respektive její verze zdarma Community Edition) je objem paměti poskytovaného clusteru, což zásadně omezuje objem analyzovaných dat a dále pak rychlost celého řešení. Tudíž postavení nějakého robustního řešení, které bude analyzovat Big Data v řádů sekund je s Community Edition Databricks nemyslitelné. Ale pro MSP, které chtějí začít s pokročilejší analýzou svých Big Data je tento nástroj ideální.

# Závěr

Cílem práce bylo najít a porovnat nástroje, které umožňují zpracování Big Data a zároveň jsou dostupné MSP. Nástroje byly vybrány na základě kritérií, které korespondují s potřebami MSP, jako je např. dostupnost nástoje zdarma. Dalším cílem bylo jeden z nástrojů vybrat, provést analýzu dat ze sociálních sítí a poskytnout MSP konkrétní příklady analýz, které by se daly využít. Nástroj byl vybrán pomocí porovnání prostřednictvím metody vícekriteriálního hodnocení, kde byla nejprve stanovena kritéria hodnocení, kterým byly přiděleny váhy podle důležitosti. Kritéria byla stanovena na základě informací z odborných publikací a potřeb MSP. Byly porovnávány čtyři nástroje a jako nejvhodnější byl zvolen nástroj Databricks. Pomocí tohoto nástroje pak byla analyzována data ze sociálních sítí Facebook a Twitter.

Cíle byly splněny a MSP mohou tedy využít řešení postavené nad nástrojem Databricks a začít analýzy využívat. Konkrétně nad sociální sítí Twitter lze využít řešení, které zahrnuje analýzu streamovaných dat, tzn. lze analyzovat data v reálném čase. Analýza dat ze sociální sítě Facebook pak poskytuje MSP možnost analyzovat konkrétní příspěvky a určovat např. pohlaví uživatelů, kteří se nějakým způsobem vyjádřili k danému příspěvku (konkrétně dali „To se mi líbí“).

Big Data jsou stále fenoménem, který se snaží podniky nějakým způsobem uchopit a vytěžit z něj to nejlepší. V budoucnu přinese efektivní zpracování všech relevantních dat pro určitý podnik výraznou změnu ve fungování a hlavně pak dramaticky zefektivní veškeré podnikové činnosti. Díky využití analýzy Big Data se zlepší každodenní rozhodování a plánování podniku, dále se pak značně zlepší cílení na zákazníka, které může mít za následek zvýšení zisku. Nicméně zatím podniky zdaleka nevyužívají potenciál dat, která jsou pro něj relevantní. S neustále vyvíjejícím se trendem „Internet věcí“ (Internet of Things) bude nárůst dat čím dál tím větší a je jen otázka firem, zda tyto datové zdroje začnou využívat.

# Seznam použité literatury

## Citace


HOLUBOVÁ, Irena, Jiří KOSEK, Karel MINAŘÍK a David NOVÁK. *Big Data a NoSQL databáze*. 1. vyd., Praha: Grada, 2015. Profesionál. ISBN 978-80-247-5466-6.

MAYER-SCHÖNBERGER, Viktor a Kenneth CUKIER. *Big Data: Revoluce, která mění způsob, jak žijeme a myslíme*. 1. vyd. Brno: Computer Press, 2014. ISBN 978-80-251-4119-9.

Evropská komise. *Uživatelská příručka: k definici malých a středních podniků* [online]. 1. Lucemburk: Úřad pro publikace Evropské unie, 2015 [cit. 2017-5-14]. ISBN 978-92-79-45316-8. Dostupné z: https://ec.europa.eu/docsroom/documents/15582/attachments/1/translations/cs/renditions/native.

LOSHIN, David. *Big data analytics: from strategic planning to enterprise integration with tools, techniques, NoSQL, and graph.* Waltham, Mass.: Academic Press, 2013. 1st ed. ISBN 978-0-12-417319-4.

O'Reilly Media, Inc. *Big Data Now: 2014 Edition*. 1st ed. California: O'Reilly Media, 2014. ISBN 978-1-491-91736-7.

KRISHNAN, Krish. *Data warehousing in the age of big data.* 1st ed. Waltham: Elsevier, 2013. ISBN 978-012-4058-910.

