VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA PODNIKATELSKÁ

FACULTY OF BUSINESS AND MANAGEMENT

ÚSTAV MANAGEMENTU

INSTITUTE OF MANAGEMENT

AUTOMATIZACE ANALÝZY DAT V OBLASTI ZÁKAZNICKÉHO ZÁŽITKU V KORPORÁTNÍ SPOLEČNOSTI POMOCÍ NÁSTROJŮ DATOVÉHO MODELOVÁNÍ

AUTOMATION OF CUSTOMER EXPERIENCE DATA ANALYSIS IN A CORPORATE COMPANY USING DATA MODELING

DIPLOMOVÁ PRÁCE

MASTER'S THESIS

AUTOR PRÁCE — Bc. Pavlína Krmelová

AUTHOR

VEDOUCÍ PRÁCE — Ing. Jiří Kříž, Ph.D.

SUPERVISOR

BRNO 2024

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA PODNIKATELSKÁ

Zadání diplomové práce

Ústav:	Ústav managementu
Studentka:	**Bc. Pavlína Krmelová**
Vedoucí práce:	**Ing. Jiří Kříž, Ph.D.**
Akademický rok:	2023/24
Studijní program:	Strategický rozvoj podniku

Garant studijního programu 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:

Automatizace analýzy dat v oblasti zákaznického zážitku v korporátníspolečnosti pomocí nástrojů datového modelování

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 práce je vybrat a naimplementovat řešení pro zefektivnění analýzy dat v oblasti zákaznického zážitku ve vybrané společnosti.

Základní literární prameny:

DEHGHANI, Zhamak, 2022. Data Mesh. 2. Dehghani. ISBN 9781492092391.

HOBERMAN, Steve, 2009. Data Modeling Made Simple. 2. Technics Publications Publication.ISBN 978-0977140060.

KIMBALL, Ralph a ROSS, Margy, 2011. Data Modeling Essentials. 1. Wiley. ISBN9781118082140.

SIMSION, Graeme; WITT, Graham a WEST, Matthew, 2015. Data Modeling Essentials. 4.Newnes. ISBN 9780123965394.

Termín odevzdání diplomové práce je stanoven časovým plánem akademického roku 2023/24

V Brně dne 6.5.2024

L. S.

doc. Ing. Vít Chlebovský, Ph.D. garant	doc. Ing. Vojtěch Bartoš, Ph.D.děkan

Abstrakt

Diplomová práce se zabývá automatizací analýzy dat v oblasti zákaznického zážitku v korporátní společnosti pomocí nástrojů datového modelování.
Práce je rozdělena do tří hlavních částí. V první části se nachází teoretická východiska pro praktickou část. Je zde popsán význam analytických aplikací, datový modeling a jednotlivé teoretické kroky pro tvorbu datového modelu. V druhé části je analyzován současný proces analýzy dat v podniku. V poslední části práce je popsána tvorba a následná implementace automatizace pomocí datového modelu.

Abstract

The diploma thesis deals with the automation of data analysis in the area of customer experience in a corporate company using data modeling tools.
The thesis is divided into three main parts. The first part provides the theoretical background for the practical part. It describes the importance of analytical applications, data modeling, and the different theoretical steps for creating a data model. The second part analyses the current data analysis process in the enterprise. In the last part of the paper, the creation and subsequent implementation of automation using a data model is described.

Klíčová slova

Data modeling, reporting, implementace datového modelu, SQL, BigQuery, Terraform, Data mining

Keywords

Data modeling, reporting, data model implementation, SQL, BigQuery, Terraform, Data mining

Bibliografická citace diplomové práce

KRMELOVÁ, Pavlína. Automatizace analýzy dat v oblasti zákaznického zážitku v korporátní společnosti pomocí nástrojů datového modelování. Brno, 2024. Dostupné také z: https://www.vut.cz/studenti/zav-prace/detail/160740. Diplomová práce. Vysoké učení technické v Brně, Fakulta podnikatelská, Ústav managementu. Vedoucí práce Jiří Kříž.

Čestné prohlášení

Prohlašuji, že předložená diplomová práce je původní a zpracovala jsem ji samostatně. Prohlašuji, že citace použitých pramenů je úplná, že jsem ve své práci neporušila autorská práva (ve smyslu Zákona č. 121/2000 Sb., o právu autorském a o právech souvisejících s právem autorským).

V Brně dne 6.5.2024

.....

podpis autora

Poděkování

Poděkování patří mému vedoucímu diplomové práce, panu Ing. Jiřímu Křížovi, Ph.D. za cenné rady, a poskytnutou podporu při psaní celé práce. Dále chci poděkovat nejmenovanému nadřízenému ve vybrané firmě, díky němuž jsem mohla získat tolik vzácných zkušeností. Ze všeho nejvíce bych ráda poděkovala své rodině za jejich neúnavnou psychickou i finanční podporu při studiu.

Obsah

1 Úvod

Datové modelování představuje kritický proces v rámci moderního podnikání, neboť se data stávají stále cennějším aktivem organizací. S exponenciálně rostoucím objemem informací, které společnosti generují a shromažďují, se zvyšuje potřeba efektivního řízení, analýzy a využití těchto dat. Aby mohly podniky tato data využít pro informované rozhodování a inovace, právě datové modelování poskytuje základní rámec pro chápání, organizaci a manipulaci a tím otevírá dveře k identifikaci nových příležitostí a zvýšení konkurenční síly na trhu.

První část diplomové práce dává teoretické pozadí pro budoucí implementaci zlepšení. Druhá část diplomové práce se zaměřuje na podrobnou analýzu současného stavu vybrané společnosti, kde hodnotíme stávající datové infrastruktury, procesy a systémy v oblasti zpracování dat zákaznického prožitku. Tento pohled nám umožní identifikovat slabiny a silné stránky v rámci existujících procesů, což je nezbytný krok k navrhování efektivnějších řešení. V třetí části bude identifikace klíčových procesů vhodných pro zlepšení a jejich následná optimalizace prostřednictvím implementace nových datových modelů a analytických nástrojů.

V průběhu této studie bude důraz kladen na přístup k navrhování právě těchto datových modelů, který bude založen na nejlepších praktikách a současných trendech v oboru. Modely budou konstruovány tak, aby byly modulární, snadno rozšiřitelné a schopné se vyvíjet spolu s rostoucími a měnícími se požadavky společnosti.

Výstupy této práce mají představovat nejen teoretickou hodnotu, ale i praktické doporučení, která povedou k reálným změnám v práci s daty uvnitř společnosti. Hlavním výstupem budou zautomatizovaná data v platformě, dle kterých budou business analytici schopni vytvořit rozhodnutí zlepšující efektivitu společnosti.

2 Cíle diplomové práce, metody a postupy zpracování

2.1 Cíle práce

Cílem práce je vybrat a naimplementovat řešení pro zefektivnění analýzy dat v oblasti zákaznického zážitku ve vybrané společnosti. Celý systém by měl být navržen a implementován tak, aby splňoval minimálně tyto požadavky:

Automatizace – Datový model by měl být postaven tak, aby automaticky plnil tabulky každý den.

Napojení na vizualizační platformu – Model by měl být propojitelný s platformou, která bude intuitivní, kontrolovatelná týmem zodpovědným za datovou kvalitu a lekce dostupná napříč společností pro možné zobrazení odkudkoliv.

Přehledné zobrazení – Prostředí aplikace musí být přehledné a informace, které aplikace poskytuje musí být jednoduše čitelné.

2.2 Vymezení problému

Cílem projektu je zefektivnit způsob zpracování konkrétních dat v rámci firmy a vytvořit automatizaci celého procesu analýzy a zjednodušit práci analytikům a zároveň zlepšit rychlost databáze. Čili problémem je chybějící automatizovaný systém pro konkrétní data.

