VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA PODNIKATELSKÁ
FACULTY OF BUSINESS AND MANAGEMENT

ÚSTAV INFORMATIKY
INSTITUTE OF INFORMATICS

NÁVRH DATOVÉHO SKLADU V SAAS SPOLEČNOSTI
DESIGN OF DATA WAREHOUSE IN SAAS COMPANY

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — **Bc. Adam Zetocha**
AUTHOR

VEDOUCÍ PRÁCE — **Ing. Jiří Kříž, Ph.D.**
SUPERVISOR

BRNO 2020

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA PODNIKATELSKÁ

Zadání diplomové práce

Ústav:	Ústav informatiky
Student:	**Bc. Adam Zetocha**
Studijní program:	Systémové inženýrství a informatika
Studijní obor:	Informační management
Vedoucí práce:	**Ing. Jiří Kříž, Ph.D.**
Akademický rok:	2019/20

Ředitel ústavu Vám v souladu se zákonem č. 111/1998 Sb., o vysokých školách ve znění pozdějších předpisů a se Studijním a zkušebním řádem VUT v Brně zadává diplomovou práci s názvem:

Návrh datového skladu v SaaS společnosti

Charakteristika problematiky úkolu:

Úvod
Cíle práce, metody a postupy zpracování
Teoretická východiska práce
Analýza současného stavu
Vlastní návrhy řešení
Závěr
Seznam použité literatury
Přílohy

Cíle, kterých má být dosaženo:

Cílem diplomové práce je návrh datového skladu na základě potřeb společnosti. Bude dosažena efektivní práce s daty a datovými zdroji včetně tvorby reportů pro vybrané oddělení společnosti.

Základní literární prameny:

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

GEMIGNANI, Zach, Chris GEMIGNANI, Richard GALENTINO a Patrick SCHUERMANN. Efektivní analýza a využití dat. Brno: Computer Press, 2015. ISBN 978-80-251-4571-5.

LABERGER, Robert. Datové sklady: Agilní metody a business inteligence. Brno: Computer press, 2012. ISBN 978-80-251-3729-1.

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

POUR, Jan, Miloš MARYŠKA a Ota NOVOTNÝ. Business Intelligence v podnikové praxi. Praha: Professional Publishing, 2012. ISBN 978-80-7431-065-2.

Termín odevzdání diplomové práce je stanoven časovým plánem akademického roku 2019/20

V Brně dne 29.2.2020

L. S.

doc. RNDr. Bedřich Půža, CSc.
ředitel

doc. Ing. et Ing. Stanislav Škapa, Ph.D.
děkan

Abstrakt

Diplomová práca obsahuje postupný návrh a budovanie dátového skladu v startupe vyvíjajúcom analytický SaaS produkt. Teoretické poznatky z oblasti dátových skladov a business intelligence sú premietnuté do návrhu a následného budovania dátového skladu predovšetkým nad marketingovými dátami s tým, že proces plnenia dátového skladu a reportingu je plne automatizovaný.

Abstract

The diploma thesis consists of design and steps leading to build of data warehouse in star tup developing SaaS product. Theoretical information about data warehouses and business intelligence are projected into design and following process of data warehouse development mainly for marketing data. Importing process of data into a data warehouse and reporting are fully automated.

Kľúčové slová

Dátový sklad, Business Intelligence, Google BigQuery, Reporting, ETL

Key words

Date warehouse, Business Intelligence, Google BigQuery, Reporting, ETL

Bibliografická citácia

ZETOCHA, Adam. *Návrh datového skladu v SaaS společnosti*. Brno: Vysoké učení technické v Brně, Fakulta podnikatelská, 2020. 91 s. Vedoucí práce Ing. Jiří Kříž, Ph.D.

Čestné prehlásenie

Prehlasujem, že predložená diplomová práca je pôvodná a spracoval som ju samostatne. Prehlasujem, že citácie použitých prameňov sú úplné, že som vo svojej práci neporušil autorské práva (v zmysle Zákona č. 121/2000 Sb., o právu autorském a o právech souvisejících s právem autorským).

V Brne dňa 17. mája 2020

……

podpis studenta

Poďakovanie

Týmto by som rád poďakoval svojmu vedúcemu práce Ing. Jiří Kříž, Ph.D. za odbornú pomoc pri vypracovávaní tejto diplomovej práce.

Zároveň by som chcel poďakovať firme Smartlook za umožnenie práce na danom projekte a mojim kolegom za ich odborné rady. V neposlednom rade by som chcel poďakovať mojej rodine za podporu v celej dĺžke môjho štúdia.

OBSAH

ÚVOD

V dnešnej technologickej ére tvoria dáta jedno z najdôležitejších aktív jednotlivých firiem, ale aj jednotlivcov. S príchodom sociálnych sietí do života obyčajných ľudí na konci prvého desaťročia 21. storočia sa kompletne pretransformoval digitálny svet. Ľudia sa začali čoraz viac deliť o svoje osobné informácie na sociálnych sieťach, nakupovanie cez internetové obchody pomaly vytláča nakupovanie v kamenných obchodoch. Tradičné doručovateľské služby čelia tlaku zo strany technologických firiem, ktoré vďaka dátam dokážu zásielky doručovať rýchlejšie a efektívnejšie s väčšou transparentnosťou. To, ako si objednávame jedlo bolo tiež veľmi rýchlo transformované, až to prišlo do bodu, kedy je možné si vďaka pár klikom objednať potraviny priamo do domu. Toto všetko je možné vďaka rozsiahlemu zbieraniu a spracovávaniu dát.

Tento trend podporil vznik úplne nových technologických odvetví zameraných predovšetkým na skladovanie, analýzu a reportovanie dát. Pre firmy to otvorilo možnosti ako lepšie porozumieť podnikateľskému prostrediu, v ktorom sa nachádzajú, svojim zákazníkom, produktom alebo službám, ktoré predávajú a poskytujú a v neposlednom rade aj svojim zamestnancom.

Vzniklo ale tiež veľké množstvo výziev pre firmy pôsobiace v takmer všetkých odvetviach a nutnosť inovovania zaužívaných stratégii a procesov bude v nasledujúcich rokoch nevyhnutná, inak tieto firmy ďalšiu dekádu nemusia prežiť.

CIEĽ A METODIKA PRÁCE

Cieľom diplomovej práce je navrhnúť dátový sklad pre potreby spoločnosti a ich interných užívateľov. Dôležité je aby bola dosiahnutá automatizácia práce s dátami, dátovými zdrojmi a výsledné vytváranie reportov pre špecifické oddelenia spoločnosti.

Zdrojom dát pre analytické pracovanie s dátami je interná databáza obsahujúca dáta z viacerých externých zdrojov. Zdrojom dát sú mimo samotného softwaru aj marketingové nástroje, nástroje používané technickým supportom, sales departmentom a nástroje spravujúce finančné dáta.

V analytickej časti bude najskôr predstavená spoločnosť Smartlook a samotný vyvíjaný produkt. Následne bude zanalyzovaná dátová infraštruktúra a v krátkosti jednotlivé komponenty infraštruktúry. Ďalšia časť analýzy bude venovaná dátovým zdrojom, spôsob, akým sa dáta ukladajú z dátových zdrojov na úložisko, v akých časových intervaloch a ako sú následne spracovávané. Zanalyzovaná bude súčasná podoba analytickej databáze, nad ktorou sa všetky procesy vykonávajú. Popísaná bude štruktúra tabuliek, jednotlivé väzby a prechod dát celou databázou.

V návrhovej časti budem následne daný navrhnutý dátový sklad budovať. Celý projekt budovania dátového skladu bude najskôr rozdelený na jednotlivé fázy. Po tom čo v analytickej časti budú zanalyzované potenciálne riešenia pre ETL proces a samotný dátový sklad, budú v návrhovej časti z týchto analyzovaných riešení vybrané dve, ktoré budú následne implementované. Navrhnutá bude dátová architektúra, na základe ktorej bude dátový sklad budovaný. Popísaná bude príprava zdrojovej databázy, tabuľky faktov a dimenzií, ktoré budú používané v dátovom sklade. Po úspešnej implementácii ETL softwaru a pripojenia na dátový sklad a zdroje dát, bude ako posledný krok popísaná implementácia s reportovacím softwarom, z ktorého budú následne pre overenie správneho vybudovania dátového skladu vypracovaný dva reporty.

1 TEORETICKÉ VÝCHODISKÁ PRÁCE

Prvá časť diplomovej práce je venovaná teoretickým východiskám, na základe ktorých bude v ďalších častiach práce vypracovaný návrh dátového sklad.

1.1 Dátový sklad

Dátový sklad (angl. data warehouse) je systém slúžiaci na zhromažďovanie, uchovávanie, organizáciu a zdieľanie predovšetkým historických dát. Dáta ukladané v dátovom sklade pochádzajú zväčša z prevádzkových systémov, ktoré tieto dáta zachytávajú a používajú v kontexte svojej prevádzkovej činnosti. V praxi plnia dátové sklady aj iné systémy, ako sú prevádzkové systémy. V projektoch návrhu dátového skladu sa ale pre zdrojové systéme univerzálne používa označenie prevádzkové systémy (angl. operational system). [1.]

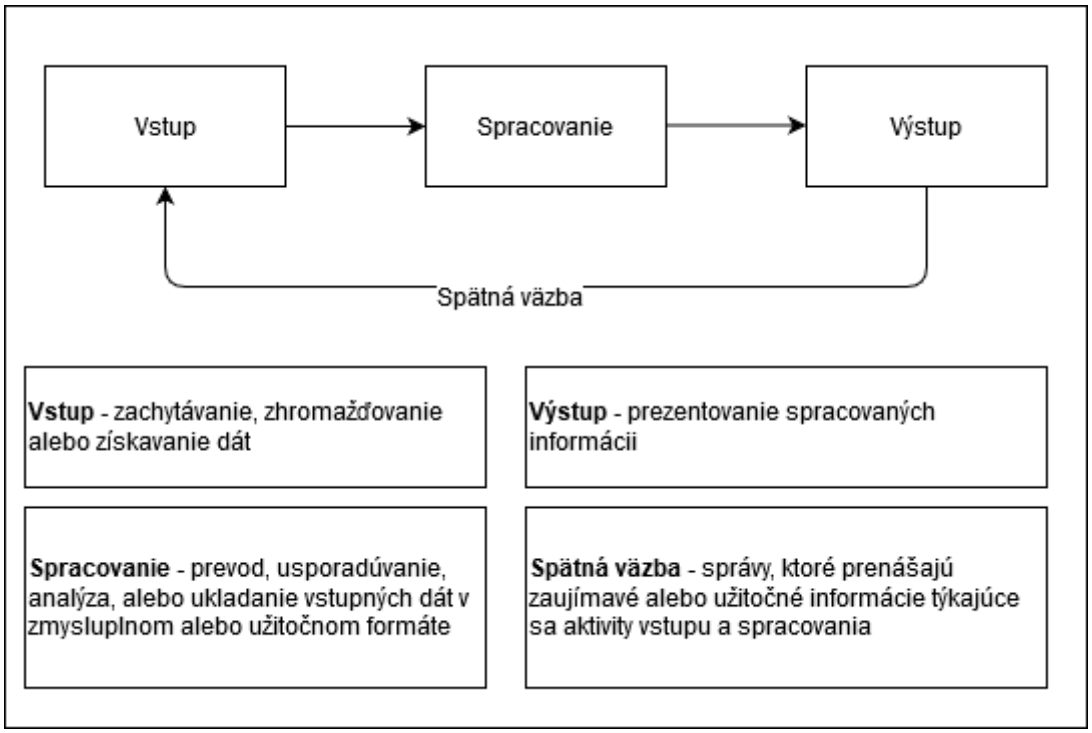
Projekty budovania dátového skladu sú bežne koncipované pre potreby celého podniku. Nebýva však neobvyklé keď je dátový sklad zameraný na špecifický obor činnosti, napríklad financie alebo marketing. Termín dátový sklad sa môže na základe rozdielnych literatúr vzťahovať v niektorých prípadoch k systému dátového skladu alebo úložisku dátového skladu. [1.]

1.1.1 Systém dátového skladu

Hlavné komponenty systému dátového skladu sú podobne ako u iných systémov tvorené vstupom, spracovaním, výstupom a spätnou väzbou. Fáza vstupu súvisí s identifikáciou a zaznamenávaním dát, pričom je kľúčová kvalita vstupných dát. Ak do systému vstupujú nesprávne dáta, môže to viesť k nepresným výstupom z dôvodu prenášania daných dát do závislých procesov, subsystémov a podnikových analýz. [1.]

V jadre celého systému sa obvykle nachádza jedna veľká databáza. Nie je ale neobvyklé, ak je jadro celého systému tvorené sadou databáz umiestnených na rozdielnych serveroch. Dôležité ale stále zostáva zachovanie centrálneho riadenia celého systému. Úlohou danej centrálnej oblasti je plnenie výkonnej role pri transformácii a uchovávaní dát. Dátová

architektúra, alebo presnejšie dátový model, špecifikuje štruktúru dát. Základné, popisné a pridružené charakteristiky dát sú navrhnuté na základe logického dátového modelu. Optimalizácia dát pre ich použitie v databáze je vykonávaná na základe fyzického dátového modelu. [1.]



Obrázok 1: Základné komponenty systému

(Zdroj: 1)

1.1.2 Architektúra dátového skladu

Architektúra dátového skladu tvorí stavebný plán dátového skladu. Pre znázornenie architektúry sa najčastejšie používajú diagramy toku dát. Architektúra v sebe zahŕňa jednotlivé fázy prechodu dát dátovým skladom a komponenty dátového skladu. Tie sa delia na [1.] :

- Zdroj dát – interné systémy, štandardné systémy, kľúčové dáta a iné
- Obstaranie dát – vrstva získavania dát, mapovanie zdrojov na cieľ
- Organizácia dát – vrstva centralizácie dát, logický model, fyzický model
- Distribúcia dát – distribučná a výkonnostná vrstva, dáta upravené k použitiu
- Výstup informácií – vrstva užívateľskej prezentácie, vykazovanie dát

1.1.3 Terminológia toku dát

V oblasti dátových skladov sa často používajú dva termíny vychádzajúce z architektúry toku dát. Zhora dole (angl. top-down) a zdola hore(angl. bottom-up). Rozdiel je zjavný, keď sa na dané termíny pozrieme z pohľadu metodík Billa Inmona a Ralpha Kimballa. Zatiaľ čo v metodike Billa Inmona sa pristupuje k dátovým skladom primárne z pohľadu dát, teda zhora dole, v metodike Ralpha Kimballa je kladený dôraz na podnikový účel a dáta slúžiace k jeho podpore. Metodika Ralpha Kimballa teda funguje na princípe zdola hore. Aj keď sú obe metodiky založené na inom prístupe, obe majú rovnaký cieľ a tým je schopnosť podniku na základe štruktúrovaných dát prijímať informované rozhodnutia. [1.]

Obrázok 2:Architektúra dátového skladu: použité informácie

(Zdroj: 1)

1.2 Dátové trhovisko

Princíp dátového trhoviska je v podstate rovnaký ako princíp, na ktorom sú založené dátové sklady. Rozdiel je v tom, že dátové trhovisko je určené iba pre potreby jedného oddelenia, divízie, pobočky alebo závodu. Dátové trhovisko môže byť interpretované ako decentralizovaný a subjektovo orientovaný dátový sklad, alebo ako základ celopodnikového dátového skladu. [4.]

V prvom prípade, v ktorom je dátové trhoviská vnímané ako decentralizovaný dátový sklad s možnosťou integrácie do celopodnikového riešenia, sú dátové trhoviská vytvárané z dôvodu skrátenia doby návratnosti investície, zníženiu celkových nákladov a výraznom

znížení rizika pri zavádzaní. Ako bolo spomenuté v kapitole vyššie, ide o takzvaný Kimballov prístup zdola hore. V druhom prípade sa jedná o Inmonov prístup zhora dole, kedy je dátové trhovisko vnímané ako dátový sklad orientovaný na pokrytie konkrétneho problému určitej špecifickej skupiny užívateľov. Umožňuje tiež vykonávanie analýz dát v požadovanom momente nad menším objemom dát. Dôležitým predpokladom je, že pred vytváraním dátového trhoviska už je vytvorený celopodnikový dátový sklad obsahujúci všetky požadované dáta, ktoré sú následne extrahované do dátového trhoviska. [4.]

1.3 OLAP databáza

OLAP databáza je vo všeobecnosti tvorená jednou alebo viacerými vzájomne súvisiacimi a poprepájanými OLAP kockami. Vo väčšine prípadov sú v OLAP kockách zahrnuté predspracované agregácie dát vytvorené na základe definovaných hierarchický štruktúr dimenzií a ich kombinácii. Postupom času sa vyvíjali nové varianty OLAP technológie [4.] :

- ROLAP – relačný OLAP
- MOLAP – multidimenzionálny OLAP
- HOLAP – hybridný OLAP
- DOLAP – desktopový OLAP
- WOLAP – web based OLAP
- RTOLAP – OLAP v reálnom čase

1.4 Business inteligence

Technológia business intelligence, v skratke BI, slúži predovšetkým potrebám riadiacich pracovníkov. Hlavnou úlohou je poskytovať dôveryhodné dáta v hodnotovom kontexte. Dáta musia byť z pohľadu danej spoločnosti dôležité a dostupné v každom okamžiku. BI umožňuje podnikovým užívateľom využívať základné dáta pri kvantifikovanom rozhodovaní a v podnikových procesoch. [1.]

Business intelligence je v podstate zastrešujúci termín, ktorý zahŕňa znalosti, technológie, procesy, aplikáciu a postupy zjednodušujúce podnikové rozhodovanie. Pracuje sa predovšetkým s historickými dátami v požadovanom kontexte tak, aby bolo čo najviac zjednodušené rozhodovanie pre budúcnosť. Najdôležitejšie dáta pochádzajú z interných zdrojov o prevádzkových aspektoch súvisiace s taktickým a strategickým plánovaním. Informácie bývajú vo väčšine prípadov usporiadané a priamo alebo nepriamo sa odvodzujú z podnikových procesov. Následný výstup z dát sa väčšinou používa pre potreby internej analýzy, ale je možné ho použiť aj pre potreby externej analýzy, kam patrí SWOT a PEST kompetitívna analýza. [1.]

Obrázok 3: Business intelligence

(Zdroj: 8)

Technológia BI má za úlohu napomôcť pri určovaní podnikových cieľov na základe počiatočných kľúčových indikátorov výkonu. Dané kľúčové indikátory výkonu môžu byť rozdelené na metriky, ktoré sú priamo odvodzované z prevádzkových systémov. V prípade, že tieto metriky nie sú odvodené, môžu byť agregované zo základných prevádzkových metrík. Najčastejšie prebieha analýza v prostredí BI na agregovanej úrovni, na úrovni inštancií to nie je až také obvyklé. Dôležitosť BI riešenia naberá na dôležitosti hlavne v prípadoch, kedy prevádzkové systémy zle zbierajú a agregujú dostupné dáta. V takýchto prípadoch sa ako centrálne miesto pre zbieranie informácií používa prostredie BI. [1.]

Systém BI by sa v ideálnom prípade mal vyznačovať nasledujúcimi vlastnosťami [1.]:

- Rýchlosť – rýchle spracovávanie požiadaviek
- Aktuálnosť – dostupné informácie sú aktuálne
- Presnosť – dostupné informácie sú správne
- Užitočnosť – dáta poskytujú určitú hodnotu

1.5 Dátové formáty

Veľká väčšina aplikácií pracujúcich vo webovom prostredí má kľúčové časti implementované ako služby. Tieto kľúčové služby sú následne využívané ďalšími časťami aplikácie, napríklad klientskou aplikáciou vyvinutou v JS(javascripte) a bežiacou v prehliadači. Dáta sú medzi jednotlivými časťami aplikácie predávané v presne definovaných formátoch. Pri komunikácii by mal byť dátový formát rovnaký ako dátový formát ukladaných dát, to zabezpečí obmedzenie réžie potrebnej na konverziu dát. [3.]

1.5.1 JSON

Skratka formátu JSON znamená Javascript Object Notation. Už z názvu je jasné, že formát vznikol ako podmnožina jazyka JS. Pôvodne bol vyvinutý za účelom predávania dát medzi klientskou časťou a serverovou časťou webovej aplikácie. Ako podmnožina jazyku Javascript umožňuje reprezentáciu základných dátových štruktúr a ich použitie vo webovom prehliadači. Formát sa ale podarilo rozšíriť do takého štádia, že existuje knižnice umožňujúce jeho použitie vo všetkých štandardne rozšírených jazykoch. JSON je možné veľmi ľahko mapovať priamo na objekty daného jazyka. [3.]

Reprezentovať môže jednoduché dátové typy – string, boolean, číslo a null. A štruktúrované dátové typy – objekt a pole. Tieto štruktúrované dátové typy môžu ešte následne obsahovať ďalšie dátové typy či už sa jedná o jednoduché, alebo štruktúrované. Flexibilita tohto formátu umožňuje reprezentáciu takmer akejkoľvek dátovej štruktúry. [3.]

1.5.2 XML

Skratka formátu XML znamená eXtensible Markup Language. Dôvodom vzniku tohto dátového formátu bola potreba zjednodušenia komplexného značkovacieho jazyka SGML. Značkovacie jazyky slúžia na doplnenie elementov do jednoduchých textov, ktoré im pridajú význam. Tieto značkovacie jazyky sa pôvodne používali predovšetkým v elektronickom publikovaní na reprezentáciu rozsiahlych dokumentov. Umožnené bolo prepojenie pomocou odkazov a vyhľadávanie. Svoje uplatnenie si našli predovšetkým v právnych textoch alebo technických dokumentáciách. [3.]

