EKONOMICKÁ FAKULTA TUL

Bakalářská práce

Zpracování marketingových dat prostřednictvím ETL nástroje Bizzflow

Studijní program: B0688P140002 Informační management
Autor práce: **Anna Ševčíková**
Vedoucí práce: Ing. Petr Weinlich, Ph.D.
Katedra informatiky

Liberec 2023

EKONOMICKÁ FAKULTA TUL

Zadání bakalářské práce

Zpracování marketingových dat prostřednictvím ETL nástroje Bizzflow

Jméno a příjmení:	**Anna Ševčíková**
Osobní číslo:	E20000188
Studijní program:	B0688P140002 Informační management
Zadávající katedra:	Katedra informatiky
Akademický rok:	2022/2023

Zásady pro vypracování:

1. Možnosti zpracování marketingových dat.
2. Vymezení relevantních databázových a Business Intelligence pojmů.
3. Představení ETL nástroje Bizzflow.
4. Analýza a transformace dat pomocí Bizzflow.
5. Zhodnocení a formulace doporučení.

Rozsah grafických prací:	
Rozsah pracovní zprávy:	30 normostran
Forma zpracování práce:	tištěná/elektronická
Jazyk práce:	Čeština

Seznam odborné literatury:

- FORTA, Ben, 2013. *Sams Teach Yourself SQL in 10 Minutes*. 4. vyd. Indianapolis: Sams Publishing. ISBN 978-06-7233-607-2.
- HARRINGTON, Jan L., 2016. *Relational Database Design and Implementation*. 4. vyd. Cambridge: Elsevier. ISBN 978-0-12-804399-8.
- PROKOP, Marek a kolektiv autorů, 2014. *Online marketing*. Brno: Computer Press. ISBN 978-80-251-4155-7.
- LAURENČÍK, Marek a Michal BUREŠ, 2018. *SQL: podrobný průvodce uživatele*. Praha: Grada Publishing. ISBN 978-80-271-0774-2.
- REEVE, April, , 2013. *Managing Data in Motion: Data Integration Best Practice Techniques and Technologies*. Waltham: Elsevier. ISBN 978-0-12-397167-8.
- PROQUEST, 2022. *Databáze článků ProQuest* [online]. Ann Arbor, MI, USA: ProQuest. [cit. 2022-09-30]. Dostupné z: http://knihovna.tul.cz/.

Konzultant: Ing. Jiří Tobolka, spoluzakladatel & partner Bizztreat

Vedoucí práce:	Ing. Petr Weinlich, Ph.D. Katedra informatiky
Datum zadání práce:	1. listopadu 2022
Předpokládaný termín odevzdání:	31. srpna 2024

L.S.

doc. Ing. Aleš Kocourek, Ph.D.
děkan

Ing. Petr Weinlich, Ph.D.
vedoucí katedry

V Liberci dne 1. listopadu 2022

Prohlášení

Prohlašuji, že svou bakalářskou práci jsem vypracovala samostatně jako původní dílo s použitím uvedené literatury a na základě konzultací s vedoucím mé bakalářské práce a konzultantem.

Jsem si vědoma toho, že na mou bakalářskou práci se plně vztahuje zákon č. 121/2000 Sb., o právu autorském, zejména § 60 – školní dílo.

Beru na vědomí, že Technická univerzita v Liberci nezasahuje do mých autorských práv užitím mé bakalářské práce pro vnitřní potřebu Technické univerzity v Liberci.

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

Současně čestně prohlašuji, že text elektronické podoby práce vložený do IS/STAG se shoduje s textem tištěné podoby práce.

Beru na vědomí, že má bakalářská práce bude zveřejněna Technickou univerzitou v Liberci v souladu s § 47b zákona č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších předpisů.

Jsem si vědoma následků, které podle zákona o vysokých školách mohou vyplývat z porušení tohoto prohlášení.

27. dubna 2023 Anna Ševčíková

Zpracování marketingových dat prostřednictvím ETL nástroje Bizzflow

Anotace

Tato bakalářská práce se zaměřuje na vysvětlení významných teoretických pojmů z oblastí online marketingu, databází, databázových systémů, SQL a Business Intelligence. V praktické části se tyto pojmy využívají při zpracování marketingových dat, jež jsou následně vizualizována v přehledném dashboardu. Cílem práce je přiblížit čtenářům významy těchto pojmů a představit řešení projektu prostřednictvím nástroje Bizzflow.

Klíčová slova

Bizzflow, Business Intelligence, databáze, ETL, GoodData, SQL, online marketing, vizualizace.

Marketing data processing via ETL tool Bizzflow

Annotation

This bachelor thesis focuses on explaining important theoretical concepts in the areas of online marketing, databases, database systems, SQL and Business Intelligence. In the practical part, these concepts are used in the processing of marketing data, which is then visualized in a clear dashboard. The aim of the thesis is to introduce the readers to the meanings of these concepts and to present the project solution through the Bizzflow tool.

Key Words

Bizzflow, Business Intelligence, databases, ETL, GoodData, SQL, online marketing, visualization.

Poděkování

Tímto bych ráda poděkovala Ing. Petru Weinlichovi, Ph.D. za metodické vedení práce a za jeho podporu, trpělivost, cenné rady a připomínky.

Ráda bych také poděkovala všem kolegům z firmy BizzTreat za praktické a hodnotné rady, předané zkušenosti a celkovou ochotu.

Obsah

Seznam obrázků

Seznam tabulek

Seznam zkratek

B2B	Business to business
BI	Business Intelligence
DDL	Data Definition Language
DMA	Datamart (datové tržiště)
DML	Data Manipulation Language
DWH	Data Warehouse (datový sklad)
ETL	Extract, Transform, Load
FK	Foreign key (cizí klíč)
IT	Information Technology (informační technologie)
KPI	Key Performance Indicator (klíčové ukazatele výkonnosti)
NF	Normal forms (normální formy)
PK	Primary key (primární klíč)
PR	Public relations
QaA	Questions and answers (otázky a odpovědi)
SEO	Search Engine Optimization (optimalizace pro vyhledavače)
SEQUEL	Structured English Query Language
SQL	Structured Query Language

Úvod

Informace se stávají značně neodmyslitelnou součástí při fungování celých podniků. Zejména odvětví marketingu je čím dál tím více datově řízené, jelikož firmy dnes chtějí znát podrobné informace o svých zákaznících. Jak se jim daří oproti konkurenci či zda chtějí, aby o jejich produktu nebo službě mělo povědomí více lidí. Právě díky analyzování, zpracování a vhodné vizualizaci dat mají firmy možnost získat cenné informace, které mohou tvořit základ celým marketingovým strategiím. S rozvojem informačních technologií přichází i nové technologie pro zpracování dat. Tato data často bývají na různých místech a v různých kvalitách. Skrze ETL nástroj Bizzflow je pak možné tato data správně zpracovat a následně využít možnosti nástrojů třetích stran k jejich prezentaci.

Tato bakalářská práce je rozdělena na dvě části. V první části, týkající se teoretických poznatků, lze nalézt úvod do online marketingu, představení sociální sítě LinkedIn jakožto zdroje dat, a také popis datové analýzy. Čtenáři jsou uvedeni do problematiky databází a základních databázových pojmů, jako jsou modely databází, jejich prvky a základy SQL jazyka. Poté jsou seznámeni s pojmy z oblasti BI, a v neposlední řadě s ETL nástrojem Bizzflow.

Druhá, praktická část, představuje potenciální návod k tomu, jak data nejen zpracovat prostřednictvím Bizzflow, ale také jak je vizualizovat nástrojem GoodData. Dále obsahuje vhled do řízení marketingu ve firmě Bizztreat, na jehož základě byla provedena analýza dat z LinkedIn, která vedla k vytvoření počátečního mockupu. V Bizzflow části je popsán postup zahrnující konfiguraci datového zdroje k projektu, očištění a transformace dat skrze SQL dotazy. Společně pak s úpravou datového modelu je popsán postup nahrání dat do nástroje GoodData. Pomocí jazyka MAQL je u vizualizace nezbytné předem zformulovat metriky, díky nimž je možné vytvořit výsledné insighty a KPI ke zkonstruování konečného dashboardu podle zadaného mockupu.

Cílem bakalářské práce je vysvětlení a pochopení teoretických poznatků z výše vyjmenovaných oblastí. Následně využít nabyté znalosti v praktické části skrze analýzu dat z LinkedIn, konfiguraci projektu v Bizzflow, psaní SQL dotazů a vizualizaci dat.

1. Možnosti zpracování marketingových dat

Marketingová data jsou v dnešní době zásadním zdrojem informací o zákaznicích či trhu. Efektivním zpracováním a následnou vizualizací těchto dat lze lépe pochopit chování zákazníků. Na základě toho poté kustomizovat nejen nabízené produkty a služby, ale i zlepšovat reklamní kampaně a dosáhnout tím většího úspěchu v oblasti marketingu. Zvláště oblast online marketingu se stává více a více populárnější, i proto bude práce specializovaná na tato data. V této kapitole bude rozebrán pojem online marketing, příklady nejpoužívanějších technik online marketingu a shrnut obecně pojem datové analýzy, jakožto krátký popis pro proces zpracování dat.

1.1 Online Marketing

Pod pojmem online marketing je možné si představit soubor aktivit prováděných přes internet, jež jsou spojené s manipulací, přesvědčováním a udržováním vztahů se zákazníky, a skrze tyto aktivity lze dosáhnout předem stanovených marketingových cílů (Burešová 2022).

Marketingový svět je dnes velice ovlivněn internetem. Lidé si mohou si zakoupit produkty, které jsou dostupné jen online, jako například antivirový program nebo placené úložiště dat. V posledních letech (hlavně kvůli covidové krizi) se na internet přesunuly i služby, u kterých předtím nebylo běžné je vykonávat online, jako kurzy cvičení či vaření (Burešová 2022).

1.1.1 Techniky online marketingu

Existuje několik způsobů, jak rozdělit online marketing do různých skupin – přes jednotlivé online kanály (web, e-shop, sociální sítě atd.), přes marketingové nástroje (reklama, PR, copywriting, video apod.) nebo dokonce do takzvaných podoborů online marketingu, mezi které patří optimalizace pro vyhledávače (dále jako SEO), sociální média, obsahový marketing, e-mailing a uživatelská zkušenost spojená s webovým designem. Pro účely práce bylo vybráno rozdělení podle podoborů online marketingu (Burešová 2022).

Sociální média

Prvním zpracovaným podoborem budou sociální média. Sociální média charakterizuje Burešová (2022, s. 182) jako *„podmnožinu médií, která využívají internetového spojení"*. Podstatou sociálních médií je tedy určitá interakci mezi lidmi. Pod slovem interakce, se v tomto případě, skrývá více činností. Lidé skrze sociální média komunikují soukromými zprávami, dále mohou reagovat na různé příspěvky takzvaným „like", u nás přeloženo jako „Líbí se mi". Mohou vkládat komentáře pod příspěvky, kterými sdílí informace, eventuálně sdělují své názory, a to nejen kladné, ale i záporné. U určitých typů sociálních medií, jako jsou různé Wiki stránky, mohou lidé editovat originální obsah. Sociální média se staly nedílnou součástí života většiny lidí. Slovo „influencer" je v dnešní době úzce spjato se sociálními médii. Pro některé je být influencerem povolání, které je uživí. Na sociálních mediích mohou interagovat s lidmi i firmy. S relativně malými náklady tak mohou komunikovat se svými nynějšími anebo budoucími zákazníky (Burešová 2022).

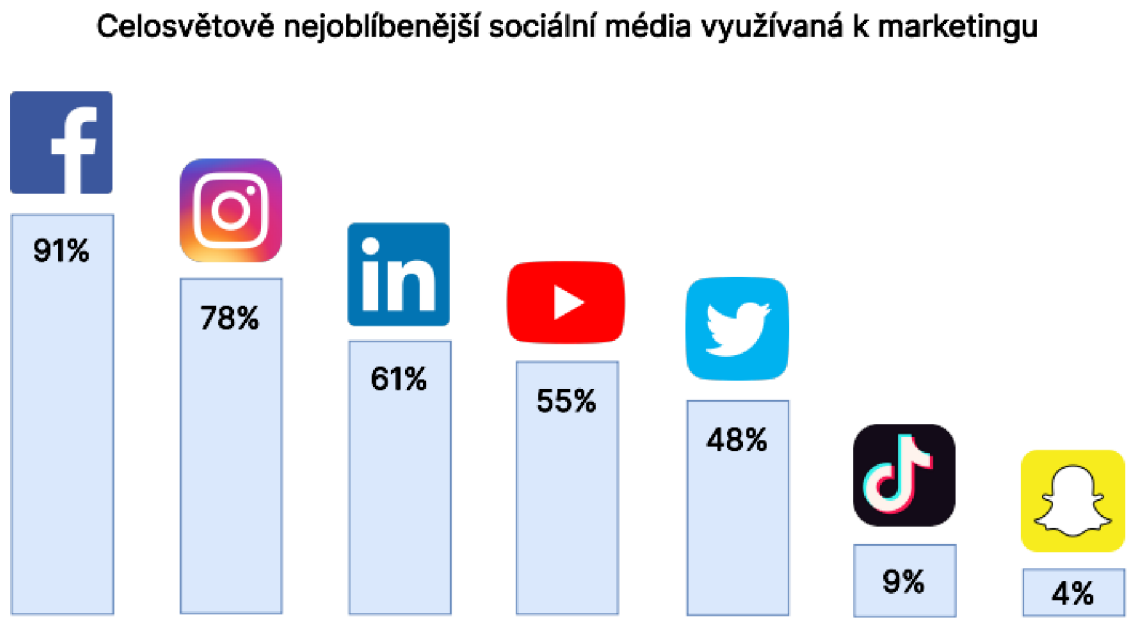
Sociální média jsou široký pojem, a tak rozhodnutí o tom, zda se jedná, případně nejedná o sociální médium, se určuje na základě několika charakteristických rysů:

- **Aktuálnost:** Sociální média jsou komunikačním kanálem, který zaznamenává a signalizuje jakékoliv změny v obsahu uživateli. Aktuálnost se v tomto případě projevuje tak, že lidé mohou neprodleně po vydání příspěvku či textu jej komentovat, editovat nebo na něj reagovat (Burešová 2022).
- **Editace:** Sociální média jsou sociálními hlavně z důvodu, že jednotlivý uživatelé sdílí svůj názor na dané téma, a tím mohou ovlivňovat názory ostatních. Editovat lze nejen vlastní obsah, ale i již vytvořený, jako je tomu na wiki systémech (Burešová 2022).
- **Validita:** Uživatelé na sociálních médiích mají možnost hodnotit obsah. Toto hodnocení pak napomáhá dalším uživatelům s určením, o jak kvalitní obsah se jedná (Burešová 2022).
- **Sdílení obsahu:** Sdílení obsahu, informací, názorů je základním rysem sociálních médií. Může se jednat o zveřejňování fotek, videí nebo článků, ale i o již zmíněné komentování, přidávání liků a přesdílení obsahu jiných uživatelů, a to skrze různé sociální média (například přesdílení videa z YouTube na Instagram) (Burešová 2022).

Sociální média je možné rozdělit několika způsoby, příklady jsou dle převažujících funkcí, dle zaměření nebo dle marketingové taktiky. Dělení nemusí být vždy zcela přesné a některá sociální média by mohla být ve více kategoriích. Představení dělení dle marketingové taktiky (Janouch 2014):

- **Sociální sítě:** Facebook, Instagram, LinkedIn, Messenger a další
- **Blogy, videoblogy, mikroblogy:** Twitter
- **Diskusní fóra, QaA portály:** Yahoo!, Answers
- **Wikis:** Wikipedia, Wikisofia
- **Sociální záložkovací systémy:** Digg, Delicious, Jagg
- **Sdílená multimédia:** YouTube, Flickr, Rajče
- **Virtuální světy:** Second Life, The Sims (Janouch 2014)

Obrázek níže představuje graficky zpracovanou skutečnost, jaká sociální média patří mezi nejoblíbenější, v roce 2022, v oblasti online marketingu (McCormick 2022).



Obrázek 1: Celosvětově nejoblíbenější sociální média využívaná k marketingu
Zdroj: vlastní tvorba, diagrams.net

Content marketing

Za content marketing, přeloženo jako obsahový marketing, je považována jakákoliv marketingová aktivita, která pomáhá prodávat produkt, službu nebo organizaci

s nabídkou bezplatných informací. Jde o účinný způsob, jak se angažovat v daném odvětví, jak oslovit potenciální zákazníky a uzavírat obchody. Marketéři v mnoha odvětvích tak vytváří cenný obsah nejen pro potenciální, ale i pro stávající zákazníky. Obsahový marketing se prolíná s marketingem na sociálních médiích. Za obsahový marketing je pokládán například příspěvek na sociálních sítích, který sděluje informace o firemní konferenci. Dále také článek v odborném časopise o trendech v oboru firmy, nebo i webinář s informacemi o nové technologii, kterou firma využívá. Dle popisu výše, nejedná se o synonymum ke slovu „reklama". Ačkoliv reklama také sděluje bezplatné informace, podstatou obsahu příspěvků content marketingu je přirozenou cestou ukázat lidem, proč je produkt či služba dané firmy lepší než jiné produkty a služby (Bly 2020).