Řešením bude sestavení datového modelu, který firma potřebuje pro automatizaci selekce dat, a také vybrání vizualizačního nástroje, který pomůže při zobrazení statistických analýz a dashboardů.

V praxi by měl systém pomoci hlavně analytikům ušetřit čas, a to především v každodenních výběru dat, následného exportu do tabulek. Pomůže přehledně zobrazovat požadované informace, a v neposlední řadě pomoci šetřit náklady vynaložené na zatížení databáze.

Jak vyplývá již z názvu práce, bude se věnovat především návrhu a implementaci nového řešení automatizace / datového modelu do firmy.

2.3 Metodika

Pro analýzu současného stavu bude využit induktivní přístup a přesněji výzkumná metoda nepřímého zprostředkovaného pozorování. Tato metoda je ideální vzhledem k cíli práce a přístupu vyhodnocení současného stavu. V této práci se bude vycházet ze znalostí procesů ve firmě, také z dokumentace vybrané firmy. Postupným popisem procesu bude popsáno jádro problému a následné navržení řešení. Bude využita SWOT analýza, která pomáhá identifikovat silné a slabé stránky společnosti, stejně jako příležitosti a hrozby v dané problematice. Pro sběr dat, co se týče performování databáze, či popisu délky procesů, bude využito manuálního sběru dat z databáze firmy, respektive bude prováděn přímý test zahlcení databáze a následné vyhodnocení pomocí popisu, či matematicko-statistických metod. Budou také identifikovány rizika a pomocí matice rizik vybrána rizika s největším dopadem a pravděpodobností. Pro tyto rizika bude vytvořen plán minimalizace.

Terénní poznámky mají tutéž funkci jako záznamových arch u standardizovaného pozorování (podrobněji viz. Hendl 2008, str. 197)

3 Teoretická východiska práce

3.1 Analytické úlohy v řízení společnosti

Analýza dat je v dnešní době základním kamenem každé společnosti pro informované rozhodování a strategické plánování. V dnešním světě založeném na datech může schopnost efektivně analyzovat a interpretovat data znamenat rozdíl mezi úspěchem a neúspěchem společnosti. Jako všechno, i analýza dat se v dnešním světě modernizuje a s množstvím informací je to jedním z nejdůležitějších aspektů. Je proto nutné mít kvalitní zabezpečená data, investovat do výzkumu a mít automatizované procesy, zvláště v korporátním světě, kde každé rozhodování se děje na základě dat. Dobrou kvalitu dat zajišťuje nejen platformní pozadí, které správně propojuje data z backendu do databáze, ale také datové modelování, který staví datové modely.

3.2 Definice datového modelování

Datový model popisuje informace uspořádaným způsobem, který umožňuje jejich efektivní ukládání a vyhledávání v relačním systému pro správu databází (RDBMS), jako je SQL Server, MySQL nebo Oracle. Model si lze představit jako způsob, jak převést logiku přesně souvisejících věcí v reálném světě a vztahů mezi nimi do pravidel, která lze dodržovat a vynucovat počítačovým kódem. Metodiky datového modelování existují od počátku počítačové techniky. Data potřebují strukturu, která poskytne počítači způsob, jak data pochopit a zpracovat jejich bity a bajty. Samozřejmě dnes pracujeme i s nestrukturovanými a polostrukturovanými daty, ale pro mě to znamená, že jsme se vyvinuli do sofistikovanějšího paradigmatu, než byla předchozí výpočetní technika. Proto datový model zůstává a tvoří základ pro budování sofistikovaných podnikových aplikací. (Kang, Yi, Liu, 2014)

Když se v souvislosti s daty setkáme s pojmy jako modelování a návrh, máme na mysli logický návrh a fyzickou implementaci databáze. Datové modelování může znamenat dokumentování návrhu softwaru a podnikových systémů. Diagramy, symboly a textové odkazy se používají k znázornění toku dat v rámci organizace nebo softwarové aplikace. Datové modelování se také podílí na rozhodovacích a koncepčních obchodních procesech celé organizace. Model určuje logickou strukturu a styl, jakým jsou data organizována, jak

se k nim přistupuje nebo jak se s nimi manipuluje. Organizace mají v současné době zájem shromažďovat velké objemy dat o svém podnikání, ukládat tato data do úložišť a používat analytiku k rozhodování. Modelovat data je možné různými způsoby. Toto modelování zcela závisí na struktuře dat aplikace a na požadavcích aplikace. Dobrý datový model umožní snadné úložiště a také umožní dobrý výkon. Požadavky uživatelů / analytiků jsou hlavním faktorem pro vývoj logického datového modelu a ten se pak promítne do fyzického datového modelu. Součástí datového modelování je často identifikace zdrojů dat.

Za poslední čtvrtstoletí prošlo datové modelování (DM) hlubokými změnami, zejména co se týče jeho cílů a metod použití. Zpočátku bylo datovém modelování zaměřeno výhradně na podporu návrhu databázových systémů, což bylo podpořeno pokrokem v databázových technologiích. V současné době je v literatuře o manažerských informačních systémech (MIS) chváleno jako osvědčený přístup ke strukturování dat a je považováno za kritický základ podnikového modelování, přičemž jeho přínosy jsou všeobecně uznávány a jen zřídka zpochybňovány (Kang, Yi, Liu, 2014).

3.2.1 Historie datových modelů

Historie datových modelů nás vede až do období rané výpočetní techniky, kdy byla data ukládána na magnetické pásky nebo diskové jednotky a vyhledávání probíhalo pomocí jednoduchých záznamů. V roce 1958 představili J. W. Young a H. K. Kent nový přístup k modelování informačních systémů. V roce 1959 byl založena komise jazyků datového systému CODASYL, což vedlo k rozvoji programovacího jazyka COBOL a konceptu "integrovaného datového skladu" (IDS) Charles Bachmana v 60. letech. Tento přístup se dále rozvinul v IDMS ve společnosti B.F. Goodrich. (Maier, 1996)

Základní přístupy ve výpočetní technice jako je "hierarchický datový model" a "síťový datový model", dominovaly po desetiletí a stále se používají. E. F. Codd a C. J. Date z IBM vyvinuli v 70. letech relační model dat, který Codd rozšířil v roce 1985 o "Dvanáct pravidel relačního modelu". Tento model zavedl koncepty jako "normalizace" a "třetí normální forma" (3NF). (Maier, 1996)

V roce 1996 Ralph Kimball představil "hvězdicové schéma", což bylo rozšíření konceptu datového skladu W. H. Inmona. Dan Linstedt v roce 2001 představil Data Vault metodiku, která kombinovala přístupy Inmona a Kimballa a zahrnovala historická data. Vylepšení tohoto přístupu, Data Vault 2.0, představené v roce 2013, řešilo výzvy spojené s velkými daty a NoSQL. Cesta datového modelování tak pokračuje, přizpůsobuje se a vyvíjí podle požadavků neustále se měnícího technologického prostředí. (Maier, 1996)

3.2.2 Fáze datového modelování

Každý datový model se skládá z několika fází, které je potřeba projít, aby datový model byl optimalizován a dobře fungující. Dwivedi ve své studii z roku 2022 definovala 4 fáze modelování. (Dwivedi, Chourasiya, 2022)

Konceptuální

Jedná se o první krok v procesu modelování, který dává teoretický řád datům. Těmto modelům se také říká doménové modely a obvykle se používají ke zkoumání doménových konceptů se zúčastněnými stranami projektu. Jaké jsou potřeby, jaká jsou data. Konceptuální modely popisují, jaké entity existují a jaké jsou jejich vztahy, a tvoří tak zaměření a rozsah komponenty datové architektury. V tradičních týmech se konceptuální datové modely často vytvářejí jako předchůdce logických datových modelů nebo jako alternativa k logickým datovým modelům. Je to prakticky rešerše před postavením datového modelu.