Veľký dopyt po formáte prepojujúcom systémy a výmenu dát medzi nimi podporil rozšírenie práve dátového formátu XML, ktorý umožňuje použitie aj na tieto činnosti. To z neho spravilo najpoužívanejší formát používaný na výmenu dát. Ako nadstavby vznikli napríklad formáty SOAP alebo WSDL, ktoré doplnili pôvodne XML o vrstvy umožňujúce autentizáciu alebo popis rozhrania. [3.]

1.5.3 YAML

Skratka formátu YAML – Ain't Markup Language. Formát YAML bol vyvinutý za účelom zapisovania dát do takej podoby, aby boli čitateľné ľuďmi, teda nebol primárne vyvinutý pre komunikáciu informačných zariadení. Aj keď je primárnym cieľom YAML formátu čitateľnosť pre ľudí, je zachované mapovanie na dátové štruktúry bežne používaných programovacích jazykov. Aj keď YAML nie je až tak rozšírený v porovnaní s predchádzajúcimi dvoma jazykmi XML a JSON, je možné sa s ním stretnúť hlavne v systémoch, kde je používaný pre zápis konfiguračných súborov. [3.]

1.5.4 CSV

Skratka formátu CSV – Comma-separated Values. Formát CSV slúži predovšetkým na ukladanie dát v tabuľkovej podobe. Dáta sú zapísané v riadkoch oddelené určitým oddeľovačom, napríklad bodkočiarkou alebo čiarkou. Pokiaľ sa exportuje z databázy, tak prvý riadok obsahuje názvy jednotlivých stĺpcov. Jednoduchosť CSV formátu ale

spôsobila to, že nebola predstavená štandardizovaná podoba. To spôsobuje napríklad problémy s rozdielnymi znakmi použitými na oddeľovanie. [3.]

1.6 Dimenzionálne modelovanie

Dimenzionálne modelovanie patrí medzi základné analytické metódy pre riešenie aplikácií BI. Hlavným cieľom dimenzionálneho modelovanie je vytvorenie základnej logiky uloženia a usporiadania dát tak, aby nad nimi bolo možné vykonávať analytické a plánovacie podnikové riadenie. Vo výsledku by mal byť k dispozícii flexibilný dátový model, podporujúci rozsah všetkých nie len súčasných, ale aj budúcich analýz. Obsahom dátového modelovania je [4.]:

- Vymedzenie všetkých dimenzií vrátane ich obsahu a vnútorných hierarchií
- Definícia sledovaných ukazovateľov – faktov
- Definícia väzieb medzi faktami a dimenziami

1.6.1 Tabuľky faktov

Tabuľky faktov obsahujú dáta o sledovaných predovšetkým ekonomických ukazovateľoch. Jednotlivé ukazovatele sú identifikované cudzími kľúčmi, ktoré sú zložené z cudzích kľúčov dimenzionálnych tabuliek. Obsahovať môžu tiež vlastné umelé primárne kľúče. Stĺpce v tabuľke faktov vo svojej podstate obsahujú kľúčové atribúty alebo hodnoty ukazovateľov. Dimenzie obsiahnuté v tabuľke faktov vyjadrujú rozsah a podrobnosti o sledovaných ukazovateľoch. Riadky predstavujú výsledky jednotlivých meraní, s tým že dáta sú zvyčajne priraďované na čo najnižšiu úroveň detailu. Jednoznačnú identifikáciu každého záznamu zabezpečuje zložený primárny kľúč tabuľky. [4.]

1.6.2 Tabuľky dimenzií

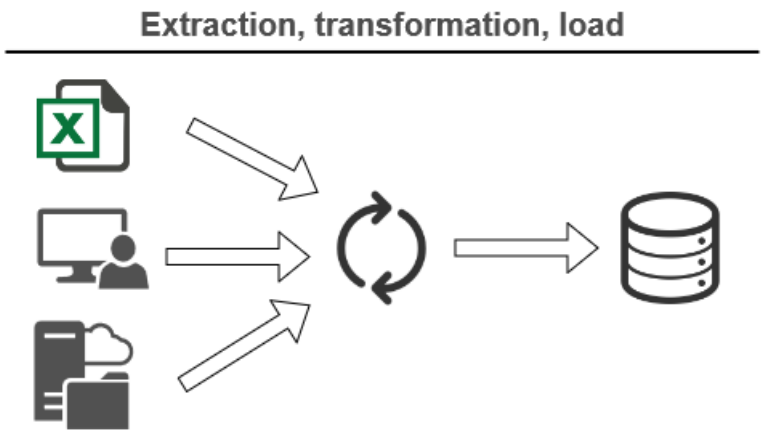
Tabuľky dimenzií obsahujú dáta o podnikových číselníkoch. Patria sem napríklad dimenzie produktov, marketingové dimenzie, dimenzie o dodávateľoch. Medzi

najzákladnejšie dimenzie, ktoré by mala obsahovať každá firemná databáza, sú časová dimenzia, dimenzia obsahujúca krajiny alebo zahraničné meny. [4.]

1.7 ETL – Extract, Transform, Load

ETL je proces označovaný tiež ako dátová pumpa. Extrakcia, transformácia a nahrávanie dát je jednou z najvýznamnejších komponent komplexu Business Intelligence. Hlavné úlohy daného procesu sú podľa názvu zrejmé. Prvým krokom procesu je extrakcia dát. Tá zahŕňa zbieranie dát zo všetkých dátových zdrojov plniacich dátový sklad alebo dátový trh. Druhá fáza takzvaná transformácia, pozostáva z úpravy dát do požadovanej formy a podoby. Posledným krokom celého procesu je load, teda nahranie získaných a očistených dát do dátových štruktúr dátového skladu alebo dátového trhoviska. [2.]

Nástroje ETL je možné použiť na prenášanie dát medzi jednotlivými databázami alebo dátovými súbormi, ako je napríklad MS Excel. Dáta sú zo zdrojových súborov vo veľkej väčšine prípadov prenášané naraz v dopredu predefinovaných intervaloch. Môžu byť teda dávkované denne týždenne alebo aj raz za mesiac, záleží na potrebe nových dát. Najviac pracovnej kapacity v procese ETL je vynakladaných na transformáciu dát, teda ich čistenie a prípravu na nahranie do DB. Táto činnosť zaberá zväčša až 60% vynaloženej pracovnej kapacity. [2.]



Obrázok 4: Extraction, transformation, load

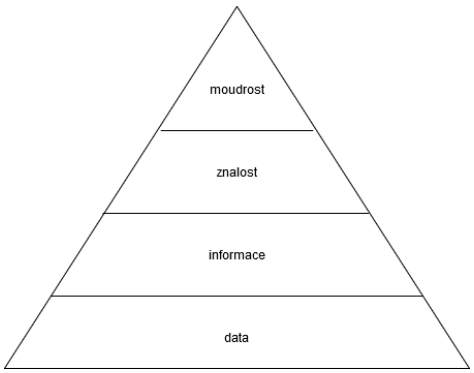
(Zdroj: Vlastná úprava)

Aby proces ETL fungoval správne, musia byť splnené niektoré charakteristiky. Zo zdrojových systém sú ťahané iba také dáta, ktoré majú analytický, plánovací a rozhodovací charakter. Dáta, ktoré sú pomocou extrakcie vytiahnuté zo zdrojových systémov, sú následne transformované do nových predpripravených dátových štruktúr analytických databáz. Tieto dátové štruktúry sú navrhnuté tak, aby odpovedali špecifickým podnikovým potrebám. Keďže dáta sú naťahované z viacerých zdrojov, ktoré môžu dané dáta obsahovať viac krát a v rôznej kvalite, veľmi dôležitou úlohou ETL procesu je dáta prichádzajúce z viacerých zdrojov očistiť od duplicít. Súčasťou konsolidácie dát je teda aj dôkladné čistenie dát tak, aby bola dosiahnutá čo najvyššia kvalita ukladaných dát do cieľových dátových štruktúr. [2.]

1.8 Informačný management

Jednou z najviac kľúčových aktivít BI je pokladaný informačný management, tiež označovaný ako BIIM vo firmách, kde je vyhradené špeciálne oddelenie venujúce sa tejto problematike. BIIM predstavuje spojenie niekoľkých rôznorodých disciplín. Zaraďujeme sem napríklad data warehousing, master data management, data mining, data quality alebo metadata management. [2.]

„Information Management je definovaný jako program, který řídi lidi, procesy a technologie v podniku, aby umožnil kontrolu nad strukturou, procesy, dodávkami a použitím informací nutných pro účely řízení a Business Intelligence. [2.]“



Obrázok 5: Pyramída DIZM

(Zdroj: 5.)

Dôležitou podmnožinou informačného managementu je proces nazývaný data management. Tento proces je predovšetkým zameraný na riadenie dát, ale rieši tiež napríklad oblasti tvorby, získavania, zdieľania alebo uchovávania dát. Rozdiel medzi data managementom a information managementom je v praxi veľmi nepatrný. [2.]

Tabuľka 1: Komponenty IM vs komponenty DM
(Zdroj: 2)

Komponenty informačného managementu	Komponenty Data managementu
Strategie Information Managementu	Data Governance
Business Intelligence a Performance mngmt.	Data Quality
Enterprise Data Management	Master Data Management
Information Assest Management	Metadata Management
Enterprise Content Management	Dátová architektúra
Content Delivery	Bezpečnosť
Architektura a technologické možnosti	Data Retention a archivácia

1.8.1 Data Governance

Pojem Data Governance v sebe zahŕňa také procesy, ktoré v spoločnosti zabezpečujú najvyššiu kvalitu a riadenie kľúčových dát tak, aby sa na dôveryhodnosť týchto dát mohol spoľahnúť každý zamestnanec spoločnosti bez obáv z nízkej kvality dát. Aj keď je Data Governance dôležitou súčasťou informačného managementu spoločnosti, časová náročnosť celého procesu môže byť príčinou toho, že si zavedenie DG môžu dovoliť iba väčšie spoločnosti, v ktorých si viacero oddelení spravuje svoje vlastné dáta. Za takéto spoločnosti môžeme napríklad považovať veľké obchodné korporácie alebo štátnu správu. [2.]

Za najdôležitejšie ciele DG môžeme považovať zvýšenie kvality dát, konzistentné využívanie dát, jasnú dátovú architektúru podporujúcu kvalitu dát, minimalizáciu

opakovaného spracovávania dát, zriadenie procesov na riadenie kvality dát a metriky umožňujúce meranie danej kvality, definovanie jednoznačnej zodpovednosti za dáta a bezpodmienečné zaistenie bezpečnosti dát. [2.]

1.8.2 Dátová kvalita

V predchádzajúcej podkapitole bol popísaný proces Data Governance. Súčasťou Data Governance je tiež kvalita dát. Keďže sú dáta v súčasnej dobe považované za jedno z najcennejších aktív spoločnosti, je kvalita dát veľmi kľúčová. Keby boli dáta používané na rozhodovanie vrcholového managementu chybné, bolo by takmer nemožné získať z daných dát hodnotné informácie vhodné pre rozhodovanie. [2.]

Nekvalitné dáta môžu negatívne ovplyvniť reputáciu spoločnosti ak sú širokému okoliu poskytované mylné informácie, napríklad pomocou webových stránok. Zlá kvalita dát môže mať tiež negatívny vplyv na riadenie spoločnosti, kedy môžu nekvalitné dáta spôsobiť mylné vyhodnotenie situácie vedúce k zlým rozhodnutiam vrcholového managementu v kľúčových otázkach. Ak sa napríklad nekvalita dát prejaví v zlom vyúčtovaní poskytovaných služieb alebo predávaných produktov, s veľkou pravdepodobnosťou hrozí strata reputácie u daného zákazníka, ktorý potenciálne môže túto informáciu rozšíriť ďalej a spôsobiť ešte väčšiu stratu z dôvodu odlákania potenciálnych zákazníkov. [2.]

Kvalitné dáta musia byť [2.]:

- Presné
- Úplné
- Jedinečné
- Konzistentné
- Dôveryhodné
- Pravdivé
- Aktuálne

Podobne ako Data Governance je aj dátová kvalita úzko spätá so správnym nastavením procesov, ktoré zodpovedajú za dlhodobé udržanie kvality dát. Veľmi dôležitú úlohu zastávajú metriky, slúžiaci k posúdeniu aktuálnej dátovej kvality. Medzi metriky zaraďujeme čísla (rozsah, presnosť), väzby medzi dátami (počet chýbajúcich cudzích kľúčov), počet hodnôt null alebo reťazce (dĺžka, vzory). S metrikami dátovej kvality sa tiež spojuje Data Profiling. Tento proces predstavuje hodnotenie informačných zdrojov a zber štatistických údajov o dátových zdrojoch. Tie sú následne hodnotené podľa prednastavených metrík. [2.]

1.8.3 Master Data Management

Master data sú v preklade do češtiny kmeňové dáta podniku, ale v dnešnej dobe sa častejšie využíva pojem hlavné dáta. Medzi tieto dáta zaraďujeme dáta o zákazníkoch, výrobkoch, produktoch alebo službách ponúkaných spoločnosťou a dodávateľoch. Hlavnou úlohou Master Data Managementu je tieto dáta spravovať. Správa kmeňových dát spočíva napríklad v pravidelnom kontrolovaní číselníkov a ich popisov tak, aby boli používané v rámci celej spoločnosti a nenastal taký prípad, keby sa budujú podobné databáze vo viacerých oddeleniach spoločnosti. [2.]

Po uvedení predchádzajúcich skutočností môžeme Master Data Managementu definovať ako správu hlavných dát z podnikových nástrojov, aplikácií a metód. Tieto nástroje, aplikácie a metódy implementujú zásady, infraštruktúru a postupy a podporujú zachytávanie, integrácie a následne zdieľané používanie presných, konzistentných a úplných dát. [2.]

Prínos Master Data Managementu je zjavný hlavne v nasledujúcich oblastiach [2.]:

- Lepšia dátová kvalita vytvára kvalitnejšie BI
- Podnikové procesy sú efektívnejšie
- Zvýšenie informovanosti o zákazníkoch spoločnosti
- Reporting na presnejšej a kvalitnejšej úrovni
- Marketing má umožnené presnejšie cielenie kampaní

1.9 Cloud computing

Cloud computing predstavuje vývoj softwarového riešenia založeného na business modely ponúkania škálovateľnej IT technológie, externým zákazníkom prostredníctvom internetu v podobe služby. Pod pojmom IT technológia je myslená kombinácia softwaru a hardwaru, ktoré môžu odpovedať platforme, infraštruktúre alebo aplikácii. Cloud computing business model sa ďalej delí na tri modely [3.]:

- **SaaS – Software as a Service –** v tomto prípade sa jedná o vývoj a poskytovanie softwaru ako služby, väčšinou na báze pravidelných automatických platieb. Do tohto modelu sa zaraďuje aj spoločnosť Smartlook.
- **PaaS – Platform as a Service –** v tomto prípade sa jedná o vývoj a poskytovanie platformy ako služby. Tieto platformy sú predovšetkým určené vývojárskym tímom, pre ktorých ponúkajú sadu nástrojov slúžiacich k vývoju aplikácií, alebo všeobecne pre prevádzkovanie softwaru. Sme môžeme zaradiť Microsoft Azure.
- **IaaS – Infrastructure as a Service –** v tomto prípade sa jedná o poskytovanie infraštruktúry ako služby. Poskytované sú väčšinou robustné, hardwarovo náročné riešenia, ktoré by neboli normálne pre firmy dostupné. AWS od firmy je Amazon je IaaS riešenie.

V prípade všetkých troch modelov koncový užívateľ neplatí za nákup samotného hardwaru alebo softwaru, ale využíva ich formou predplatného, teda platí za ich používanie. Platí sa za využívané obdobie a vo väčšine prípadov na báze opakujúcich sa platieb či už za mesačné, alebo ročné a viacročné predplatné. Okrem platby za dĺžku využívania služby sa účtuje aj za veľkosť uložených alebo spracovaných dát. Cloudové riešenia delíme z hľadiska cieľov množiny užívateľov ešte na privátne a komunitné. [3.]

Nižšie je uvedených niekoľko výhod cloudových riešení [3.]:

- Možnosť využívania cloudovej služby takmer z hocijakého miesta s prístupom na internet.
- Zákazník službu prakticky využíva bez nutnosti správy danej technológie.

- Robustné riešenia môžu byť na základe možností individuálne škálovateľné, čo predstavuje na jednej strane priestor pre zákazníkov na prispôsobiteľnosť na základe špecifických potrieb, ale aj pre vývojárov možnosť zvýšenia hodnoty produktu.
- Keďže sú dáta uložené v cloudovom priestore, dostupnosť týchto dát je umožnená vybraným užívateľom.

S cloudovými riešeniami sa samozrejme nespájajú iba výhody, ale je tu aj celá rada očividných nevýhod a potenciálnych problémov. Patria sem predovšetkým nasledovné [3.]:

- Všetky dáta uložené vo verejnom cloude, sú ukladané na miesto, nad ktorým nemá daná firma priamu kontrolu. Preto je dôležité dbať na vybratie poskytovateľa zaručujúceho vysokú bezpečnosť uložených dát.
- Prílišné napojenie firemných procesov na určitú službu môže spôsobiť proprietárne uzamknutie, spôsobujúce náročnú migráciu. Toto je nevýhoda najmä z pohľadu zákazníka. Z pohľadu firiem ponúkajúcich službu je ale toto brané ako výhoda a ideálny stav.

1.10 Funkcie Smartlook softwaru

Základnou funkciou Smartlook je nahrávanie sessions návštevníkov prichádzajúcich na web. Inštalácia je v celku jednoduchá a rýchla záležitosť v prípade, že sa jedná o nahrávanie webových stránok. Po vytvorení užívateľského účtu si zákazník vytvorí projekt, ktorý vygeneruje javascript kód s jedinečným identifikátorom projektu. Kód môže byť na web nasadený manuálne do hlavičky stránky medzi <head> </head> tagy, alebo pomocou pluginu napríklad do rozšíreného a populárneho CMS Wordpress. [23.]

```
<!--Smartlook script-->
<script type="text/javascript">
  window.smartlook||(function(d) {
  var o=smartlook=function(){ o.api.push(arguments)},h=d.getElementsByTagName('head')[0];
  var c=d.createElement('script');o.api=new Array();c.async=true;c.type='text/javascript';
  c.charset='utf-8';c.src='https://rec.smartlook.com/recorder.js';h.appendChild(c);
  })(document);
  smartlook('init', '72887ba94222ffb01c37c587bc8bf4464c4205b4');
</script>
```

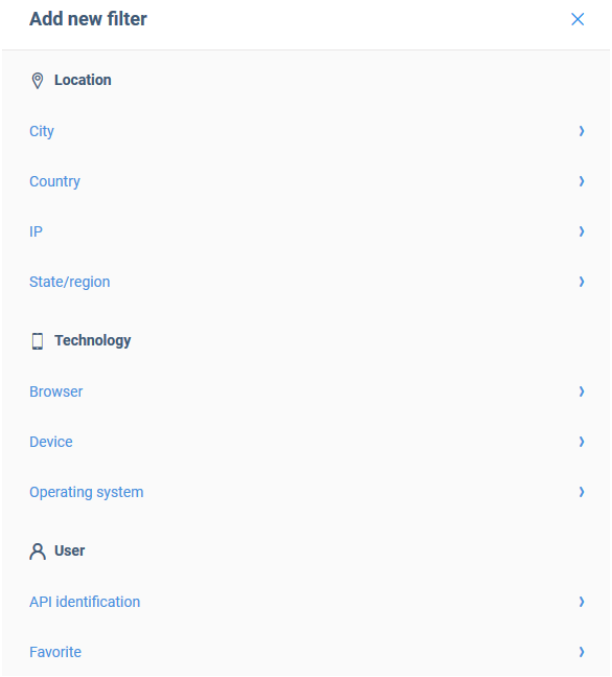
Obrázok 6: Smartlook script

(Zdroj: Vlastná úprava)

Hneď po nasadený je Smartlook schopný nahrávať prichádzajúcich návštevníkov na web. Samostatné nahrávky, hlavne vo veľkých počtoch, sú veľmi náročne uchopiteľným nástrojom, ktorý sám o sebe prináša firme malú pridanú hodnotu a predstavuje veľkú časovú náročnosť na analýzu. Preto boli do Smartlooku pridané rozšírené analytické funkcie, zvyšujúce pridanú hodnotu celého produktu a samotného nahrávania. [23.]

1.10.1 Segmenty

Základným nástrojom pre prácu s nahrávkami sú segmenty. Tie poskytujú možnosť filtrovania nahrávok na základe veľkého počtu podmienok. K dispozícii sú segmenty, filtrujúce napríklad na základe geografickej polohy, z ktorej prichádza návštevník webu. Ďalej sú k dispozícii filtre na filtrovanie nahrávok iba zo špecifických zariadení alebo filtrovanie na základe operačných systémov zariadení používaných na prezeranie webu. Je možnosť filtrovať nahrávky na základe dĺžky času stráveného na webe alebo zdroja, ktorý priviedol návštevníka na web, čím môžu byť rôzne marketingové kampane alebo organické návštevy. Návštevníci sa dajú tiež filtrovať na základe toho, či sú vracajúcimi, alebo novými návštevníkmi na webe. [23.]