MADISON, Michael et al. NoSQL Database Technologies. Journal of International Technology and Information Management [online]. 2015, vol. 24, no. 1, s. 1-I. ISSN 15435962.Dostupné z: https://search.proquest.com/docview/1757727794?accountid=17116

GUPTA, Amarnath. *Characteristics of Big Data - Variety*. In: Coursera [online]. [cit. 2017-02-28]. Dostupné z: https://www.coursera.org/learn/big-data-introduction/lecture/oVg4p/characteristics-of-big-data-variety

DEAN, Jeffrey a Sanjay GHEMAWAT. *MapReduce: Simplified Data Processing on Large Clusters*. In: Sixth Symposium on Operating System Design and Implementation [online]. San Francisco, 2004 [cit. 2017-02-28]. Dostupné z: https://static.googleusercontent.com/media/research.google.com/cs//archive/mapreduce-osdi04.pdf

**Bibliografie**

GHEMAWAT, Sanjay, Howard GOBIOFF a Shun-Tak LEUNG. *The Google File System* [online]. [cit. 2016-12-30]. Dostupné z: https://static.googleusercontent.com/media/research.google.com/cs//archive/gfs-sosp2003.pdf

WALKER, Ben. *EVERY DAY BIG DATA STATISTICS – 2.5 QUINTILLION BYTES OF DATA CREATED DAILY* [online]. In: 2015 [cit. 2017-01-22]. Dostupné z: http://www.vcloudnews.com/every-day-big-data-statistics-2-5-quintillion-bytes-of-data-created-daily/

VORHIES, William. *How Many "V's" in Big Data? The Characteristics that Define Big Data* [online]. **2014** [cit. 2017-01-22]. Dostupné z: http://www.datasciencecentral.com/profiles/blogs/how-many-v-s-in-big-data-the-characteristics-that-define-big-data

MILOŠ, Marek. *Nástroje pro Big Data Analytics*. Praha, 2013. Diplomová práce. Vysoká škola ekonomická v Praze. Vedoucí práce Doc. Ing. Jan Pour, CSc.

LINHART, Ondřej. *Využití dat ze sociálních sítí pro BI*. Praha, 2013. BAKALÁŘSKÁ PRÁCE. Vysoká škola ekonomická v Praze. Vedoucí práce Ing. PhDr. Antonín Pavlíček, Ph.D.

FACEBOOK. Graph API Reference. Developers.facebook.com [online]. 2015 [cit. 2017-03-21]. Dostupné z: https://developers.facebook.com/docs/graph-api/reference

Hortonworks Team. Hortonworks Data Platform. [online] 2013 [cit. 2017-03-21]. Dostupné z: http://hortonworks.com/products/hdp/

Cloudera Team. Cloudera's Distribution of Hadoop. [online] 2013 [cit. 2017-03-21]. Dostupné z: http://www.cloudera.com/content/cloudera/en/products/cdh.html

Databricks Team. Databricks Platform. [online] 2015 [cit. 2017-03-21]. Dostupné z: https://databricks.com/product/unified-analytics-platform

Evropská komise. *2016 SBA Fact Sheet - Czech Republic* [online]. 1. Europe: Europe Comission, 2016 [cit. 2017-12-14]. Dostupné z: https://ec.europa.eu/growth/smes/business-friendly-environment/performance-review_en#sba-fact-sheets

Asociace malých a středních podniků a živnostníků České republiky (AMSP ČR). *Investice do IT a práce s daty ve firmách* [online]. 1. Praha: AMSP, 2014 [cit. 2017-12-14]. Dostupné z: http://amsp.cz/uploads/Pruzkumy/Vysledky_pruzkumu_Investice_do_IT_a_prace_s_daty_ve_firmach.pdf

ČESKÝ STATISTICKÝ ÚŘAD. *Využívání informačních a komunikačních technologií v podnikatelském sektoru - v roce 2015* [online]. 1. Praha: Odbor statistik rozvoje společnosti, 2015 [cit. 2017-12-14]. Dostupné z: https://www.czso.cz/documents/10180/37556244/062005-15.pdf/004ef709-90ed-4cec-9d07-a0f685148dad?version=1.2

LANDROCK, Holm, Oliver SCHONSCHEK a Prof. Dr. Andreas GADATSCH. *Big Data Vendor Benchmark 2015* [online]. In: . Mnichov, Německo, 2015, s. 91 [cit. 2017-05-04]. Dostupné z: https://www.t-systems.com/solutions/big-data-vendor-download/1298900_1/blobBinary/Big+Data+Vendor_Download-ps.pdf

LOSHIN, David. *Comparing the leading big data analytics software options* [online]. 2015 [cit. 2017-05-04]. Dostupné z: http://searchbusinessanalytics.techtarget.com/feature/Comparing-the-leading-big-data-analytics-software-options