SEO

Search Engine Optimization, zkratkou SEO, představuje, podle Ungra (2014), optimalizaci nalezitelnosti v pravém slova smyslu. Jinými slovy, jakékoliv vyhledávání je možné optimalizovat. Může se jednat o hledání v telefonním seznamu, na sociálních sítích nebo ve fulltextových internetových vyhledávačích. Pro současné SEO platí 4 základní pravidla – je unikátní, obsahově zajímavé, má virální potenciál, a v neposlední řadě je sociální neboli je provázané se sociálními sítěmi a komunikuje tak, že člověk má pocit, jako by komunikoval s opravdovým člověkem (kolektiv autorů 2014).

Jak již bylo zmíněno výše, existuje více podoborů online marketingu, ovšem pro účely práce je dostačující představení těchto tří.

1.2 LinkedIn

LinkedIn je označován podle Náplavové (2014, s. 146) za *„sofistikovanou profesionální sociální síť a zároveň nástroj, který lze hojně využít pro navazování vysoce odborných profesních kontaktů nejen pro oblast HR."* (kolektiv autorů 2014).

LinkedIn byl založen roku 2002, oficiálně byl ovšem spuštěn až v roce 2003. Za spoluzakladatele je považován Reid Hoffman, americký podnikatel, který mimo

LinkedIn založil i finanční technologickou společnost, PayPal. Dnes má LinkedIn více než 850 milionů členů z více než 200 zemí a oblastí po celém světě (LinkedIn c2023).

Lidé si zde vytváří osobní či firemní profil a skrze něj komunikují s ostatními uživateli LinkedInu, přidávají příspěvky do konkrétně orientovaných skupin a spravují svůj profil. LinkedIn je v dnešní době velice využíván jako veřejně dostupný životopis. Při založení osobního profilu si lidé mohou vyplnit dosažené vzdělání, konkrétní pozici, na které momentálně pracují i veškeré předchozí pozice, obory, které je zajímají, a mnoho dalšího. Mimo jiné, mohou psát příspěvky na svou zeď. Pokud si účet na LinkedInu založí firma, profil funguje jako zdroj informací s možností vkládat nejnovější informace formou statusů na zeď (kolektiv autorů 2014).

Dnes LinkedIn tvoří více než polovinu veškerého provozu na sociálních sítích na B2B webech a blozích. Asi 80 % potenciálních zákazníků z oblasti B2B pak pochází ze sítě LinkedIn. Pokud by se množství profilů rozdělilo na odvětví, přibližně 4 % všech profilů na LinkedIn jsou z oblasti IT. Dále by se zařadily profily z oblastí zdravotnictví, stavebnictví, maloobchodu a vzdělávání (McCormick 2022).

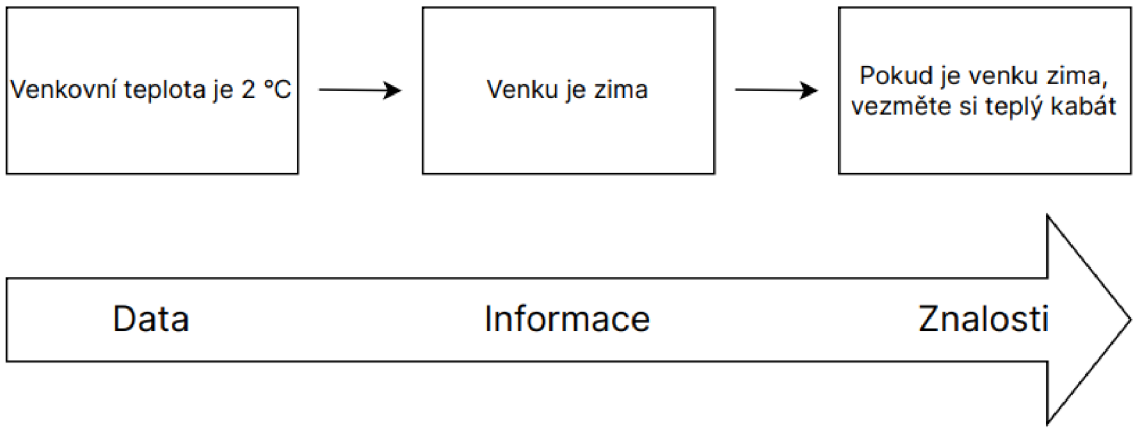
1.3 Proces datové analýzy

Datová analýza je proces sběru, transformace, analýzy a interpretace dat s cílem získat užitečné informace za účelem dosažení konkrétních předem stanovených cílů či vyřešení stanovených problémů (Cuesta 2015). Tento proces je často využíván v oblasti marketingu, kde je třeba pochopit chování zákazníků, trhů a procesů. V kapitole bude představen rozdíl mezi daty, informacemi, znalostmi a následně bude obecně popsán proces datové analýzy.

1.3.1 Informace, data, znalosti

Před začátkem této podkapitoly je důležité rozlišit pojmy data, informace a znalosti. Data jsou představovaná pouhými čísly, bez většího smyslu. Z dat se stanou informace ve chvíli, kdy je jim přiřazen určitý význam. Poslední fází je znalost. Znalost vznikne ve chvíli, kdy z dat a informací vznikne sada pravidel, dle kterých je možné

učinit rozhodnutí. Grafický příklad převedení dat na znalosti je uveden níže (Cuesta 2015).



Obrázek 2: Převod z dat na znalosti
Zdroj: vlastní tvorba, diagrams.net

1.3.2 Proces

Definice datové analýzy zní podle Cuesty (2015, s. 27): „Datová analýza je procesem, v jehož rámci se surová data seřadí a roztřídí, aby je bylo možné použit v metodách, které pomáhají interpretovat minulost a předvídat budoucnost.“. Právě v důsledku datové analýzy neexistují jen čísla, ale i informace (Cuesta 2015).

Samotný proces se skládá z pěti úkonů, a to z definování problému, přípravy dat, průzkumu dat, prediktivního modelování a vizualizace výsledků (Cuesta 2015).

Definování problému

Při definici problému je důležité, aby zadavatel jasně formuloval problém, otázku nebo požadavek a aby vykonavatel jasně porozuměl zadanému úkolu. Otázky bývají obecné jako kolik bude stát zlato příští měsíc, nebo jak lze sledovat rozdíly v chování různých skupin zákazníků. Porozumění předpokladů a cílů tvoří základ tomu, že bude projekt z hlediska datové analýzy úspěšný (Cuesta 2015).

Příprava dat

Přípravou dat se rozumí jejich očištění od nevalidních vstupů, normalizace a transformace na požadovanou podobu. Tato fáze může trvat dlouho, avšak se nesmí uspěchat, jelikož výsledky ze špatně připravených dat mohou být nepřesné. Kvalitní data splňují rysy kompletnosti, koherence, jednoznačnosti, správnosti, standardizace a multiplicitnosti (Cuesta 2015).

Průzkum dat

Při prozkoumávání dat se v podstatě jen nahlíží na data z grafické nebo statistické stránky, hledají a zkoumají se vztahy mezi nimi (Cuesta 2015).

Prediktivní modelování

Pod pojmem prediktivní modelování je možné si představit celý proces vytváření odpovídajícího statistického modelu, který předpoví s největší pravděpodobností správný výsledek. Takových modelů je několik, avšak pro účely práce není potřebné je rozebírat. Pro zjištění vhodnosti modelu existují dva způsoby validace (Cuesta 2015):

Křížová validace umožňuje nejen zjištění velikosti modelu, ale také jeho porovnání s výkonem více modelů. Postup vypadá tak, že se data rozdělí do dvou stejně velkých souborů a otestují se tak, aby bylo jasné, jak budou fungovat v praxi. Referenční validací se datová sada náhodně rozpadne na tři podmnožiny, jednou z nich bude zkušební sada, druhou validační sada a poslední bude testovací sada (Cuesta 2015).

Vizualizace výsledků

Poslední fází je vizualizace. V této fázi je třeba rozhodnout, jak budou data vizualizována a především kde. Možností je několik. Data lze prezentovat prostřednictvím tabulek, grafů, řídících panelů či infografiky. Toto rozhodnutí závisí na tom, kam vizualizace bude nasazena a zda bude v mobilním zařízení, na webu, na plakátě, na billboardu, a tak dále (Cuesta 2015).

2. Vymezení relevantních databázových a Business Intelligence pojmů

Vymezení těchto pojmů je důležité pro následné pochopení aplikovaného řešení v práci. Z oblasti databází budou vysvětleny pojmy jako jsou databáze a modely, dále jejich klíčové stavební prvky, vztahy mezi tabulkami a určitá pravidla, která je třeba dodržovat. Bude představen SQL jazyk vhodný pro práci s tabulkami a na konci kapitoly budou vymezeny relevantní BI pojmy.

2.1 Databáze a databázový systém

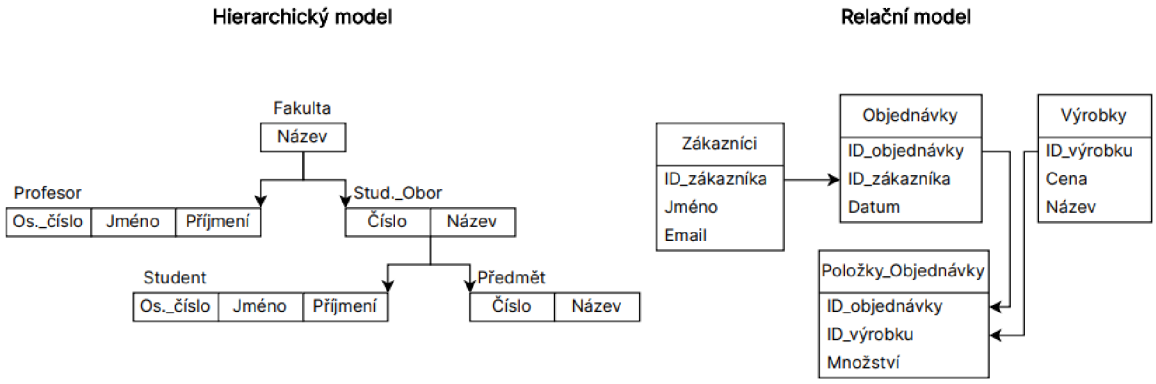
Pro uchování zpracovaných dat jsou využívány databáze. Je důležité rozlišit pojmy databáze a databázový systém. Pod pojmem databáze je možné si představit uspořádanou soustavu dat. Při požadavku uspořádanosti je zároveň vyžadováno, aby se soustava dat dala snadno upravovat, doplňovat o nové informace či odstraňovat již nadbytečné, případně nesprávné informace. Tím bude databáze udržována co nejaktuálnější. Tyto úpravy se provádí skrze databázový systém, což je software, jenž se využívá pro práci s daty (Laurenčík a Bureš 2018).

2.2 Databázové modely

Databázový model tvoří hierarchii databáze a databázový systém podle této hierarchie ukládá objekty a tvoří mezi nimi vztahy. Odlišnost v těchto modelech je pouze ve způsobech toho, jak jsou data ukládaná do databází. V dnešní době existuje mnoho databázových modelů, avšak pro tuto práci postačí představení tří z nich (Laurenčík a Bureš 2018).

Prvním vzniklým modelem byl hierarchický model. Tento model využívá hierarchické uspořádání rodič – potomek, ovšem byl poměrně rychle nahrazen síťovým modelem, a to kvůli své nedostatečnosti vystihnutí všech možností ve vztazích mezi daty. Síťový model je pouze rozšířen o vztah „více ku více“, jinak řečeno potomek může mít více rodičů (Laurenčík a Bureš 2018).

Posledním představeným a dnes nejvíce využívaným konceptem modelu je relační. Relační model uchovává data ve formě tabulek, která jsou uspořádána do řádků a sloupců. Dojde-li k tomu, že jsou data v některých sloupcích společná s daty ve sloupcích v jiných tabulkách, mají tyto dvě tabulky určitý vztah (Laurenčík a Bureš 2018).



Obrázek 3: Rozdíl mezi hierarchickým a relačním modelem databáze
Zdroj: vlastní tvorba, diagrams.net

2.3 Prvky databází

Každá databáze je složena z několika tabulek. Samotné tabulky jsou poté složeny ze sloupců, představujících definici dat, která se ve sloupci vyskytují. Řádky naopak představují jednotlivé záznamy tabulky. Mezi tabulkami vznikají specifické vztahy, které jsou vyjadřovány skrze takzvané klíče. Koncem této kapitoly budou představena pravidla, která je třeba dodržet, aby data, jež jsou v databázi uložena, byla kvalitní a neznehodnocená.

2.3.1 Tabulka

Veškerá data jsou v databázi uložena ve formě tabulek. Aby byly tabulky pro databázový systém čitelné, je důležité splnění určitých formalit (Laurenčík a Bureš 2018).

Každá tabulka je vnímána jako souvislá dvourozměrná struktura složená z řádků a sloupců. Nevyskytují se v ní žádné prázdné řádky, aby systém mohl určit rozsah,

kde jsou data umístěna. Každý sloupec tabulky představuje atribut, který má odlišný název a data v daném sloupci musí odpovídat jednotnému formátu dat, jenž bude popsán kapitole 2.3.2. V neposlední řadě musí mít každá tabulka atribut či kombinaci atributů, které jednoznačně identifikují daný řádek tabulky (Coronel a Morris 2016).

2.3.2 Typy sloupců

Údaje uložené ve sloupcích je možné rozdělit na textové, číselné, datumové a logické. Ovšem takto jednoduché rozdělení systému nestačí, a tak je nutné rozlišovat další podskupiny. Je třeba také dodat, že každý software může využívat řadu datových typů, jež nejsou v práci popsány, jelikož budou představeny jen obecně známé datové typy, využívané většinou softwarů (Beaulieu 2020).

Číselné typy

Nejčastěji využívané číselné typy jsou tři. Prvním je Integer (INT), který představuje celé číslo o rozsahu ± 2 147 483 647. Druhým typem je FLOAT, reálné číslo s rozsahem přibližně ± $3{,}4 \cdot 10^{38}$. Posledním představeným typem je DOUBLE, opět reálné číslo s přibližným rozsahem ± $1{,}8 \cdot 10^{308}$ (Beaulieu 2020).

Textové typy

Textové typy se rozlišují v zásadě tři. Typ CHAR, jenž reprezentuje text s pevně zadanou délkou a typ VARCHAR, textový typ s proměnlivou délkou. Rozdíl je v tom, že při pevné délce se jako hodnota uloží text a počet mezer chybějících do stanovené délky, zatímco při proměnlivé délce je uložen pouze text. Maximální délka typu CHAR je 255 bytů, typ VARCHAR můžu dosahovat délky až 65 535 bytů. Pokud je třeba uložit delší text, jako jsou emaily nebo různé popisy, využívá se datový typ TEXT (Beaulieu 2020).

Datumové a časové typy

V datumových a časových typech se rozlišují čtyři typy. Jedním z nich je DATETIME, jenž udržuje hodnotu data i času. Ve chvíli, kdy je zapsáno do tohoto pole pouze datum, čas bude ve tvaru „00:00:00“. DATE je čistě datumový typ. Pokud bude

zapsán do tohoto pole datumový i časový údaj, čas se smaže. Dalším typem je TIME. Analogicky k typu DATE, pole s tímto formátem udržuje údaj pouze o čase, datum by se smazalo. Poslední typ, TIMESTAMP, představuje časové razítko. Tento typ se velice využívá pro zaznamenání aktuálního data a času při vložení nového záznamu (Beaulieu 2020).

Logické typy

Logický typ je jen jeden, nazývaný BIT. Jeho hodnotami jsou 0 (False) nebo 1 (True) (Beaulieu 2020).

Tyto datové typy a další jsou využívané při psaní v BigQuery, MS SQL či MySQL.

2.3.3 Klíče

Databázové klíče jsou prvkem v návrhu a správě relačních databází. Slouží k jedinečné identifikaci záznamu v tabulce, proto je díky tomuto prvku jednoduché v tabulkách vyhledávat a upravovat je. Existuje několik druhů klíčů, pro potřeby práce postačí představení primárního a cizího klíč. Klíče jsou také důležité pro zachování integrity dat a optimalizaci výkonu databázových operací.