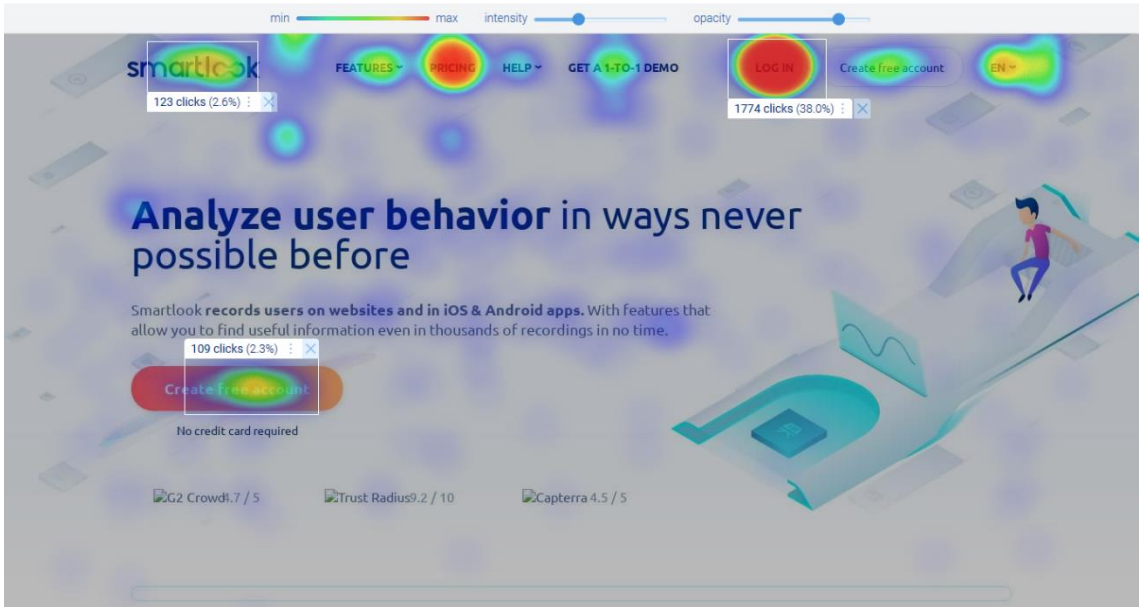


Obrázok 7: Filter menu
(Zdroj: 9)

1.10.2 Teplotné mapy

Teplotné mapy (angl. Heatmapy) slúžia na princípe zachytávania agregovaných dát zobrazujúcich miesta, na ktoré ľudia na stránke najviac klikajú, ako sa po stránke pohybujú myšou. [23.]

Heatmapa sa vyhotovuje na špecifickej URL, ktorá má byť analyzovaná. Z jednej z nahrávok ľudí, ktorý danú stránku navštívili, sa vyhotoví snímka obrazovky a na základe udalostí od všetkých návštevníkov, ktoré boli na danej URL zachytené, sa vyhotovujú takzvané horúce body zobrazujúce intenzitu interakcií s jednotlivými elementami. Heatmapy sa vyhotovujú vo formáte desktopových zariadení, mobilných zariadení a tabletov. [23.]



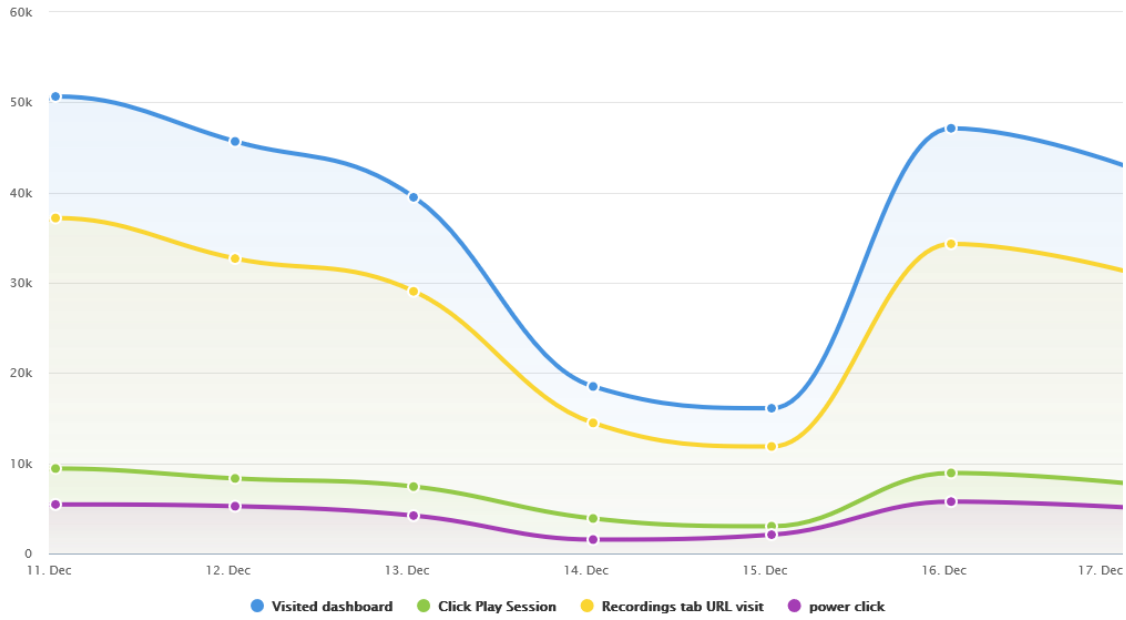
Obrázok 8: Heatmapa
(Zdroj: 9)

1.10.3 Udalosti

Za udalosť sa pokladá každá akcia vykonaná návštevníkom na webe. Smartlook zachytáva 4 základné druhy udalostí [23.] :

- Navštívená URL
- Kliknutie na text
- Kliknutie na CSS selektor
- Napísaný text

Udalosti sa zachytávajú automaticky od momentu nasadenia Smartlook scriptu na web. Užívateľovi sa pri filtrovaní udalostí následne ukazujú retroaktívne. Zmyslom udalostí je filtrovanie iba tých nahrávok, ktoré danú udalosť obsahujú. Tým sa enormne zefektívňuje celý proces analýzy a práce s dôležitými nahrávkami z celkového počtu nahraných sessions. Výskyt jednotlivých udalostí je možné zobrazovať v grafickom rozhraní. [23]



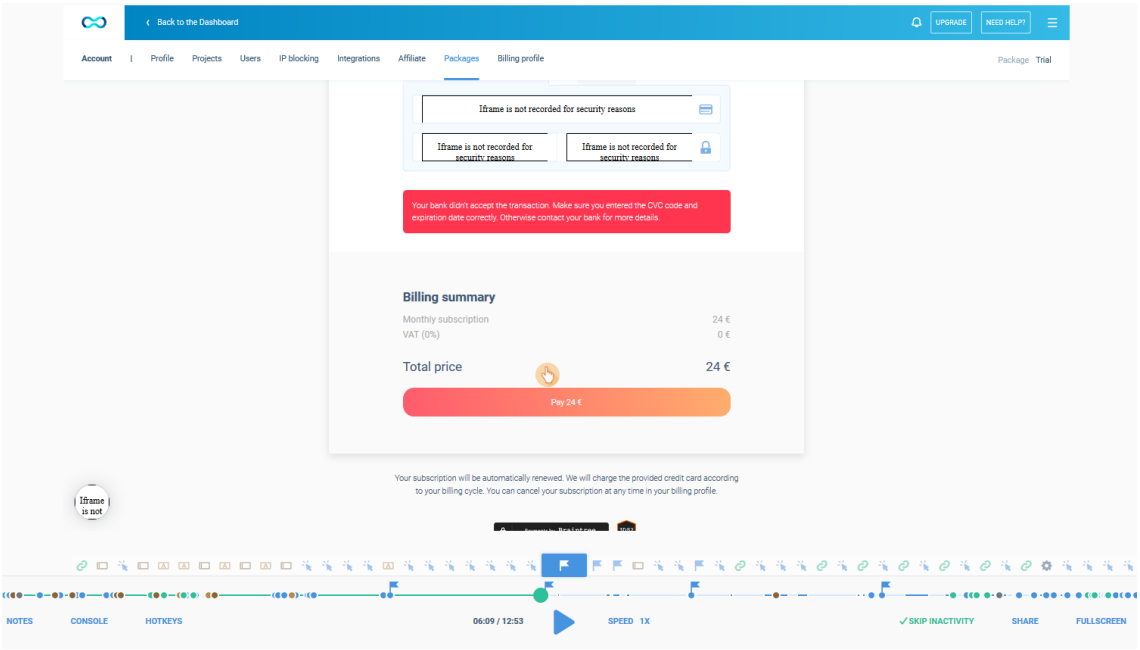
Obrázok 9: Eventy graf
(Zdroj: 9)

1.10.4 Funnely

Funnel je anglický názov pre lievik. Funkciou funnelu je zobraziť konverziu prechodu jednotlivými krokmi obsiahnutými vo funnely. Medzi jednotlivými krokmi platí podmienka **AND.** Na vytvorenie funnelu sa využívajú udalosti, ktoré bezpodmienečne nasledujú jedna za druhou. [23.]

Obrázok 10: Funnel
(Zdroj: 9)

Najlepším príkladom využitia funnelu je sledovanie prechodu zákazníka jednotlivými krokmi platobného procesu. Vďaka nahrávkam si vie užívateľ presne pozrieť dôvody toho, prečo návštevník nepokračoval na ďalší krok procesu. Na obrázku nižšie je vidieť prípad, kedy sa zákazníkovi kvôli chybe na platobnej bráne nepodarilo zaplatiť. Takáto informácia má veľkú výpovednú hodnotu a môže internetovému obchodníkovi zachrániť inak stratený zisk. Správne nastavenie funnela má potenciál odhaliť množstvo chýb, ktoré môžu vo výsledku spôsobiť veľké škody danému businessu. [23.]



Obrázok 11: Interface nahrávky
(Zdroj: 9)

2 ANALÝZA SÚČASNÉHO STAVU

V tejto časti práci detailne rozoberiem spoločnosť, v ktorej som zamestnaný, a v ktorej daný dátový sklad vytváram. Popísaná bude podnikateľská činnosť spoločnosti, objasnené pojmy spojené s názvom firmy a samotnými produktami v ponuke. Zanalyzované dve potenciálne riešenia ETL procesu a dve riešenia dátového úložiska, ktorú sú momentálne na trhu. Z technického hľadiska budú zanalyzované jednotlivé dátové zdroje, analytickú databázu, do ktorej sa následne dáta zo zdrojov ukladajú a finálny výstup informácií vo forme reportov.

2.1 Popis spoločnosti

Spoločnosť Smartsupp je technologickým startupom založeným v Brne v roku 2014. Pôvodne začala firma vývojom a sprostredkovaním rovnomenného živého chatu predovšetkým pre eshopy na domácom českom trhu. Postupom času sa ale podarilo preraziť na medzinárodný trh a firma začala stabilne rásť. V roku 2016 bola vyvinutá príplatkovou funkciou do živého chatu. Jednalo sa o doplnkovú službu založenú na nahrávaní jednotlivých sessions návštevníkov prichádzajúcich na stránku, na ktorej bol nasadený živý chat. Nahrávané sessions poskytovali lepší náhľad do aktivít, ktoré návštevník na stránke vykonával. V prípade, že sa návštevník stretol s komplikáciami, ktoré sa následne snažil vyriešiť s agentom na živom chate, mal agent k dispozícii nahrávku, na základe ktorej bol lepšie schopný porozumieť problému a vyriešiť ho efektívnejšie.

Obrázok 12: Logo Smartsupp
(Zdroj: 6)

Doplnková služba nahrávania návštevníckych sessions mala úspech a firma sa rozhodla vyvinúť samostatný produktom, ktorý bol následne nazvaný Smartlook. Od roku 2016 teda firma Smartsupp s.r.o. vyvíja dva rozdielne produkty.

- **Smartsupp –** živý chat pre weby a eshopy
- **Smartlook –** analytický nástroj nahrávajúci návštevy zákazníkov na webe

2.2 Vývoj Smartlooku

Smartlook dostal od roku 2016 vlastný vývojový tím a z pôvodného doplnku k živému chatu sa stal samostatný produkt. Od začiatku bol Smartlook spustený ako bezplatný nástroj na nahrávanie zákazníkov na webe. V tej dobe neobsahoval okrem nahrávania žiadne pokročilé analytické funkcie. Na základe požiadaviek zákazníkov bola ako prvá pridaná funkcia teplotných máp. Funkcii teplotných máp bude po zvyšok času na základe priemyselného štandardu referované ako heatmaps. Nová funkcia zvýšila potenciál produktu a získala väčší záujem užívateľov.

Obrázok 13: Logo Smartlook
(Zdroj: 6)

Firemný business model funguje na princípe software as a service, v skratke SaaS. Užívatelia mesačne platia za využívanie ponúkaného produktu s tým, že predplatné majú možnosť kedykoľvek zrušiť. V ponuke je možnosť zaplatiť si mesačné predplatné alebo si za zvýhodnených podmienok predplatiť ročné predplatné.

Keďže postupný vývoj Smartlooku vyžadoval väčší developerský tím a užívateľská základňa sa dostala na úroveň približne 50 000 aktívnych užívateľov, firma sa rozhodla zaviesť platené balíčky aj na tento produkt. Do tej doby boli k dispozícii iba predplatné pre Smartsupp. Zachovala sa možnosť bezplatného používania služby ale s príchodom platených balíčkov sa na bezplatné účty začali vzťahovať obmedzenia.

V priebehu niekoľkých mesiacov sa podarilo motivovať na prechod z bezplatného využívania služby približne 1500 platených užívateľov. Postupným vývojom produktu sa začala rysovať ideálna cieľová skupina zákazníkov a smerovanie produktu sa začalo odkláňať od pôvodného B2C trhu na B2B. To inicializovalo v roku 2018 vývoj dvoch

nových funkcií – udalostí (ang. Events) a lievikov (ang. Funnels). V tretom a štvrtok kvartáli roku 2018 sa k nahrávaniu webov pridal vývoj nahrávania mobilných aplikácií. Ku koncu roku pracovalo v Smartlook tíme približne 30 ľudí na plný úväzok.

V roku 2019 bol potvrdený produktový - market fit. Obrat firmy sa vyrovnal nákladom, pri čom mesačný obrat firmy činí približne 2,4 milióna korún a užívateľská základne je na úrovni 250 000 užívateľov.

2.3 ETL integračné služby

Na trhu je dostupné veľké množstvo riešení poskytujúcich služby prenosu dát z prevádzkových systémov do finálnych destinácií. V tejto kapitol zanalyzujem dve dostupné riešenia na trhu.

2.4 Stitch

Stitch je cloudové riešenie, založené na open source platforme pre rýchle presúvanie dát. Jednoduchá a výkonná ETL služba, prepája všetky dátové zdroje z databáz ako je napríklad MySQL alebo MongoDB, na SaaS aplikácie ako sú napríklad Salesforce a Zendesk. Tieto dáta sú následne replikované do destinácií podľa výberu užívateľa.

2.4.1 Destinácie

Pri replikácií dát Stitchom, sa tieto dáta ukladajú do destinácie alebo dátového skladov podľa vlastného výberu. V súčasnej dobe dovoľuje Stitch pripojiť iba jednu destináciu k jednému účtu. Preto je dôležité správne posúdiť všetky dostupné možnosti tak, aby sa predišlo nutnosti zmeny destinácie alebo re-replikácie všetkých dát.

Zoznam dostupných destinácii:

	Release status	Stitch availability	Fully managed?
Amazon S3	Released	All Stitch plans	No
Google BigQuery	Released	All Stitch plans	Yes
Google BigQuery	Released	All Stitch plans	Yes
data.world	Released	All Stitch plans	Yes
Microsoft Azure SQL Data Warehouse	Released	All Stitch plans	Yes
Panoply	Released	All Stitch plans	Yes
PostgreSQL	Released	All Stitch plans	Depends
Amazon Redshift	Released	All Stitch plans	No
Snowflake	Released	All Stitch plans	Yes

Obrázok 14: Stitch destinácie
(Zdroj: 10)

Pri voľbe destinácií je dôležité dávať si pozor na kompatibilitu s dátovými zdrojmi. Niektoré integrácie môžu byť čiastočne alebo plne nekompatibilné s destináciami, ktoré má Stitch k dispozícii. Problém s kompatibilitou môže nastať v podpore ukladania viacerých dátových typov v jednom stĺpci. Ak napríklad integrácia pošle stĺpce s rozdielnymi dátovými typmi, niektoré destinácie môžu tieto dáta odmietnuť replikovať.

Pri nahrávaní dát si treba dať pozor na to, ako sú dáta nahrávané do destinácie. Špecificky sa jedná o to, ako sa aktualizujú už existujúce stĺpce obsahujúce dáta. Stitch v tomto prípade podporuje dva typy správania pri nahrávaní dát. Viac o nahrávaní dát bude popísaných v neskoršej podkapitole.

2.4.2 Integrácie

Integrácia s dátovými zdrojmi je vykonávaná pomocou Singer, open source štandardu pre ETL, ktorý povoľuje replikáciu dát z akýkoľvek dátových zdrojov. Stinger integrácie, ku

ktorým sa tiež referuje aj ako kohútikom, umožňujú užívateľovi prenechanie orchestrácie, bezpečnosti a dostupnosti dátových kanálov na Stitch. Proces integrácie je nasledovný:

1. Vytvorenie Singer kohútiku.
2. Vytvorenie Import API integrácie.
3. Konfigurácia Singer kohútiku pre posielanie dát do Stitch destinácie.
4. Spustenie kohútiku a odoslanie dát do Stitchu.
5. Stitch dáta obdrží a zprocesuje.

2.4.3 Extrakcia dát

Extrakcia nastáva po tom, čo bola úspešne napojená integrácia a bola vykonaná štruktúrovaná synchronizácia. Po extrakcii dát z tabuliek a stĺpcov sa automaticky replikujú všetky dát. Ak to integrácia dovoľuje, tak sa užívateľ môže rozhodnúť, ktoré špecifické tabuľky a stĺpce sa replikujú. Pri replikácií dát je tiež treba vybrať metódu replikácie dát, ktorá bola popísaná v predchádzajúcej kapitole.

Replikačné kľúče sú stĺpce, ktoré Stitch používa pre identifikáciu nových a aktualizovaných dát pre replikáciu. Ak je replikačný kľúč nesprávne nastavený, môže to spôsobiť dátovú nezrovnalosť, zvýšenú latenciu, a vysoký počet riadkov. Preto je dôležité tieto kľúče nastaviť správne. Okrem replikačný kľúčov je dôležité dbať aj na primárne kľúče. Na rozdiel od replikačných kľúčov, ktoré sú používané počas extrakčnej fáze replikačného procesu, sú primárne kľúče používané počas posledného kroku replikačnej fáze na nahrávania dát do potrebnej destinácie. Primárne kľúče identifikujú unikátne riadky v tabuľke a zabezpečujú, že iba najaktuálnejšie dáta sú prenesené do destinácie. Replikovaná tabuľka musí obsahovať aspoň jeden z nasledujúcich stĺpcov a dátových typov.

Data type	Available for	Notes
DATETIME	All integrations	
INTEGER	All integrations	Includes `BIGINT` and `MEDIUMINT`
TIMESTAMP	All integrations	
FLOAT	MongoDB v1+ integrations	
INT64	MongoDB v1+ integrations	
NUMBER	Oracle v1+ integrations	
OBJECTID	MongoDB v1+ integrations	
UUID	MongoDB v1+ integrations	

Obrázok 15: Stitch dátové typy

(Zdroj: 11)

2.4.4 Replikačné metódy

Na replikáciu tabuliek, Stitch využíva jednu z troch replikačných metód.

- Log-based incremental method
- Key-based incremental method
- Full table replication

Log-based incremental method je replikačná metóda počas ktorej Stitch identifikuje modifikácie záznamov vrátane selectov, insertov a mazaní na základe DB binárnych záznamov. Táto metóda je k dispozícii iba pre Amazon DynamoDB, Microsoft SQL Server, MySQL, Oracle, a PostgreSQL databáze.

Key-based incremental method je replikačná metóda, ktorá identifikuje nové a aktualizované dáta pomocou stĺpca, ktorému sa referuje ako k replikačnému kľúču.

Full table replication je metóda, v ktorej sa replikujú všetky riadky tabuľky vrátane nových, aktualizovaných a existujúcich. Táto metóda sa využíva vtedy, ak tabuľka

neposkytuje stĺpec vhodný pre Key-based incremental alebo Log-based incremental nie je k dispozícii.

2.4.5 Podporované dátové typy

Stitch podporuje dve kategórie dátových typov pri replikácii. **Bežné** dátové typy, ktoré sú podporované všetkými integráciami a **špecifické** dátové typy naviazané na konkrétnu integráciu.

As of **June 19, 2018**, the data types in the table below are supported for all integrations.