Logické

Na základě struktury vytvořené v konceptuální fázi se proces logického modelování pokouší zavést řád vytvořením klíčových hodnot a vztahů v logické struktuře. Jedná se o popis tabulek, prvotní návrh struktury sloupů, dokument popisující více konkrétně potřeby a následné řešení těchto potřeb třetích stran, pro které se datový model buduje. Může se do této fáze přiřadit a terraformování tabulek, které slouží k definici tabulek, vytvoření indexů a sjednocenosti formátu s ostatními modely v databázi.

Fyzické

Tento krok rozděluje data do skutečných tabulek, clusterů a indexů (partitioning) potřebných pro datové úložiště. Fyzické datové modely slouží k návrhu vnitřního schématu databáze, zobrazují datové tabulky, datové sloupce těchto tabulek a vztahy mezi tabulkami. V této fázi se definuje DAG, který spouští skripty v nastavenou hodinu.

Ačkoli logický datový model a fyzický datový model znějí velmi podobně a ve skutečnosti tomu tak je, úroveň detailů, které modelují, se může výrazně lišit. Je to proto, že cíle každého diagramu jsou odlišné – logický datový model je možné použít k prozkoumání koncepcí domény se zúčastněnými stranami a fyzický datový model k definování návrhu databáze. Složitost se zvyšuje od konceptuálního modelu přes logický model až po fyzický model. Proto vždy začínáme konceptuálním datovým modelem, abychom na vysoké úrovni pochopili, jaké jsou jednotlivé entity v našich datech a jak spolu souvisejí. Poté přejdeme k logickému datovému modelu, abychom pochopili detaily našich dat, aniž bychom se starali o to, jak budou skutečně implementovány, a nakonec k fyzickému datovému modelu, abychom přesně věděli, jak náš datový model implementovat ve zvolené databázi. (Dwivedi, Chourasiya, 2022)

Existuje mnoho možných vizuálních reprezentací datových modelů, ale primární, která se dnes při návrhu databází používá, je klasický model entit a vztahů. Jedná se jednoduše o schéma políček, která popisují entity s jejich doprovodnými datovými body uvnitř, a čar mezi políčky, které popisují vztahy mezi entitami. (Dwivedi, Chourasiya, 2022)

3.3 Trendy v datovém modelování

Datové modelování je rychle rozvíjející se odvětví, proto za posledních pár let vnikla řada trendů, na které se budou dle Michelle Knight v roce 2024 data-driven společnosti zaměřovat.

3.3.1 Rozšíření modelů specifických pro dané odvětví

Datové modely se zabývají tím, co je třeba vytvořit a jak. Aby se společnosti dostaly k relevantnímu řešení, zdokonalují se v modelování svých doménových dat. Znalostní obory, jako jsou finance, medicína nebo právo, vyjadřují různé potřeby v oblasti datového modelu (Aiken, 2023). Každé odvětví má navíc specifickou, konzistentní terminologii a pojmy nezbytné pro podnikání. Společnosti potřebují tyto jemnosti zachytit prostřednictvím hotových datových modelů a šablon, které jsou snadno dostupné pro použití v komponentách datové architektury. Tímto způsobem organizace ušetří čas, který by musely věnovat opětovnému vytváření standardních entit a vztahů datového modelu, které jsou nedílnou součástí jejich obchodních odvětví. Místo toho mohou věnovat více času pochopení a odsouhlasení modelování svých konkrétních služeb a definování svých obchodních pravidel. Tento trend směrem k modelům pro konkrétní odvětví bude v roce 2024 a v dalších letech rychle narůstat, protože společnosti chtějí efektivnější způsob přístupu k datům bez zbytečné práce navíc. (Knight, 2023)

3.3.2 Větší využívání konceptuálního modelování

Se zaměřením na doménové datové modelování a s oživením zájmu o zlepšování kvality dat se budou organizace stále více obracet ke konceptuálním modelům vysvětlené v kapitole 3.2.2 Fáze datového modelování. Prostřednictvím konkretizace koncepčního datového modelu se budou obchodní a technologické týmy vzájemně zapojovat do vytváření společného slovníku a sladění toho, jakou datovou infrastrukturu aktualizovat nebo vybudovat. V ideálním případě budou firmy navazovat na logický datový model, aby formalizovaly implementaci dohodnuté konceptualizace; vzhledem k tlaku na plnění úkolů se však mnoho firem bude snažit v roce 2024 logické modely vynechat nebo jim věnovat o něco méně času.

3.3.3 Větší obliba znalostních grafů

Přestože datové modelování má mnoho formátů, přičemž oblíbené jsou stále diagramy vztahů mezi entitami, relační diagramy a diagramy toku dat, je očekávené, že na vrcholu seznamu se objeví znalostní grafy, vizualizace entit a jejich vztahů. Společnosti se potýkají se zkráceným časovým rámcem pro získání použitelných datových modelů, chtějí okamžité poznatky a pracují se stále více nestrukturovanými daty. Znalostní grafy poskytují nástroje pro zvládnutí všech tří požadavků. Znalostní grafy, které jsou zvláště skvělé pro konceptuální modelování dat, umožňují pochopit řešení, které je třeba vytvořit, související a relevantní faktory, které je třeba vzít v úvahu, a zahrnují metadata a kontext kolem dat. Znalostní graf navíc sleduje změny v závislosti na vývoji dat a metadat.

3.3.4 Lepší samoobslužné platformy

S rozšířením modelů specifických pro dané odvětví a zvýšeným používáním konceptuálních datových modelů budou lidé z byznysu využívat a vyžadovat lepší samoobslužné platformy, aby mohli experimentovat s datovými modely prostřednictvím interaktivních vizualizací a aktivně se zapojit do rozhovorů s technologickými týmy. Snadná dostupnost datových sad prostřednictvím cloud computingu a tlak na včasnější a informovanější rozhodování na základě dat v reálném čase navíc přiměje podniky k aktualizaci a vytváření datových modelů za běhu bez předchozí konzultace s technologickými týmy.

3.4 Data mining

Data mining neboli dolování dat v překlady, je novodobý proces, který používá většina společností s velkým množstvím dat čili i vybraná firma. Data mining využívá algoritmy a různé další techniky k převodu velkých souborů dat na užitečné výstupy. Pro dosažení maximální efektivity v procesu dolování dat, datoví analytici obvykle dodržují specifickou sérii kroků. Tento strukturovaný přístup pomáhá předcházet problémům, které by mohly být jinak v průběhu analýzy přehlédnuty. První krok spočívá ve zvážení kontextu projektu a cílů, kterých chce společnost dosáhnout pomocí dat. Po definování obchodního problému následuje zkoumání dostupných datových zdrojů, jejich zabezpečení, uložení a metody sběru. Tento krok také zahrnuje posouzení limitů dat a toho, jak tyto limity ovlivní celkový proces. Ve třetím kroku se data shromažďují, nahrávají a extrahují. Dále se data čistí, standardizují, odstraňují se odlehlé hodnoty, kontrolují se na přítomnost chyb a hodnotí se jejich přiměřenost. Tento krok zahrnuje také zvážení velikosti dat, jelikož příliš velká množství dat mohou zpomalit analýzu. S čistým souborem dat analytici používají různé metody dolování dat pro identifikaci vzorců, trendů a asociací. Data mohou být také začleněna do prediktivních modelů pro hodnocení budoucích trendů. Vyhodnocení výsledků bude dalším krokem, který zahrnuje sumarizaci, interpretaci a prezentaci zjištění z datových modelů pro rozhodovací orgány, které byly dříve z procesu vyloučeny. Zde se organizace rozhoduje, jak na základě těchto zjištění jednat. Závěrečný krok spočívá v realizaci opatření na základě výsledků analýzy. Společnost může provést strategické změny nebo určit, že zjištění nejsou dostatečně relevantní. Vedení pak hodnotí dopady na podnik a identifikuje nové obchodní výzvy nebo příležitosti pro budoucí cykly dolování dat (1. Anon, 2023).