Primární klíč

Jak již bylo zmíněno, každá tabulka musí mít jeden atribut, respektive kombinaci atributů, které jednoznačně identifikují každý záznam v dané tabulce. Tento identifikátor je označován dále jako primární klíč (primary key). Veškeré hodnoty, uložené jako primární klíč, musí být unikátní a nesmí nabývat hodnoty NULL. Hodnota NULL je obecně považována za problematickou v relačních databázích. Nepředstavuje to stejné, jako kdyby byla uložena do buňky 0. V buňce se uloží hodnota NULL, pokud do ní nebude vložen jakýkoliv datový záznam. Z tohoto důvodu NULL hodnota představuje buď neznámou hodnotu atributu, známou, ale chybějící, anebo nepoužitelnou hodnotu. V praxi je možné si představit primární klíč jako IČO firmy nebo rodné číslo osoby (Coronel a Morris 2016).

Cizí klíč

Cizí klíč (foreign key) je takový klíč, který odkazuje na primární klíč z jiné tabulky. Pro cizí klíč naopak platí, že hodnoty NULL může nabývat. Toto je možné si převést na příklad z reálného života. Existuje tabulka produktů a tabulka prodejců. Každý produkt má jako primární klíč dané své identifikační číslo. Aby bylo možné říci, jaké produkty každý prodejce prodává, na základě vazby se přidá do tabulky prodejců identifikační čísla produktů jako cizí klíč. Ve chvíli, kdy prodejce neprodává žádný produkt, který je uveden v tabulce produktů, cizí klíč nabyde hodnoty NULL (Coronel a Morris 2016).

2.3.4 Vztahy mezi tabulkami

Pokud dojde k propojení dvou tabulek na základě vztahu, je třeba tento vztah specifikovat. Existují tři možnosti, jak mohou být tabulky propojeny, a to vztah jedna ku více (1:M), jedna ku jedné (1:1) a více ku více (M:N).

Vztah 1:M

Tento vztah je ideální pro relační modelování. Je to nejčastěji využívaný vztah, hlavně proto, že nejlépe vystihuje většinu vztahů v reálném životě (Coronel a Morris 2016). Představuje takový vztah, kdy jeden řádek odpovídá jednomu, více nebo žádnému řádku z jiné tabulky. Je možné si to popsat na příkladu. Na škole jsou studenti, kteří si zapisují volitelné předměty. Škola omezila přihlašování tak, že každý student si může zapsat pouze jeden volitelný předmět. Při pohledu z druhé strany, na jednom předmětu bude zapsáno více studentů. Existuje možnost, že dojde k situaci, že si volitelný předmět nevybere ani jeden student nebo právě jeden, z tohoto důvodu je správně využita relace 1:M.

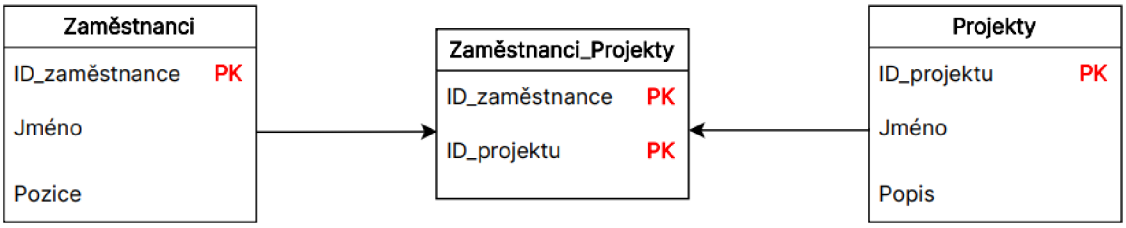
Vztah 1:1

Tento vztah je v relačních databázích spíše ojedinělý, a to z důvodu, že povětšinou je reálné mít všechny údaje v jedné tabulce a není třeba rozdělovat data do dvou tabulek. Jiným důvodem k rozdělení dat je, že tabulka bude příliš velká, zpravidla 255 sloupců. Tímto vztahem se rozumí tomu, kdy k právě jednomu záznamu je přiřazen právě jeden záznam (Laurenčík a Bureš 2018). Příkladem vztahu by mohl být

prezident a stát. Každý stát má právě jednoho prezidenta a každý prezident úřaduje právě v jednom státě.

Vztah M:N

Vztah M:N vyjadřuje takový vztah, kdy je k více řádkům přiřazeno více záznamů z jiné tabulky. Konkrétně takovýto vztah, ve své podstatě, v relačních databázích nelze realizovat, a proto se rozkládá na dva vztahy 1:M. Děje se tak přes speciální tabulku zvanou vazební tabulka. Všechny její atributy jsou tvořeny pouze primárními klíči dvou tabulek navázaných na ni. Situaci lze uvést na příkladu se zaměstnanci (Zaměstnanci) pracujícími na projektech (Projekty). V tomto případě je tedy situace, kdy jeden zaměstnanec může pracovat na více projektech a zároveň na jednom projektu může pracovat více zaměstnanců. Je vhodné vytvořit ve vazební tabulce primární klíč složený z obou atributů (ID_zaměstnanci a ID_projektu), a to hlavně z důvodu, že bude zajištěna unikátní existence kombinací zaměstnanec – projekt, protože primární klíč musí splňovat podmínku jedinečnosti.



Obrázek 4: Úprava vazby M:N na dvě vazby 1:M pomocí vazební tabulky
Zdroj: vlastní tvorba, diagrams.net

Kardinalita a parcialita

Termíny kardinalita a parcialita specifikují vztahy mezi tabulkami o něco víc. Pojem kardinalita vyjadřuje minimální a maximální počet výskytu objektu spojeného s jiným objektem. Ku příkladu již zmíněných studentů a volitelných předmětů. Student si může zapsat více předmětů, jeden nebo žádný. Parcialita je pojem, který specifikuje volitelnost účasti objektu v návaznosti na jiný. Opět s příkladem student – volitelný předmět. Je rozdíl, zda si student může zapsat volitelný předmět či musí si zapsat volitelný předmět.

Referenční integrita

Ke vztahům mezi tabulkami v databázi existují i určitá omezení pro práci s daty v těchto tabulkách. Jsou to taková omezení, která brání tomu, aby se v související tabulce vyskytli hodnoty cizího klíče, které neodkazují na žádnou hodnotu primárního klíče v tabulce hlavní. Ku příkladu, existuje banka, ve které není veden jediný bankovní účet, ovšem není možné, aby existoval bankovní účet bez banky. Tyto omezovací podmínky jsou souhrnně nazvány jako referenční integrita (Laurenčík a Bureš 2018).

Zpravidla rozdělujeme dva typy referenční integrity, restriktivní a kaskádovou. Restriktivní omezení neumožňují upravovat záznamy, pokud se vyskytuje hodnota primárního klíče v tabulce související, jako hodnota cizího klíče. Do související tabulky nelze přidat hodnotu, která se nevyskytuje v tabulce hlavní. Kaskádové omezovací podmínky umožňují změny v související tabulce takzvaně v kaskádě neboli veškeré změny, které se dějí v hlavní tabulce, se dějí i v související. Upraví-li se primární klíč, stejně tak bude upraven i cizí klíč, pokud v související tabulce existuje. Analogicky tomu tak je i při mazání záznamů (Laurenčík a Bureš 2018).

2.3.5 Normalizace

Redundance neboli nadbytečnost dat je také jeden z velice problémových jevů v databázích. Pro minimalizaci redundance, a tím pádem i snížení výskytu anomálií v datech, existuje proces, jímž je vyhodnocována a upravována strukturu dat, ale i tabulek. Normalizace funguje prostřednictvím řady stupňů nazývaných normální formy. V základu se rozlišují tři normální formy, dále ale existuje také 4NF a rozšíření k 3NF neboli Boyce-Coddova normální forma. Ze strukturálního hlediska logika řady funguje tak, že 2NF je lepší než 1NF a 3NF je lepší než 2NF. Obecně platí, že nejvyšší stupeň je ten nejžádanější. Na druhou stranu, s vyšším stupněm normalizace potřebuje databázový systém více prostředků pro komunikaci s koncovými uživateli. Proto je v některých případech žádoucí i denormalizace, přechod z vyššího stupně na nižší (Coronel a Morris 2016).

2.4 SQL (Structured Query Language)

Složitost relačních databází vyžadovala nástroj za účelem ulehčení práce s daty. Proto v 70. letech 20. století byl za tímto účelem vytvořen dotazovací jazyk SQL, dříve pojmenovaný jako SEQUEL. Prvním databázovým systémem, který využíval jazyk SQL, byl Oracle, uvedený na trh v roce 1977 firmou Relational Software (Laurenčík a Bureš 2018). Dotazovací jazyk umožňuje vytvářet struktury databází a tabulek, provádět základní úkony správy dat, jako je přidávání, mazání, úprava záznamů a využívat složité příkazy určené k transformaci surových dat na užitečné informace. S jeho rozšířením bylo třeba vytvořit určitý standard, aby byl co nejvíce uživatelsky přívětivý. Kvůli jeho implementaci v řadě databázových systémů je potřeba určitá míra přenositelnosti, aby se uživatel nemusel pořád dokola učit základy, a lehké naučitelnosti. Funkce jazyka SQL spadají do dvou větších kategorií, DDL (Data Definition Language) a DML (Data Manipulation Language) (Coronel a Morris 2016).

2.4.1 DDL

Pod pojem DDL spadají funkce definování dat. Jedná se o vytváření databázových objektů jako jsou tabulky – příkazem `CREATE TABLE`, vztahy mezi tabulkami a s nimi související primární a cizí klíče či definování datových typů sloupců. Dále do této kategorie patří i definování přístupových práv k těmto objektům (Coronel a Morris 2016).

2.4.2 DML

Data Manipulation Language pokrývá funkce jazyka SQL, které se využívají pro manipulaci s daty, jak z názvu vyplývá. Pomocí těchto příkazů je možné vkládat záznamy do tabulky – příkazem `INSERT`, vypisovat záznamy – příkazem `SELECT`, upravovat záznamy – příkazem `UPDATE`, nebo tyto záznamy mazat – příkazem `DELETE` (Coronel a Morris 2016).

Syntaxe příkazů

Zvláštností jazyka SQL od jiných jazyků, např. programovacích, je necitlivost jazyka rozlišovat malá a velká písmena. Zažitou konvencí při psaní příkazů je, že klíčová slova se píšou velkými písmeny a názvy tabulek a sloupců naopak malými písmeny. Další vlastností SQL je, že při provádění dotazu je ignorován přechod na nový řádek. Pro přehlednost je proto lepší psát každý další logický krok na nový řádek a úplný konec dotazu zakončit středníkem. Syntaxe, bez které nelze příkaz spustit v některých databázových systémech, je již zmíněné zakončení dotazu středníkem. Patří sem i psaní textových konstant do apostrofů, případně naopak psaní číselných konstant samostatně, bez apostrofů. (Laurenčík a Bureš 2018). Například v BigQuery není vyžadováno zakončit dotaz středníkem, ale je vyžadovaná syntaxe při psaní textových a číselných konstant.

SELECT

Příkaz `SELECT` je nejvyužívanějším typem příkazu v SQL. Využívá se k vypsání obsahu tabulky. Jeho celkové znění je: `SELECT` název_sloupce `FROM` název_tabulky. Klauzule `FROM` je povinným klíčovým slovem při `SELECT` dotazu, aby databázový systém věděl, z které tabulky má sloupce vypsat. Příkaz `SELECT` lze rozšířit i o další klíčová slova, kterými jsou přidána různá kritéria, které záznamy je vyžádáno vypsat. Skrze klíčové slovo `WHERE` je vypsán seznam podmínek oddělených logickými či speciálními operátory (`AND`, `OR`, `NOT` atd.). Slovem `ORDER BY` je výčet záznamů seřazen dle specifického atributu buď sestupně nebo sestupně. Dále existuje skupina slov nazývaná jako agregační funkce. Těmito funkcemi je možné vypočítat ze sloupce počet záznamů (`COUNT`), součet záznamů (`SUM`), průměrnou hodnotu ze záznamů (`AVG`) a další. S agregačními funkcemi je nutné využití klauzule `GROUP BY`, která rozdělí seznam záznamů podle určitého atributu, například celkový počet odpracovaných hodin rozdělí přes jednotlivé pracovníky. V neposlední řadě, slovo `HAVING` se pojí s klauzulí `GROUP BY`. Funguje na podobném principu, jako spojení `SELECT` a `WHERE`, tedy využívá se k vymezení souhrnů podle definované podmínky (Coronel a Morris 2016). Obecný příklad dotazu může vypadat takto:

```
SELECT          seznam_sloupců
FROM            název_tabulky
[WHERE          seznam_podmínek]
[GROUP BY       seznam_sloupců]
[HAVING         seznam_podmínek]
[ORDER BY       seznam_sloupců  ASC/DESC]
```

Obrázek 5: Obecný příklad SQL dotazu
Zdroj: vlastní tvorba, diagrams.net

2.5 Datový sklad (Data Warehouse)

Datový sklad definuje Gála a kol. (2015, s. 116) jako „integrovaný, konsolidovaný, subjektově orientovaný, stálý a časově rozlišený souhrn dat, uspořádaný pro podporu potřeb managementu“. Jedná se o speciální typ relační databáze. Při rozvedení definice je možné říct, že data v datovém skladu jsou rozdělena dle jejich typu, a ne na základě aplikace, kde vznikla. Po rozdělení z různých zdrojů, struktur a forem jsou data ukládána v rámci celého podniku do jednotného centrálního datového skladu, nikoli v jednotlivých útvarech. Po uložení jsou práva na data povětšinou pouze pro čtení, tudíž se zde žádná nová data nevytváří ani neupravují. V neposlední řadě, jelikož jsou data v datovém skladu časově rozlišena, existuje zde dimenze času, na základě které se ukládá historie těchto dat (Pour a kol. 2018).

Na informace vznikající z datového skladu jsou kladeny určité požadavky, které lze formulovat v několika bodech (Pour a kol. 2018):

- Je třeba, aby byly lehce pochopitelné a dostupné v nejrůznějších kombinacích v co nejkratší době odezvy (Pour a kol. 2018).
- Musí být věrohodné, což obnáší celý proces shromáždění, kontroly a čistění. Data jsou k dispozici až po adekvátní úpravě (Pour a kol. 2018).
- Je vyžadováno, aby tvořily základ pro zlepšování řídících a rozhodovacích procesů podniku (Pour a kol. 2018).
- Z důvodů možných změn musí být DWH adaptabilní k uživatelským požadavkům, prostředí podniku, datovým zdrojům i novým technologiím (Pour a kol. 2018).

- Informace mohou mnohdy být citlivého charakteru, proto je předpokladem řízený přístup do DWH (Pour a kol. 2018).

V dnešní době bývají data v datovém skladu normalizovaná a k případné denormalizaci dochází až na úrovni datového tržiště (Pour a kol. 2018).

2.6 Datové tržiště (Datamart)

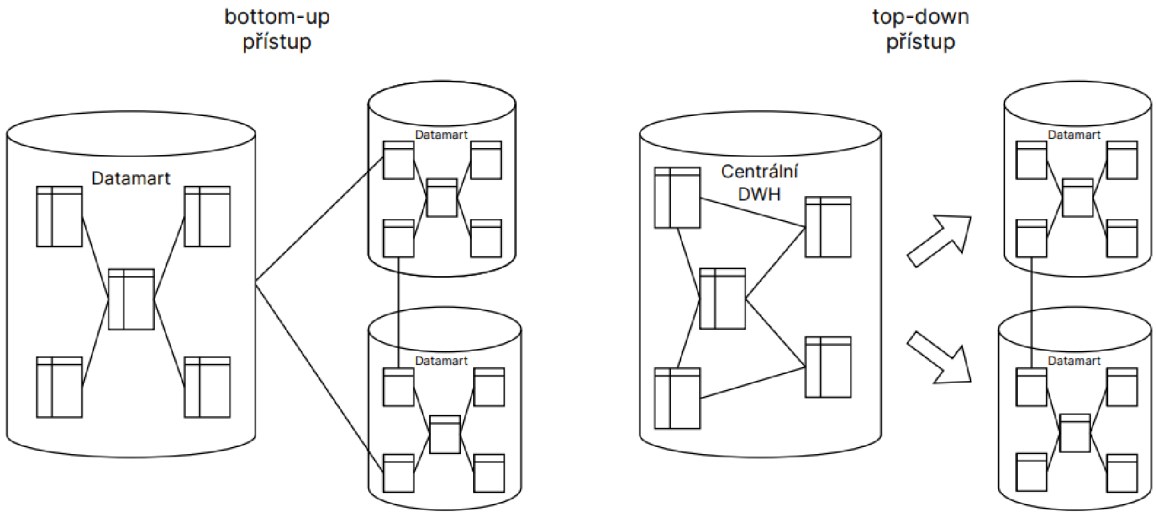
Datamarty mají analogický princip k datovým skladům s rozdílem, že jsou určena pro specifický okruh uživatelů, jako oddělení či pobočku větší firmy. Zpravidla se rozlišují dva typy interpretací datových tržišť (Gála a kol. 2015):

1. decentralizovaný, subjektově orientovaný datový sklad, nebo
2. základ celopodnikového datového skladu (Gála a kol. 2015).