BIGINT	BIT	CHAR	DATE
DECIMAL	DOUBLE	FLOAT	INTEGER
LONGVARCHAR	LONGNVARCHAR	NCHAR	NVARCHAR
NUMERIC	REAL	SMALLINT	TIME
TIMESTAMP	TINYINT	VARCHAR	

Obrázok 16: Stitch podporované dátové typy 2

(Zdroj: 11)

2.4.6 Nahrávanie dát

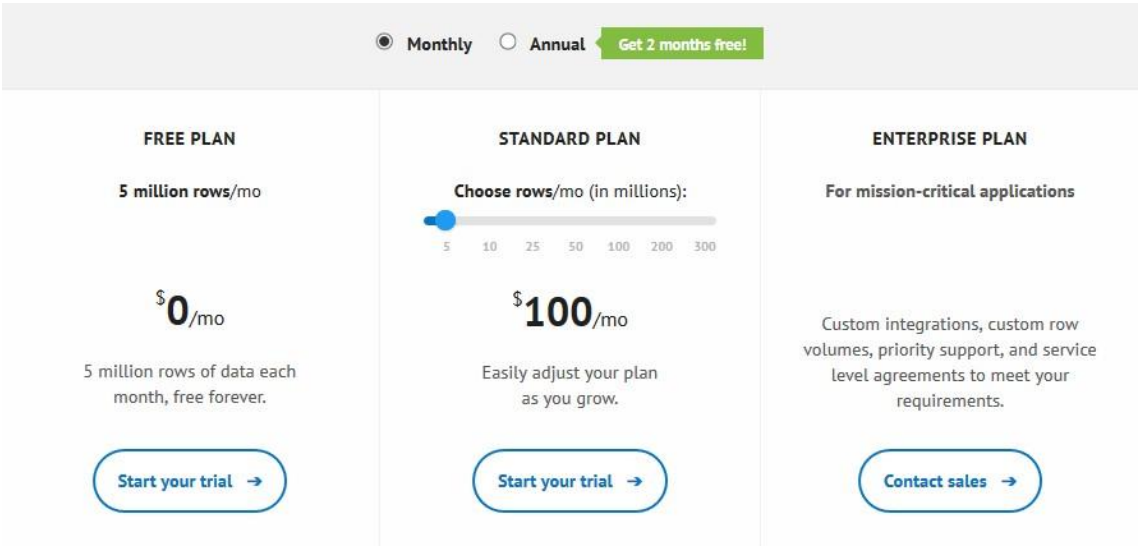
Pre nahrávanie dát do cieľovej destinácie, Stitch používa jeden z dvoch nahrávacích spôsobov:

- **Upsert –** keď sú dáta nahrávané pomocou Upsert, existujúce riadky sú aktualizované v tabuľkách s definovanými primárnymi kľúčmi. Stitch na základe primárnych kľúčov, deduplikuje záznamy predtým ako sú nahrané. To znamená že je vždy k dispozícii iba jedna verzia existujúcich dát.
- **Append only –** keď sú dáta nahrávané pomocou Append only, nové záznamy sú pripojené na konci tabuľky ako nové riadky. Existujúce riadku nie sú

aktualizované ani v prípade definície primárneho kľúča. To znamená, že môže existovať viacero verzií jedného riadku v tabuľke.

2.4.7 Cenník

K dispozícii sú plány založené na počte prenášaných riadkov z dátových zdrojov. Do 5 miliónov riadkov mesačne je Stitch k dispozícii zadarmo.



Obrázok 17: Stitch cenník

(Zdroj: 12)

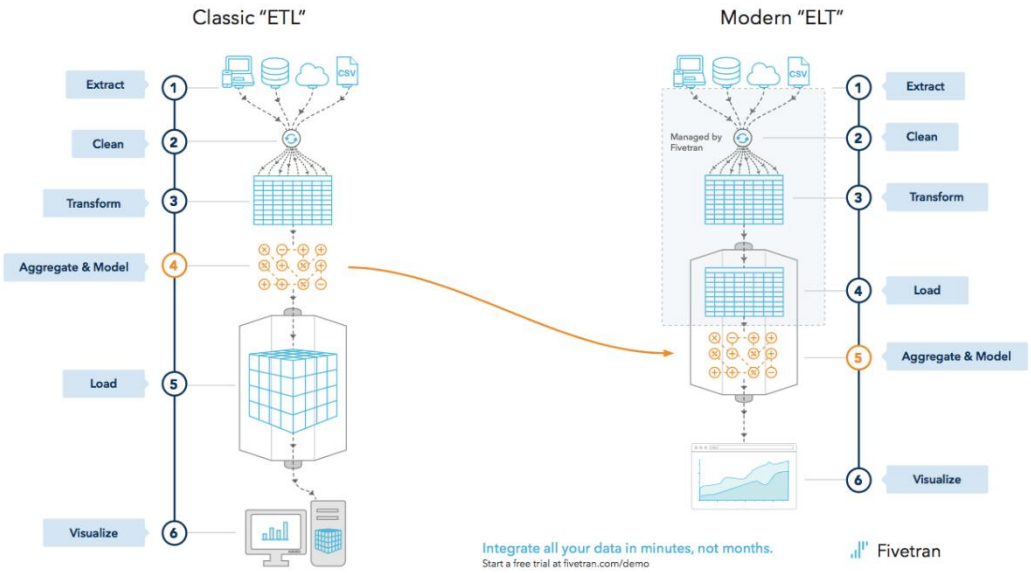
2.5 Fivetran

Fivetran je služba, zabezpečujúca centralizáciou dát z rôznych zdrojov, ktoré môžu byť spravované priamo z prehliadača. Extrahované dáta zo zdrojov sú následne nahrané do dátového skladu. Stavebným blokom dátovej organizácie sú tabuľky a schémy. Tabuľky sú prirodzene zložené zo stĺpcov a riadkov. Schémy sú záložky obsahujúce viaceré tabuľky. Každý Fivetran konektor vytvára a riadi svoju vlastnú schému.

2.5.1 ETL vs ELT

Drahé dátové úložiská a poddimenzované dátové sklady do nedávna znamenali, že prístup k dátam vyžadoval budovanie a udržovanie labilných ETL kanálov. Vo Fivetrane

sa k tomu rozhodli pristúpiť rozdielne. Lacné cloudové úložiská a výkonnejšia dátové infraštruktúra umožnila zmenu zaužívaného ETL procesu na modernejšiu dátovú architektúru ELT – extrahovanie a ukladanie surových dát do dátového skladu prebieha najskôr, transformácia je vykonávaná nasledovne. Poskytuje to zvýšenú všestrannosť a použiteľnosť.



Obrázok 18: Fivetran architektúra

(Zdroj: 13)

2.5.2 Dátové typy

Fivetran podporuje všetky tradičné dátové typy pre destinácie v ponuke. V prípade potreby dátového skladu alebo databázy, môžu byť niektoré netradičné dátové typy zo zdrojov transformované na akceptovaný formát.

DATA TYPE
BOOLEAN
SMALLINT
INTEGER
BIGINT
REAL
DOUBLE
DECIMAL
DATE
TIMESTAMP
TEXT
JSON

Obrázok 19: Fivetran dátové typy

(Zdroj: 14)

2.5.3 Konektor

Jeden účet je tvorený niekoľkými konektormi, ktoré ťahajú dáta z viacerých dátových zdrojov do jednej alebo viacerých destinácií. Tieto konektory sa delia na dve fundamentálne kategórie: pull a push.

Pull konektor aktívne načítava alebo ťahá dáta zo zdroja. Tieto dáta sa zo zdroja získavajú v pravidelných frekvenciách. Používané je SSL - zašifrované pripojenie na zdrojový systém. Na ťahanie dát z databáz sa používa predovšetkým ODBC/JDBC metóda. Na ťahanie dát z webových služieb sa používa API založené na REST a SOAP.

Push konektor prijíma dáta z dátových zdrojov, ktoré sú mu zaslané alebo posunuté. Tieto dáta zo zdrojových súborov sú posielané ako eventy. Po obdržaní týchto eventov ich Fivetran ukladá ako JSON súbory.

2.5.4 Príprava a nahrávanie dát

Po tom čo sú dáta prijaté, sú normalizované, čistené, triedené a odstránené duplicity. Počas procesu spracovávania môžu byť záznamy dočasne buffernuté na disk a zašifrované za použitia tajného kľúča.

Finalizované záznamy sú nahrané do „file storage bucket“, kde sú znova zašifrované pomocou špeciálneho kľúča, známeho iba konkrétnemu procesu vykonávajúcemu zápis. Dočasný súbor je do 24 hodín vymazaný.

Dáta sú následne nakopírované do stagingovej schémy v destinácií dátového skladu. Kľúč použitý na zašifrovanie dátového prenosu je odoslaný do dátového skladu tak aby dáta mohli byť následne rozšifrované. Tento proces sa opakuje podľa naplánovaného rozvrhu.

Zoznam podporovaných databáz:

- DynamoDB
- MariaDB
- MongoDB
- MySql
- Oracle
- PostgreSQL
- SQL Server

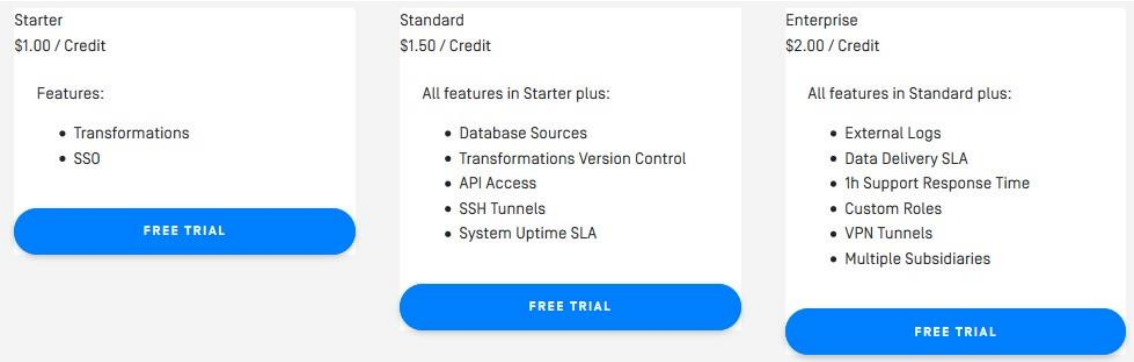
Zoznam podporovaných destinácií:

- Azure Data Explorer
- Azure Synapse
- BigQuery
- Databricks
- MySQL
- PostgreSQL

- Redshift
- Snowflake
- SQL Server

2.5.5 Cenník

Fivetran má k dispozícii 14 dňový trial účet, po ktorom treba kontaktovať obchodné oddelenie spoločnosti a dosť kvótu na základe špecifických požiadaviek. Na rozdiel od Stitchu nie je k dispozícii Free plan.



Obrázok 20: Fivetran cenník

(Zdroj: 15)

2.6 Dátové sklady

V tejto podkapitole budú zanalyzované a popísané dve riešenia dátových trhov, ktoré sú momentálne na trhu a majú potenciál splnenia požiadaviek pre vybudovania dátového skladu v Smartlooku. Prvé riešenie je od spoločnosti Google a druhé od spoločnosti Amazon.

2.7 BigQuery

BigQuery data je vysoko škálovateľné, bez serverové a nákladovo efektívne cloudové riešenie od spoločnosti Google. Prístup do BigQuery je k dispozícii cez Cloudovú konzolu, klasické webové rozhranie, pomocou príkazového riadku alebo pomocou volaní cez REST API za použitia rôznych knižníc(Java, .Net, Python) Pre vizualizáciu dát je k dispozícii množstvo nástrojov tretích strán.

BigQuery jobs sú akcie vykonávaná samotným dátovým skladom pre načítavanie, exportovanie, dotazovanie a kopírovanie dát. Pri použití cloudovej konzole, webového rozhrania alebo CLI na jednu z vyššie spomenutých činností, je BigQuery job source automaticky vytvorený, naplánovaný a spustený. Z dôvodu potenciálneho veľkého časového rozsahu procesu sú BigQuery jobs vykonávané asynchronne.

Datasety sú kontajnery na najvyššej úrovni používané na organizáciu a kontrolu tabuliek a pohľadov. Tieto datasety sú obsiahnuté v špecifickom projekte. Keďže tabuľky a pohľady musia patriť do konkrétneho datasetup, musí byť vytvorený minimálne jeden pred nahraním dát do BigQuery.

Schémy tabuliek je možné špecifikovať pri načítaní dát do tabuliek alebo pri vytváraní prázdnych tabuliek. Alternatívne je možné využiť autodetekciu schém pre podporované dátové formáty. Každá tabuľka má na základe schémy definované názvy stĺpcov, dátové typy a ostatné informácie. Podporované sú natívne tabuľky, externé tabuľky a pohľady.

Podporované dátové typy:

Name	Data type	Description
Integer	INT64	Numeric values without fractional components
Floating point	FLOAT64	Approximate numeric values with fractional components
Numeric	NUMERIC	Exact numeric values with fractional components
Boolean	BOOL	TRUE or FALSE (case insensitive)
String	STRING	Variable-length character (Unicode) data
Bytes	BYTES	Variable-length binary data
Date	DATE	A logical calendar date
Date/Time	DATETIME	A year, month, day, hour, minute, second, and subsecond
Time	TIME	A time, independent of a specific date
Timestamp	TIMESTAMP	An absolute point in time, with microsecond precision
Struct (Record)	STRUCT	Container of ordered fields each with a type (required) and field name (optional)
Geography	GEOGRAPHY	A pointset on the Earth's surface (a set of points, lines and polygons on the WGS84 reference spheroid, with geodesic edges)

Obrázok 21: BigQuery podporované dátové typy

(Zdroj: 16)

Pre dotazovanie nad dátami je treba do dátového skladu najskôr dáta nahrať. Dáta môžu byť do dátového skladu nahrané z cloudového úložiska, Google služieb ako sú napríklad Google Ads, streamovaním jednotlivých záznamov, používaním DML(data manipulation language) výrazov vykonávajúcich bulk vloženie alebo za použitia BigQuery I/O konektoru v Dataflow kanále. Nahrávanie priamo z Google Drivu momentálne nie je k dispozícii.

Dáta môžu byť nahrávané v rôznych formátoch. Tieto dáta sú následne konvertované do špeciálneho BigQuery columnar formátu. Ak sú dáta zašifrované, je podporované UTF-8 šifrovanie pre vnorené alebo opakované zápisy dát. Ak sa nahrávajú dáta z .CSV súboru, je podporované ISO-8859-1 šifrovanie. Pre nahrávanie komprimovaných a nekomprimovaných dát je preferovaný Avro binárny formát.

2.7.1 Cenník

Cenník je nastavený tak, aby bol škálovateľný a flexibilný s ohľadom na technické požiadavky a rozpočet. Fakturácia môže byť na princípe mesačných poplatkov za uložené dáta v posledných 90 dňoch alebo dlhodobá fakturácia pre uložené dáta, ktoré neboli modifikované viac než 90 dní. Dotazovanie môže byť fakturované na základe rozsahu jednotlivých dotazov alebo na základe dopredu dohodnutej sadzby.

Tabuľka 2: Google BigQuery cenník
(Zdroj: 17)

Operácie	Cenník	Detail
Active storage	$0.020 per GB	The first 10 GB is free each month.
Long-term storage	$0.010 per GB	The first 10 GB is free each month.
BigQuery Storage API	$1.10 per TB	The BigQuery Storage API is not included in the free tier.

Súčasťou Google Cloud Free tier sú služby, ktoré poskytujú určitý limit zdrojov zadarmo.

Tabuľka 3: Google BigQuery cenník 2

(Zdroj: 17)

Zdroj	Mesačný limit zadarmo
Storage	The first 10 GB per month is free.
Queries (analysis)	The first 1 TB of query data processed per month is free.
BigQuery ML CREATE MODEL queries	The first 10 GB of data processed by queries that contain CREATE MODEL statements per month is free.

2.8 Redshift

Redshift je dátový sklad od spoločnosti Amazon patriaci do produktovej rady AWS. Základom dátového skladu Redshiftu je PostgreSQL. Tento open sourcový operačno-relačný databázový systém povoľuje prácu pomocou rozšíreného SQL dotazovacieho jazyku. Celý dátový sklad je umiestnený v cloude a dovoľuje obrovskú škálovateľnosť od desiatok gigabytov až po petabyty a vyššie.

Redshift dátový sklad je kolekciou výpočtových zdrojov nazývaných uzly. Tieto uzly sú organizované do skupín nazývaných klastre. Každý jeden klaster má priradený hlavný uzol a viacero výpočtových uzlov. Hlavný uzol prijíma dotazy od klientskej aplikácie, analyzuje dotazy a vytvára exekučné plány. Exekučný plán je následne rozdelený medzi výpočtové uzly. Výsledky sú následne agregované hlavným uzlom a odoslané do klientskej aplikácie.

Pri spúšťaní klastru je nutné určiť typy uzlov. Jednotlivé uzly majú stanovú CPU, RAM, kapacitu úložiska a úložný disk. Typy uzlov sú:

- **RA3 –** pri tomto type uzla je k dispozícii škálovateľnosť úložiska a výkonu podľa potreby používateľa.

Tabuľka 4: Amazon Redshift RA3 cluster

(Zdroj: 18)

Node Size	vCPU	RAM (GiB)	Slices Per Node	Managed Storage Per Node	Node Range	Total Capacity
ra3.16xlarge	48	384	16	64 TB	2–128	8.192 PB

- **DS2 –** tieto uzly sú optimalizované pre veľkú dátovú záťaž a používajú HDD úložiská.

Tabuľka 5: Amazon Redshift DS2 cluster

(Zdroj: 18)

Node Size	vCPU	RAM (GiB)	Slices Per Node	Managed Storage Per Node	Node Range	Total Capacity
ds2.xlarge	4	13	2	2 TB HDD	1–32	64 TB
ds2.8xlarge	36	244	16	16 TB HDD	2–128	2 PB

- **DC2 –** tieto uzly sú optimalizované pre pracovné zaťaženie náročné na výkon. Keďže používajú SSD úložiská, poskytujú rýchlejší zápis a výpis dát ako DS uzly.

Tabuľka 6: Amazon Redshift DC2 cluster

(Zdroj: 18)

Node Size	vCPU	RAM (GiB)	Slices Per Node	Managed Storage Per Node	Node Range	Total Capacity
dc1.large	2	**15**	**2**	160 GB SSD	1–32	5.12 TB
dc1.8xlarge	32	244	32	2.56 TB SSD	2–128	326 TB
dc2.large	2	15.25	2	160 GB NVMe-SSD	1–32	5.12 TB
dc2.8xlarge	32	224	16	2.56 TB NVMe-SSD	2–128	326 TB

Hlavný uzol riadi všetku komunikáciu s klientskymi programami a výpočtovými uzlami. Analyzuje a vytvára exekučné plány na vykonanie databázových operácií, predovšetkým sériu krokov potrebných k dosiahnutiu výsledku z komplexných dotazov. Na základe exekučného plánu hlavný uzol skomponuje kód, rozdistribuuje skomponovaný plán medzi výpočtové uzly a priradí rozsah dát, ktorý bude spracovávaný jednotlivými uzlami. SQL príkazy sú vždy spracovávané na hlavnom uzle okrem prípadu, kedy sa príkaz dotazuje na tabuľky uložené na výpočtovom uzle.

Klaster môže obsahovať jednu alebo viacero databáz. Dáta sú uložené na výpočtových uzloch. SQL klient komunikuje s hlavným uzlom, ktorý koordinuje vykonanie dotazu s výpočtovými uzlami. Amazon Redshift je relačný databázový správcovský systém (RDBMS) takže umožňuje kompatibilitu s ostatnými RDBMS aplikáciami. Hoci poskytuje rovnakú funkcionalitu, ako typické RDBMS vrátane OLTP funkcii ako sú vkladanie a mazanie dát, je optimalizovaný predovšetkým pre analýzy vyžadujúce veľký výkon a reportovanie nad veľkými dátovými sadami.

Aj keď je Amazon Redshift založený na PostgreSQL 8.0.2, obsahuje niektoré zásadné rozdiely, na ktoré treba pri budovaní dátového skladu myslieť. Bol navrhnutý pre OLAP a BI aplikácie, ktoré vykonávajú komplexné dotazovanie nad veľkými dátovými sadami. Keďže adresuje veľmi rozdielne požiadavky, špecializované dátové schémy a engine na vykonávanie dotazov, ktorý Amazon Redshift používa, sú kompletne odlišné od klasickej PostgreSQL implementácie.

2.8.1 Cenník

Amazon poskytuje tri možnosti financovania dátového skladu.

- **On-demand cenník –** typ financovanie, pri ktorom sa platí za reálne využívané zdroje na hodinovej báze
- **Spectrum cenník –** pri tomto druhu financovania sa platí za spúšťanie SQL dotazov. Účtuje sa na základe naskenovaných bytov, zaokrúhlených na megabyty, s minimálnou veľkosťou 10MB za dotaz. Za príkazy manipulujúce tabuľky CREATE/ALTER/DROP TABLE, oddiely a chybné dotazy sa neplatí.

- **Rezervované inštancie –** typ financovania určený pre ustálené pracovné zaťaženie. Oproti on-demand poskytuje výraznú zľavu ak je využité v dlhodobom horizonte 1 až 3 rokov. Tento typ financovania zahrňuje uskladnenie dát na klastroch uzlov a zálohu v Amazon S3. Platba je k dispozícii na mesačnej báze, zaplatenie časti sumy dopredu a doplatenie zvyšku sumy v priebehu využívania služby alebo zaplatenie celej sumy dopredu.

Tabuľka 7: Amazon Redshift cenník

(Zdroj: 19)

	Dĺžka záväzku	dc2.large	dc2.8xlarge	ds2.xlarge	ds2.8xlarge
On-demand	-	$0.324/h	$6.048/h	$1.026/h	$8.208/h
Spectrum	-	$5.00/TB of data			
No upfront payment	1 rok	$0.260	$4.830	$0.820	$6.560
	3 roky	-	-	-	-
Partialy upfront payment	1 rok	$0.213	$4.180	$0.678	$5.420
	3 roky	$0.129	$2.070	$0.306	$2.440
All upfront payment	1 rok	$0.208	$4.100	$0.664	$5.310
	3 roky	$0.121	$1.940	$0.286	$2.280

2.9 Dátová infraštruktúra

Firma Smartlook spravuje a pracuje s obrovským množstvom dát. Ako mladá a vznikajúca firma si nemohla dovoliť stavbu a prevádzkovanie vlastnej dátovej infraštruktúry v takom rozsahu, aby stíhala držať krok s nárastom dát, ktoré spravuje

a využívajú tisícky zákazníkov každý deň. Okrem obrovskej kapacity je dôležitá aj bezpečnosť dát a ich samotná dostupnosť.