3.5 Terraform

Pro ideální definici tabulek, což je prvním krokem pro sestavení datového modelu je ideální použití terraformu. Terraform je populární open-source nástroj pro správu infrastruktury jako kódu. Mezi hlavní výhody používání terraformu patří například umožnění týmům definovat a spravovat infrastrukturu pomocí správy verzí, což usnadňuje spolupráci více lidí a práci na stejné kódové základně. To může pomoci zlepšit spolupráci a snížit riziko chyb. Další výhodou je automaticky udržování historie verzí infrastruktury, což usnadňuje návrat k předchozím verzím v případě potřeby. To může pomoci chránit před chybami a zajistit rychlé zotavení z chyb. Umožňuje také definovat infrastrukturu pomocí konfiguračního jazyka vysoké úrovně, což znamená, že můžete konzistentním a předvídatelným způsobem specifikovat požadovaný stav infrastruktury. Je možné také definovat infrastrukturu jako modulární komponenty, které lze snadno opakovaně použít při více nasazeních. To může pomoci omezit duplicitu a usnadnit správu infrastruktury v měřítku (2.Anon, 2023).

Celkově terraform nabízí řadu výhod pro správu infrastruktury jako kódu. Stal se povinným nástrojem pro nastavení reprodukovatelné multicloudové infrastruktury. Níže na obrázku č.1 je zobrazena struktura terraformování tabulky, která bude implementována v řešení.


```
[
  {
    "name": "survey_fwid",
    "mode": "REQUIRED",
    "type": "INTEGER",
    "description": "Unique identifier of the fact table row"
  },
  {
    "name": "event_id",
    "mode": "NULLABLE",
    "type": "STRING",
    "description": "The unique identifier of the log in event_logs_nested."
  },
  {
    "name": "visitor_id",
    "mode": "NULLABLE",
    "type": "STRING",
    "description": "The unique identifier of the cookie (identifier of a browser)"
  },
  {
    "name": "session_id",
    "mode": "NULLABLE",
    "type": "STRING",
    "description": "The unique identifier of the front-end session. Session is created everytime customer reloads the page or tab."
  },
```


Obrázek 1: Příklad struktury terraformu (vlastní zpracování)

3.6 BigQuery

BigQuery je datový sklad bez serveru od Google Cloud, speciálně navržený pro rozsáhlou analýzu dat. Je to výkonná platforma, která uživatelům umožňuje analyzovat rozsáhlé datové soubory, často v řádu miliard řádků, pomocí známého jazyka ANSI SQL s působivou rychlostí. Tento nástroj se využívá především pro datové sklady a nabízí centralizované úložiště pro různé formy dat. Výraznou vlastností BigQuery je integrace možností strojového učení prostřednictvím BigQuery ML, která uživatelům umožňuje vytvářet a spouštět modely přímo v databázi pomocí dotazů SQL v konzoli. BigQuery navíc podporuje analýzu v reálném čase, což umožňuje okamžitou analýzu streamováních dat pro získání aktuálních poznatků, což je užitečné zejména pro podniky, které potřebují dělat rychlá rozhodnutí založená na datech. (dokumentace Google Cloud, 2023)

Jednou z hlavních výhod BigQuery je jeho bezserverová architektura, která odstraňuje potřebu správy serverů a zjednodušuje tak infrastrukturu. Tato bezserverová povaha spolu s vysokou škálovatelností znamená, že se dokáže automaticky přizpůsobovat měnícím se objemům dat a zatížení dotazů, takže je vysoce efektivní pro zpracování velkých datových souborů bez nutnosti rozsáhlého plánování kapacity. Další klíčovou vlastností je bezpečnost, která spočívá v šifrování dat v klidovém stavu i při přenosu. Jeho bezproblémová integrace s dalšími službami Google Cloud navíc zvyšuje jeho schopnosti jako komplexního cloudového analytického řešení. (dokumentace Google Cloud, 2023)

V kapitole o fyzické fázi datového modelování byly zmíněné indexy, clustering a partitioning a to je hlavní výhodou BigQuery. Tento datový sklad je možné propojit s Terraformem, kde jsme schopni udělat clusterování, partitioning a indexování, a to představuje integraci pokročilých technik správy dat a automatizace cloudové infrastruktury. Pochopení jednotlivých konceptů umožňuje pochopit, jak společně zlepšují zpracování a správu dat v BigQuery.

Terraform lze v BigQuery použít k automatizaci nastavení a správy zdrojů BigQuery. To zahrnuje vytváření datasetů, tabulek a správu oprávnění. Pomocí Terraformu mohou týmy aplikovat řízení verzí na svou datovou infrastrukturu, zajistit konzistentní nastavení v

různých prostředích (vývojové, stagging, produkční) a zefektivnit proces správy a škálování zdrojů BigQuery. (dokumentace Terraform, 2023)

Clustering označuje uspořádání dat v tabulkách BigQuery na základě obsahu jednoho nebo více sloupců. Toto uspořádání může zlepšit výkonnost dotazů a snížit náklady tím, že umožňuje nástroji BigQuery efektivně prohledávat pouze relevantní podmnožiny dat. Shlukování je výhodné zejména u velkých tabulek a lze jej kombinovat s rozdělením (partitioning) pro ještě efektivnější správu dat. Když je tabulka shlukována, data se automaticky seřadí na základě zadaných sloupců, což může být velmi efektivní pro filtrování a agregační dotazy. Typické je cluterování podle ID tabulky.(3. Anon, 2023)

Paritioning neboli rozdělení zahrnuje rozdělení tabulky na segmenty, tzv. oddíly, které zefektivňují správu dat a dotazování. V BigQuery lze tabulky rozdělit na základě konkrétního sloupce (typická jsou časového razítka) nebo podle rozsahu. Rozdělení na oddíly je nezbytné pro správu velkých datových sad, protože umožňuje efektivnější dotazy díky tomu, že se místo celé tabulky prohledávají pouze příslušné oddíly, čímž se snižují náklady a zvyšuje výkon. (Dokumentace Google BigQuery, 2023)

Je důležité zmínit také velkou výhodu BigQuery, a to jsou sloty. Sloty jsou jednotky výpočetní kapacity v BigQuery. Při spuštění dotazu přidělí BigQuery tyto sloty k provedení dotazu. Počet slotů určuje rychlost, jakou se dotaz provede. Ve sdíleném účtu mají dotazy každého týmu přístup k proměnlivému počtu slotů v závislosti na celkové poptávce. BigQuery nabízí také rezervované sloty, které jsou vyhrazeny pro například datové modely, které mají prioritu, což zajišťuje předvídatelný výkon a izolaci od zátěže ostatních uživatelů. Efektivní správa slotů je klíčem k optimalizaci výkonu a nákladů na dotazy v BigQuery. (dokumentace Google BigQuery, 2023)

Každá z těchto součástí hraje zásadní roli při optimalizaci využití BigQuery pro rozsáhlou analýzu dat. Terraform automatizuje a spravuje infrastrukturu, clustering a partitioning optimalizují výkon dotazů a správu dat a sloty řídí výpočetní zdroje přidělené těmto dotazům. Společně tvoří komplexní rámec pro efektivní a škálovatelnou analýzu dat v cloudu.