Případ decentralizovaného datového skladu je výhodný z hlediska kratší doby pro návratnost investic, nižších nákladů a nižšího rizika při jejich implementaci. Postupně se vzniklé datamarty integrují do celopodnikového řešení. Tento přístup je nazýván jako Kimballův, respektive bottom-up (Pour a kol. 2018).

Druhým přístupem je vnímání datamartu jako problémově orientovaného datového skladu. Jeho účelem je specializace na konkrétní problematiku zúženého okruhu uživatelů s následnou možností flexibilní analýzy dat dané problematiky nad menším objemem dat. Předpokladem pro realizaci přístupu je již existující centrální DWH, ze kterého se data extrahují do jednotných specializovaných datamartů. Jde o Inmonův přístup, přezdívaný top-down (Pour a kol. 2018).

Koncept datamartů i datových skladů je realizován v prostředí relačních databázových systémů, jako jsou BigQuery, MS SQL, ale i Oracle (Pour a kol. 2018).



Obrázek 6: Příklady přístupů k datamartům
Zdroj: vlastní tvorba, diagrams.net

2.7 ETL proces

Jednou z nejdůležitějších komponent celého BI řešení je právě ETL, označovaná také jako data pipeline (datová pumpa). Proces se skládá ze tří částí. V data pipeline nejprve dojde k extrahování dat ze zdrojových systémů (Extract), která jsou následně upravena a uspořádána do žádané formy (Transform) a poté jsou uložena do předem určených datových struktur, tedy do schémat složených z datových skladů a datamartů. Proces je vhodným nástrojem pro přenos dat mezi více databázemi či datovými soubory, ať už se jedná o tabulkové nebo textové soubory. Protože ETL nástroje pracují hlavně v časových intervalech, kdy jsou data přenášena najednou, transformace dat v souvislosti s ETL jsou pracovně, finančně i časově nákladnější, a proto je na ně vynaloženo až 60 % dostupných pracovních kapacit. I přes tuto skutečnost, pro fungující BI řešení představuje ETL nezbytný předpoklad (Gála a kol. 2015).

Pro ETL je charakteristické, že obsahuje pouze taková zdrojová data, jež jsou nezbytná pro plánovací, analytické a rozhodovací aktivity podniku. Další charakteristikou je transformování dat do předem navržených a vytvořených datových struktur, které budou odpovídat potřebám celopodnikového řízení. To souvisí s možnostmi toho, jak moc detailní jsou datové soubory (granularitou) a na tom, přes kolik úrovní agregace je možné data sledovat (multidimenzionalita). Jak již

bylo zmíněno, data pochází povětšinou z více zdrojových databází, může se tedy stát, že stejná data jsou uložena na více místech, a to i v různých kvalitách. Tento jev je problémový z důvodu, že do analytických databází mohou povětšinou data vstoupit pouze jednou a ve správné formě pro následnou práci s nimi. Řešením je konsolidace a očištění dat v transformační vrstvě pro vyloučení multiplicit a chyb (Pour a kol. 2018).

Alternativní variantou ETL je ELT řešení. Prostřednictvím ELT se procesy spojené s plněním DWH dají rozvrhnout na jednotlivé části. Proces začíná extrakcí zdrojových dat, která jsou ale na rozdíl od ETL rovnou nahrána do DWH a teprve poté dochází k jejich čistění a transformaci. Přístup je o to flexibilnější z pohledu dostupnosti dat, na druhou stranu časově náročnější z důvodu nahrávání veškerých zdrojových dat, která by v ETL byla jinak zúžena na potřebná a očištěna od nevalidních (Gála a kol. 2015).

2.8 Prezentace a vizualizace dat

Množství informací se odvíjí i od jejich interpretace. Pokud jsou data vizualizována nesprávným způsobem, jejich hlavní porozumění může zaniknout (Pour a kol. 2018).

Vizualizovat data je dosažitelné několika způsoby. Formou tabulek se data prezentují především při vytváření reportů. Pokud ovšem prezentujeme velké množství dat, tabulky bývají nepřehledné a těžko se v nich hledají souvislosti nebo případné odchylky. Pro takové množství dat se využívá grafické znázornění (Pour a kol. 2018).

Způsob vizualizace se liší od účelu. Pokud se připravuje infografika pro marketingovou kampaň, musí být kreativní, atraktivní, někdy až extravagantní, aby zákazníky zaujala. Ve chvíli, kdy se připravuje grafika v rámci analytických úloh v SSBI, je kladen důraz na to, aby data byla jednoznačná a aby byl na první pohled jasný jejich význam. Grafika je pak jednodušší, méně osobní a bez zbytečných rušivých elementů jako jsou piktogramy či nadbytečné efekty (Pour a kol. 2018).

Při rozhodování, jakou grafiku vybrat, je třeba dbát na to, že data jsou různého charakteru a dle toho se liší jejich účel i způsob zpracování. Hodnoty určené k reprezentaci lze rozdělit do tří kategorií (Pour a kol. 2018):

- **Kvantitativní:** Číselné hodnoty je možné měřit a kumulovat. Pro zaměstnance tak můžeme vypočítat například počet odpracovaných hodin, výši mzdy, a tak dále (Pour a kol. 2018).
- **Kategorické:** Nominální, nebo také kvalitativní, hodnoty mají abstraktní charakter. Nelze s nimi počítat, ale lze je použít pro rozdělení kumulovaných počtů hodnot ze seznamu, respektive rozdělit hodnoty podle kategorií. Ku příkladu u zaměstnance jsou kategorie jméno, příjmení, pohlaví, profese, a jiné (Pour a kol. 2018).
- **Ordinální:** Kvalitativní hodnoty s možným určením pořadí. Může se jednat o stupeň vzdělání, mzdovou tarifní třídu či seřazení lidí dle data narození (Pour a kol. 2018).

Příklady využívaných vizualizačních nástrojů jsou GoodData a Microsoft Power BI.

3. Představení ETL nástroje Bizzflow

Bizzflow představuje celkové ETL řešení pro jakýkoliv use case. Je postavené na standardních cloudových službách a je implementované pro tři nejvyužívanější cloudová prostředí, mezi která patří Google Cloud Platform, Amazon AWS a MS Azure (Bizzflow [2023]).

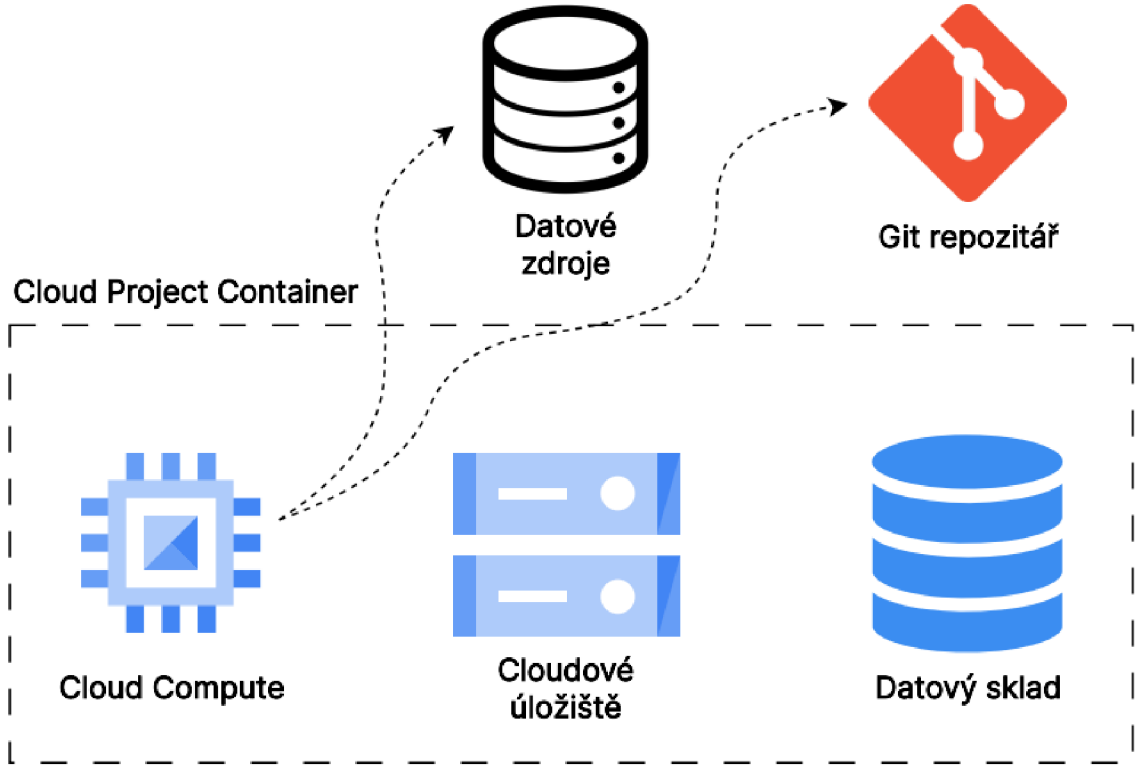
3.1 Koncept

Celý koncept Bizzflow lze rozdělit na tři části, mezi které patří datové zdroje, git repozitář a takzvaný Cloud Project Container. Jak již bylo v práci zmíněno, v datových zdrojích se vyskytují veškerá data, která se plánují extrahovat a zpracovávat. Může se jednat o Google Spreadsheets, o tabulku z databáze či o endpoint z API. V Bizzflow již existuje řada předdefinovaných extraktorů jako je extraktor z Google Sheetů, z HubSpotu, z MS SQL, z Facebooku a další. Celý projekt je uložen v Git repozitáři, který z pravidla musí obsahovat veškeré SQL soubory, JSON nebo YAML konfiguraci a projektovou dokumentaci (Bizzflow [2023]).

Poslední část představuje Cloud Project Container, v překladu jako cloudový kontejner. Jedná se o samostatný a spustitelný balíček obsahující všechny potřebné komponenty pro spuštění programu a skládá se vždy nejméně ze tří částí – Cloud Compute, cloudového uložiště a z analytického skladu (Bizzflow [2023]).

Cloud Compute je služba poskytující virtuální stroje. Bizzflow pracuje s dvěma z nich. První Airflow, vyvinutý společností Apache, je virtuální stroj poskytující UI pro sledování úloh a s možností spouštění požadavků. Kromě vynuceného spuštění, Airflow scheduler každých pár sekund kontroluje, zda neexistují požadavky, které mají být spuštěny. Druhým strojem je worker. Ten je spuštěn při náročných úlohách. Cloudové uložiště je další částí kontejneru a slouží k ukládání CSV souborů a protokolů. Analytický datový sklad je poslední částí a pravděpodobně tou, se kterou se nejvíce interaguje. Zde se píšou a probíhají samotné SQL transformace. V rámci datového skladu lze vytvořit i datamart, ze kterého se zpřístupní omezené množství dat dalším programům třetích stran (povětšinou vizualizačním nástrojům). Datové

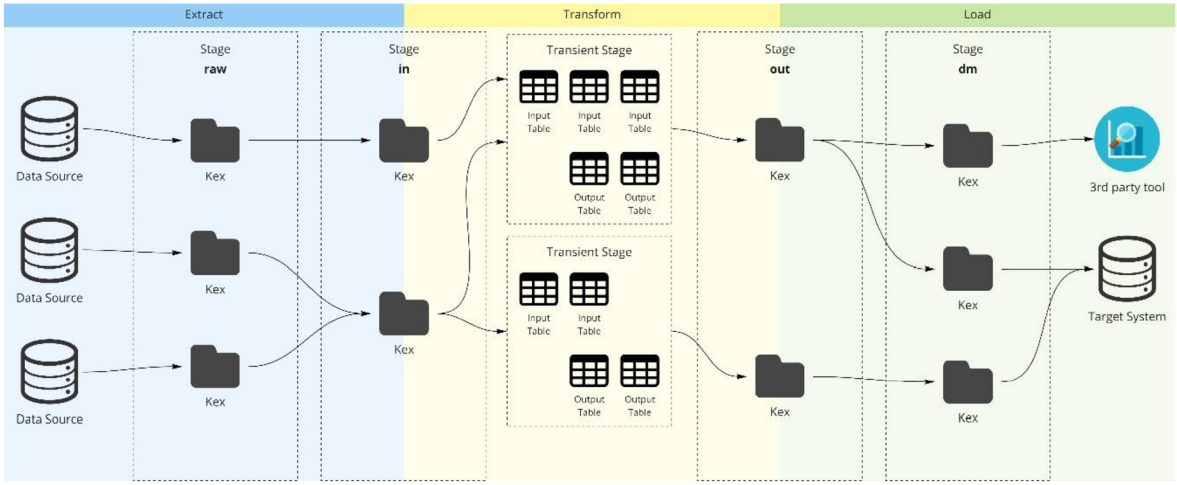
sklady, se kterými Bizzflow funguje, jsou BigQuery, Snowflake a MS SQL Server (Bizzflow [2023]).



Obrázek 7: Koncept Bizzflow
Zdroj: vlastní tvorba, diagrams.net

3.2 Struktura datového skladu

Bizzflow využívá stejnou strukturu DWH pro všechny tři možnosti datových skladů. Struktura je rozdělena na několik ETL fází, ve kterých z nedostačujících či nevalidních dat, jako jsou NULL hodnoty, data pouze za poslední měsíc nebo prázdné znaky v buňkách, vytvoří data použitelná například pro vizualizaci. Na obrázku s titulkem Obrázek 8: Struktura datového skladu v Bizzflow je zobrazena celá struktura skladu (Bizzflow [2023]).



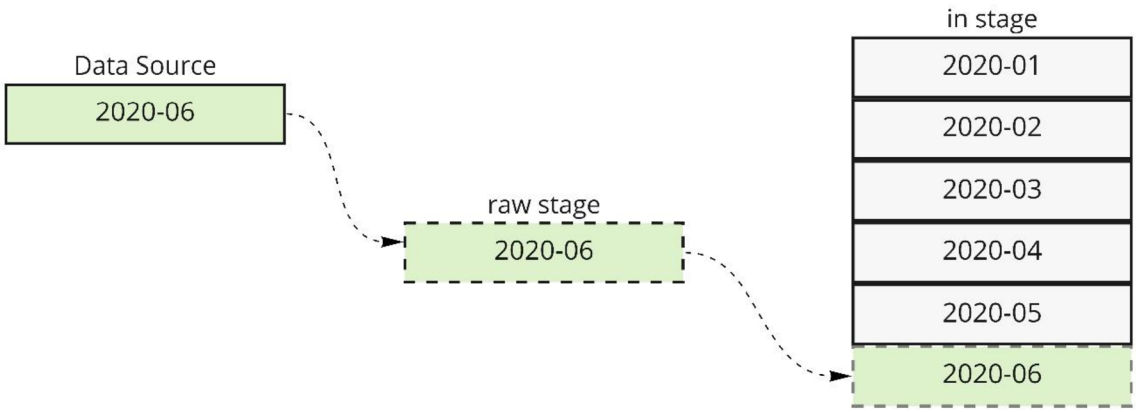
Obrázek 8: Struktura datového skladu v Bizzflow
Zdroj: Bizzflow

raw

Raw fáze obsahuje data v přesně takovém stavu, ve kterém byly extrahována. Jakékoliv větší změny se dějí až v dalších fází. Je to z důvodu, že pokud se vyskytne jakákoliv chyba v data pipeline, je možné se v rámci projektu podívat, jak data vypadala před jakýmkoliv zpracováním (Bizzflow [2023]).

in

Fáze in vždy obsahuje data, která jsou připravena k použití v transformacích. Dochází zde k různým úpravám. Pokud jsou stažená data časově omezená, například pouze měsíc dozadu, tak Bizzflow projekt obsahuje soubor zvaný Step.json, v jehož konfiguraci existuje řádek pro nastavení inkrementace. Protože není třeba extrahovat stejná data, která již byla extrahována, stáhnou se jen nově přidaná data. V raw fázi sice bude uložen jen měsíční přírůstek dat, ale v in fázi budou data kompletní za celou historii. Tento jev, nazývaný incremental snapshotting, vyobrazuje následující obrázek. Kromě nastavení inkrementu jsou data v in fázi také očištěná (Bizzflow [2023]).