Smartlook spravuje všetky svoje dáta a prevádzkuje virtuálne serveri pod záštitou Amazon Web Services. Toto riešenie umožňuje Smartlooku platiť iba za toľko výkonu a úložiska, koľko reálne využíva. Firma je tiež schopná pružnejšie sa prispôsobovať nárastu zákazníkov a množstvu ich dát.

Obrázok 22: Amazon web services logo

(Zdroj: 19)

Keďže dáta od zákazníkov majú charakter citlivých informácii a obsahujú častokrát osobné údaje, bezpečnosť týchto dát je nevyhnutnosť. Dáta sú preto zašifrované za použitia algoritmu 256-bit Advanced Encryption Standard(AES-256). K ochrane všetkých dát je použité SSL/TLS šifrovanie. Amazon Web Services úložiská sú certifikované podľa normy ISO 27001.

Obrázok 23: Certifikácie

(Zdroj: 9)

2.10 Zdroje dát

Mimo zbierania a spracovávania dát pre zákazníkov samotným Smartlookom softwarom, zbierame a pracujeme s viacerými službami poskytujúcimi dáta pre marketing, vrcholový management, sales a support.

2.10.1 Udalosti

Všetky udalosti, ktoré užívatelia vytvoria sú zaznamenané ako definície. Tieto definície eventov sa ad hoc nahrávajú do analytickej databázy. Smartlook je nasadený aj na našich vlastných stránkach. Nahrávky využíva na svoju prácu viacero tímov, preto má každý tím vytvorené svoje vlastne udalosti filtrujúce nahrávky a dáta.

Veľké využitie majú vlastné udalosti sledujúce najmä využívanie rôznych tlačidiel v užívateľskom rozhraní, prechody platobným procesom, využívanie upgrade a downgrade možností. Tieto dáta sa automaticky nahrávajú do analytickej databázy z Amazon ES databázy do tabuľky SRC_Event.

2.10.2 Google adwords

Google adwords je služba od spoločnosti Google pracujúca na princípe reklamného systému. Inzeranti prihadzujú sumy za kľúčové slová, ktoré používatelia vyhľadávajú pomocou Google vyhľadávača. Najvyššie umiestnenie pri výslednom vyhľadávaní získava inzerant s najkvalitnejšie nastavenou reklamou a čo najvyššou sumou ponúknutou za jeden preklik.

V Smartlooku je rozbehnutých niekoľko CPC kampaní, ktoré mesačne generujú veľké množstvo dát najmä pre marketingové oddelenie. Tieto dáta sú následne vyhodnocované a vypovedajú o efektivite zvolených marketingových stratégií v jednotlivých regiónoch sveta. Zobrazujú náklady za príchod potenciálnych zákazníkov a trhy, na ktorých je Smartlook silnejší ako konkurencia.

Obrázok 24: Logo Google Ads

(Zdroj: 20)

Tieto dáta sa v súčasnej dobe exportujú priamo zo služby Google adwords na konci mesiaca vo forme .CSV exportu. Kampane sú rozdelené do 4 skupín podľa regiónov, preto sa vykonáva export dát zo všetkých 4 zdrojov. Následne sú dáta manuálne ukladané do analytickej databázy pomocou dotazu:

```
LOAD DATA LOCAL INFILE
'C:\\Path of the document\\Smartlook\\Data processing\\Google ads
reporty\\1_2020\\MonthlyExport_West.csv'
INTO TABLE analytics.SRC_Adwords
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
IGNORE 3 LINES;

LOAD DATA LOCAL INFILE
'C:\\Path of the document\\Smartlook\\Data processing\\Google ads
reporty\\1_2020\\MonthlyExport_Central.csv'
INTO TABLE analytics.SRC_Adwords
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
IGNORE 3 LINES;

LOAD DATA LOCAL INFILE
'C:\\Path of the document\\Smartlook\\Data processing\\Google ads
reporty\\1_2020\\MonthlyExport_East.csv'
INTO TABLE analytics.SRC_Adwords
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
IGNORE 3 LINES;

LOAD DATA LOCAL INFILE
'C:\\Path of the document\\Smartlook\\Data processing\\Google ads
reporty\\1_2020\\MonthlyExport_South.csv'
INTO TABLE analytics.SRC_Adwords
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
IGNORE 3 LINES;

LOAD DATA LOCAL INFILE
'C:\\Path of the document\\Smartlook\\Data processing\\Google ads
reporty\\1_2020\\MonthlyExport_Display.csv'
INTO TABLE analytics.SRC_Adwords
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
IGNORE 3 LINES;
```

2.10.3 Google analytics

Google analytics je webová analytická služba ponúkaná spoločnosťou Google. Sleduje a reportuje traffic na webovej stránke, na ktorej je nasadený. Na danej stránke následne vypracováva komplexné analýzy návštevníkov stránky. Medzi základné informácie, ktoré sú z GA dostupné, patrí denný počet všetkých návštevníkov rozpracovaný na vracajúcich sa a nových návštevníkov, priemernú dĺžku návštevy na stránke, bounce rate a priemerný počet stránok prezretých počas jednej návštevy. Medzi tie komplexnejšie štatistiky patria demografické dáta, behaviorálne dáta, geografické dáta alebo technologické dáta.

Obrázok 25: Google Analytics

(Zdroj: 20)

V súčasnej dobe sú z GA exportované dáta zobrazujúce pôvod návštevníka, zdroj návštevy, kampaň cez ktorú prišiel a v akých počtoch návštevníci prichádzajú na stránky Smartlooku v časových intervaloch. Tieto dáta sa rovnako ako pri Google adwords manuálne exportujú na konci mesiaca a manuálne nahrávajú do analytickej databázy pomocou dotazu:

```
LOAD DATA LOCAL INFILE
'C:\\Path of the document\\Smartlook\\Data processing\\Google
Analytics\\GA_Visitors_20191201.csv'
INTO TABLE analytics.SRC_GA_Visitors
FIELDS TERMINATED BY ';'
LINES TERMINATED BY '\n'
IGNORE 15 LINES;
```

2.10.4 Intercom

Intercom je služba poskytujúca online živý chat slúžiaci najmä pre účely support departmentu a čiastočne pre sales a product team. Ako multifunkčný nástroj umožňuje používanie botov na získavanie leadov, cielené oslovovanie zákazníkov používajúcich aplikáciu v reálnom čase a v neposlednom rade poskytuje živú podporu pre zákazníkov spolu s rozsiahlou help section priamo dostupnou z chatu. V Smartlooku sú zbierané dáta najmä o počte živých chatov za mesiac, kategórie chatových diskusií, výsledkoch cielených správ, efektivitu onboardingu a návštevnosť jednotlivých sekcií podpory.

Obrázok 26: Logo Intercom

(Zdroj: 21)

Tieto dáta sa exportujú v .CSV formáte do Google Sheet dokumentu, z ktorého sa následne nahrávajú do databázy podľa potreby marketingového oddelenia za použitia dotazu:

```
LOAD DATA LOCAL INFILE
'C:\\Path of the document\\Smartlook\\Data processing\\Google
Analytics\\IntercomExport_20191201.csv'
INTO TABLE analytics.SRC_IntercomExport
FIELDS TERMINATED BY ';'
LINES TERMINATED BY '\n'
IGNORE 15 LINES;
```

2.10.5 Marketingové náklady

Všetky marketingové výdaje sú zaznamenávané na mesačnej báze v Google dokumente marketingovým manažérom. Zaznamenávané sú výdaje na:

- **Online marketingové kampane** – AdWords, Facebook, Linkedin, Quora Ads
- **Akcie a eventy** – organizácia SaaS movementu v Brne, účasť na eventoch v Poľsku, Maďarsku, v Írsku a v USA kde Smartlook vystavoval

Nové dáta v Google dokumentu sú následne exportované vo formáte CSV. Dáta z exportovaného súbor sú potom nahraté do analytickej databázy pomocou dotazu:

```
TRUNCATE TABLE analytics.SRC_MarketingCostsMonthly;
LOAD DATA LOCAL INFILE
'C:\\Path of the document\\Smartlook\\Data processing\\Google
Analytics\\MarketingCosts.csv'
FIELDS TERMINATED BY ';'
LINES TERMINATED BY '\n'
IGNORE 1 LINES;
```

2.10.6 Chartmogul

Chartmogul je služba poskytujúca detailný pohľad na kľúčové finančné ukazovatele dôležité najmä pre SaaS firmy. V Smartlooku sa pomocou Chartmogulu zbierajú dáta o trende a tempe rastu obratu, priemernej celoživotnej hodnote zákazníka a hodnote churnu. Churn je pre firmy v oblasti SaaS priemysle kľúčový ukazovateľ vypovedajúci o priemernej mesačnej strate zákazníkov. Tieto dáta sa následne na mesačnej báze rovnako ako dáta z Google Adwords a Google Analytics manuálne ukladajú do databázy pomocou dotazu:

```
INSERT INTO analytics.SRC_ChartMogul_LTV (YearMonth,
CustomerLifetimeValue, AverageRevenuePerAccount, CustomerChurnRate)
VALUES (201911, 793.00, 67.17, 8.47);
```

2.11 Zhodnotenie analýzy

Z vykonanej analýzy je jasné že momentálne používané riešenie je funkčné a postačuje, ale z dlhodobého hľadiska je potrebná zmena. Väčšina dátových zdrojov používaných predovšetkým marketingovým oddelením sa do databázy nahráva pomocou manuálnych procesov na pravidelnej báze. Nie sú nastavené automatické dátové pumpy. Niekoľko dátových zdrojov, ako je napríklad pipedrive využívaný sales tímom, nie sú absolútne napojené na databázu a preto sa dáta z tejto platformy nedajú analyzovať. Pre efektívnejšie využívanie zbieraných dát je nevyhnutná potreba zlepšenia celého procesu práce s dátami. Vybudovanie dátového skladu v tejto situácii je dobrá časová investícia.

3 NÁVRHOVÁ ČASŤ

V tejto časti diplomovej práce bude popísaný postup budovania dátového skladu.

3.1 Plán projektu

Fyzickému budovaniu dátového skladu predchádza naplánovanie celého projektu. Tento plán zahŕňa definíciu cieľa a požiadaviek budovania dátového skladu. Tieto požiadavky musia byť definované vedením spoločnosti a budúcimi používateľmi dátového skladu. Na základe definovaných požiadaviek sa vypracuje plán budovania dátového skladu tak, aby finálne výstupy spĺňovali zadefinované ciele projektu.

3.1.1 Cieľ projektu

Cieľ projektu je vypracovaný na základe požiadaviek predovšetkým koncových užívateľov a odsúhlasený vedením spoločnosti. Hlavným cieľom je vylepšené vykazovanie a reportovanie dát pre účely jednotlivých oddelení spoločnosti s ohľadom na škálovateľnosť riešenia, s potenciálom rozšírenia vybudovaného riešenia v budúcnosti. Dátový sklad a reportovanie z neho sa v tejto fáze bude vypracovávať najmä pre dáta z marketingových zdrojov.

3.1.2 Požiadavky projektu

Keďže dátový sklad sa buduje pre použitie koncových užívateľov, je dôležité zistiť od koncových užívateľov, aká je ich predstava o výstupe. V rozsahu, v ktorom je dátový sklad budovaný, bude slúžiť okrem managementu spoločnosti a jednotlivých oddelení aj zamestnancom spoločnosti. Po tom čo bol stanovený cieľ celého projektu, bol tento cieľ s koncovými užívateľmi prekonzultovaný detailnejšie tak, aby výsledok splňoval požiadavky koncových užívateľov.

3.1.3 Sekvencia úloh

Po zistení cieľa a jednotlivých požiadaviek projektu nasleduje plán budovania a nasadenia dátového skladu. Najskôr je potrebné zanalyzovať zdrojové systémy, ktoré vykazujú dáta použité v dátovom sklade. Následne je treba pripraviť dáta už v existujúcej databáze na export do nového dátového skladu. Po analýze dostupných riešení bolo rozhodnuté, že ako ETL služba sa použije software od spoločnosti Stitch a samotný dátový sklad bude vybudovaný v platforem Google BigQuery. Tieto systémy treba pripraviť a nasadiť. Stitch je treba napojiť na zdrojové systémy a následne na Google BigQuery. Ako posledný krok bude nasledovať napojenie dátového skladu na Google Data Studio a vytvorenie reportu.

3.1.4 Rozpočet

Dátový sklad sa v tejto fáze buduje pre obmedzený počet dát. Databázy Smartlooku, obsahujú obrovské množstvá dát a migrácia celého systému by bola nákladná a zdĺhavá. Dátový sklad sa v tejto podobe buduje najmä kvôli zlepšeniu reportovania pre marketingové oddelenie nad dátami z marketingových produkčných systémov. V rozsahu, v ktorom je dátový sklad budovaný stačia riešenia od Stitch aj Google BigQuery, ktoré sú k dispozícii bezplatne. Celý dátový sklad sa teda buduje ako pilotný projekt s rozpočtom pokrývajúcim predovšetkým prácu povereného zamestnanca.

3.2 Plán budovania dátového skladu

Pre dosiahnutie stanoveného cieľa a splnenie požiadaviek bude budovanie dátového skladu rozdelené do niekoľkých fáz. Každá fáza má svoju časovú náročnosť, preto je dôležité sledovať úspešné vykonanie jednotlivých fáz nasadzovania. Celý proces budovania dátového skladu bol rozdelený na tieto fázy:

- Analýza zdrojov.
- Návrh dátového modelu.
- Príprava zdrojovej databázy.
- Vytvorenie dátového modelu v Google BigQuery.
- Export dát zo zdrojovej databázy a import do Google BiqQuery.

- Napojenie Stitch ETL na zdroje a na destinácie v Google BigQuery.
- Napojenie Google BigQuery na Google Data Studio.

3.3 Zdrojové systémy

Základná analýzu zdrojových systémov bola vypracovaná už v analytickej časti. V návrhovej časti budú nasledovné podkapitoly sústredené na popísanie štruktúry dát a dátových exportov zo zdrojových systémov.

3.3.1 Google Adwords

Marketingové oddelenie má ročne pridelený rozpočet na rôzne marketingové kampane. Jednu z najväčších položiek tvoria PPC kampane spustené na platforme Google. Inak tomu nie je ani v Smartlooku. Na platforme Google Adwords je rozbehnutých viacero PPC kampaní, ktorých výkon sa sleduje na mesačnej báze. Sledujú sa konkrétne parametre ako je krajina, na ktorú je reklama mierená, prednastavená skupina reklamy, špecifická kampaň v danej skupine, náklady na konkrétnu kampaň a mena, v ktorej sú náklady uvedené.

Reporty sú vždy na začiatku nového mesiaca vyhotovované za predchádzajúci mesiac a sú k dispozícii v samotnej platforme Google Adwords. Proces nahrávania dát do databázy prebiehal manuálnym exportovaním dát z reportov v Google Adwords a následným manuálnym importovaním do zdrojovej databáze. Výsledný mesačný report obsahuje nasledovné dáta.

#	Názov	Dátový typ
1	Date	DATE
2	Country	TEXT
3	Adgroup	TEXT
4	Campaign	TEXT
5	Currency	TEXT
6	Cost	DECIMAL

Obrázok 27: Google Adwords atribúty tabuľky

(Zdroj: Vlastný návrh)

3.3.2 Google Analytics

Pomocou Google Analytics v Smartlooku sú sledované počty ľudí, ktorý sa dostanú na stránky Smartlooku. Mimo základnej informácii o tom, koľko ľudí prišlo na stránky GA poskytuje aj informácie o tom, z akých zemí títo ľudia prichádzajú, aké zariadenia používajú alebo priemerný čas, ktorý strávili na našej stránke. Tieto dáta v kombinácii s Google Ads pomáhajú cieliť reklamu na určité svetové regióny, kategórie používateľov alebo firiem.

Sledujú sa krajiny, z ktorých návštevníci prichádzajú, zdroj alebo médium návštevy, skupina Adwords kampane, konkrétna kampaň v skupine, celkový počet návštevníkov a počet nových návštevníkov. Export týchto dát prebieha rovnako ako pri Google Ads na mesačnej báze a tiež bol doteraz vykonávaný manuálnym ukladaným dát do databázy. Výsledný mesačný report obsahuje nasledovné dáta.

#	Názov	Dátový typ
1	Date	DATE
2	Country ISO Code	TEXT
3	Source / Medium	TEXT
4	AdWords Ad Group	TEXT
5	Campaign	TEXT
6	Users	INT
7	New Users	INT

Obrázok 28: Google Analytics atribúty tabuľky

(Zdroj: Vlastný návrh)

3.3.3 Intercom

Intercom je nástroj používaný predovšetkým oddelením technickej podpory, ale využíva sa aj na marketingové účely. Dáta z tohto nástroja po spracovaní poskytujú informácie o strednej hodnote prvej odpovede, úplnom počte chatov, najčastejších problémoch. Reporty tiež obsahujú informácie o zákazníkoch kontaktujúcich oddelenie podpory. Marketing pomocou Intercomu onboarduje nových zákazníkov, oznamuje novinky v Smartlooku alebo oznamuje prebiehajúce akcie.

3.3.4 Chartmogul

Je nástroj určený na sledovanie ekonomických parametrov so zameraním na spoločnosti pôsobiace v SaaS priemysle. V Smartlooku je využívaný na sledovanie vývoja troch metrík.

- Customer lifetime value
- Average revenue per account
- Customer churn rate

Výsledný mesačný report obsahuje nasledovné dáta.

#	Názov	Dátový typ
1	YearMonth	INT
2	CustomerLifetimeValue	DECIMAL
3	AverageRevenuePerAccount	DECIMAL
4	CustomerChurnRate	DECIMAL

Obrázok 29: Chartmogul atribúty tabuľky

(Zdroj: Vlastný návrh)

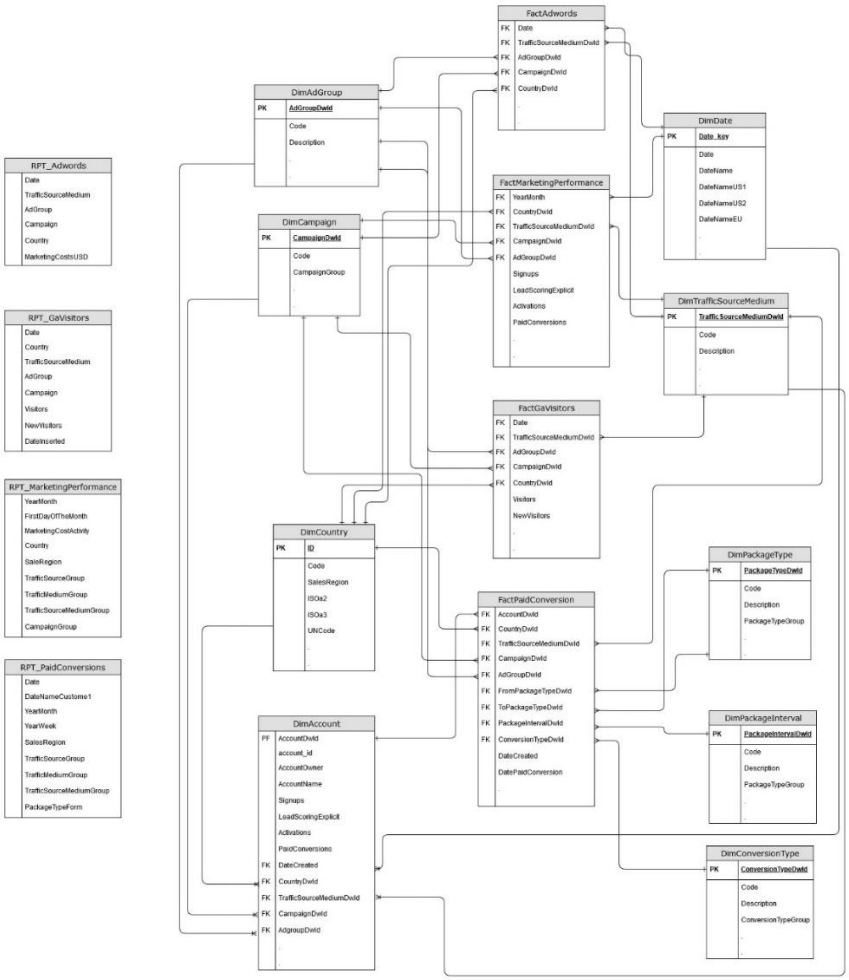
3.4 Návrh dátového modelu

Po charakterizovaní zdrojových systémov, dát a exportov dát z nich, bude v tejto kapitole popísaný návrh dátového modelu. Dátový model bude vizualizovaný pomocou ER diagramu. Techniky a metodológie dátového modelovania sú použité pre modelovanie dát v štandardizovanej, konzistentnej a predvídateľnej podobne z dôvodu správneho spravovania zdrojových systémov. V širokej praxi sú uznávané dve modelovacie metodológie.

- Metodika zdola hore – táto metodika sa predovšetkým zameriava na dáta v organizácii, vykazovanie a vypracovanie analýz z dát je až druhotné. Z hľadiska toku dát sa táto metodológia zameriava predovšetkým na centralizované získavanie a ukladanie dát.
- Metodika zhora dole – táto metodika sa viac ako na dáta v organizácii špecializuje na vytvorenie riešenia podporujúceho vykazovanie a analýzu dát. Dôraz je kladený na získanie požiadaviek užívateľov na základe ktorých, je DW vybudovaný.