3.7 Dagy a jejich úloha při automatizaci

V oblasti modelování a automatizace dat, zejména v prostředích využívajících nástroje jako například Airflow nebo Astronomer, hrají klíčovou roli směrované acyklické grafy, Directed Acyclic Graphs (DAG). DAG je v podstatě konfigurační nástroj, který vymezuje řadu úloh a jejich vzájemných závislostí v rámci automatizovaného pracovního postupu, graficky znázorněných pomocí uzlů symbolizujících úlohy a hran označujících závislosti mezi nimi. Je většinou definován jazykem Python. Tyto DAGy mají zásadní význam pro orchestraci složitých pracovních postupů a umožňují datovým inženýrům přesně definovat pořadí a logiku provádění skriptů a plnění tabulek, čímž zajišťují, že jsou pečlivě dodržovány všechny závislosti. (dokumentace Airflow, 2023)

DAGy jsou nedílnou součástí plánování a automatizace úloh, protože umožňují jejich provádění v předem stanoveném čase nebo za specifických podmínek, což je nezbytné pro automatizaci opakujících se úloh zpracování a analýzy dat. Zahrnují také mechanismy pro zpracování chyb a opakování, které přidávají vrstvu robustnosti do celého procesu zpracování dat tím, že určují akce v případě selhání úlohy, jako je pokus o opakování, odeslání upozornění nebo přeskočení úlohy. (dokumentace Airflow, 2023)

Souhrnně lze říci, že v kontextu datového modelování a automatizace jsou DAGy nezbytné pro definování, plánování a monitorování složitých datových pracovních postupů, které zajišťují efektivní a systematické zpracování a analýzu velkých objemů dat automatizovaným a řízeným způsobem. Níže je příklad definice DAGu.

```
dag = DAG(
    dag_id="survey_model_1d",
    default_args={
        "owner": "název týmu",
        "email": "sem se zadají členové týmu, kteří chtějí dostávat
                  upozornění v případě selhání DAGu",
        "email_on_failure": True,
        "email_on_retry": False,
        "depends_on_past": False,
        "start_date": datetime(2023, 12, 1, 0, 0, 0),
        "sla": timedelta(hours=6),
        "retries": 10,
        "retry_delay": timedelta(minutes=15),
    },
    schedule_interval=timedelta(days=1),
    max_active_runs=1,
```

```
template_searchpath=retrieve_sql_path(__file__),
template_undefined=jinja2.Undefined,
params={
    "src_gcp_project_id": var("src_gcp_project_id"),
    "dst_gcp_project_id": var("dst_gcp_project_id"),
    "src_bq_project": var("src_bq_project"),
    "dst_bq_project": var("dst_bq_project"),
},
```

3.8 Analytické aplikace

Hlavní výhodou datového modelování je následné napojení na analytické aplikace, které pomáhají vizualizovat data. V dnešní době existuje mnoho platforem, které jsou uzpůsobené tak, aby si každý datový analytik mohl zobrazit cokoliv napříč mnoha modely. Každá statistika lze následně dodat i v grafickém znázornění. Vybraná firma využívá platformu Looker Studio, dříve známý jako Google Data Studio. Je to nástroj od Google pro vizualizaci dat, který umožňuje uživatelům vytvářet informační a vizuální dashboardy pro analýzu a sdílení dat.

Mezi hlavní alternativy Looker Studia patří Klipfolio, DashThis, Tableau, DataBox, PowerBI, AgencyAnalytics a Geckoboard. Klipfolio nabízí méně integračních možností (300) ve srovnání s Looker Studiem (600), ale poskytuje funkce white-label a specializované balíčky pro agentury. DashThis umožňuje připojení neomezeného počtu zdrojů a vytváření white-label reportů, ale má méně integračních možností (34) než Looker Studio. Tableau nabízí real-time data a interaktivní dashboardy s pokročilými vizualizačními a analytickými možnostmi, ale vyžaduje expertní znalosti pro efektivní používání. (dokumentace Looker. 2023)

DataBox podporuje mobilní aplikace a nabízí denní a týdenní scorecardy, které Looker Studio nenabízí. (dokumentace DataBox. 2023) PowerBI od Microsoftu se vyznačuje atraktivními vizualizacemi a automatickým obnovením dat z cloudových zdrojů (dokumentace Microsoft. 2023). AgencyAnalytics je vhodný pro SEO a klíčová slova a nabízí více než 350 šablon (dokumentace AgencyAnalytics. 2022). Geckoboard je jednoduchý a snadno použitelný nástroj pro monitorování obchodních a marketingových dat, vhodný pro malé firmy (dokumentace Geckoboard. 2023).

Looker Studio se vyznačuje svou integrační schopností a širokým spektrem šablon, což usnadňuje vytváření reportů. Jeho hlavní výhodou je snadné připojení datových zdrojů od Google, zatímco připojení třetích stran může vyžadovat použití placených konektorů. Looker Studio je efektivní nástroj pro uživatele, kteří hledají komplexní a přizpůsobitelnou platformu pro vizualizaci dat. A vzhledem k tomu, že vybraná firma využívá BigQuery pro uložení dat, je Looker ideální, protože obě platformy jsou od Googlu a jejich propojení je velmi snadné.

Každá z uvedených alternativ má své specifické silné a slabé stránky, a výběr nejlepšího nástroje samozřejmě závisí na konkrétních potřebách firmy a preferencích v oblasti vizualizace a analýzy dat.

3.9 Řízení rizik

Smejkal a Rais vnímají riziko jako nezbytný prvek každého podnikání a organizační aktivity. Podle nich riziko nemusí být vždy spojeno s negativními důsledky, ale může také představovat příležitost pro rozvoj organizace. V rámci analýzy rizika rozlišují dva hlavní přístupy: kvantitativní a kvalitativní, v závislosti na tom, jakým způsobem jsou rizika vyjadřována. (Smejkal, Rais, 2013)

3.9.1 Kvalitativní metody

Kvalitativní metody se zaměřují na posouzení vážnosti možných dopadů a pravděpodobnosti výskytu rizikových událostí. Tyto metody se vyznačují tím, že rizika jsou vyjádřena ve formě rozsahů nebo slovně, a jsou obvykle založeny na kvalifikovaných odhadech, což je subjektivnější přístup oproti kvantitativním metodám. Kvalitativní metody se často uplatňují tam, kde jsou dostupné číselné údaje pro kvantitativní metody nedostatečné z hlediska kvality nebo kvantity. (Smejkal, Rais, 2013)

Model McKinsey 7S popisuje sedm klíčových faktorů, které jsou důležité pro úspěšný provoz organizace: strategii, strukturu, systémy, styl, spolupracovníky, schopnosti a sdílené hodnoty. Tento model je užitečný pro analýzu rizik v organizaci, jelikož změny v jakémkoli z těchto faktorů mohou ovlivnit celkovou výkonnost a rizika, kterým organizace čelí. (Fotr, 2012)

SWOT analýza, nazvaná podle anglických termínů Strengths, Weaknesses, Opportunities, Threats, se zaměřuje na interní a externí aspekty firmy. Silné a slabé stránky se vztahují k internímu prostředí, zatímco příležitosti a hrozby se vztahují k externímu prostředí. SWOT je flexibilní metoda používaná v různých oblastech analýzy, přizpůsobená konkrétnímu předmětu analýzy. (Fotr, 2012)

3.9.2 Kvantitativní metody

Kvantitativní metody pak zahrnují matematické výpočty rizika založené na frekvenci výskytu hrozby a jejím dopadu. Tyto metody poskytují číselné hodnocení pravděpodobnosti vzniku incidentu a jeho dopadu, jsou přesnější než kvalitativní metody, ale vyžadují více času a finančních prostředků pro jejich provedení. Nabízejí finanční vyjádření rizik, což je pro jejich řízení považováno za výhodné. (Smejkal, Rais, 2013)