Obrázek 9: Incremental snapshotting
Zdroj: Bizzflow

tr

V tomto případě se nejedná tak úplně o fázi z důvodu, že existuje pouze dočasně. Pokud existují SQL skripty pro transformaci dat z in fáze, tak Bizzflow tabulky s jiným prefixem než in_ nebo out_ vytvoří dočasně, provede veškeré transformace s nimi, a nakonec vypíše pouze tabulky, jež obsahují prefix out_ do out fáze (Bizzflow [2023]).

out

V out fázi jsou poté jen zpracovaná data, která jsou již připravená pro to být v produkčním prostředí (Bizzflow [2023]).

dm

Jak již bylo v práci uvedeno, v datamartech jsou pouze data, která jsou specifická například pouze pro jedno oddělení firmy. Příkladovou situací je, že oddělení prodeje potřebuje data z marketingu, aby mohlo optimalizovat své prodeje. Smíchat veškerá data z marketingu s veškerými daty z prodeje může mít za následek například únik citlivých dat. Proto je dobré si vytvořit speciální datamart přímo pro tento účel (Bizzflow [2023]).

3.3 Správa Bizzflow

Ačkoliv je Bizzflow znamenitým nástrojem pro správu projektů, obsahuje určitá omezení. Ze samotného Airflow není možné provádět běžné úkony, jakožto mazání či úpravu tabulek. Dále chybí mechanismus, jenž by umožnil se bezpečně připojit k databázi a testovat transformace mimo produkční prostředí. Tyto problémy byly vyřešeny skrze kustomizované prostředí zvané Flow UI.

3.3.1 Flow UI

Flow UI je webový frontend, zatím v Beta verzi, jenž umožňuje zobrazovat a spravovat zdroje v rámci projektu Bizzflow. Jedná se o úložiště datového skladu, Sandboxy, Credentials (Bizzflow [2023]).

V úložišti lze provádět jednoduché akce, jako vyhledat si potřebné kexy a tabulky, zobrazit je a mazat, upravovat a kopírovat tabulky (Bizzflow [2023]).

Sandbox je prostředí, které může využít každý uživatel k vývoji transformací bez zásahu do produkčního prostřední. Skrze kex a libovolný databázový nástroj, jako je například DBeaver, je možné se k sandboxu připojit a spouštět dotazy, aniž by došlo ke změnám v datovém skladu (Bizzflow [2023]).

Aby bylo možné se k sandboxu připojit, je potřeba mít credentials. V podstatě to jsou přihlašovací údaje, které jsou automaticky generovány při prvním načtení sandboxu. Tyto údaje jsou zabezpečeny tak, aby si je mohla zobrazit jen ta osoba, které patří. Kromě credentials využívaných pro připojení k sandboxu, existují také project credentials, generované uživatelem pro používání v konfiguračních souborech Bizzflow projektu, nebo speciální credentials, které Bizzflow potřebuje pro interní použití (Bizzflow [2023]).

3.4 Shrnutí teoretické části

Zpracování dat je rozsáhlé téma zaštiťující hned několik významných pojmů. Tyto pojmy byly vysvětleny za účelem lepší orientace v dalších částech práce. Datová analytika se skládá z pochopení databází a jednotlivých principů a postupů, jež je vhodné praktikovat a dodržovat za cílem získání relevantních výsledků. Jednou z oblastí, ve které lze aplikovat datovou analýzu, je online marketing. Bylo vysvětleno, co pojem představuje a do oblastí zaštiťující tento pojem patří například SEO, content marketing či marketing na sociálních sítích. Znalost pojmů jako jsou tabulky, atributy, vztahy mezi tabulkami a normalizace jsou důležité pro následnou práci ve vybraných nástrojích. Další velice důležitou částí je dotazování přes příkazy v SQL jazyce, proto jsou v práci vypsány jednotlivá klíčová slova využívaná v tomto jazyce a jak je třeba s nimi nakládat. Postup je složen z několika částí, jedna z nich je vybrání správného BI řešení. V případě práce byl zvolen ETL nástroj Bizzflow. Byla představena struktura projektu tvořeného v Bizzflow a bylo dále popsáno několik podstatných pojmů z oblasti Business Inteligence, které zároveň jsou podstatné pro pochopení této struktury. Po zvolení řešení postup pokračuje získáním dat, následným očištěním a transformováním a končí samotnou vizualizací vhodným nástrojem.

4. Analýza a transformace dat pomocí Bizzflow

Analýza a transformace dat představují klíčové procesy v oblasti zpracování dat. Skrze analýzu dat lze nalézt vztahy a trendy mezi daty pro následné vytvoření užitečných informací. Transformací dat, zahrnující čištění a specifické transformace, je dosaženo přídavných informací. Obsah této kapitoly se zaobírá problematikou pochopení vedení marketingu ve firmě BizzTreat, což je nezbytné k tomu, aby byl zrealizován co nejvíce vyhovující a praktický dashboard, jenž bude vytvořen na základě předem sestaveného mockupu. Dle výsledného mockupu jsou data transformována prostřednictvím ETL nástroje Bizzflow do chtěné podoby vyhovující vizualizačnímu nástroji GoodData, v němž bude tvořen dashboard. Pro samotný dashboard je ale třeba nejprve upravit již existující datový model a vytvořit pomocné metriky pro tvorbu insightů.

4.1 Marketing v BizzTreat

Marketing je ve firmě BizzTreat veden převážně externími zaměstnanci, a to přibližně od léta roku 2022. Hlavním kanálem je sociální sít LinkedIn, dále se ale přidávají příspěvky i na sociální síť TikTok. Skrze příspěvky na těchto sítích je poukázáno, že v současné době mít BI řešení je velice výhodné, i přes skutečnost, že prvotní investice může být vyšší. Je tím cíleno na budoucí potenciální zákazníky bez, respektive s nevyhovujícím BI řešením.

Na LinkedIn je prováděno několik aktivit. Jednou z nich je přidávání příspěvků. Za pomoci externích zaměstnanců byl vytvořen částečný plán přidávání příspěvků, kdy přibližně 3x týdně je přidán příspěvek ve formě textu a přibližně 1x týdně ve formě videa. Samozřejmě jednotliví ambasadoři mohou přidávat i příspěvky vlastní tvorby. Budoucí zákazníci sami na základě těchto příspěvků kontaktují ambasadory a mají zájem se dozvědět více. Druhou aktivitou je oslovování potenciálních zákazníků, ať už na základě jejich aktivit na příspěvcích ambasadorů nebo na základě filtrace požadavků, ze strany BizzTreat, jako například počet zaměstnanců firmy či peněžní obrat, a následného oslovení. Ambasadoři opět mohou využít připravených zpráv externisty, anebo mohou zvolit více individuální přístup a zjistit si informace o dané

firmě osobně a následně s nimi pracovat. K tomuto oslovování dochází přes mailing, cold calling a přes samotnou sociální síť LinkedIn.

4.2 Analýza dat z LinkedIn

Před samotným zpracováním je třeba rozhodnout, která data je nezbytné získat pro dané řešení. Po dohodě s firmou BizzTreat se rozhodlo, že vzhledem k firemnímu marketingu, bude analýza sestavena nad daty týkající se přidávání příspěvků. Z následné diskuse vzešlo, jaká data by bylo dobré sledovat a jaké nové informace získat.

Budou sledovány aktivity jednotlivých ambasadorů BizzTreat, kteří jsou v současné době ve firmě tři. Za aktivity je považováno přidávání příspěvků a sdílení příspěvků. U samotných příspěvků lze sledovat několik metrik i atributů. První skupinou jsou počty, mezi které patří počet liků, sdílení, komentářů a impresí. Do druhé skupiny je možné zařadit sledování toho, zda jednotlivé příspěvky obsahují nějaký text a zda je k nim přidán obrázek, video nebo odkaz. Aby mohly být příspěvky zpracovány nástroji, jako je Bizzflow, musí na každý příspěvek odkazovat primární klíč v podobě identifikačního kódu. Dále by bylo užitečné vidět, kdo daný příspěvek přidal, případně sdílel a kdy k tomu došlo. Na základě těchto požadavků byl sestaven seznam potřebných atributů k sestavení řešení:

- Email ambasadora,
- ID příspěvku,
- Datum aktivity,
- ID sdíleného příspěvku: Tento atribut obsahuje čísla, z nichž každé odkazuje na ID cizího příspěvku. Vysvětleno na příkladu, pokud ambasador 1 sdílí příspěvek ambasadora 2, ve sloupci ID sdíleného příspěvku pro záznam ambasadora 1 bude ID příspěvku ambasadora 2.
- Text příspěvku,
- Počet liků na příspěvku,
- Počet impresí příspěvku,
- Počet komentářů na příspěvku,
- Počet sdílení příspěvku,

- Sloupec obsahující informaci, zda je u příspěvku obrázek,
- Sloupec obsahující informaci, zda je u příspěvku video, a poslední
- Sloupec obsahující informaci, zda je u příspěvku odkaz.

4.2.1 Mockup

Po vyřešení otázky, která data je potřeba získat, je možné přejít na návrh řešení. Výsledkem práce bude dashboard, neboli dynamický nástroj, jenž slouží k prezentaci klíčových informací ohledně daného tématu. Aby se zamezilo co možno nejvíce různým komunikačním šumům a zbytečným chybám, je vhodné v takovém případě vytvořit mockup. Mockup představuje detailní návrh vytvořený za účelem znázornění řešení s cílem vyhovět požadavkům zadavatele a co nejvíce ulehčit práci vykonavateli. Je patřičné poznamenat, že mockup není finální řešení, tudíž může být v průběhu práce pozměněn tak, aby bylo možné pro vykonavatele řešení vytvořit, ovšem pod podmínkou spokojenosti zadavatele.

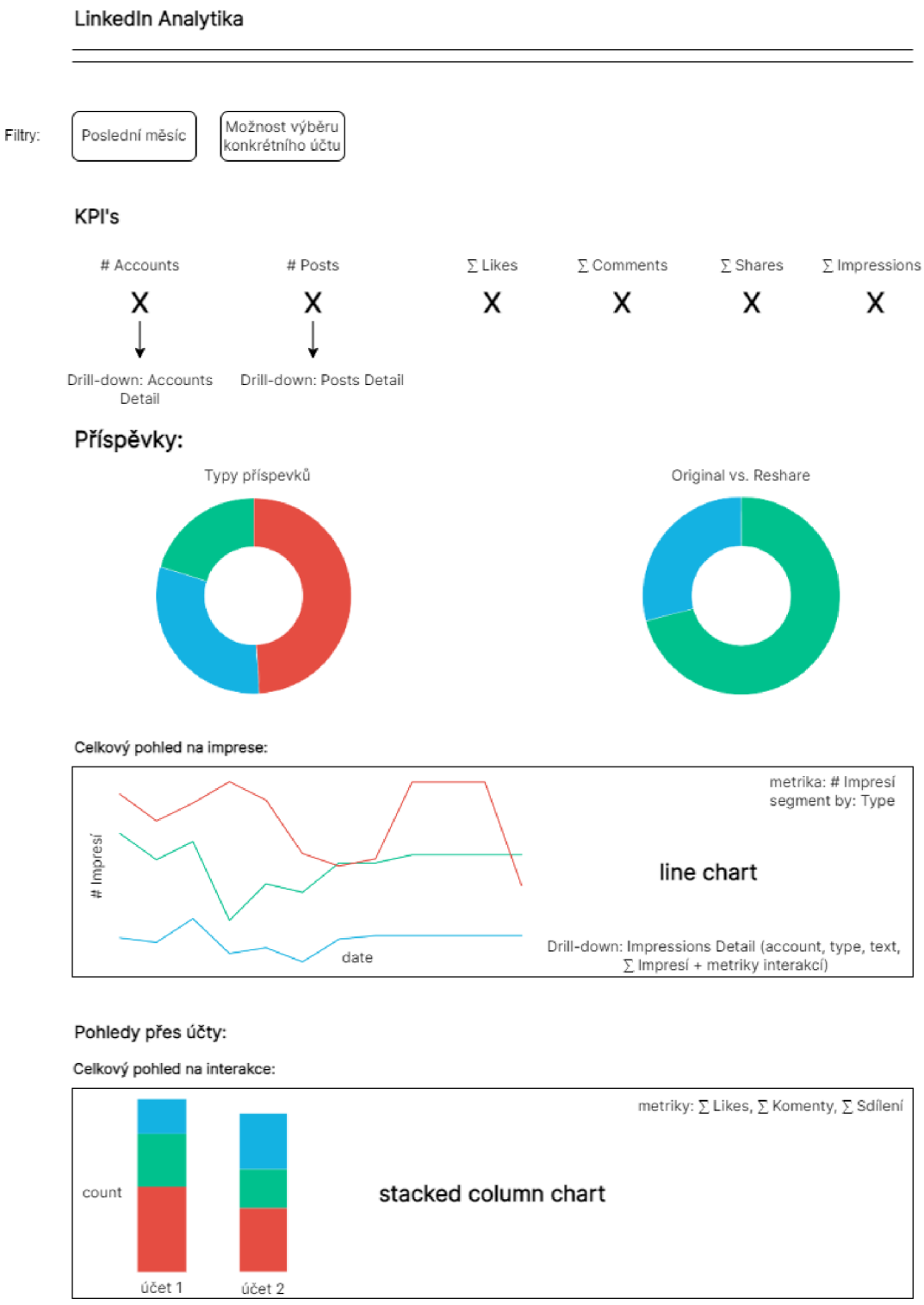
Níže uvedený mockup byl vytvořen prostřednictvím platformy Diagrams.net. Mockup je rozdělen na čtyři sekce. V první sekci jsou zobrazeny klíčové indikátory, neboli KPI. Tyto indikátory představují obecná čísla, která jsou společná pro celý dashboard a patří mezi ně celkový počet účtů, kolik příspěvků celkově přidali či sdíleli a metriky týkající se těchto příspěvků, tedy počet liků, komentářů, sdílení a impresí.

Do druhé sekce, zvané Příspěvky, lze zařadit insighty, jež ukazují obecné informace o příspěvcích. Prvním typem insightu je Donut chart, skrze který jsou prezentovány informace o typech příspěvků, příkladem kolik procent příspěvků jsou příspěvky pouze s obrázkem, kolik pouze s videem, atd. a informace o tom, kolik příspěvků je přidaných a kolik sdílených. Druhým typem insightu v této sekci je Line chart, pomocí nějž je vykreslená metrika o počtu impresí v čase a ta je sledována přes jednotlivé typy příspěvků.

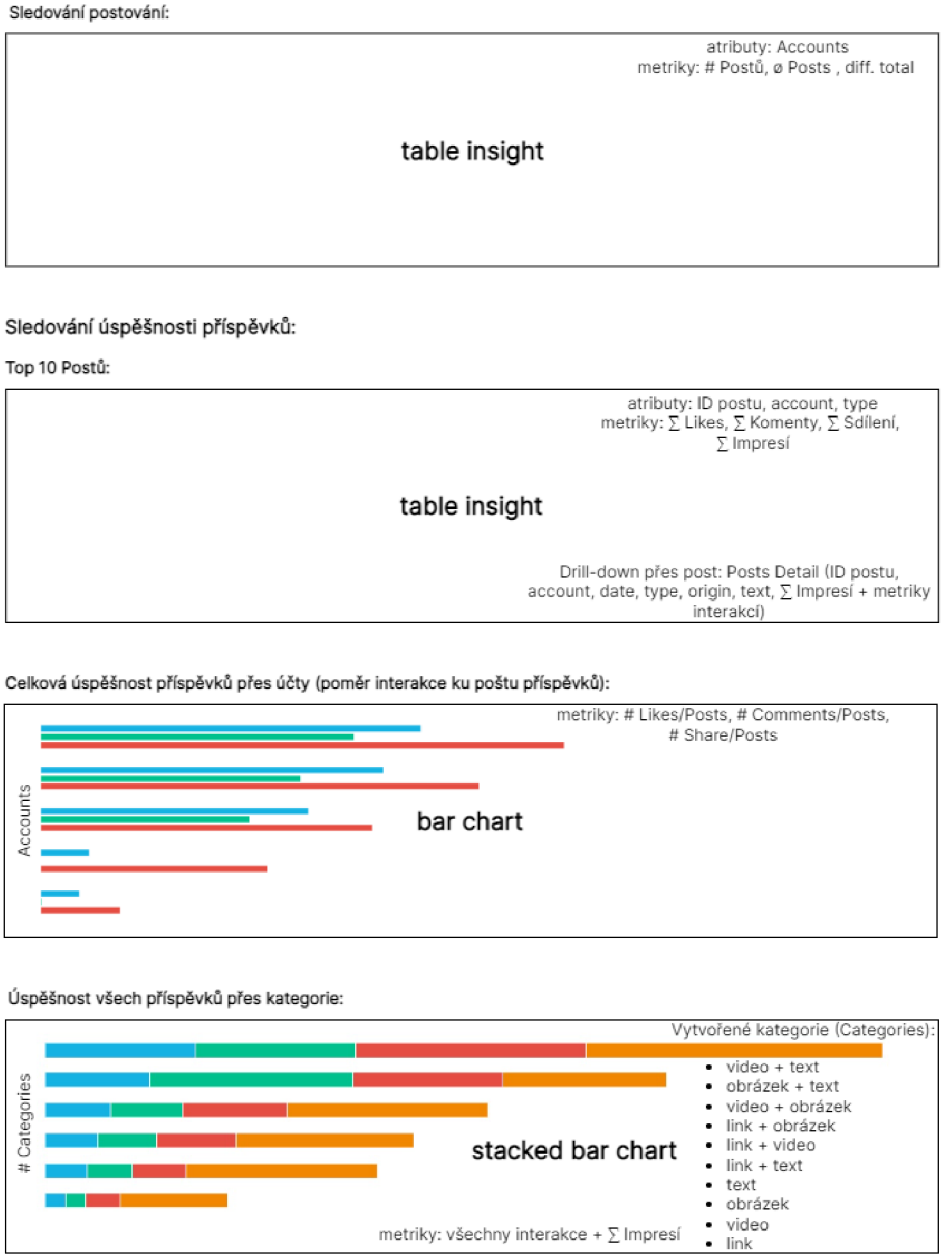
Předposlední sekce představuje pohledy přes účty. Insight Celkové pohledy přes interakce se zaměřuje na jednotlivé účty a k nim přiřazené počty liků, komentářů a sdílení k jejich příspěvkům. Druhý insight slouží ke sledování přidávání příspěvků. Počet příspěvků v měsíci je porovnáváno s průměrným celkovým počtem příspěvků a pro lepší přehlednost představuje poslední sloupec rozdíl mezi těmito hodnotami.