Požadované sú výstupy na úrovni marketingových dát, pochádzajúcich predovšetkým z marketingových nástrojov. Keďže cielime výhradne na zákazníkov podnikajúcich v online priestore, drvivá väčšina marketingových nákladov ide práve do PPC kampaní. Hlavnou požiadavkou marketingového oddelenia je prístup k informáciám a vizualizácii nákladov na jednotlivé PPC kampane podľa regiónov.

Na základe tejto požiadavky bol vyhotovený ER diagram zobrazujúci požadovanú štruktúru dát a tabuliek, ktoré budú vytvorené a importované do dátového skladu.



Obrázok 30: ER diagram

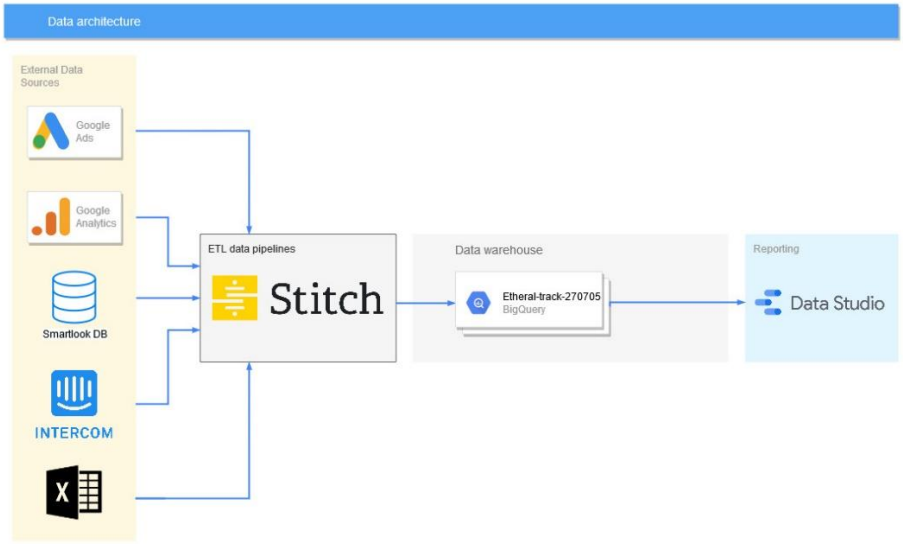
(Zdroj: Vlastný návrh)

3.5 Návrh dátovej architektúry

Po definícii dátového modelu založenej na metodike zdola hore, je dôležité definovať skutočný tok dát celým systémom dátového skladu. Dátová architektúra sa odvodzuje od dát, ktoré sú k požadovanému výstupu potrebné. Pri prístupe zameranom na získavanie informácií z dát, teda prístupe zameranom na výstup je dôraz kladený na dátové trhy, ktoré zjednodušene odpovedajú potvrdeným dimenziám. V tomto prípade je dátový model postavený na princípe hviezdy.

Veľké množstvo nástrojov BI v súčasnosti uprednostňuje jednoduché schémy hviezdy, to znamená, že z pohľadu vykazovania a OLAP daného riešenia sa musí použiť hviezdicový formát dátových trhov. Zdrojové dátové trhy môžu mať ľubovoľný návrh.

Nasledujúci koncept vykresľuje dátový tok jednotlivými komponentami dátového skladu. Externé dátové zdroje predstavujú zdroje dát, ktoré boli popísané v niektorej z predchádzajúcich kapitol. Patria sem Google Analytics, Google Adwords, Intercom, Google docs a dáta zo zdrojovej databázy Smartlook. Spracovávanie dát bude zabezpečovať ETL dátový kanál Stitch, ktoré následne dáta zo zdrojových systémov bude ukladať v požadovanom tvare v samotnom dátovom sklade Google BigQuery. Výsledné reportovanie a vizualizáciu dát bude zabezpečovať Google Data Studio.



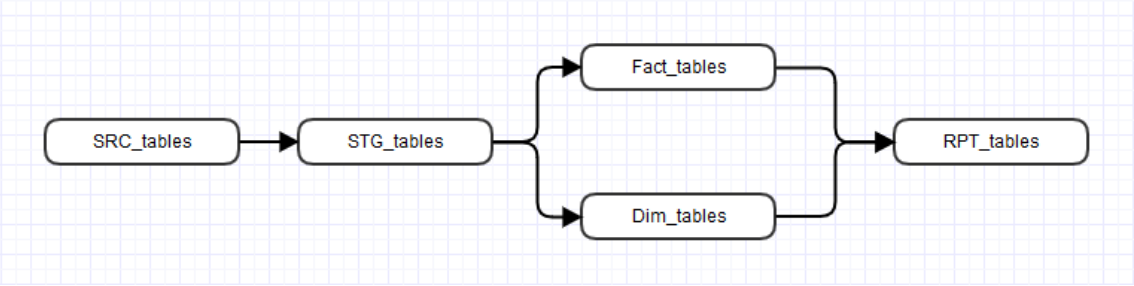
Obrázok 31: Dátová architektúra

(Zdroj: Vlastný návrh)

3.6 Príprava zdrojovej databázy

Dáta v analytickej databáze tvoria základ pre dátový sklad. Analytická databáza obsahuje aktuálne dáta o všetkých zákazníkoch, vytvorených definíciách eventov, funnelov, heatmap, kompletný zoznam transakcií, faktúr, marketingových kampaní a ďalších. Tabuľky v analytickej databáze boli s ohľadom na budovanie dátového skladu rozdelené na:

- Source tabuľky
- Staging tabuľky
- Tabuľky faktov
- Tabuľky dimenzií
- Reportovacie tabuľky



Obrázok 32: Prechod dát tabuľkami

(Zdroj: Vlastný návrh)

3.6.1 Source tabuľky

Všetky dáta, nahrávané do databázy, sa najprv ukladajú do pripravených source tabuliek. Pre každý zdroj, z ktorého sú dáta importované bola vytvorené špecifická source tabuľka s označením SRC_. Ak dáta nie je potrebné spracovávať ďalej, v daných tabuľkách zostávajú zapísané a nijako sa s dátami ďalej nemanipuluje.

Väčšinu dát vstupujúcich do databázy je ale potreba spracovať pre ďalšiu prácu. Pre následne spracovanie dát zo source tabuliek boli pripravené ETL procedúry. Dané ETL procedúry zabezpečujú extrakciu, čistenie a presun dát medzi tabuľkami. Dáta ukladané

do source tabuliek na dennej báze sú iniciované každý deň v ranných hodinách a to z dôvodu najmenšieho zaťaženie prenosových kanálov v ranných hodinách.

3.6.2 Staging tabuľky

Pre potreby čistenia dát z source tabuliek boli vytvorené staging tabuľky, ktoré slúžia ako dočasné tabuľky na prípravu dát pre plnenie tabuliek faktov a dimenzii. Dáta sa v staging tabuľkách:

- Deduplikujú
- Čistia
- Normalizujú do viacerých tabuliek
- Denormalizujú z viacerých tabuliek do jednej tabuľky
- Extrapolujú

V žiadnej zo staging tabuliek dáta nezostávajú zapísané dlhšiu dobu a tabuľky sú čistené po každej vykonanej procedúre. Tieto tabuľky je teda možné považovať za dočasnú zbernicu dát, ktorá slúži na ich vyčistenie.

3.6.1 Tabuľky dimenzií

Tabuľky dimenzii obsahujú deskriptívne charakteristiky faktov za pomoci atribútov, ktoré sú vo väčšine prípadov v textovom formáte. Kombinácia dát z tabuliek dimenzií a faktov sa používajú na skomponovanie reportov. Tabuľky dimenzií sa pravidelne rozširujú novými dátami zo source tabuliek. V zdrojovej databáze sú tabuľky dimenzií označované Dim_.

- **DimAccount –** táto dimenzia obsahuje informácie o všetkých účtoch, ktoré boli historicky v Smartlooku vytvorené. Každý zákazník má svoje vlastné špecifické ID, ktoré slúži ako jeho jedinečný identifikátor. Primárny kľúč je teda account_id. Mimo číselného identifikátoru majú jednotlivé účty atribút AccountName, ktorý obsahuje emailovú adresu priradenú k danému účtu. Subscription, ktoré v tom momente daný účet má aktívne, je pod atribútom AccountType. Ďalej táto tabuľka

obsahuje dáta o tom, či je klient stále aktívny a používa Smartlook, kto je vlastník daného účtu a aký jazyk používa. Z pohľadu tohto daného projektu sú ale dôležité marketingové dáta. Ku každému účtu je priradené médium, cez ktoré bol zákazník privedený, z akej zeme zákazník pochádza a konkrétna marketingová kampaň a reklamná skupina, ktorá zákazníka priviedla. Tieto dáta sú ku každému účtu priradené pomocou cudzích kľúčov daných atribútov. Tieto atribúty sú CountryDwId, TrafficSourceMediumDwId, CampaignDwId, AdgroupDwId. Mimo tieto atribúty obsahuje tabuľka aj niekoľko ďalších, tie ale nie sú dôležité výstup projektu, tak sa nimi nebudem zaoberať.

- **DimDate –** táto dimenzia obsahuje dátumy na úrovni granularity kalendárneho dňa. Primárnym kľúčom jednej entity je atribút date_key. Ďalej je daný kalendárny deň zapísaný v americkej a európskej forme v atribútoch DateNameUS1, DateNameUS2 a DateNameEU. Ktorý konkrétny deň v týždni, v mesiaci, v roku a aký je konkrétny názov dňa, sú pod atribútmi DayOfWeek, DayOfMonth, DayOfYear a DayNameOfWeek. K dispozícii je aj granularita na úrovni týždňa, mesiaca a kvartálu v roku pod atribútmi YearWeek, YearMonth a YearQuarter.
- **DimCountry –** táto dimenzia obsahuje všetky krajiny sveta. Primárnym kľúčom je atribút CountryDwId. Názov krajiny je pod atribútom Code. Ku krajinám sú priradené aj kódy podľa ISOa2, ISOa3, UN číselný kód a atribút SalesRegion.
- **DimCurrency -** táto dimenzia obsahuje svetové meny. Primárnym kľúčom je atribút CurrencyDwId. Skratka názvu danej meny je pod atribútom Code a približný kurz je pod atribútom ExchangeRate. K dispozícii je 8 svetových mien podľa abecedy – BRL, CNY, CZK, DKK, EUR, HUF, PLN, RUB, USD. **Nemusí tu byť táto dimenzia**
- **DimPackageType –** táto dimenzia v sebe zahŕňa všetky druhy balíčkov, ktoré boli historicky k dispozícii.. Primárnym kľúčom je atribút PackageTypeDwId. Názov balíčku je pod atribútom Cod a k dispozícii je ešte skupina, do ktorej daný balíček patrí. Táto skupina je pod atribútom PackageTypeGroup.
- **DimPackageInterval –** táto dimenzia definuje rôzne dĺžky platnosti jednotlivých balíčkov. Primárnym kľúčom je atribút PackageIntervalDwId.

- **DimCampaign –** táto dimenzia obsahuje historicky všetky marketingové kampane, ktoré boli za existencie Smartlooku spustené. Kampaň je základnou jednotkou pod zdrojom návštevy. Na úrovni kampaní sa rieši geografické cielenie, jazyk kampaní alebo celkový budget. Primárnym kľúčom tejto dimenzie je CampaignDwId. Názov samotnej kampane je následne pod atribútom Code a kampaňová skupina je pod atribútom CampaignGroup
- **DimAdGroup –** táto dimenzia obsahuje skupiny reklám. AdGroup je vlastne podskupinou kampane, vďaka ktorej je možné definovať cielenie kampaní presnejšie. AdGroupy sa vytvárajú aj kvôli prehľadnosti. Primárnym kľúčom tejto dimenzie je atribút AdgroupDwId a názov adgroupy je po atribútom Code.
- **DimTrafficSourceMedium -** táto dimenzia obsahuje všetky kombinácie médií a zdrojov užívateľských návštev na našom webe. Médiom môže byť priama návšteva, organická návšteva alebo návšteva privedená pomocou cpc kampane. Zdrojom sú platformy, ktoré návštevnosť prinášajú. Sem patrí napríklad Google, Facebook, LinkedIn atď. Primárny kľúčom tejto dimenzie je atribút TrafficSourceMediumDwId. Spojený názov je pod atribútom code a skupiny, do ktorých sa dané médiá a zdroje zaraďujú, sú pod atribútom TrafficSourceMediumGroup. Táto dimenzia je plnená dvom ďalšími tabuľkami, keďže ide o kombinácie zdrojov a médií, obsahuje cudzie kľúče z dimenzií DimTrafficSource a DimTrafficMedium.
- **DimConversionType –** táto dimenzia obsahuje tipy konverzií. Tie máme zadefinované iba dve a to 1st Paid Conversion a Reactivation. Primárny kľúč tejto tabuľky je ConversionTypeDwId.

3.6.2 Tabuľky faktov

Tabuľky faktov obsahuje merania, metriky a fakty business procesu. Bývajú väčšinou lokalizované v strede schémy a obsahujú cudzie kľúče k tabuľkám dimenzií. Primárny kľúč tabuľky faktov býva skomponovaný z viacerých cudzích kľúčov obsiahnutých v tabuľke faktov. Väčšina tabuliek faktov obsahuje časové pozorovanie určitej udalosti. V zdrojovej databáze sú tabuľky faktov označované predponou Fact.

- **FactAdwords –** Do tejto tabuľky faktov vstupujú dimenzie DimDate, DimTrafficSourceMedium, DimAdGroup, DimCampaign a DimCountry. Cudzie kľúče spomenutých 5 dimenzii tvoria primárny kľúč tejto tabuľky.

Date	TrafficSourceMediumDwId	AdgroupDwId	CampaignDwId	CountryDwId	MarketingCostsUSD	DateInserted
2020-01-20	588	4	3 589	150	0,000	2020-02-04 06:01:06
2020-01-02	588	5	3 589	150	1,472	2020-02-04 06:01:06
2020-01-05	588	7	3 589	150	0,000	2020-02-04 06:01:06
2020-01-22	588	8	3 589	150	2,584	2020-02-04 06:01:06
2020-01-14	588	8	3 589	150	0,000	2020-02-04 06:01:06
2020-01-17	588	9	3 589	150	0,000	2020-02-04 06:01:06
2020-01-08	588	9	3 589	150	0,000	2020-02-04 06:01:06
2020-01-19	588	9	3 589	150	0,000	2020-02-04 06:01:06
2020-01-06	588	9	3 589	150	0,000	2020-02-04 06:01:06
2020-01-03	588	10	3 589	150	0,000	2020-02-04 06:01:06

Obrázok 33: Fact Adwords tabuľka

(Zdroj: Vlastný návrh)

- **FactGAVisitors –** Do faktovej tabuľky FactGAVisitor vstupujú rovnaké dimenzie ako v predchádzajúcej tabuľke FactAdwords. Nepočítajú sa návštevy zo Smartlook dashboardu.

Date	CountryDwId	TrafficSourceMediumDwId	AdgroupDwId	CampaignDwId	Visitors	NewVisitors	DateInserted
2020-01-13	235	12 983	-1	-1	1	0	2020-02-05 06:02:01
2020-01-14	235	12 983	-1	-1	1	0	2020-02-05 06:02:01
2020-01-13	235	13 599	-1	-1	1	0	2020-02-05 06:02:01
2020-01-14	235	13 599	-1	-1	1	0	2020-02-05 06:02:01
2020-01-12	235	24 124	-1	-1	2	0	2020-02-05 06:02:01
2020-01-13	235	24 124	-1	-1	1	0	2020-02-05 06:02:01
2020-01-12	235	24 284	-1	-1	1	1	2020-02-05 06:02:01
2020-01-12	235	25 367	-1	-1	2	1	2020-02-05 06:02:01
2020-01-13	235	25 367	-1	-1	2	0	2020-02-05 06:02:01
2020-01-14	235	25 367	-1	-1	3	0	2020-02-05 06:02:01

Obrázok 34: Fact Google Analytics tabuľka

(Zdroj: Vlastný návrh)

- **FactMarketingPerformance -** Do faktovej tabuľky zaznamenávajúcej výkon marketingový kampaní vstupujú rovnaké dimenzie ako do predchádzajúcich dvoch faktových tabuliek. Jeden riadok tejto tabuľky predstavuje súčet nových návštevníkov a prihlásení za mesiac. Všetky relevantné metriky sú priradené k dátumu registrácie, čo znamená, že celý marketing je de fact kohortová analýza.

YearMonth	CountryDwId	TrafficSourceMediumDwId	CampaignDwId	AdGroupDwId	NewVisitors	Signups	LeadScoringExplicit	LeadScoringImplicit	Activations
201 802	208	140	-1	-1	1	1	20	5	0
201 802	208	472	-1	-1	6	2	20	13	0
201 802	208	475	-1	-1	0	5	65	25	1
201 802	208	475	587	-1	11	2	35	5	0
201 802	208	475	596	-1	526	97	1 410	606	30
201 802	208	475	599	-1	27	7	150	42	1
201 802	208	475	2 118	-1	5	1	25	5	0
201 802	208	482	-1	-1	28	17	265	92	4
201 802	208	491	-1	-1	1	1	15	5	1
201 802	208	588	58	36	19	1	10	5	0

Obrázok 35: Fact marketing performance tabuľka

(Zdroj: Vlastný návrh)

- **FactPaidConversion –** Jeden riadok reprezentuje jednu zmenu v predplatnom z trial alebo free účtu na platený. Sem vstupujú rovnaké dimenzie ako do predchádzajúcich faktových tabuliek.

AccountDwId	account_id	CountryDwId	TrafficSourceMediumDwId	AdgroupDwId	CampaignDwId	DateCreated	DatePaidConversion	FromPackageTypeDwId
274 698	317 107	166	592	-1	-1	2019-04-05	2019-04-19	4
274 679	317 205	143	592	-1	-1	2019-04-06	2019-05-29	4
274 905	317 347	201	1	-1	-1	2019-04-08	2019-09-23	4
275 167	317 595	84	592	-1	-1	2019-04-09	2019-04-29	4
275 165	317 610	77	1 607	-1	-1	2019-04-09	2019-04-24	4
275 399	317 805	61	592	-1	-1	2019-04-10	2019-04-30	4
275 585	317 899	215	1	-1	-1	2019-04-11	2019-04-24	4
275 650	318 072	201	1	-1	-1	2019-04-12	2019-09-17	4
275 659	318 146	31	1 045	-1	-1	2019-04-12	2019-04-12	4
275 663	318 361	235	24 031	-1	-1	2019-04-15	2019-10-02	4

Obrázok 36: Fact paid conversion tabuľka

(Zdroj: Vlastný návrh)

3.6.3 Reportovacie tabuľky

Reportovacie tabuľky zlučujú dáta z tabuliek faktov a dimenzií na základe ich cudzích kľúčov. Keďže jedna tabuľka faktov môže obsahovať dáta z viacerých tabuliek dimenzií, tieto dáta sú v tabuľke faktov zapísané vo forme cudzích kľúčov, ktoré majú vo väčšine prípadov numerickú podobu a to ID danej entity. Aby ale bolo možné získať použiteľné informácie z týchto tabuliek, je potrebné tieto numerické hodnoty zobraziť vo forme ich pridruženej hodnoty. Na toto preto slúžia reportovacie tabuľky. Pre každú tabuľku faktov som vypracoval reportovaciu tabuľku.

- **RPT_Adwords –** výstupom tabuľky RPT_Adwords je celková cena za kampane v jednotlivých krajinách, rozdelená podľa konkrétneho zdroja a média, na ktorom bola kampaň rozbehnutá a do ktorej reklamnej skupiny kampaň patrila.

Date	Traffic Source Medium	AdGroup	Campaign	Country	MarketingCostsUSD
2017-12-07	google / cpc	UX_testing	Asia - broad	Korea South	0,115
2017-09-05	google / cpc	Website_Visitor_tracking	Asia - broad	Korea South	0,172
2018-02-13	google / cpc	Session_Replay	Asia - broad	Korea South	0,092
2018-03-22	google / cpc	Heatmaps	Asia - broad	Korea South	0,103
2018-07-06	google / cpc	Heatmaps	Asia - broad	Korea South	0,229
2018-06-08	google / cpc	Heatmaps	Asia - broad	Korea South	0,092
2018-02-28	google / cpc	Remarketing_Banner_standart	Asia - remarketing	Korea South	0,069
2018-10-18	google / cpc	Remarketing_Banner_standart	Asia - remarketing	Korea South	0,275
2018-07-11	google / cpc	Website_recording	RU - broad	Russian Federation	0,229
2018-05-28	google / cpc	Website_recording	RU - broad	Russian Federation	0,172
2016-11-09	google / cpc	Brand	RU - broad	Russian Federation	0,195

Obrázok 37: Reportovacia tabuľka Google Adwords

(Zdroj: Vlastný návrh)

- **RPT_GAVisitors –** výstupom tabuľky RPT_GAVisitors sú počty nových a vracajúcich sa návštevníkov na stránky Smartlooku, podľa jednotlivých krajín rozdelených na zdroje a médiá návštev a kampane, ktoré ich na stránky Smartlooku priviedli. Táto tabuľka je vo výsledku podobná ako tabuľka RPT_Adwords s tým rozdielom, že v tejto tabuľke sa sleduje celkový počet privedených ľudí a nie náklady na privedenie návštevnosti.