4 Analýza současného stavu

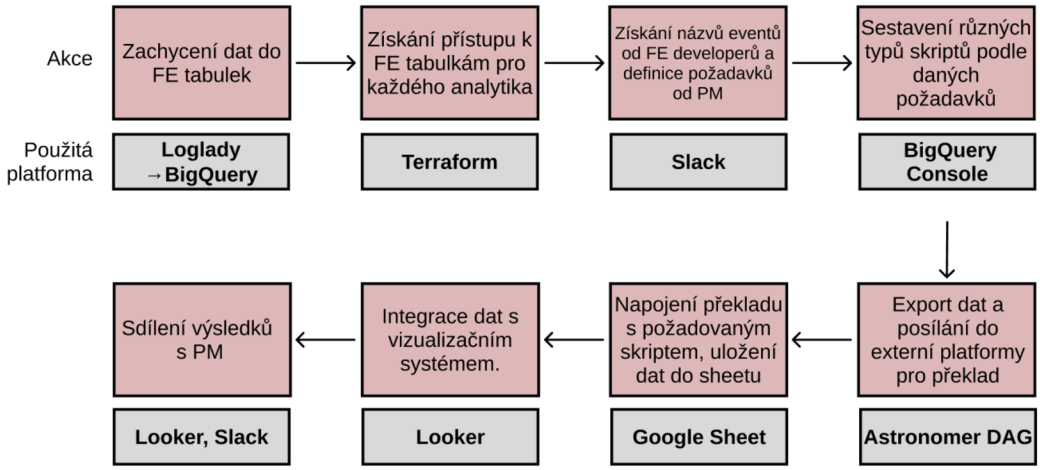
Tato kapitola se zaměří na stručné představení společnosti, rozebere aktuální způsob datové analýzy frontendových dat v čisté raw podobě. Bude také provedena SWOT analýza zaměřená na identifikace silných a slabých stránek, hrozeb a příležitostí při implementaci datového modelu. A v poslední řadě bude zhodnocení celé analýzy a návrh řešení.

4.1 Stručné představení společnosti

Vybraná firma se zaměřuje na poskytování inovativních cestovních služeb. Jejich platforma umožňuje uživatelům kombinovat lety od více než 500 leteckých společností, včetně nízkonákladových a tradičních přepravců. Společnost také spolupracuje s dalšími firmami z oboru cestovního ruchu. Vybraná firma je známá svým důrazem na technologii a inovace. Vyvíjí vlastní software a algoritmy, které jsou základem jejího vyhledávače letů. Tyto technologie umožňují firmě nabídnout unikátní kombinace letů, které nejsou běžně dostupné na tradičních cestovních platformách. Vybraná firma představuje příklad úspěšné inovace v oblasti cestovního ruchu. Její schopnost adaptovat se na měnící se trh a zaměření na technologický vývoj ji řadí mezi přední hráče v oblasti online cestování.

4.2 Aktuální způsob analýzy dat v podniku

Pro pochopení procesu celkové analýzy dat od sběru dat po prezentování výsledků analýz je na obrázku č.1 vytvořena procesní mapa, kde jsou zaznamenány postupné provedené akce a platformy k těmto akcím použité. Každý z těchto kroků bude v této kapitole detailněji popsán.



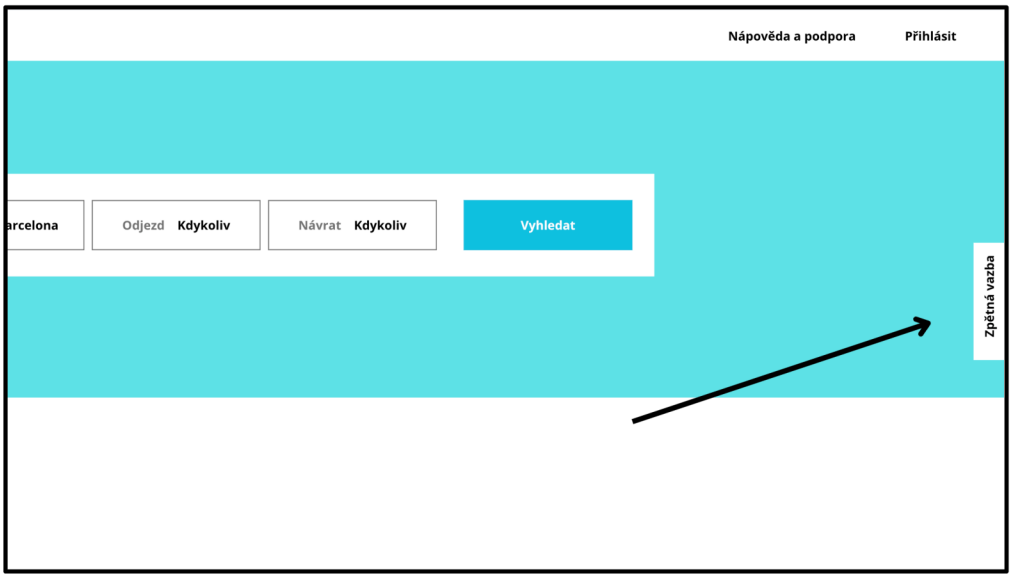
Obrázek 2: Procesní mapa aktuálního zpracování dat ve vybraném podniku (vlastní zpracování)

4.2.1 Procesní analýza současného stavu

Aby byl důkladně pochopen proces analýzy surových dat shromážděných z frontendu a odeslaných do BigQuery (BQ) prostřednictvím nástroje pro logování, jako je Loglady, bude rozebrána celá cesta krok za krokem. První je sběr dat na frontendu. Frontend aplikace nebo webové stránky sleduje různé aktivity a interakce uživatelů, jako jsou kliknutí, zobrazení stránky, odeslání formuláře atd. Kromě uživatelských akcí se sledují také systémové události, jako jsou protokoly chyb, metriky výkonu a volání API. Tato data jsou následně formátována do strukturovaného formátu (většinou JSON), který je vhodný pro protokolování. Dalším krokem je přenos dat prostřednictvím aplikace Loglady. Loglady zachycuje strukturovaná data z frontendu a tento nástroj může odfiltrovat nepotřebná data nebo obohatit data o další kontext (například ID relace, časové značky). Data jsou díky této aplikaci bezpečně přenášena do BigQuery. To může zahrnovat šifrování a dodržování

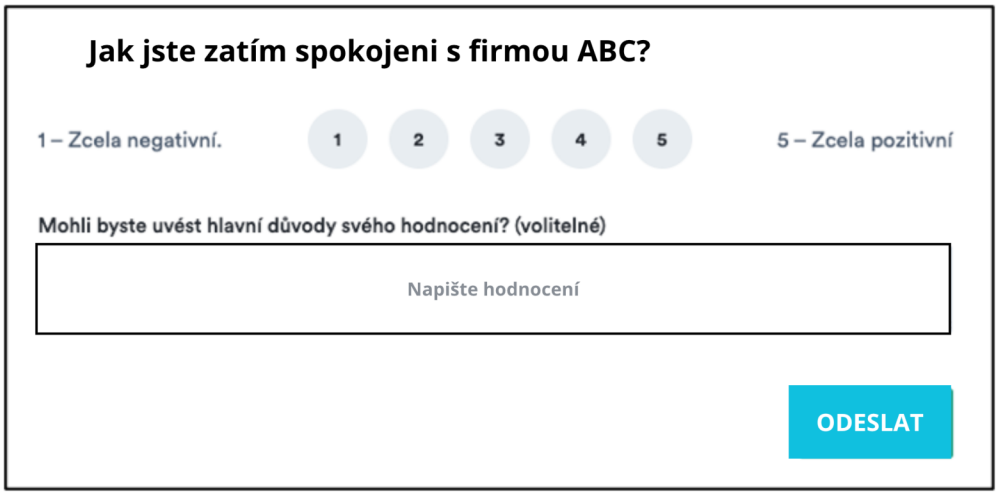
předpisů o ochraně osobních údajů. Data jsou uložena v tabulkách. Pro strukturování dat jsou definována schémata, která mohou zahrnovat pole jako ID uživatele, typ události, časové razítko. Kvůli efektivitě mohou být data rozdělena podle data nebo indexována na základě často dotazovaných polí. Tomuto aktu se říká Partition a clustering.