Poslední a nejdelší sekce umožňuje sledovat příspěvku ve větším detailu, a to jejich úspěšnost. První insight bude jednoduchá tabulka přesně s deseti záznamy o nejúspěšnějších příspěvcích s informacemi, kdo příspěvěk přidal, o jaký příspěvek se jedná a jednotlivé metriky k němu. Pro druhý insight v této sekci jsou vytvořeny metriky, které znázorňují úspěšnost jednotlivých interakčních metrik ku celkovýmu počtu příspěvků daného účtu. Opět tyto nové metriky jsou sledovány přes účty ambasadorů. Posledním insightem celého dashboardu je Stacked Bar Chart. Tento graf slouží k rozdělení příspěvků na typy a k následnému vyzobrazení interakčních metrik. Díky tomuto grafu lze přehledně rozeznat, zda se vyplatí některé typy příspěvků vůbec přidávat, důvodem může být dosahování nízkého počtů liků či komentářů u daného typu.

Za cíl je kladeno sledovat, jak si jednotlivé aktivity vedou. Zjistit například, zda příspěvky s textem a obrázkem mají vyšší dosah než pouze příspěvky s čistým textem. Dalším bodem k sledování je, jak si vedou jednotlivý ambasadoři. Jak moc jsou aktivní a jak se jim v těchto aktivitách daří, případně co je třeba zlepšit.



Obrázek 10: Navržený mockup dashboardu 1/2
Zdroj: vlastní tvorba, diagrams.net



Obrázek 11: Navržený mockup dashboardu 2/2
Zdroj: vlastní tvorba, diagrams.net

4.3 Datový zdroj a konfigurace projektu

Data, jež byla získána a poskytnuta firmou BizzTreat, byla uložena do CSV souborů. Tyto soubory je třeba nahrát do prostředí, ze kterého budou extrahována do BigQuery databáze. Jako prostředí byly vybrány Google Spreadsheets, neboli Google Tabulky. K tabulkám je možné přistupovat skrze Google Drive a jelikož se jedná o cloudový

nástroj, výhodou je možnost sdílet libovolné tabulky mezi více uživateli, tím pádem jejich správa není závislá pouze na lidech, kteří mají přístupové údaje k účtu. Tabulka, jež je v práci využívaná, má strukturu třech listů, z nichž každý představuje jednoho ambasadora BizzTreat. Data v nich jsou rozdělena do sloupců, které byly definovány v kapitole 4.2. Je důležité, aby byla dodržována stejná struktura pro všechny listy, v opačném případě by extrakce nebyla efektivní. Posledním krokem prováděným v tabulce je udělení přístupu k servisnímu účtu. Ten je udělován za účelem samotné extrakce.

Ve chvíli, kdy je připravený zdroj dat, je třeba přejít ke konfiguračním souborům pro Bizzflow. Prvním konfiguračním souborem je soubor pro nastavení extraktoru. Zde je specifikováno, že v případě této práce se jedná o Google Spreadsheet extraktor, a proto je nutné vložit url adresy listů, jaký servisní účet je zde napojen a kam budou data uložena.

```
{
    "type": "ex-google-spreadsheet",
    "config": {
      "spreadsheets":
          {
              "url-adresa-1-listu": [1],
              "url-adresa-2-listu": [2],
              "url-adresa-3-listu": [3]
          },
          "service_account": "#!#:google_drive_svc",
          "scope":
["https://spreadsheets.google.com/feeds","https://www.googleapis.com/aut
h/drive"],
          "output_folder": "/data/out/tables/",
          "debug": true,
          "sort_columns": true
    }
}
```

Druhým konfiguračním souborem je soubor step.json, který je do projektu generován standardně. Protože listy jsou v podstatě totožné, ale s jinými daty, nejprve se pomocí konvence UNION definuje, kde jsou uložené extrahované tabulky či jednotlivé listy, a následně se určuje inkrement, primární klíč, výsledný kex, název tabulky a projektu.

```
    "whitelist": {},
    "union": {
```

```
  "bizztreat-
internal.raw_ex_google_spreadsheet_linkedin_posts_stats.linkedin_posts_s
tats_1": {
    "sources": [
      {
        "kex": "raw_ex_google_spreadsheet_linkedin_posts_stats",
        "table": "linkedin_posts_stats_2",
        "project": "bizztreat-internal"
      },
      {
        "kex": "raw_ex_google_spreadsheet_linkedin_posts_stats",
        "table": "linkedin_posts_stats_3",
        "project": "bizztreat-internal"
      }
    ],
    "distinct": false
  }
    },
"filter": {},
"copy": {
"bizztreat-
internal.raw_ex_google_spreadsheet_linkedin_posts_stats.linkedin_posts_s
tats_1":{
            "incremental": true,
            "primary_key": ["__timestamp","post_urn"],
            "mark_deletes": false,
            "destination": {
                "kex": "in_00_linkedin",
                "table": "linkedin_posts_stats",
                "project": "bizztreat-internal"
            }
        }
    }
```

4.4 Zpracování dat

V této podkapitole bude rozepsán a popsán proces zpracování dat pomocí SQL dotazů. Strukturu projektu lze rozdělit na několik vrstev, které existují z důvodu organizace a přehlednosti. První vrstva je značena jako 00 vrstva, do které vstupují data v takové podobě, jaké byla extrahována a jsou zde očištěna. Poté následuje 10 vrstva, do které vstupují data již očištěná z 00 vrstvy a jsou zde zpracovávána taková data, jež budou využívána v celém projektu víckrát. Další transformační vrstvou je 20 vrstva, do které vstupují data z 00 vrstvy, v případě existence tabulek

v 10 vrstvě tak z 10 vrstvy, a dochází zde ke specifickým transformacím jednotlivých tabulek. Poslední fází je datamart, představující pouze výčet dat, jež jsou dále nahrávána do nástrojů třetích stran, v případě této práce, nástroje GoodData.

4.4.1 Očištění dat

První vrstvou, která byla pro práci vytvořena, je 00 LinkedIn vrstva. V této vrstvě dochází k očištění dat od nevalidních vstupů, tedy data vstupují do této vrstvy v takzvané raw podobě a mnohdy tato data nejsou připravena k pokročilým transformacím. Jedním takovým nevalidním vstupem je NULL hodnota. Jak již bylo zmíněno, NULL hodnota znázorňuje prázdnou buňku a takovými buňkami se velice špatně pracuje. Kvůli NULL hodnotám by bylo jakékoliv vyhledávání nebo filtrovaní tabulky komplikované a zpomalilo by to procesy prováděné s tabulkou. S NULL hodnotami by mohlo dojít také porušení integrity tabulek, normalizace, případně k nejednoznačné interpretaci záznamů. NULL hodnoty je tedy třeba nahradit adekvátním ekvivalentem k datovému typu sloupce. V BigQuery se rozlišují 4 datové typy, textové, číselné, datumové a logické. Logický typ bude převeden na textový, aby v případě prázdné buňky byla dosazena hodnota '--empty--' a ne jedna z hodnot TRUE či FALSE.

Vyskytuje-li se hodnota NULL ve sloupci textového typu, je nahrazena hodnotou '--empty--'. Jednoznačně tak označíme buňku jako prázdnou, přesto obsahuje hodnotu. Příklad očištění v takovém případě vyobrazuje následující obrázek:

```
CASE WHEN `post_id` IS NULL OR TRIM(`post_id`) = '' THEN '--empty--'
ELSE CAST(TRIM(`post_id`) as STRING) END as `post_id`
```

V případě číselné hodnoty je vhodné použít nahrazení NULL hodnoty hodnotou 0. Pokud jde o INTEGER, hodnota bude 0, ale v případě FLOAT bude nahrazovací hodnota 0.0:

```
CASE WHEN `comments` IS NULL OR TRIM(`comments`) = '' THEN 0 ELSE
CAST(TRIM(`comments`) as Int64) END as `num_comments`
```

Posledním typem je datumový typ. NULL hodnotu v tomto případě je více než vhodné nahradit datumem, který je vzdálený od dat, která se v tabulce vyskytují. V případě

této práce je takovou hodnotou '1970-01-01 00:00:00' pro typ TIMESTAMP, v případě typu DATE je hodnotou datum bez času, tedy '1970-01-01'. Příklad části dotazu k očistění datumových typů pak vypadá takto:

```
CASE WHEN `__timestamp` IS NULL OR TRIM(`__timestamp`) = '' THEN
CAST('1970-01-01 00:00:00' as TIMESTAMP) ELSE CAST(TRIM(`__timestamp`)
as TIMESTAMP) END as `timestamp`
```

Dalším nevalidním vstupem je špatný formát. Mnohdy se tak děje například u datumových typů, kdy oddělovač může být kolmá čára či tečka místo spojovníku. Špatný formát je třeba převést na správný za pomoci těchto funkcí:

```
FORMAT_TIMESTAMP('%Y-%m-%d %H:%M:%S', CAST(PARSE_TIMESTAMP('%m/%d/%y
%H:%M', `time_added`) AS TIMESTAMP)) AS `time_added`
```

4.4.2 Transformace dat

Jsou-li data očištěná, lze je ve 20 LinkedIn vrstvě transformovat a obohacovat o nové informace, a to podle předem vytvořeného mockupu. Výstupní tabulka obsahuje pouhý výčet níže vypsaných sloupců popsané viz Tabulka 1: Obsah LinkedIn Posts Stats. Jediný přidaný atribut, který není již obsažen v extrahované tabulce, je jméno ambasadora. To je přidáno s cílem uživatelsky přívětivějšího zobrazení. Skrze sloupec `name` z tabulky `in_owners` je tedy zajištěno zobrazení jména ambasadora. K tomu je třeba dodat, že tabulka `in_owners` neobsahuje výčet ambasadorů, nýbrž obsahuje jména a emaily lidí z firmy BizzTreat. Pro účely projektu by bylo ale zbytečné zakládat nový dataset pro jména ambasadorů, proto byla využita tabulka `in_owners`.

```
CREATE OR REPLACE TABLE `tr`.`out_linkedin_posts_stats` AS
SELECT
--PK
    CONCAT(`post_id`, '_', a.`snapshot_date`) AS `post_id`,
-- dates
    a.`snapshot_date`,
    `time_added`,
-- attributes
    c.`name`,
    a.`email`,
    `user_text`,
```

```
    `reshare_id`,
    `link_post`,
    `image_post`,
    `video_post`,
-- facts
    `num_comments`,
    `num_impressions`,
    `num_likes`,
    `num_shares`,
    b.`num_posts_total_in_month`
FROM `tr`.`cln_linkedin_posts_stats_snapshot` a
LEFT JOIN `tr`.`in_owners` c
    ON a.`snapshot_date` = c.`snapshot_date` AND a.`email` = c.`email`
;
```

Za pomoci druhé popsané transformace je zajištěno vyřazení takzvaných znovu zveřejněných příspěvků. Tyto příspěvky lze poznat tak, že mají stejné ID jako originální příspěvky, počet impresí na tomto příspěvku je nula, ačkoliv jsou zaznamenané počty nasbíraných likes, sdílení, komentářů a posledním identifikačním bodem je, že mají hodnotu sloupce `reshare_id` '--empty--'. Ve chvíli, kdy jsou všechny tyto tři podmínky splněny, je záznam vyřazen. Pokud ambasador znovu zveřejní příspěvek jiného ambasadora, ID příspěvku se v tabulce vyskytuje dvakrát. Proto bylo vytvořeno ID složené z ID příspěvku a emailu účtu, jenž příspěvek znovu zveřejnil. Vytvořila se tedy tabulka, ve které jsou tyto ID zapsány a v následné transformaci jsou vyřazeny prostřednictvím klíčových slov `NOT IN`. Důvodem vyřazení těchto záznamů je zkreslování skutečnosti. Dále je v transformaci vytvořen sloupec s `snapshot_date`, který obsahuje datum extrakce.

```
CREATE OR REPLACE TABLE `tr`.`tmp_linkedin_posts_stats_republished_posts` AS
SELECT DISTINCT
    `post_id`||'_'||`email` AS `post_id_email`
FROM `tr`.`in_linkedin_posts_stats`
WHERE `num_impressions` = 0 AND `reshare_id` = '--empty--'
;

CREATE OR REPLACE TABLE `tr`.`cln_linkedin_posts_stats_snapshot` AS
SELECT
    `post_id`,
    `email`,
    b.`snapshot_date`
FROM `tr`.`in_linkedin_posts_stats` a
INNER JOIN (
    SELECT
```

```
    EXTRACT(DATE FROM `timestamp` AT TIME ZONE 'Europe/Prague') AS
`snapshot_date`,
    MAX(`timestamp`) AS `max_timestamp`
    FROM `tr`.`in_linkedin_posts_stats`
    GROUP BY 1
) b ON a.`timestamp` = b.`max_timestamp`
WHERE
    `post_id`||'_'||`email` NOT IN (SELECT `post_id_email` FROM
`tr`.`tmp_linkedin_posts_stats_republished_posts`)
;
```

Třetí transformace představuje jeden celý výsledný dataset a je vytvořena za účelem sledování metrik přes jednotlivé elementy příspěvků. `CREATE` dotaz obsahuje klauzuli `UNION ALL`, přes kterou je možné vytvořit tabulku pomocí určitých podmínek a záznamy následně spojit. Podmínkou jednotlivých `SELECT` dotazů jsou názvy elementů, které jsou obsaženy u příspěvků, tedy text, video, image a link. Pro vytvoření jedinečného ID je použita funkce `CONCAT`, jež spojuje záznamy ze dvou sloupců ve tvaru, `` `post_id`'_'`snapshot_date` ``. Je tak zajištěno, že při dalším případném stažení dat se lze jednoduše odkázat na daný záznam s daným datem stažení.

```
CREATE OR REPLACE TABLE `tr`.`out_linkedin_posts_stats_elements` AS
SELECT
    CONCAT(`post_id`, '_', `snapshot_date`) AS `post_stats_id`,
    'image' AS `element`
FROM `tr`.`cln_linkedin_posts_stats_snapshot`
WHERE `image_post` = 'TRUE'

UNION ALL

SELECT
    CONCAT(`post_id`, '_', `snapshot_date`) AS `post_stats_id`,
    'video' AS `element`
FROM `tr`.`cln_linkedin_posts_stats_snapshot`
WHERE `video_post` = 'TRUE'

UNION ALL

SELECT
    CONCAT(`post_id`, '_', `snapshot_date`) AS `post_stats_id`,
    'link' AS `element`
FROM `tr`.`cln_linkedin_posts_stats_snapshot`
WHERE `link_post` = 'TRUE'
```

```
UNION ALL

SELECT
    CONCAT(`post_id`, '_', `snapshot_date`) AS `post_stats_id`,
    'text' AS `element`
FROM `tr`.`cln_linkedin_posts_stats_snapshot`
WHERE `user_text` != '--empty--'

;
```

Poslední transformací v SQL kódu je dotaz s využitím agregační funkce `COUNT` pro vypsání počtu příspěvků přidaných nebo sdílených daným účtem v každém měsíci. Za pomoci příkazu `LEFT JOIN` je tato tabulka napojena na výstupní tabulku za účelem získání sloupce `` `num_posts_total_in_month` `` do výstupní tabulky. Kromě agregační funkce je v dotazu využita i funkce `DATE TRUNC`, která slouží ke zkrácení data, definované prvním parametrem funkce, na určitý časový interval, který je specifikován ve funkci druhým parametrem. V případě tohoto dotazu je nutné využít současně funkci `CAST`, jež převede datový typ hodnoty na jiný typ, definovaný druhým parametrem funkce. Příklad dotazu vypadá následovně:

```
CREATE OR REPLACE TABLE `tr`.`tmp_linkedin_posts_stats_posts` AS
SELECT
    `email`,
    `snapshot_date`,
    CAST(DATE_TRUNC(`time_added`, MONTH) AS DATE) AS `month_added`,
    count(`post_id`) AS `num_posts_total_in_month`
FROM `tr`.`cln_linkedin_posts_stats_snapshot`
GROUP BY 1,2,3
;
```

Jsou vytvořeny tedy dvě výstupní tabulky, první LinkedIn Posts Stats a druhá LinkedIn Posts Stats Element, jejichž atributy jsou znázorněny v tabulkách níže.