Date	Country	TrafficSourceMedium	Adgroup	Campaign	Visitors	NewVisitors	DateInserted
2019-01-01	Argentina	(direct) / (none)	(not set)	(not set)	8	1	2019-04-29 11:17:38
2019-01-01	Argentina	google / organic	(not set)	(not set)	7	2	2019-04-29 11:17:38
2019-01-01	Argentina	smartsupp.com / referral	(not set)	(not set)	1	0	2019-04-29 11:17:38
2019-01-01	Australia	(direct) / (none)	(not set)	(not set)	8	2	2019-04-29 11:17:38
2019-01-01	Australia	google / organic	(not set)	(not set)	8	2	2019-04-29 11:17:38
2019-01-01	Australia	quora.com / referral	(not set)	(not set)	1	0	2019-04-29 11:17:38
2019-01-01	Australia	smartsupp.com / referral	(not set)	(not set)	1	0	2019-04-29 11:17:38
2019-01-01	Australia	feng.com / referral	(not set)	(not set)	1	0	2019-04-29 11:17:38
2019-01-01	Austria	(direct) / (none)	(not set)	(not set)	8	1	2019-04-29 11:17:38
2019-01-01	Austria	at.search.yahoo.com / referral	(not set)	(not set)	1	0	2019-04-29 11:17:38

Obrázok 38: Reportovacia tabuľka Google Analytics

(Zdroj: Vlastný návrh)

- **RPT_MarketingPerformance -** výstupom tejto tabuľky je finančné hodnotenie marketingového výkonu všetkých kampaní. Kampane sú rozdelené rovnako ako v predošlých dvoch tabuľkách podľa jednotlivých zemí, médií a zdrojov návštev. Zaznamenaný je počet užívateľov, ktorých dané kampane priviedli, ale aj to, koľko z tých užívateľov si aktivovalo Smartlook a koľký z nich prešli na platený balíček. Zaznamenané sú náklady danej kampane, preto je možné vďaka tejto tabuľke zistiť, aký jednotlivé kampane vytvorili čistý finančný zisk alebo naopak stratu.

YearMonth	FirstDateOfTheMonth	MarketingCostsActivity	Country	SalesRegion	TrafficSourceGroup	TrafficMediumGroup	TrafficSourceMediumGroup	CampaignGroup
201 902	2019-02-01	(not set)	Belarus	Rest of Europe	Other	Referral	Other/ Referral	(not set)
201 902	2019-02-01	(not set)	Belarus	Rest of Europe	Smartlook	(none)	Smartlook/ (none)	(not set)
201 902	2019-02-01	(not set)	Belgium	EU	Direct	(none)	Direct/ (none)	(not set)
201 902	2019-02-01	(not set)	Belgium	EU	Facebook	CPC	Facebook/ CPC	Facebook
201 902	2019-02-01	(not set)	Belgium	EU	Facebook	Referral	Facebook/ Referral	(not set)
201 902	2019-02-01	(not set)	Belgium	EU	Google	Organic	Google/ Organic	(not set)
201 902	2019-02-01	(not set)	Belgium	EU	Google	Referral	Google/ Referral	(not set)
201 902	2019-02-01	(not set)	Belgium	EU	Linkedin	CPC	Linkedin/ CPC	Linkedin Mobile
201 902	2019-02-01	(not set)	Belgium	EU	Other	CPC	Other/ CPC	Capterra Campaigns
201 902	2019-02-01	(not set)	Belgium	EU	Other	Integration	Other/ Integration	(not set)
201 902	2019-02-01	(not set)	Belgium	EU	Other	Other	Other/ Other	Smartlook Newsletters & Promos

Obrázok 39: Reportovacia tabuľka marketing performance

(Zdroj: Vlastný návrh)

- **RPT_PaidConversions -** výstupom tejto tabuľky sú všetky platené konverzie. Dostupné sú dáta o jednotlivých zákazníkoch a dáta o celej konverzii. K ID zákazníka sú priradené všetky vyššie spomínané marketingové dáta, teda kampaň,

médium a zdroj návštevy. Ďalej je k dispozícii druh balíčku a typ konverzie, ktorý vykonal z dimenzie DimPackageType. Jednotlivé konverzie sú brané z časového hľadiska.

Date	DateNameCustom1	YearMonth	YearWeek	SalesRegion	TrafficSourceGroup	TrafficMediumGroup	TrafficSourceMediumGroup	PackageTypeFrom
2019-03-18	Monday, Mar 18 2019	201 903	201 912	EU	Google	CPC	Google/ CPC	free
2019-03-18	Monday, Mar 18 2019	201 903	201 912	EU	Google	Organic	Google/ Organic	free
2019-03-19	Tuesday, Mar 19 2019	201 903	201 912	EU	Direct	(none)	Direct/ (none)	trial
2019-03-19	Tuesday, Mar 19 2019	201 903	201 912	EU	Google	Organic	Google/ Organic	free
2019-03-19	Tuesday, Mar 19 2019	201 903	201 912	EU	Google	Organic	Google/ Organic	free
2019-03-19	Tuesday, Mar 19 2019	201 903	201 912	EU	Google	Organic	Google/ Organic	free
2019-03-19	Tuesday, Mar 19 2019	201 903	201 912	EU	Smartlook	(none)	Smartlook/ (none)	free
2019-03-19	Tuesday, Mar 19 2019	201 903	201 912	Middle East	Quora	Referral	Quora/ Referral	free
2019-03-20	Wednesday, Mar 20 2019	201 903	201 912	EU	Smartlook	(none)	Smartlook/ (none)	free
2019-03-20	Wednesday, Mar 20 2019	201 903	201 912	Latin America & Carribean	Google	Organic	Google/ Organic	free

Obrázok 40: Reportovacia tabuľka paid conversion

(Zdroj: Vlastný návrh)

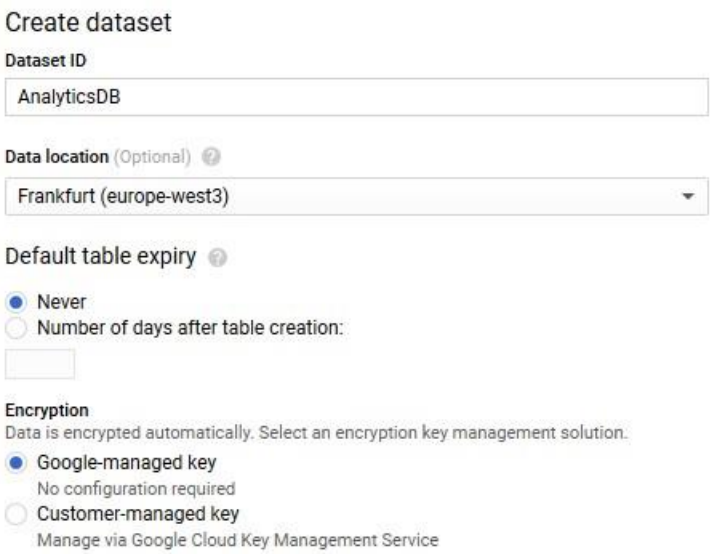
3.7 Tvorba dátového modelu v Google BigQuery

Prvým krokom vytvárania dátového modelu je vytvorenie účtu v Google cloud platform. Táto platforma v sebe zlučuje všetky cloudové služby, ktoré Google má v súčasnosti v ponuke. K dispozícii sú výpočtové nástroje, nástroje na skladovanie dát, správu sieťovej prevádzky, Big Data nástroje alebo nástroje pre prácu s umelou inteligenciou. Google BigQuery patrí medzi nástroje z kategórie Big Data.

Po zaregistrovaní účtu sa zobrazí centrálna konzola Google cloud platform. Pre použitie a zobrazenie dát v konzole je potrebné vytvorenie projektu, tento krok vyžaduje zadanie názvu projektu a zvolenie organizácie na základe domény, v mojom prípade smartlook.com. Po vytvorení projektu sa otvorí centrálny dashboarde, v ktorom sa nachádzajú základné informácie o projekte, využívaných zdrojoch, použitej pamäti na SQL serveroch a fakturačné údaje. Po zoznámení sa s celou platformou prichádza práca v samotnom BigQuery nástroji.

Prvý krokom pri vytváraní dátového modelu je vytvorenie zdroja. Nazvaný bol ako etheral-tract-270715. Zdroj v Google BigQuery zastupuje po projekte druhú najvyššiu entitu. Jednotlivé zdroje v sebe združujú dátové sety. Nasledujúcim rokom je vytvorenie dátového setu, ktorý bol nazvaný AnalyticsDB. Dátový set predstavuje v Google BigQuery alternatívu ku klasickej relačnej databáze. Pri vytváraní nového dátového setu je potreba nastaviť lokalitu, kde budú dáta skladované. Predvolenou lokalitou pre dátový

set je US multiregion, to však z hľadiska umiestnenia našich dát je nevyhovujúce. Väčšina našich súčasných dáta je skladovaných na serveroch vo Frankfurte, preto bol pre dátový set AnalyticsDB z hľadiska čo najmenšej latencie vybraná lokalita vo Frankfurte, ktorá patrí do regiónu europe-west3. Dané umiestnenie je následne použité aj pre všetky tabuľky vytvorené v tomto dátovom sete. Mimo regiónu je potrebné nastaviť predvolenú dobu, po ktorej budú tabuľky v danom dátovom sete vymazané. Keďže je cieľom projektu vybudovanie dátového skladu, ktorý v sebe bude ukladať historické dát, tabuľkám nie je nastavená doba premazávania dát. Všetky dáta sú šifrované automaticky, v poslednom kroku je ale treba zvoliť metódu správy šifrovacích kľúčov. Na výber je správa šifrovania pomocou Google-managed key alebo Custom-managed key. V tomto prípade bola zvolená správa šifrovania pomocou Google-managed key.



Obrázok 41: Google BQ tvorba datasetu

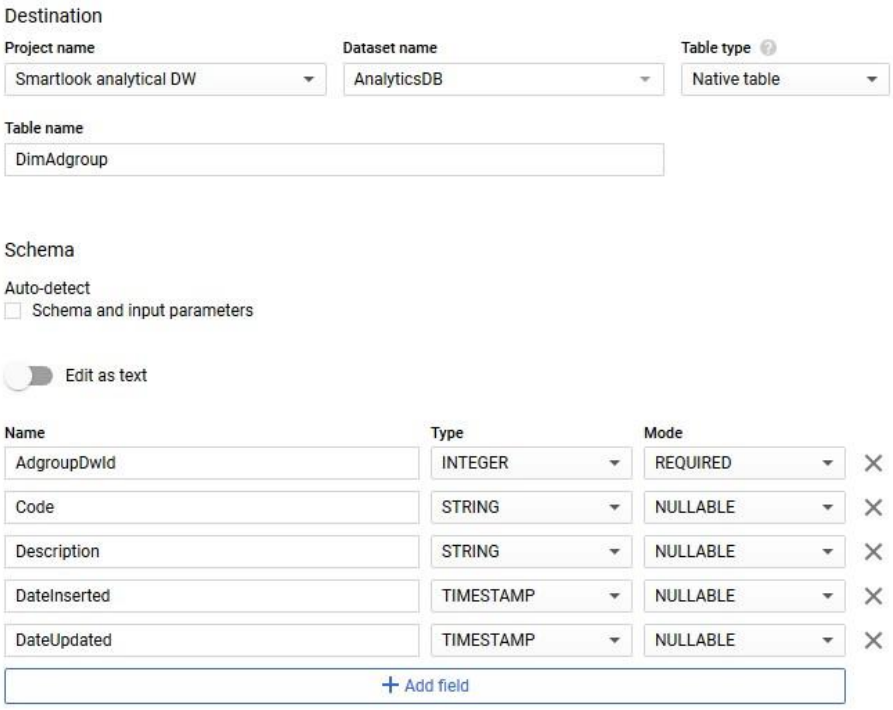
(Zdroj: 21)

Po vytvorení dátového setu prichádza na radu tvorba jednotlivých tabuliek podľa vyhotoveného ER diagramu.

3.7.1 Definovanie tabuliek

V predchádzajúcej kapitole bol popísaný postup prípravy zdrojovej databázy, kde boli tabuľky a dáta v nich pripravené na export do dátového skladu. V dátovom sete AnalyticsDB preto nasleduje definovanie jednotlivých tabuliek.

Pri tvorbe tabuľky treba definovať názov projektu a dátový set do ktorého daná tabuľka bude patriť. Názvy a štruktúra jednotlivých tabuliek sa budú držať dátového slovníku zdrojovej databázy. Atribúty a entity tabuľky je možné definovať dvoma spôsobmi: create dotazom alebo manuálnym definovaním jednotlivých atribútov a ich dátových typov. K dispozícii je tiež automatické zisťovanie štruktúry tabuľky, ktoré ale pre zaistenie bezchybovosti nebolo využité. Na obrázku nižšie je vidieť príklad definovania tabuľky DimAdgroup.



Obrázok 42: Google BQ tvorba tabuľky

(Zdroj: 21)

3.7.2 Export a plnenie tabuliek

Všetky dáta v predpripravených tabuľkách zo zdrojovej databázy boli vyexportované v CSV formáte bez zahrnutia riadku definujúceho názov jednotlivých atribútov. Keďže veľkosť niektorých exportovaných tabuliek dosahuje desiatky až stovky MB, rozhodol som sa najskôr všetky exportované tabuľky uložiť v rámci platformy. Súčasťou Google cloude platformy je služba Storage, ktorá slúži výhradne na uskladňovanie dát potrebných pre ďalšiu prácu v rámci platformy.

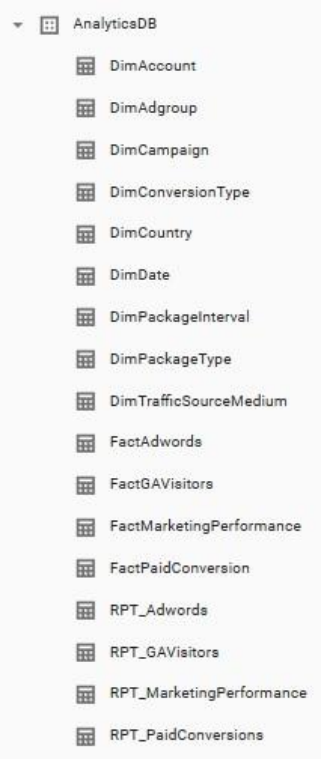
Základným elementom celej služby je takzvaný bucket v rámci ktorého sa ukladajú všetky súbory. Pri vytváraní bucketu je potrebné určiť, mimo názvu podobne ako pri dátovom sete lokalitu, kde budú dáta uložené. Voľba lokalizovania dát má vplyv na celkovú cenu úložiska, výkon a dostupnosť. Na výber sú tri druhy umiestnení dát.

- Region – najnižšia latencia v rámci jedného regiónu
- Multi-region – najvyššia dostupnosť v rámci veľkej oblasti
- Dual-region – veľká dostupnosť a malá latencia v rámci dvoch regiónov

Keďže bol dátový sklad, ktorý bude plnený dátami z exportovaných tabuliek, umiestnení v regióne europe-west3 (Frankfurt), bolo zvolené rovnaké regionálne umiestnenie aj pre bucket nazvaný Tables_DBanalytics. Okrem lokality je potrebné zvoliť štandardnú triedu úložiska. Dôležité je dbať na to, ako často bude k daným dátam pristupované a po akú dlhú dobu je nevyhnutné ich mať k dispozícii. K dispozícii sú nasledovné 4 triedy:

- Standard – určené pre krátkodobé ukladanie a frekventovaný prístup k dátam
- Nearline – určené pre backup dát a pre dát, s ktorými sa pracujeme menej ako raz mesačne
- Coldline – určené pre potreby krízového prístupu k dátam a pre dáta, s ktorými sa pracuje menej ako raz za kvartál
- Archive – dobré pre dlhodobé uchovávanie dát, s ktorými sa pracuje menej ako raz ročne

Vytvorený bucket Tables_DBanalytics je na záver naplnený všetkými vyexportovanými tabuľkami vo formáte CSV. Plnenie tabuliek dátami následne znova prebieha v dashboarde Google BigQuery, kde je každá pripravená tabuľka naplnená dátami z tabuliek uložených v buckete. Výsledný dátový set vyzerá nasledovne.



Obrázok 43: Google BQ tabuľky

(Zdroj: 21)

3.7.3 Procedúry

Na zabezpečenie konštantného plnenia tabuliek faktov, dimenzií a reportovacích tabuliek sú potrebné procedúr. Tieto procedúry budú uložené v rámci dátového skladu a spúšťané v prednastavených intervaloch. V Google BigQuery spadajú procedúry pod BigQuery jobs. Keďže môže vykonanie týchto jobs zabrať väčšie množstvo času, sú vykonávané asynchrónne. Kratšie akciem ako je napríklad načítavanie zoznamu zdrojov alebo načítavanie metadát, nie je vykonávané pomocou BigQuery jobs. V tejto podkapitole budú rozobrané procedúry, ktoré zabezpečujú plnenie jednotlivých faktových tabuliek.

- **ETL2B_FactAdwords** – procedúra zabezpečujúca plnenie faktovej tabuľky FactAdwords. Základné dáta sa najskôr vyťahujú zo zdrojovej tabuľky obsahujúcej dáta z pravidelného mesačného reportu služby Google ads. Keďže sú výsledky jednotlivých kampaní počítané v USD, všetky výsledky v iných menách sú prepočítavané na USD. Zvyšné dáta sú pomocou inner a left joinu spájané na základe cudzích kľúčov z tabuliek DimAdGroup, DimCampaign a DimCountry.

```
INSERT INTO ethereal-tract-270715.AnalyticsDB.FactAdwords
(
      Date
      ,TrafficSourceMediumDwId
      ,AdgroupDwId
      ,CampaignDwId
      ,CountryDwId
      ,MarketingCostsUSD
)
SELECT
      stg.Date
      ,588 AS TrafficSourceMediumDwId -- Google / CPC
      ,ag.AdgroupDwId
      ,cam.CampaignDwId
      ,c.CountryDwId
      ,CAST((stg.Cost * cur.exchangeRate / cur2.exchangeRate) AS
      DECIMAL(12,3))
FROM ethereal-tract-270715.AnalyticsDB.STG_Adwords stg
INNER JOIN ethereal-tract-270715.AnalyticsDB.SRC_currency cur
      ON stg.Currency = cur.code
INNER JOIN ethereal-tract-270715.AnalyticsDB.SRC_currency cur2
      ON cur2.code = 'USD'
INNER JOIN ethereal-tract-270715.AnalyticsDB.DimAdgroup ag
      ON stg.Adgroup = ag.Code
INNER JOIN ethereal-tract-270715.AnalyticsDB.DimCampaign cam
      ON stg.Campaign = cam.Code
LEFT JOIN ethereal-tract-270715.AnalyticsDB.DimCountry c
ON COALESCE(MapToCode, stg.Country) = c.Code;
```

- **ETL2B_FactGaVisitors** – procedúra zabezpečuje rovnako ako v predchádzajúcom prípade plnenie tabuľky FactGaVisitors dátami z pravidelných mesačných reportov. V tabuľke FactGaVisitors ale neprebiehajú výpočty ceny kampane, ale počty nových a vracajúcich sa užívateľov.

```
INSERT INTO ethereal-tract-270715.AnalyticsDB.FactGAVisitors
(
      Date,
      CountryDwId,
      TrafficSourceMediumDwId,
      AdgroupDwId,
      CampaignDwId,
      Visitors,
      NewVisitors
)
SELECT
```

```
        stg.Date
        ,c.CountryDwId
        ,tsm.TrafficSourceMediumDwId
        ,ag.AdgroupDwId
        ,cam.CampaignDwId
        ,`Users` AS Visitors
        ,`New Users` AS NewVisitors
FROM ethereal-tract-270715.AnalyticsDB.STG_GA_Visitors stg
LEFT JOIN ethereal-tract-
270715.AnalyticsDB.DimTrafficSourceMedium tsm
        ON stg.`Source / Medium` = tsm.Code
LEFT JOIN ethereal-tract-270715.AnalyticsDB..DimCountry c
        ON TRIM(stg.`Country ISO Code`) = c.ISOa2
LEFT JOIN ethereal-tract-270715.AnalyticsDB..DimAdgroup ag
        ON stg.`AdWords Ad Group` = ag.Code
LEFT JOIN ethereal-tract-270715.AnalyticsDB..DimCampaign cam
        ON stg.Campaign = cam.Code
WHERE stg.`Source / Medium` NOT IN ('dashboard.smartsupp.com /
referral');
```

- **ETL2B_FactMarketingPerformance –** všetky relevantné metriky sú pomocou tejto procedúry priraďované k dátumu vytvorenia účtu v Smartlooku daným užívateľom. Celoživotná hodnota zákazníka je vypočítaná ako súčet všetkých zaplatených faktúr + projekcia toho, že zákazník bude platiť ďalších 9 až 10 mesiacov. Pri každom načítaní dát do tabuľky prebieha rekalkulácia za posledných 6 mesiacov.
- **ETL2B_FactPaidConversion** – procedúra napĺňa tabuľku FactPaidConversion dátami o platených konverziách. Jeden riadok predstavuje jednu zmenu v subscription z trial alebo free účtu na platený. Dôležité je vylúčiť z platobných konverzií zmeny v rámci trial a demo účtov, ktoré systémovo môžu vyzerať ako konverzia, ale sú to len interné operácie na účtoch potenciálnych klientov.

3.8 Implementácia Stitch ETL

Táto kapitola bude venovaná konfigurácii nástroja Stitch, ktorý bude slúžiť ako dátový kanál medzi dátovými zdrojmi a dátovým skladom v Google BigQuery. Tento nástroj bol vybratý z dôvodu lepšej cenovej dostupnosti, väčšej škály natívnych integrácií na zdrojové systémy a väčšej škály destinácii dát. K dispozícii je pre testovanie 14 dňový trial účet, ktorý umožňuje plnohodnotné odskúšanie celého nástroju. Po založení účtu a vyplnení všetkých povinných informácii nasledujú tri fázy:

- Napojenie integrácie.
- Napojenie destinácie.
- Vykonanie prenosu.