# Seznam příloh

# Příloha A - Podniky využívající sociální sítě, leden 2017

*podíl na celkovém počtu firem v dané velikostní a odvětvové skupině (v %)*

| Odvětví (ekonomická činnost) – CZ NACE | Velikost firmy (počet zaměstnanců) | | | |
|---|---|---|---|---|
| | 10–49 | 50–249 | 250+ | Celkem |
| **Zpracovatelský průmysl – C (10–33)** | **22,3** | **41,2** | **60,2** | **29,8** |
| Potravinářský, nápojový a tabákový průmysl (10–12) | 34,2 | 53,5 | 72,6 | 40,9 |
| Textilní, oděvní, kožedělní a obuvnický průmysl (13–15) | 32,7 | 32,7 | 38,7 | 32,9 |
| Dřevozpracující a papírenský průmysl (16–18) | 21,1 | 40,8 | 48,6 | 25,5 |
| Chemický, farmaceutický, gumárenský a plastový průmysl; Průmysl skla a stavebních | 13,9 | 34,7 | 64,6 | 25,3 |
| Výroba kovů, hutních a kovodělných výrobků (24–25) | 13,4 | 37,2 | 58,1 | 20,8 |
| Výroba počítačů, elektronických a optických přístrojů a zařízení (26) | 25,7 | 50,9 | 71,4 | 36,8 |
| Výroba elektrických zařízení, výroba strojů a zařízení j. n. (27–28) | 26,2 | 46,8 | 66,2 | 36,3 |
| Automobilový průmysl a výroba ostatních dopravních prostředků (29–30) | 27,8 | 31,0 | 53,6 | 36,9 |
| Výroba nábytku; Ost. zpracovatelský průmysl; Opravy a instalace strojů a zařízení (31–33) | 26,1 | 45,0 | 53,5 | 31,0 |
| **Výroba a rozvod energie, plynu, vody, tepla a činn. související s odpady – D, E (35–39)** | **17,2** | **41,9** | **56,4** | **26,3** |
| **Stavebnictví – F (41–43)** | **20,7** | **26,2** | **55,5** | **21,6** |
| **Velkoobchod a maloobchod; opravy a údržba motorových vozidel – G (45–47)** | **41,6** | **72,4** | **76,7** | **46,0** |
| Velkoobchod, maloobchod a opravy motorových vozidel (45) | 48,2 | 71,8 | 86,7 | 52,0 |
| Velkoobchod, kromě motorových vozidel (46) | 40,0 | 72,8 | 75,5 | 44,6 |
| Maloobchod, kromě motorových vozidel (47) | 41,3 | 72,0 | 75,8 | 45,8 |
| **Doprava a skladování – H (49–53)** | **19,7** | **37,8** | **73,5** | **25,0** |
| **Ubytování, stravování a pohostinství – I (55–56)** | **58,0** | **78,4** | **95,5** | **59,8** |
| Ubytování (55) | 76,0 | 92,9 | 100,0 | 78,5 |
| Stravování a pohostinství (56) | 51,8 | 64,4 | 93,2 | 52,7 |
| **Informační a komunikační činnosti – J (58–63)** | **67,2** | **83,2** | **93,1** | **71,4** |
| Činnosti v oblasti vydavatelství, filmu, videozáznamů a televizních programů (58–60) | 78,8 | 94,3 | 91,0 | 82,2 |
| Telekomunikační činnosti (61) | 69,6 | 82,5 | 100,0 | 72,8 |
| Činnosti v oblasti informačních technologií; Informační činnosti (62–63) | 64,4 | 81,1 | 92,6 | 68,9 |
| **Činnosti v oblasti nemovitostí – L (68)** | **23,2** | **39,4** | **.** | **24,6** |
| **Profesní, vědecké a technické činnosti – M (69–75)** | **38,6** | **40,3** | **88,1** | **39,9** |
| **Administrativní a podpůrné činnosti – N (77–82)** | **30,5** | **37,6** | **53,9** | **34,2** |
| Činnosti cestovních agentur a kanceláří (79) | 83,5 | 83,6 | . | 84,2 |
| Ostatní administrativní a podpůrné činnosti (77–78, 80–82) | 24,4 | 36,4 | 52,0 | 30,1 |
| **Celkem** | **32,7** | **46,8** | **65,6** | **36,5** |

Zdroj:https://www.czso.cz/csu/czso/vyuzivani-informacnich-a-komunikacnich-technologii-v-podnikatelskem-sektoru-2016-2017