Předmětem analyzovaných dat a následné hledání nejlepšího řešení bude provedeno na datech z implementovaného formuláře pro zpětnou vazbu na webu společnosti. Jedná se přesnější o tlačítko Feedback na obrázku č. 2



Obrázek 3: Snímek tlačítka zpětné vazby (vlastní zpracování)

Po rozkliknutí tohoto formuláře se objeví 2 otázky. A to přesněji otázka: „Jak jste spokojeni se společností ABC“ a druhá otázka je pro zanechání komentáře. Na obrázku č.3 lze vidět vizuální provedení obou otázek. Toto tlačítko je jen pro demonstraci vyfoceno ze stránky vyhledávání letů, který je pod vedení doménového týmu Search.



Obrázek 4: Rozhraní tlačítka zpětné vazby (vlastní zpracování)

Každý zákazník, který klikne na tlačítko zpětná vazba, už má automaticky vytvořen záznam této akce, které se říká log. I kdyby se zákazník rozhodl neposkytnout zpětnou vazbu a rozmyslel si to, pak je stejně log vytvořen, protože otevřel toto proklikávací tlačítko zpětné vazby. Každý tento log má svá specifika, které je nutné znát, aby kdokoliv mohl tyto data v databázi najít. Pro každou otázku, ať už číselnou nebo komentář je odpověď zaznamenána zvlášť. V tomto případě by bylo nutné získat hned 3 atributy. Každý event je složen z modulu, kategorie a akce. Kategorie nese název "nps feedback", ten je stejný ať už je akce dokončena, nebo ne. Modul se liší podle domény a týmu, zodpovědný za daný formulář. Tudíž pokud se daný formulář nachází na stránce vyhledávání letů, pak je za daný formulář zodpovědný search tým. Modul tedy bude „Search". Stejně tak to bude platit i u „booking", „mmb" (Manage My Booking), „app", a dalších doménových týmů. Každý tým toto nové rozhraní implementoval v jiný časový okamžik, tudíž je nutné, aby produktoví manažeři, nebo tým zodpovědný za data komunikoval sjednocení názvů těchto eventů, pro budoucí přehlednost. Atribut akce se liší podle typu otázky: bude to například „modal opened", tento log byl uložen, jakmile zákazník klikl na tlačítko zpětné vazby. Ale nenese žádnou informaci z otázek, ani skóre, ani komentář. „Email saved" je log ukazující, že email zákazníka byl uložen. „Rate saved" - zákazník ohodnotil číselně první otázku, stejně tak „comment saved" zase druhou otázku.

Každý zákazník má přiděleno pár unikátních identifikačních klíčů (ID). První ID je event ID, unikátní pro každý log. Session ID je unikátní pro každou nově otevřenou kartu v prohlížeči. Se zavřením této karty a znovuotevřením se vytvoří nové session ID. Visitor ID má nejvyšší granularitu, a to v podobě jednoho ID pro jedno zařízení, které zákazník

používá. Nebo také pro jedno Cookies. Pomocí tohoto ID je možné sledovat zákazníkovo chování napříč dny a měsíci.

Další nutností je samotná úprava výsledků z dotazníku, které jsou sice uložené pod danými logy, s daným ID, ale jsou ve formátu JSON. Pomocí datového typu JSON je možnost do BigQuery načíst polostrukturovaný JSON, aniž by bylo nutné předem zadávat schéma pro data. To umožňuje ukládat a dotazovat se na data, která se ne vždy řídí pevnými schématy a datovými typy. Při načítání dat datového typu JSON může BigQuery kódovat a zpracovávat každé pole JSON zvlášť. Na hodnoty polí a prvků pole v datech JSON se je pak možné se dotazovat pomocí speciálních operátorů (Oficiální dokumentace Google Cloud, 2023*)*. Tudíž pro analytiky, kteří budou chtít pracovat s těmito daty je nutností se nejprve obeznámit s těmito základními informacemi. Netřeba zdůraznit, že je nutné také umět programovat v jazyky SQL. BigQuery má dvě vestavěné možnosti pro extrakci JSON z polí a to JSON_EXTRACT a JSON_EXTRACT_SCALAR. Tyto funkce jsou neefektivní při extrakci více hodnot z jednoho JSON pole, protože JSON pole je analyzováno jednou pro každou extrahovanou hodnotu. Výkon se exponenciálně zhoršuje s větším počtem extrahovaných sloupců, protože se zvyšuje jak délka pole JSON, tak počet volání JSON_EXTRACT_SCALAR. Vzhledem k tomu, že je to jediná možnost, jak extrahovat potřebná data, není efektivní, aby byl skript častokrát spouštěn, a to právě kvůli zpomalování databáze, kterou používá celá firma.

Po rozebrání technické stránky uložení dat a následné analýzy by bylo potřebné stanovit si základní kroky, které jsou nutné pro dodání raw dat k analytikům až po dodání finálních analýz k produktovým manažerům do všech potřebných týmů.

4.2.2 Přístup ke zdrojovým tabulkám

Momentálně jako první krok by bylo nutné vytvořit přístup do zdrojové tabulky každému analytikovi, produktovému manažerovi, či komukoliv, kdo bude pracovat se zmíněnými daty. V této tabulce se nachází veškerá data o chování zákazníků na stránkách, a nelze rozdělit právo přístupu jen na určitá data, tudíž jako hlavní problém vyvstává, že kdokoliv si má právo si zobrazit všechny informace, což u některých citlivých údajů není ideální.

4.2.3 Definování si cíle analýzy a následná struktura dat

Druhým krokem je nutnost definování si cíle, co od daných dat chtějí analytici zjistit, jaké informace jsou primární a jaké sekundární. Jaké jsou požadavky produktových manažerů. Podle toho je potřeba vždy sestavit skript. Pro každou skupinu produktových manažerů budou tyto priority jiné. Například produktoví manažeři pro Booking budou chtít analyzovat data z Bookingu, a to hlavně informace týkající se webového rozhraní a spokojenosti se stránkami. Bylo by nutné odfiltrovat stížnosti například se zákaznickou linkou. Pro každý tým s jinými prioritami by bylo nutné sestavit speciální skript a následný export dat do excelu. Případně práci s excelem a filtrování informací v této platformě.

Níže je demonstrovaný skript, který není zaměřený na konkrétní doménu. Skript vybere jako základní identifikátory pro každou událost event_id, visitor_id, session_id a bid. Z objektu JSON rekvizit získá informace o jazykových preferencích a e-mailu. Skóre se určuje na základě akce provedené uživatelem (například "hodnocení uloženo" nebo "kliknutí na tlačítko odeslat") a obsahu objektu props JSON, konkrétně hodnoty přiřazené klíči "question". Stejně tak na základě kategorie události a konkrétních akcí nebo otázek skript klasifikuje typ dotazníku pod názvem sloupce survey_type ("ease_of_use", "mmb_experience", "refund_survey" atd.). Dále vytáhne komentáře buď z polí "text" nebo "comment" v rámci objektu JSON props. Skript extrahuje také informace o typu zařízení, platformě, modelu a podrobnosti o operačním systému. Dalším atributem, který je pro analýzu atraktivní je konkrétní krok v cestě uživatele (například "bookingStep") a příznaky jako "isLoggedIn", "isPartialOffer" a další.