3.8.1 Integrácia s Google Ads

Google Ads integrácia replikuje dáta za použitia Google AdWords API (v201809). Na použitie danej API je potrebné, aby boli všetky reklamné účty pripojené na „My Client Center account“, v skratke MCC. Tento účet zaisťuje správu niekoľkých účtov pod jedným prihlásením.

Po vybratí integrácie zo zoznamu dostupných zdrojových systémov, je potrebné zadať meno integrácia, zvoliť históriu synchronizovania dát a frekvenciu replikácie dát. Dĺžka histórie synchronizovania dát určuje, za akú dlhú dobu sa budú dáta zo zdrojovej integrácie s Google Ads replikovať do dátového skladu. Štandardná prednastavená dĺžka histórie je jeden rok. V prípade dát zo Smartlook bola ale zvolená 2 ročná história synchronizovania dát. Frekvencia replikácie určuje to, ako často bude Stitch replikovať dáta z integrácie do dátového skladu. Štandardná prednastavená dĺžka frekvencie replikácie je 24 hodín, čo je pre potreby daného dátového skladu vyhovujúce. Z toho dôvodu daná dĺžka frekvencie replikácie nebude menená. Po základných nastaveniach nasleduje voľba všetkých účtov v Google Ads, ktorých dáta majú byť replikované do dátového skladu. Ak test spojenia prebehne úspešne, nasleduje voľba tabuliek a atribútov, ktoré budú replikované.

Po úspešnej konfigurácii integrácie Google Ads, nasleduje napojenia Stitchu na dátový sklad. K dispozícii je natívna integrácia s Google BigQuery. Podmienkou pre úspešné prepojenie je mať BigQuery account s administratívnymi právami a projekt, v ktorom je povolená fakturácia.

Prvým krokom je vytvorenie Service account, v Google cloud platform konzole. Pri vytváraní účtu je dôležité zvoliť rolu účtu správne -> BigQuery Admin. Ďalším krokom je voľba povolení pre účty, ktoré budú zapojené v procese spracovávania dát. Posledným

krokom je vytvorenie bezpečnostného kľúča vo formáte JSON. Tento kľúč je potrebný pre finalizáciu prepojenia dátového skladu a dátového kanálu.

3.8.2 Integrácia s Google analytics

Google analytics integrácia využíva na replikáciu dát Google Analytics Reporting API v4. Pre túto integráciu platí limit na maximálne 10 metrík a 7 dimenzií v jednom reporte. História synchronizácie dát je štandardne nastavená na 30 dní a frekvencia replikácií je štandardne nastavená na 6 hodín. V prípade daného dátového skladu sú tieto štandardné hodnoty vyhovujúce, preto nebudú menené. Ďalším krokom konfigurácie je výber metrík a dimenzií, ktoré budú exportované a následne importované pomocou Stitchu do dátového skladu.

Na základe tabuľky, do ktorej sa budú importovať dáta zo zdroja boli vybrané dve metriky (GA: new visitors a GA: visitors). Z pomedzi dimenzii sa budú exportovať dimenzie date, sourcemedium, adwordsAdGroupID, campaign a countryISO. Týmto krokom je integrácia Google analytics dokončená.

3.8.3 Integrácia služby Intercom

Intercom integrácia replikuje dáta za použitia Intercom REST API (v1.0). Na rozdiel od Google analytics a Google Ads, táto integrácia nespadá do free plánu, ale je treba mať zakúpenú licenciu. História synchronizácie dát je štandardne nastavená na 1 rok, frekvencia replikácií je štandardne nastavená na hodinovej báze. Obe tieto hodnoty sú vyhovujúce, preto sa pokračuje s štandardne predvolenými hodnotami. Konfigurácia pozostáva z voľby mena a vloženia Intercom app ID hodnoty, ktorá je obsiahnutá v javascriptovom kóde nasadzovanom na web. Posledným krokom je potvrdenie povolenia prístupu ku všetkým požadovaným operáciám v Intercome.

3.8.4 Integrácia so zdrojovou databázou

Poslednou integráciou je napojenie Stitch ETL na niektoré dôležité tabuľky zo zdrojovej databázy, ako je napríklad tabuľka obsahujúca dáta o klientoch. Zdrojová databáza beží na platforme Amazon S3. Export dát prebieha vo forme CSV súborov z tabuliek v databáze. Zo základných informácií je potrebné definovať bucket, v ktorom sú dáta, potrebné k exportu a AWS account ID.

Po vyplnení základných informácií je potreba konfigurácie spracovávaných tabuliek. Je treba zadefinovať názov tabuľky, z ktorej sa bude vykonávať export, aké všetky atribúty a entity je potrebné z tabuliek exportovať, ktorý atribút tvorí primárny kľúč a aký je znak oddeľovača jednotlivých záznamov v riadku. Frekvencia replikácií je štandardne stanovená na jednu hodinu a história synchronizácií je štandardne stanovaná na jeden rok.

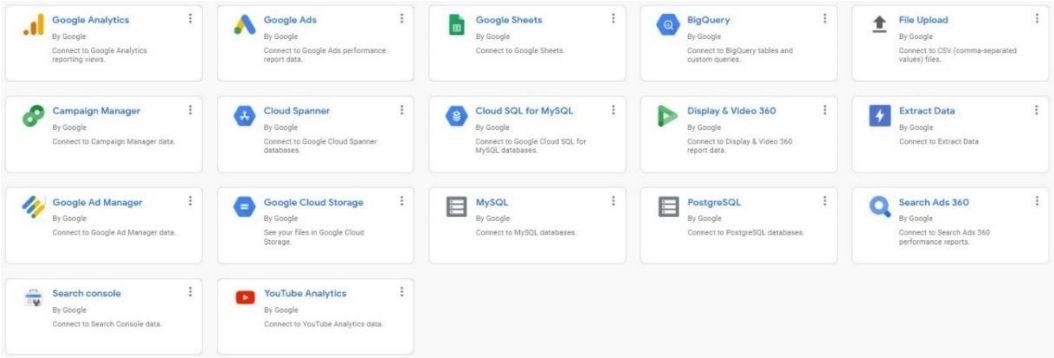
3.9 Integrácia Google BigQuery s Google Data studio

Takmer posledným krokom procesu budovania dátového skladu je napojenie reportovacieho nástroja. Z integračných dôvodov som zvolil nástroj z portfólia firmy Google.

Google Data Studio je nástroj umožňujúci vytváranie reportov pomocou vizualizácie dát. Poskytuje možnosť pripojenia rôznych dátových zdrojov, vďaka čomu sa eliminuje potreba plánovania periodického obnovovania dát pre reporty. Ďalším benefitom je možnosť voľby časového rozsahu vizualizovaných dát. K dispozícii sú nasledovné dátové pripojenia:

- Google Analytics
- Google Adwords
- MySQL
- PostgreSQL

Dáta sa do dátového štúdia môžu tiež nahrávať priamo z dátových zdrojov alebo z analytickej databázy. K dispozícii je natívna integrácia s Google BigQuery.



Obrázok 44: Google BigQuery dátové zdroje

(Zdroj: 21)

Pokiaľ jeden účet spravuje aj dátový sklad v Google BigQuery, aj reportovací nástroj Google Data Studio, proces prepojenia je veľmi jednoduchý a interaktívny. Po zvolení integrácie je potrebné zvoliť správny projekt, v mojom prípade Smartlook analytics DW, následne je potrebné vybrať dátový set AnalyticsDB a posledným krokom je voľba tabuľky, nad ktorou sa bude report vypracovávať. Potom čo je dátový zdroj napojený, je potrebné vybrať entity, ktoré budú v reporte zahrnuté. Keďže reportovacie tabuľky pozostávajú z už skomponovaných entít, nad dátovým zdrojom je možné vytvoriť prvý report.

3.10 Tvorba reportu

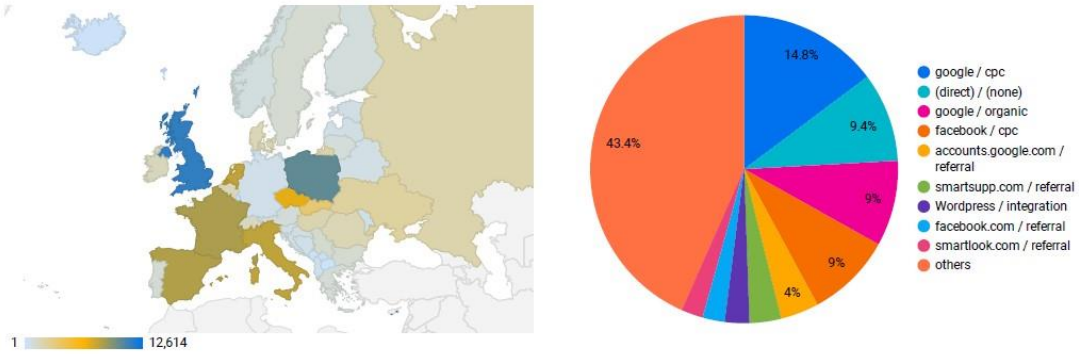
Posledným krokom celého procesu je tvorba finálneho reportu pre potvrdenie funkčnosti celej štruktúry a kompletnosti dátového skladu. Pre demonštráciu bol vytvorený report v Google data studiu nad dátami z Google Adwords a Google analytics. Po napojení všetkých dátových zdrojov, teda v našom prípade reportovacích tabuliek v dátovom sklade AnalyticsDB, je na výber z veľkého množstva grafických prvkov pre zobrazenie dát.

Prvý report obsahuje náklady za vedenie marketingových kampaní rozdelených podľa skupiny reklám a krajiny, na ktorú bola kampaň zacielená. Vo forme koláčového grafu sú zobrazené jednotlivé zdroje návštevností s top 6timi zobrazenými v percentách. Vedľa koláčového grafu je viditeľná mapa európskych krajín, z ktorých pochádzal najväčší počet návštev. Čím je krajina tmavšia, tým viacej návštev z danej krajiny pochádzalo.

Marketing CPC campaign performance

	Date	TrafficSourceMedium	Campaign	Adgroup	Country	MarketingCostsUSD
1.	29 Jan 2019	google / cpc	UK - Display CIA Mobile	CIA - PlaytestCloud	United Kingdom	160.02
2.	16 Jan 2019	google / cpc	UK - Display CIA Mobile	CIA - G2Crowd Mobile App...	United Kingdom	96.24
3.	3 Feb 2019	google / cpc	UK - Display CIA Mobile	CIA - Unity Mobile	United Kingdom	86.52
4.	2 Feb 2019	google / cpc	UK - Display CIA Mobile	CIA - Unity Mobile	United Kingdom	79.41
5.	16 Jan 2019	google / cpc	DE - Display CIA Mobile	CIA - G2Crowd Mobile App...	Germany	77.22
6.	23 Jan 2019	google / cpc	UK - Display CIA Mobile	CIA - Login Chartboost	United Kingdom	75.16
7.	19 Mar 2019	google / cpc	UK - Display CIA Mobile	CIA - Unity Mobile	United Kingdom	72.68
8.	16 Jan 2019	google / cpc	UK - Display CIA Mobile	CIA - Login UXCam	United Kingdom	71.35
9.	24 Jul 2019	google / cpc	Global - Competitors - Website Anal...	Inspectlet	Canada	70.74
10.	24 Jan 2019	google / cpc	UK - Display CIA Mobile	CIA - Login Chartboost	United Kingdom	69.92

1 - 100 / 197137



Obrázok 45: Google Data Studio report

(Zdroj: 22)

Druhý report obsahuje počty nových a vracajúcich sa návštevníkov z jednotlivých zemí ďalej delených na dimenzie podľa zdroja návštevy, konkrétnej kampane a skupiny reklám. Pod tabuľkou je zobrazený graf zobrazujúci krajiny s najväčším počtom nových a vracajúcich sa návštevníkov.

Obrázok 46: Google Data Studio report 2

(Zdroj: 22)

3.11 Zhodnotenie projektu

Cieľom projektu bolo vytvorenie dátového skladu primárne pre dáta z marketingových aktivít a pre potreby marketingu s ohľadom na výsledné reportovanie. Tento cieľ bol dosiahnutý. Počiatočným krokom bolo pripravenie zdrojovej databázy, ktorá obsahuje všetky dáta. Po analýze nástrojov vytvárajúcich dátové kanály a služieb dátových skladov bol vybratý ETL nástroj Stitch na zabezpečovanie spracovávania dát a Google BigQuery ako služba dátového skladu. Následné pripojenie dátových zdrojov cez ETL Stitch na Google BigQuery prebehlo bez problémov. Po pripojení dátového skladu na Google data studio boli vytvorené dva reporty na demonštráciu funkčnosti celej dátovej štruktúry.

Z časového hľadiska trvala činnosť budovania spolu s analýzou riešení a prípravou dát a zdrojovej databázy zhruba dva mesiace. V rozsahu, v ktorom bol dátový sklad budovaný nebolo potrebné na vybudovanie využiť žiadne finančné zdroje a teda jediný náklad, ktorý budovanie dátového skladu sprevádzalo bol náklad na ľudskú prácu. Pri

využití ETL riešenia Stitch sa počet riadkov pohyboval pod 5 miliónov za mesiac, čo spadá v cenníku danej služby do plánu poskytovaného zadarmo. Čo sa týka využívania kapacít v Google BigQuery, tam bolo tiež možné pri budovaní využiť zdroje, ktoré táto služba poskytuje zadarmo.

Ak by bolo rozhodnuté o tom, že vybudované riešenie sa bude rozširovať pre potreby ďalších oddelení spoločnosti, čím by sa zvýšil objem spracovávaných dát, tak by bolo potrebné na projekt vyčleniť finančné zdroje. Kalkulácia bez presnej definície dát a integrovaných dátových zdrojov nie je možná a projekt by bolo treba značne rozšíriť a rozpracovať. Pri integrácii väčšej časti dátovej infraštruktúry do daného dátového skladu by bolo tiež treba dátového inžiniera. Tento projekt je teda možné pokladať za funkčný prototyp s potenciálom ďalšieho rozvoja k vybudovaniu kompletného riešenia na celofiremnej báze.

ZÁVER

Cieľom tejto diplomovej práce bol návrh a implementácia dátového skladu v startupovej spoločnosti vyvíjajúcej SaaS produkt. Keďže v spoločnosti už existovala základná dátová infraštruktúra, projekt návrhu a implementácie dátového skladu bol predovšetkým zameraný na dáta pre marketingové oddelenie spoločnosti s potenciálom ďalšieho rozširovania pre zvyšné oddelenia spoločnosti. Spracovávanie, transformácia a ukladanie dát bolo zautomatizované pomocou ETL nástroja Stitch. Samotný dátový sklad bol vytvorený na platforme Google BigQuery. Finálne reportovanie zabezpečuje Google Data Studio.

V prvej časti diplomovej práce som popísal teoretické východiská celej práce, tak aby boli čitateľovi jasné výrazy a princípy používané v analytickej a návrhovej časti. Teoreticky boli popísané základy dátových skladov, Business Intelligence, ETL proces spolu s najčastejšie používanými dátovými formátmi. Popísaný bol tiež proces informačného managementu, základy cloud computingu a v rýchlosti boli zhrnuté základné funkcionality Smartlook nástroja.

V analytickej časti som sa venoval analýze súčasného stavu dátovej infraštruktúry v spoločnosti. Keďže boli pri návrhu a následnej implementácii dátového skladu používané existujúce riešenia, či už pre ETL proces, ale aj pre samotný dátový sklad a reportovanie, venovaná bola značná časť analýzy možných alternatívnych riešeniam. V skratke bola tiež opísaná samotná spoločnosť a dáta spolu s dátovými zdrojmi, ktorými bol následne dátový sklad plnený.

Návrhová časť bola následne zameraná na implementáciu dátového skladu. Popísaná bola navrhovaná architektúra dátového skladu, jednotlivé tabuľky faktov a dimenzií, ktoré tvoria dátový sklad. Bol popísaný proces tvorby dátového modelu v Google BigQuery, na ktorý boli napojené dátové pumpy ETL softwaru Stitch. Konfigurácia všetkých nástrojov na ETL software bola zakončená prenosom požadovaných dát do dátového skladu. Finálna vizualizácia bola vykonaná pomocou nástroja Google Data Studio. Ako demonštrácia funkčnosti celého riešenia boli vypracované dva reporty nad dátami v dátovom sklade.

ZOZNAM POUŽITÝCH ZDROJOV

(1) LABERGER, Robert. *Datové sklady: Agilní metody a business inteligence*. Brno: Computer press, 2012. ISBN 978-80-251-3729-1.

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

(3) HOLUBOVÁ, Irena, Jiří KOSEK, Karel MINAŘÍK a David NOVÁK. *Big Data a NoSQL databáze*. Praha: Grada Publishing, 2015. ISBN 978-80-247-5466-6.

(4) POUR, Jan, Miloš MARYŠKA a Ota NOVOTNÝ. *Business Intelligence v podnikové praxi*. Praha: Professional Publishing, 2012. ISBN 978-80-7431-065-2.

(5) GEMIGNANI, Zach, Chris GEMIGNANI, Richard GALENTINO a Patrick SCHUERMANN. *Efektivní analýza a využití dat*. Brno: Computer Press, 2015. ISBN 978-80-251-4571-5.

(6) *About us* [online]. Brno, 2018 [cit. 2020-01-15]. Dostupné z: https://www.smartlook.com/about/

(7) AMAZON, Web Services. *Amazon Simple Storage Service: Developer Guide* [online]. API Version 2006-03-01. [cit. 2020-01-15].

(8) Business Intelligence Tools Market. In: *Bandera County Courier* [online]. San Francisco, 2020 [cit. 2020-05-11]. Dostupné z: https://www.bccourier.com/how-business-intelligence-tools-market-will-affect-the-growth-by-2028-with-top-leading-players-competitors-app-cluvio-manageengine-manta-board-adjust-answerdock-birch-grove-software/

(9) *Smartlook.com* [online]. Brno, 2020 [cit. 2020-05-11]. Dostupné z: https://app.smartlook.com

(10) Choosing a Destination. *Stitch Docs* [online]. Redwood City, 2020 [cit. 2020-05-11]. Dostupné z: https://www.stitchdata.com/docs/destinations/choosing-a-stitch-destination

(11) Data Typing. *Stitch Docs* [online]. Redwood City, 2020 [cit. 2020-05-11]. Dostupné z: https://www.stitchdata.com/docs/replication/data-typing

(12) Pricing. *Stitch* [online]. Redwood City, 2020 [cit. 2020-05-11]. Dostupné z: https://www.stitchdata.com/pricing/

(13) Core Concepts. *Fivetran* [online]. Oakland, 2020 [cit. 2020-05-11]. Dostupné z: https://fivetran.com/docs/getting-started/core-concepts

(14) Destinations. *Fivetran* [online]. Oakland, 2020 [cit. 2020-05-11]. Dostupné z: https://fivetran.com/docs/destinations

(15) Pricing. *Fivetran* [online]. Oakland, 2020 [cit. 2020-05-11]. Dostupné z: https://fivetran.com/pricing

(16) Specifying a schema. *Google Cloud* [online]. Mount View, 2020 [cit. 2020-05-11]. Dostupné z: https://cloud.google.com/bigquery/docs/schemas

(17) BigQuery pricing. *Google Cloud* [online]. Mount View, 2020 [cit. 2020-05-11]. Dostupné z: https://cloud.google.com/bigquery/pricing

(18) Amazon Redshift clusters. *Amazon Redshift* [online]. Seattle, 2020 [cit. 2020-05-11]. Dostupné z: https://docs.aws.amazon.com/redshift/latest/mgmt/working-with-clusters.html

(19) AWS Architecture Icons. *Amazon Redshift* [online]. Seattle, 2020 [cit. 2020-05-11]. Dostupné z: https://aws.amazon.com/architecture/icons/

(20) Logos List. *Google Brand Permissions* [online]. Mount View, 2020 [cit. 2020-05-11]. Dostupné z: https://www.google.com/permissions/trademark/logos-list/

(21) Google Cloud Platform. *BigQuery* [online]. Mount View, 2020 [cit. 2020-05-11]. Dostupné z: https://console.cloud.google.com/bigquery

(22) Marketing data report. *Google Data Studio* [online]. Mount View, 2020 [cit. 2020-05-11]. Dostupné z: https://datastudio.google.com/u/2/reporting

(23) Smartlook Knowledge Base. *Smartlook* [online]. Brno, 2020 [cit. 2020-05-11]. Dostupné z: https://help.smartlook.com/en/

ZOZNAM POUŽITÝCH SKRATIEK A SYMBOLOV

SaaS – Software as a Service

PaaS – Platform as a Service

IaaS – Infrastructure as a Service

OLAP – Online analytical processing

OLTP – Online transaction processing

BI – Business intelligence

JS – JavaScript

ETL – Extraction, transform, load

DB – Database

DG – Data governance

CMS – Content management system

URL – Uniform Resource Locator

CSS – Cascading Style Sheets

B2C – Business to customer

B2B – Business to business

SSL – Secure Sockets Layer

RDBMS – Relational Database Management System

API – Application programming interface

CLI – Command-line interface

CPU – Central processing unit

RAM – Random-access memory

HDD - Hard disk drive

SSD – Solid-state drive

ZOZNAM OBRÁZKOV

ZOZNAM TABULIEK