# Příloha B - Podniky využívající sociální média ke zlepšování obrazu firmy či uvádění produktů na trh

*podíl na celkovém počtu firem v dané velikostní a odvětvové skupině (v %)*

| Odvětví (ekonomická činnost) – CZ NACE | Velikost firmy (počet zaměstnanců) | | | |
|---|---|---|---|---|
| | 10–49 | 50–249 | 250+ | Celkem |
| **Zpracovatelský průmysl – C (10–33)** | **18,4** | **34,4** | **49,4** | **24,6** |
| Potravinářský, nápojový a tabákový průmysl (10–12) | 29,7 | 48,1 | 67,5 | 36,2 |
| Textilní, oděvní, kožedělní a obuvnický průmysl (13–15) | 26,6 | 27,7 | 34,9 | 27,2 |
| Dřevozpracující a papírenský průmysl (16–18) | 19,7 | 40,8 | 48,6 | 24,4 |
| Chemický, farmaceutický, gumárenský a plastový průmysl; Průmysl skla a stavebních | 12,1 | 28,3 | 57,2 | 21,5 |
| Výroba kovů, hutních a kovodělných výrobků (24–25) | 8,5 | 27,6 | 50,1 | 14,7 |
| Výroba počítačů, elektronických a optických přístrojů a zařízení (26) | 19,3 | 38,1 | 48,2 | 27,1 |
| Výroba elektrických zařízení, výroba strojů a zařízení j. n. (27–28) | 21,0 | 42,9 | 53,5 | 30,8 |
| Automobilový průmysl a výroba ostatních dopravních prostředků (29–30) | 24,0 | 17,0 | 34,4 | 25,0 |
| Výroba nábytku; Ost. zpracovatelský průmysl; Opravy a instalace strojů a zařízení (31–33) | 23,1 | 36,3 | 46,3 | 26,7 |
| **Výroba a rozvod energie, plynu, vody, tepla a činn. související s odpady – D, E (35–39)** | **11,0** | **33,5** | **52,7** | **19,7** |
| **Stavebnictví – F (41–43)** | **16,4** | **22,9** | **49,2** | **17,4** |
| **Velkoobchod a maloobchod; opravy a údržba motorových vozidel – G (45–47)** | **38,1** | **67,8** | **71,6** | **42,3** |
| Velkoobchod, maloobchod a opravy motorových vozidel (45) | 43,5 | 67,0 | 86,7 | 47,3 |
| Velkoobchod, kromě motorových vozidel (46) | 36,8 | 65,9 | 71,9 | 40,9 |
| Maloobchod, kromě motorových vozidel (47) | 37,8 | 72,0 | 69,0 | 42,5 |
| **Doprava a skladování – H (49–53)** | **13,8** | **30,1** | **65,6** | **18,7** |
| **Ubytování, stravování a pohostinství – I (55–56)** | **51,4** | **73,7** | **91,0** | **53,3** |
| Ubytování (55) | 68,0 | 90,3 | 100,0 | 71,3 |
| Stravování a pohostinství (56) | 45,7 | 57,5 | 86,3 | 46,6 |
| **Informační a komunikační činnosti – J (58–63)** | **57,7** | **75,7** | **87,8** | **62,4** |
| Činnosti v oblasti vydavatelství, filmu, videozáznamů a televizních programů (58–60) | 72,3 | 82,8 | 83,7 | 74,8 |
| Telekomunikační činnosti (61) | 66,0 | 72,6 | 87,5 | 68,0 |
| Činnosti v oblasti informačních technologií; Informační činnosti (62–63) | 53,3 | 74,7 | 88,9 | 59,0 |
| **Činnosti v oblasti nemovitostí – L (68)** | **21,1** | **38,0** | **.** | **22,4** |
| **Profesní, vědecké a technické činnosti – M (69–75)** | **35,4** | **32,6** | **61,1** | **35,6** |
| **Administrativní a podpůrné činnosti – N (77–82)** | **24,7** | **29,8** | **46,3** | **27,7** |
| Činnosti cestovních agentur a kanceláří (79) | 75,7 | 83,6 | . | 77,4 |
| Ostatní administrativní a podpůrné činnosti (77–78, 80–82) | 18,9 | 28,5 | 44,1 | 23,6 |
| **Celkem** | **28,3** | **40,4** | **56,0** | **31,5** |

Zdroj:https://www.czso.cz/csu/czso/vyuzivani-informacnich-a-komunikacnich-technologii-v-podnikatelskem-sektoru-2016-2017