Tabulka 1: Obsah LinkedIn Posts Stats

Název sloupce	Datový typ	Délka	Popis
post_id	STRING	65 535	identifikační číslo záznamu
snapshot_date	DATE	10	datum stažení
time_added	TIMESTAMP	26	datum přidání
name	STRING	65 535	jméno ambasadora

email	STRING	65 535	email ambasadora
user_text	STRING	65 535	text příspěvku
reshare_id	STRING	65 535	ID sdíleného příspěvku
link_post	STRING	65 535	informace, zda obsahuje příspěvek link
image_post	STRING	65 535	informace, zda obsahuje příspěvek obrázek
video_post	STRING	65 535	informace, zda obsahuje příspěvek video
num_comments	INT64	19	počet komentářu
num_impressions	INT64	19	počet impresí
num_likes	INT64	19	počet liků
num_shares	INT64	19	počet sdílení
num_posts_total_in_month	INT64	19	počet příspěvků

Zdroj: vlastní tvorba

Tabulka 2: Obsah tabulky LinkedIn Posts Stats Elements

Název sloupce	Datový typ	Délka	Popis
post_stats_id	STRING	65 535	ID záznamu
element	STRING	65 535	hodnoty: text, video, image, link

Zdroj: vlastní tvorba

Skrze tyto transformace nebylo dosaženo veškerých metrik potřebných pro výsledný dashboard, chybějící metriky budou vytvořeny ve vizualizačním nástroji GoodData prostřednictvím jazyka MAQL. V případě, že by byly veškeré metriky vytvořeny v transformační vrstvě, ve výsledné tabulce by byl nadbytečný počet sloupců.

Chybějící metriky není potřeba mít jako samostatný sloupec, jelikož jejich využití nebude tak komplexní, jako je tomu u jiných sloupců. Zároveň ale platí, že veškeré složité transformace, jimž není MAQL uzpůsobený, je nezbytné vytvořit při zpracování v Bizzflow. Samotný MAQL zvládá jednoduché agregační funkce a filtrování dle určitých kritérií.

4.4.3 Datamart

Dalším krokem je konfigurace datamartu. Protože pro tuto práci byly vytvořeny nové transformační vrstvy, 00 LinkedIn a 20 LinkedIn, je třeba vytvořit i datamart. Bizzflow projekt obsahuje další konfigurační soubor, datamarts.json, v němž jsou vypsány všechny datamarty. Pro konfiguraci nového datamartu je potřeba specifikovat ID datamartu, out kex a dm kex. Jelikož bude využíván již existující datový model, data budou přesměrována do datamartu Delivery, jež je využíván pro datový model, který bude upraven. Konfigurace nového datamartu je napsána následovně:

```
    {
        "id": "linkedin",
        "out_kex": "out_20_linkedin",
        "dm_kex": "dm_delivery"
    }
```

4.4.4 Nastavení transformací

Jako poslední část, která je prováděna přímo v Bizzflow za pomocí kódu, je konfigurace transformací. Aby bylo možné si v Airflow otestovat funkčnost napsaných transformací, je třeba vytvořit transformace v souboru transformations.json. První vytvořenou vrstvou byla 00 vrstva. Postupně je v kódu specifikován název zdroje, id transformace, vstupní a výstupní kexy. Konfigurace pro tuto vrstvu vypadá následovně:

```
[
  {
    "type": "sql",
    "source": "00_linkedin",
    "id": "00_linkedin",
    "in_kex": "in_00_linkedin",
```

```
    "out_kex": "out_00_linkedin",
    "input_tables": [],
    "query_timeout": 600,
    "transformation_service_account": null
  },
```

Druhou vytvořenou transformační vrstvou byla vrstva s název 20_linkedin. V kódu je tentokrát specifikován vstupní kex, jenž představuje veškeré výstupní tabulky z vrstvy 00_linkedin. Jelikož v 20_linkedin je pracováno i s tabulkou Owners, která existuje v jiné transformační vrstvě, než 00_linkedin, bylo třeba ji zahrnout do vstupních tabulek transformace. Konfiguraci 20_linkedin vrstvy lze vidět níže:

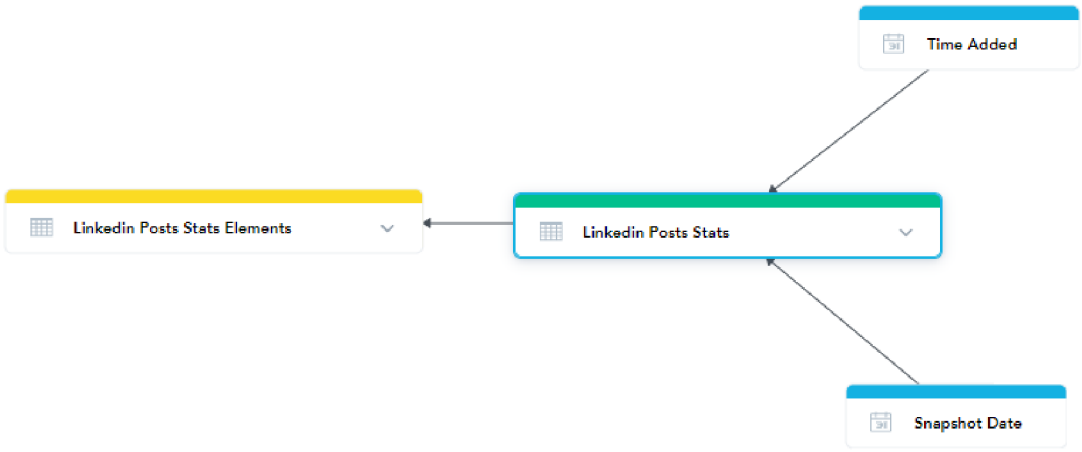
```
  {
        "type": "sql",
        "source": "20_linkedin",
        "id": "20_linkedin",
        "in_kex": [
            "out_00_linkedin"
        ],
        "out_kex": "out_20_linkedin",
        "input_tables": [
            "bizztreat-internal.out_10_common_data.owners"
        ],
        "query_timeout": 600,
        "transformation_service_account": null
    }
]
```

4.5 Datový model a vizualizace dat

Aby byl proces kompletní, připravené výstupní tabulky je třeba také vizualizovat. Pro tyto účely byla vybrána platforma GoodData. Tato platforma poskytuje rozhraní k vytváření real-time dashboardů. Před samotným tvořením dashboardu je nutné provést několik dalších kroků. Po vytvoření transformací s výstupními tabulkami se vytvoří nový či se upraví stávající datový model a data mohou být nahrána do platformy. Pomocí nově vytvořených faktů a atributů jsou tvořeny metriky pro následné sestavení insightů, z nichž se skládá výsledný dashboard.

4.5.1 Datový model a nahrání dat

Aby mohla být data nahrána do platformy, musí být předem specifikováno kam a která přesně. V modelu je tedy zapotřebí vytvořit zcela nové datasety, představované již zmíněnými výstupními tabulkami, LinkedIn Posts Stats a LinkedIn Posts Stats Elements. Dataset LinkedIn Posts Stats obsahuje dva datumové sloupce, snapshot_date a time_added, které je nutné vytvořit v modelu jako oddělené datumové dimenze. Tyto dimenze poté mají vazbu na dataset. Druhý dataset, LinkedIn Posts Stats Element, je napojen vazbou N:1 na dataset LinkedIn Posts Stats.



Obrázek 12: Využívaná část modelu
Zdroj: vlastní tvorba, GoodData Platform

K nahrání dat do platformy slouží speciální komponenty zvané writers. Data do writers přichází z datamartů, a jelikož výstupní data jsou nastavena tak, aby byla uložena v již existujícím Delivery datamartu, který je napojen na writer pro prací využívaný model, není třeba nastavovat nový writer.

4.5.2 Metriky

Prostřednictvím nově vytvořených atributů a faktů lze sestavit metriky potřebné pro výsledný dashboard. Metriky jsou vytvářeny jazykem MAQL, vyvinutý společností

GoodData, právě pro tyto účely. Dalo by se říct, že MAQL představuje zjednodušenou podobu SQL jazyka, například není zde potřeba uvádět zdrojovou tabulku dat. Naopak, pokud se vytváří metriky, vždy musí obsahovat fakt s agregační funkcí či již vytvořenou metriku. Funguje zde také podmínka `WHERE`, kterou je využívána obdobným způsobem jako v SQL. Různé příklady využití budou popsány níže.

Metriky jsou rozděleny opět na několik skupin. Do první skupiny patří metriky, které zobrazují sumy. Sumy jsou počítány z faktů `Num Likes (LinkedIn Posts Stats)`, `Num Shares (LinkedIn Posts Stats)`, `Num Impressions (LinkedIn Posts Stats)`, `Num Comments (LinkedIn Posts Stats)`. Příkladem takové metriky je `∑ Comments`. Tato metrika má za úkol vypočítat sumu všech komentářů, které jsou obsaženy v záznamech s maximální hodnotou `Date (Snapshot Date)`. Do metriky bylo třeba také specifikovat, přes jaký atribut bude suma počítána, a to přes klíčové slovo `BY`. Důvodem je datový model. Mockup obsahuje několik grafů, ve kterých je třeba metriku rozdělit přes elementy. Jelikož fakt je obsažen v tabulce LinkedIn Posts Stats, v metrice je specifikován primární klíč tabulky LinkedIn Posts Stats, jenž je zároveň hlavně cizí klíč tabulky LinkedIn Posts Stats Elements. Po této úpravě je možné metriku rozdělit při vizualizaci přes jednotlivé elementy, v opačném případě by grafy nemohly být realizované.

```
SELECT SUM(SELECT SUM(Num Likes (LinkedIn Posts Stats)) BY Post ID
(LinkedIn Posts Stats)) WHERE Date (Snapshot Date) = (SELECT
MAX(SELECT Date (Snapshot Date) BY Records of Linkedin Posts Stats
Elements) BY ALL IN ALL OTHER DIMENSIONS)
```

Další skupinou metrik jsou metriky počtu, tedy takové, jež využívají agregační funkci `COUNT`. To je využíváno u faktů jako je `Post ID (LinkedIn Posts Stats)` nebo `Email (LinkedIn Posts Stats)`. Pomocí následující metriky lze spočítat počet přidaných postů, opět s podmínkou maximálního `Date (Snapshot Date)`.

```
SELECT COUNT(Post ID (LinkedIn Posts Stats)) WHERE Date (Snapshot
Date) = (SELECT MAX(SELECT Date (Snapshot Date) BY Post ID (LinkedIn
Posts Stats)) BY ALL IN ALL OTHER DIMENSIONS)
```

Pro skupinu metrik, které počítají úspěšnost jednotlivých příspěvků, ať už přes počet liků, sdílení či komentářů, je využit způsob sestavení z již existujících metrik. Výhodou

je, že není třeba opět vypisovat veškeré podmínky, které jsou již zahrnuty v předem vytvořených metrikách. Naopak nevýhodou je, že ve chvíli, kdy je využitá metrika smazána, stane se nefunkční i tato metrika. Tento způsob je využíván i u metriky, jež počítá rozdíl mezi celkovým počtem příspěvků a průměrným počtem příspěvků. Metrika tím zobrazuje, jak si daný ambasador vede v daném období.

```
SELECT SUM(# Posts - Med. Posts) WHERE Date (Snapshot Date) = (SELECT
MAX(SELECT Date (Snapshot Date) BY Post ID (LinkedIn Posts Stats)) BY
ALL IN ALL OTHER DIMENSIONS)
```

U formátu metriky je dále nastaveno speciální zobrazení. Ve chvíli, kdy je hodnota metriky < 0, číslo bude červené, naopak ve chvíli, kdy je hodnota kladná, číslo bude zelené. Při hodnotě 0 bude hodnota mít černou barvu. Formát takovéto metriky je napsán následovně:

```
[<0][red]-#,#.##;[>0][green]#,#.##;[=0][black]#,#0
```

Metrika výše využívá již postavenou metriku zvanou `Med. Posts`. Poslední představovanou metrikou bude právě tato. Tato metrika využívá dvě speciální, ještě nepředstavené, funkce. První z nich je agregační funkce `MEDIAN`. Jak již název napovídá, pomocí funkce lze získat prostřední hodnotu seřazeného souboru hodnot. Druhá speciální funkce je využívaná za účelem ignorování specifikovaných filtrů, které jinak platí pro celý dashboard. Příkladem, níže uvedená metrika počítá medián z faktu `Num Posts Total In Month (LinkedIn Posts Stats)`, ale budou na ni fungovat veškeré filtry dashboardu kromě filtru časového období skrze datumovou dimenzi `Date (Time Added)`.

```
SELECT (SELECT MEDIAN(Num Posts Total In Month (LinkedIn Posts Stats))
WITH PF EXCEPT Date (Time Added)) WHERE Date (Snapshot Date) = (SELECT
MAX(SELECT Date (Snapshot Date) BY Post ID (LinkedIn Posts Stats)) BY ALL
IN ALL OTHER DIMENSIONS)
```

4.5.3 Tvorba dashboardu

Finální dashboard, vytvořen platformou GoodData, se skládá přesně ze šesti KPI ukazatelů, z nichž dva jsou definovaný jako insight Headline, aby bylo možné prokliknout na tabulku s větším detailem. Tyto dva ukazatele zobrazují počet ambasadorů a počet zveřejněných příspěvků. Proklikem na čísla se zobrazí tabulky s větším detailem, tedy podrobnějšími informacemi. Zbylá KPI prezentují celkový počet nasbíraných liků, komentářů, sdílení a impresí.

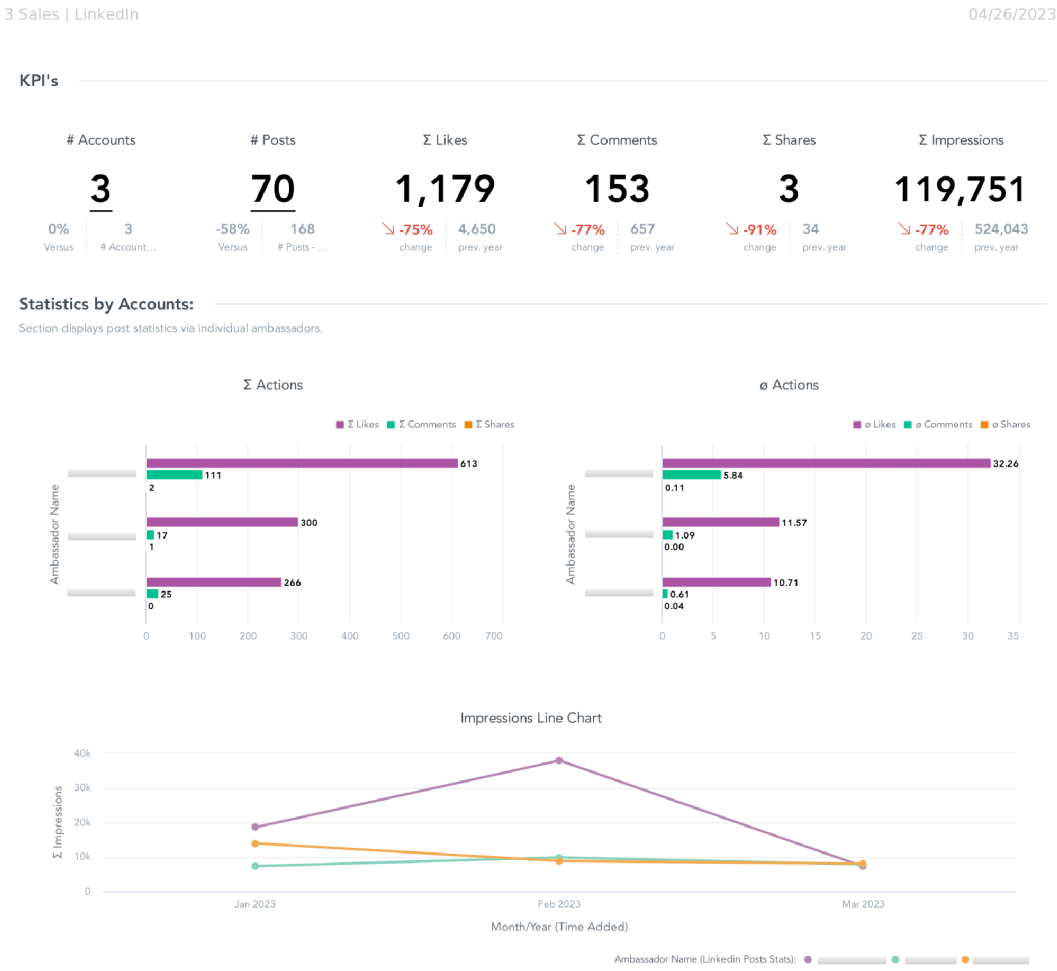
Pod klíčovými indikátory se vyskytuje sekce Statistics by Account, která představuje pohledy na příspěvky přes účty ambasadorů. Pomocí grafů Bar Chart je prezentováno, jak si jednotlivé účty vedou, co se akcí týče. V grafu nalevo jsou opět využity metriky pro sumu liků, sumu sdílení a sumu komentářů a tyto fakty jsou sledovány přes účty. Graf napravo je funkčně analogický ke grafu nalevo, avšak využívané metriky představují průměry těchto faktů. Pod grafy se vyskytuje insight Impressions Line Chart zobrazuje celkovou sumu impresí nasbíraných v čase. Insight je sledován také přes jednotlivé účty.