# Příloha C - Podniky využívající sociální média k získávání názorů/otázek od zákazníků

*podíl na celkovém počtu firem v dané velikostní a odvětvové skupině (v %)*

| Odvětví (ekonomická činnost) – CZ NACE | Velikost firmy (počet zaměstnanců) | | | |
|---|---|---|---|---|
| | 10–49 | 50–249 | 250+ | Celkem |
| **Zpracovatelský průmysl – C (10–33)** | **12,4** | **22,4** | **29,7** | **16,2** |
| Potravinářský, nápojový a tabákový průmysl (10–12) | 27,7 | 37,6 | 60,8 | 31,8 |
| Textilní, oděvní, kožedělní a obuvnický průmysl (13–15) | 18,9 | 18,8 | 34,9 | 19,5 |
| Dřevozpracující a papírenský průmysl (16–18) | 9,9 | 23,0 | 26,9 | 12,8 |
| Chemický, farmaceutický, gumárenský a plastový průmysl; Průmysl skla a stavebních | 6,8 | 20,9 | 35,3 | 14,0 |
| Výroba kovů, hutních a kovodělných výrobků (24–25) | 6,1 | 14,9 | 31,4 | 9,2 |
| Výroba počítačů, elektronických a optických přístrojů a zařízení (26) | 12,1 | 26,7 | 22,3 | 17,1 |
| Výroba elektrických zařízení, výroba strojů a zařízení j. n. (27–28) | 11,0 | 22,5 | 28,7 | 16,2 |
| Automobilový průmysl a výroba ostatních dopravních prostředků (29–30) | 17,9 | 12,3 | 15,1 | 15,2 |
| Výroba nábytku; Ost. zpracovatelský průmysl; Opravy a instalace strojů a zařízení (31–33) | 15,0 | 29,8 | 25,3 | 18,4 |
| **Výroba a rozvod energie, plynu, vody, tepla a činn. související s odpady – D, E (35–39)** | **8,6** | **21,5** | **32,7** | **13,6** |
| **Stavebnictví – F (41–43)** | **7,8** | **14,2** | **27,9** | **8,7** |
| **Velkoobchod a maloobchod; opravy a údržba motorových vozidel – G (45–47)** | **25,2** | **56,0** | **61,9** | **29,6** |
| Velkoobchod, maloobchod a opravy motorových vozidel (45) | 29,3 | 57,4 | 69,7 | 33,7 |
| Velkoobchod, kromě motorových vozidel (46) | 24,5 | 52,6 | 52,4 | 28,4 |
| Maloobchod, kromě motorových vozidel (47) | 24,5 | 62,0 | 66,7 | 30,0 |
| **Doprava a skladování – H (49–53)** | **8,2** | **22,4** | **51,1** | **12,4** |
| **Ubytování, stravování a pohostinství – I (55–56)** | **43,5** | **68,3** | **85,5** | **45,7** |
| Ubytování (55) | 64,7 | 83,2 | 83,8 | 67,3 |
| Stravování a pohostinství (56) | 36,2 | 53,8 | 86,3 | 37,5 |
| **Informační a komunikační činnosti – J (58–63)** | **34,3** | **60,4** | **67,6** | **40,6** |
| Činnosti v oblasti vydavatelství, filmu, videozáznamů a televizních programů (58–60) | 55,0 | 74,8 | 91,0 | 60,6 |
| Telekomunikační činnosti (61) | 49,1 | 58,3 | 75,0 | 51,5 |
| Činnosti v oblasti informačních technologií; Informační činnosti (62–63) | 27,6 | 57,7 | 60,7 | 34,8 |
| **Činnosti v oblasti nemovitostí – L (68)** | **13,8** | **32,3** | **.** | **15,2** |
| **Profesní, vědecké a technické činnosti – M (69–75)** | **21,6** | **26,9** | **42,1** | **22,7** |
| **Administrativní a podpůrné činnosti – N (77–82)** | **15,6** | **20,1** | **30,6** | **18,0** |
| Činnosti cestovních agentur a kanceláří (79) | 67,3 | 76,0 | . | 67,8 |
| Ostatní administrativní a podpůrné činnosti (77–78, 80–82) | 9,7 | 18,7 | 29,3 | 13,9 |
| **Celkem** | **18,8** | **29,7** | **38,3** | **21,4** |

Zdroj:https://www.czso.cz/csu/czso/vyuzivani-informacnich-a-komunikacnich-technologii-v-podnikatelskem-sektoru-2016-2017