Tabulka Posting View nabízí pohled na počet příspěvků každého účtu. Toto číslo je porovnáváno, prostřednictvím rozdílu v posledním sloupci, s druhým sloupcem, jenž reprezentuje medián souboru příspěvků. Díky rozdílu lze pozorovat, zda daný ambasador je v daném měsíci nad svým mediánem nebo pod ním. Pro přehlednost jsou kladná čísla zelená, záporná červená a hodnota nula je černá.

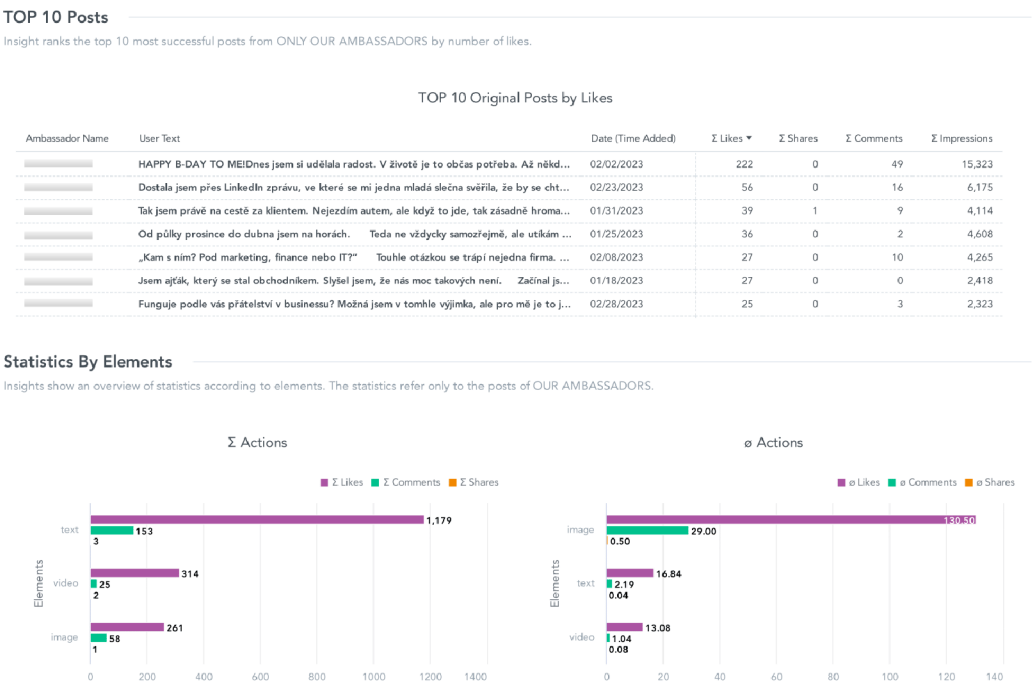
Následující tabulka s názvem TOP 10 Posts, již podle názvu, představuje deset nejúspěšnějších příspěvků publikovaných ambasadory. V tabulce jsou důležité informace o jménu ambasadora, popisu příspěvku, datumu zveřejnění a opět jednotlivých sumách akcí. Při kliknutí na hodnoty ve sloupci User Text se zobrazí tabulka s detailnějšími informacemi o příspěvcích, jako tomu je u druhého KPI.

Poslední grafy jsou obsaženy v sekci s názvem Statistics By Elements. Insight nalevo, typu Bar Chart, vykresluje znovu celkové sumy liků, sdílení a komentářu, tentokrát ovšem rozdělené na jednotlivé elementy obsažené v příspěvcích. Mezi tyto elementy patří text, obrázek, video a link. K tomu stejnému dochází i v grafu napravo, obdobné rozdělení přes elementy, ale pro průměrný počet liků, sdílení a komentářů.

Výsledný dashboard se liší v několika případech. Je to z důvodu, že mockup uvedený v kapitole 4.2.1 vyjadřuje všechny různé možnosti zobrazení získaných dat. Po přesném vypracování bylo rozhodnuto, že některé grafy jsou nepotřebné a zbytečně by zvětšovaly dashboard. Došlo k zúžení jen na grafy, které jednoznačně odpovídají na otázky jako „Jaké příspěvky mají největší dosah?“, „Jsou úspěšnější příspěvky s obrázky než jen pouhé texty?“ či „Proč jeden ambasador je úspěšnější než ten druhý?“.



Obrázek 13: Finální dashboard vytvořený platformou GoodData 1/2
Zdroj: vlastní tvorba, GoodData Platform



Obrázek 14: Finální dashboard vytvořený platformou GoodData 2/2
Zdroj: vlastní tvorba, GoodData Platform

5. Zhodnocení a formulace doporučení

Tato kapitola slouží ke shrnutí veškerých úspěchů i nedostatků, ke kterým došlo a k sepsání možných potenciálních zlepšení na projektu. Po shrnutí následuje formulace doporučení, komu a k čemu by mohla být tato bakalářská práce prospěšná.

Ačkoliv mockup sloužil jako výchozí bod pro finální dashboard, výsledné řešení se v některých bodech liší. První odlišností je celá sekce Statistics by Accounts, jež měla původně zobrazovat informace o příspěvcích. Zadavatelem bylo zhodnoceno, že tato sekce neseděla do myšlenky celého dashboardu. Ten by měl totiž zobrazovat takové informace, na základě kterých budou dělat ambasadoři rozhodnutí ohledně přidávání příspěvků. Tudíž informace v grafu Donut Chart ohledně počtu příspěvků s obrázky, texty a videi jsou nadbytečné, stejně tak jako graf prezentující, kolik příspěvků je sdílených a kolik naopak publikovaných ambasadory. Proto tyto grafy byly nahrazeny pohledy na jednotlivé účty a jejich dosahy. Zadavatel také zhodnotil, že poslední insight s informací o kategoriích příspěvků je nepřehledný v tomto rozložení. Z tohoto důvodu byl insight předělán z typu Stacked na klasický horizontální Bar Chart, a navíc se vytvořil ještě jeden graf, analogický k tomuto, ale s metrikami ukazující průměry akcí na příspěvcích. Mockup také popisoval, bez grafických předloh, takzvané drill down reporty. Pro ty bylo třeba vybrat vhodný typ grafu. Dashboard byl také graficky doladěn, například jakékoliv metriky týkající se liků jsou zobrazeny fialovou barvou atd.

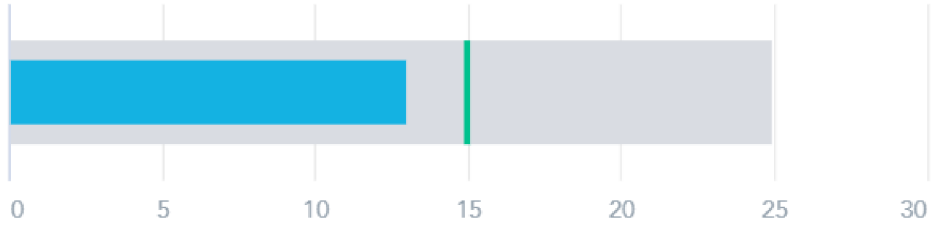
Výsledek lze považovat za úspěšný z důvodu spokojenosti ze strany zadavatele, tak ze strany autorky projektu a bude velice často využíván ambasadory i v budoucnu. Byl dodržen předepsaný pracovní postup a časový horizont.

5.1 Budoucnost projektu

V současné době není proces aktuálnosti dashboardu zcela automatizovaný. Vše závisí na datech, která jsou obsažená v google tabulkách. V budoucnu je v plánu projekt rozšířit o automatickou extrakci dat ze sítě LinkedIn, a tím bude celkový proces zautomatizovaný. Mimo tuto změnu bude třeba upravit i kód a přizpůsobit ho této úpravě.

V případě, že by došlo k rozšíření počtu ambasadorů, bylo by možné do logického datového modelu přidat zcela novou dimenzi nazvanou Ambassadors. Pravděpodobně by to byla jednoduchá tabulka, jež by obsahovala atributy jako email ambasadora, jméno ambasadora a vazbu na datumovou dimenzi Snapshot Date. Došlo by tím k větší přehlednosti, protože u nynějšího řešení je jméno ambasadora bráno z dimenze Owners, ačkoliv se ve své podstatě jedná jen o ambasadora.

Posledním možným rozšířením je nastavení plánů. Pro každého ambasadora by byl nastaven měsíční limit přidání příspěvků. Poté by se na dashboardu vytvořil insight typu Bullet Chart a bylo by možné sledovat skutečnost, kolik příspěvků musí ambasador ještě přidat, aby splnil limit.



Obrázek 15: Příklad grafu typu Bullet Chart
Zdroj: vlastní tvorba, GoodData Platform

5.2 Doporučení

Tato bakalářská práce je vhodná pro kohokoliv, v jehož zájmu je pracovat s Bizzflow. Praktická část obsahuje ve své podstatě celý postup, jak pracovat s projektem v Bizzflow. Jediné, co je třeba provést a není v práci popsáno, je zprovoznění tohoto nástroje. Je zde ale popsáno, jak sepsat konfiguraci extraktoru, jak nastavit step.json soubor a také jak vytvořit jednotlivé transformace a datamart pro práci v Airflow.

Dále může být doporučena studentům oboru Informatika a oborům podobného zaměření, kteří mají zájem o oblast datové analytiky. Práce využívá ETL proces ke zpracování dat s názornými ukázkami dotazů v SQL.

Závěr

Obsah bakalářské práce přibližuje čtenářům pochopení problematiky datové analýzy a její aplikovatelnosti na marketingová data. Cílem práce bylo vysvětlit významné pojmy z oblastí online marketingu, databází, Business Intelligence a vizualizace dat. Následně byla znalost těchto pojmů převedena na nově získané dovednosti při vytváření celého řešení projektu.

V první části lze nalézt literární rešerši zaštiťující termíny z oborů online marketingu, datové analýzy, databází a databázových systémů, se kterými je úzce spojen pojem SQL, a Business Inteligence. Mimo vyjmenované oblasti byl představen v této části také ETL nástroj Bizzflow, v němž probíhá proces zpracování dat.

Následuje praktická část, v níž se jedná o zpracování marketingových dat. Bylo třeba krátce představit marketingovou strategii firmy BizzTreat, pro niž je projekt zpracováván. Dále bylo nezbytné provést analýzu dat ze sociální sítě LinkedIn, jejíž výsledem je mockup. Na analýzu navazuje proces zpracování dat, zahrnující kroky získání dat, jejich očištění, transformaci a souběžně s tímto procesem bylo třeba konfigurovat jednotlivé části projektu pro jeho funkčnost. Celý proces končí úpravou logického datového modelu, jenž je vyhovující podmínkám dat, jež jsou nahrána do platformy GoodData a připravena k vizualizaci. Finální dashboard je tvořen přehlednými grafy i klíčovými ukazateli, ale především vyhovuje požadavkům zadavatele.

Dashboard odpovídá na otázky, díky nimž je možné optimalizovat marketing firmy. Jednoduše vyobrazuje klíčová čísla týkající se příspěvků, graficky znázorňuje celkové sumy a průměry akcí provedených na příspěvcích, i v časovém horizontu několika měsíců. Také lze sledovat úspěšnost jednotlivých příspěvků při rozdělení na elementy příspěvků, jako jsou popisek, obrázek, video či text. V neposlední řadě je pomocí malé tabulky zachycena skutečnost, kolik příspěvků bylo ve vybraný měsíc zveřejněno daným účtem s porovnáním k mediánu počtu příspěvků. Na základě těchto grafů mohou jednotliví ambasadoři sledovat svou aktivitu, jaký dosah mají jejich příspěvky a zda se jim vyplatí přidávat například více příspěvky s videem nežli bez něj.

V budoucnu bude konečné řešení rozšířeno o automatizaci stahování dat na denní bázi, a to za účelem zobrazení nejnovějších dat. Projekt byl také sestaven tak, aby po menších úpravách byl připraven na možná rozšíření v podobě více ambasadorů, nových elementů příspěvků i plánů.

Při vypracování bakalářské práce nabyla autorka mnohých poznatků spojených s literární rešerší a nových zkušeností při samotném vytváření projektu. Přínosem je také ponaučení z chyb, které vznikly při realizaci projektu a skutečnost, že komunikace je základním stavebním kamenem. Projekt úspěšně slouží ke sledování statistik ambasadorských příspěvků.

Seznam použité literatury

BEAULIEU, Alan, 2020. *Learning SQL: Generate, Manipulate, and Retrieve Data.* THIRD EDITION. Sebastopol, Canada: O'Reilly Media. ISBN 978-1-492-05761-1.

BIZZFLOW, [2023]. Bizzflow *wiki* [online]. Prague, Czech Republic: BizzTreat [cit. 2022-10-04]. Dostupné z: https://www.bizzflow.net/

BLY, Robert W., 2020. *The Content Marketing Handbook: How to Double The Results of Your Marketing Campaigns*. Irvine, California, USA: Entrepreneur Media. ISBN 978-1-61308-417-5.

BUREŠOVÁ, Jitka, 2022. *Online marketing: od webových stránek k sociálním sítím.* Praha: Grada Publishing. ISBN 978-80-271-1680-5.

CORONEL, Carlos and Steven MORRIS, 2016. *Database Systems: Design, Implementation, and Management*. 12th Edition. Boston, MA, USA: Cengage Learning. ISBN 978-1-305-62748-2.

CUESTA, Hector, 2015. *Analýza dat v praxi*. Brno: Computer Press. ISBN 978-80-251-4361-2.

GÁLA, Libor, Jan POUR a Zuzana ŠEDIVÁ, 2015. *Podniková informatika: Počítačové aplikace v* podnikové *a mezipodnikové praxi*. 3., aktualizované vydání. Praha: Grada Publishing. ISBN 978-80-247-5457-4.

JANOUCH, Viktor, 2014. *Internetový marketing*. 2. vydání. Brno: Computer Press. ISBN 978-80-251-4311-7.

kolektiv autorů, 2014. *Online marketing*. Brno: Computer Press. ISBN 978-80-251-4155-7.

LAURENČÍK, Marek a Michal BUREŠ, 2018. *SQL: podrobný průvodce uživatele*. Praha: Grada Publishing. ISBN 978-80-271-0774-2.

LINKEDIN, c2023. *O společnosti LinkedIn* [online]. Mountain View, CA, USA: Microsoft [cit. 2023-01-18]. Dostupné z: https://about.linkedin.com/cs-cz?trk=homepage-basic_directory_aboutUrl&lr=1

MCCORMICK, Kristen, *2022. The 6 Biggest, Baddest, Most Popular Social Media Platforms* [online]. Boston, MA, USA: WordStream [cit. 2023-02-07]. Dostupné

z: https://www.wordstream.com/blog/ws/2022/01/11/most-popular-social-media-platforms

POUR, Jan, Miloš MARYŠKA, Iva STANOVSKÁ a Zuzana ŠEDIVÁ, 2018. *Self Service Business Intelligence: Jak si vytvořit vlastní analytické, plánovací a reportingové aplikace*. Praha: Grada Publishing. ISBN 978-80-271-0616-5.