Skript provádí vnitřní spojení s jinými tabulkami, aby vybraná data obohatil o další dimenze, jako je modul, akce a název události. Klauzule WHERE filtruje data na určitý rozsah časových razítek a zajišťuje, že interakce nepocházejí od botů, tím, že kontroluje, zda je trafficCategory nulová.

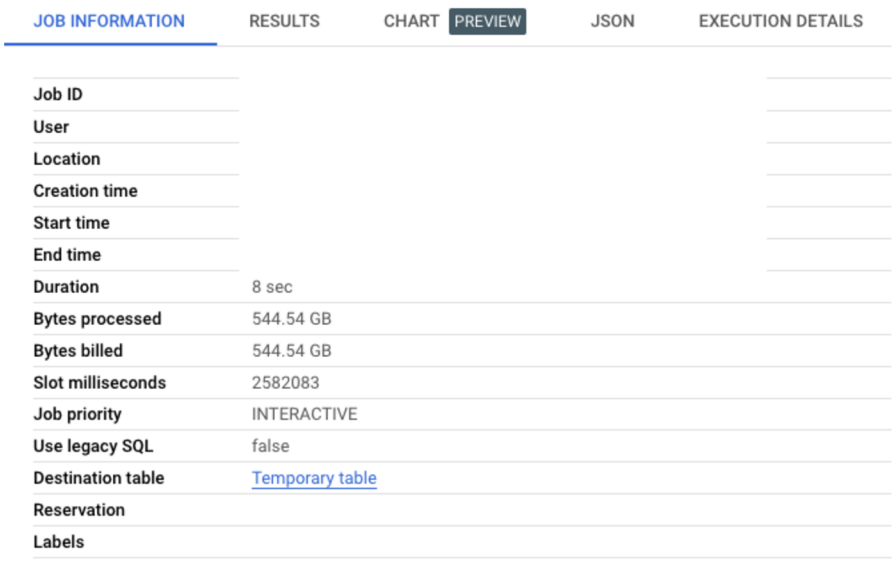
Tento skript je složitý dotaz, který slouží jako součást datového potrubí pro účely reportování nebo analýzy. Je strukturován tak, aby zpracovával vnořená data JSON, pracoval s více spojenými tabulkami a aplikoval podmíněnou logiku pro extrakci a transformaci dat do užitečnějšího formátu pro analýzu. Níže je velmi zkrácený příklad.

```
SELECT
  event_id,
```

```
  visitor_id,
  session_id,
  NULLIF(bid, 0) AS bid,
  language,
  JSON_EXTRACT_SCALAR(props, '$.email') AS email,
  CASE
      WHEN props IS NOT NULL
        AND action = 'rate saved'
        AND JSON_EXTRACT_SCALAR(props, '$.question')= 'experience'
      THEN JSON_EXTRACT_SCALAR(props, '$.value')

      WHEN props IS NOT NULL
        AND action = 'rate saved'
        AND JSON_EXTRACT_SCALAR(props, '$.question')= 'manage your bookings'
      THEN JSON_EXTRACT_SCALAR(props, '$.value')

      WHEN props IS NOT NULL
         AND action = 'submit button clicked'
      THEN JSON_EXTRACT_SCALAR(props, '$.rating')

      WHEN props IS NOT NULL
        AND nested.action = 'rate saved'
      THEN JSON_EXTRACT_SCALAR(props, '$.value')
END AS score,
  …
FROM `frontend.logs` AS nested
WHERE server_timestamp >= '2023-12-01' AND nested.serverTimestamp < '2023-12-02'
```

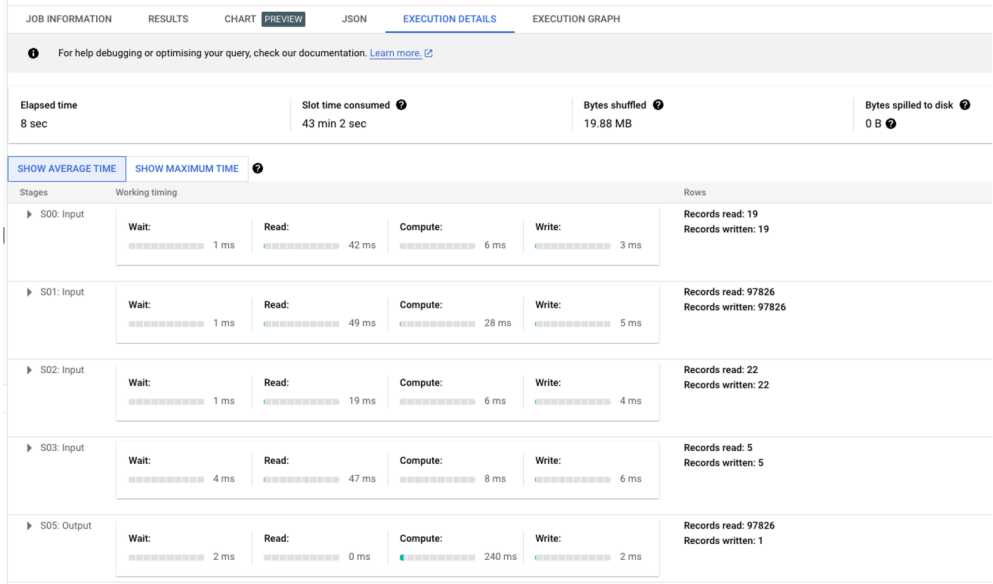
Celkový čas potřebný ke spuštění dotazu byl 8 sekund. Jedná se o čas podle UTC hodin od okamžiku odeslání dotazu po jeho dokončení. Spotřebovaný čas slotu je 43 minut a 2 sekundy času ve slotu. Jedná se o měřítko výpočetních prostředků, které dotaz využil. V BigQuery se práce provádí ve "slotech", což jsou jednotky výpočetní kapacity. Mezi jednotlivými fázemi provádění dotazu bylo promícháno přibližně 19,88 MB dat. K promíchání může dojít, když je třeba data přerozdělit mezi různé příkazy, aby mohli provést operace, jako je JOIN nebo agregace. Na disk bylo přeneseno 0 bajtů, což je dobré, protože to znamená, že dotaz byl schopen pracovat v rámci dostupné paměti, aniž by bylo nutné zapisovat mezivýsledky na disk, což by zpomalilo proces. Pro uchování dat by bylo nutné extrahovat vše například do Excelu, či upravit skript, aby uložil data do tabulky, což už je krokem k datovému modelu.

Níže je snímek obrazovky, kde jsou zaznamenány fáze provádění dotazu a jejich performace.



JOB INFORMATION	RESULTS	CHART PREVIEW	JSON	EXECUTION DETAILS
Job ID				
User				
Location				
Creation time				
Start time				
End time				
Duration	8 sec			
Bytes processed	544.54 GB			
Bytes billed	544.54 GB			
Slot milliseconds	2582083			
Job priority	INTERACTIVE			
Use legacy SQL	false			
Destination table	Temporary table			
Reservation				
Labels				

Obrázek 5: Informace o náročnosti dotazu (vlastní zpracování)



Obrázek 6: Detailnější zobrazení vytíženosti příkazů (vlastní zpracování)

Celkově ukazatele výkonnosti naznačují, že se jedná o dobře optimalizovaný dotaz. Časy čtení a zápisu jsou většinou krátké a výpočetní časy jsou také přiměřené. Skutečnost, že nedochází k úniku dat na disk, naznačuje, že dotaz byl proveden efektivně v rámci paměťových zdrojů.

Dotaz zpracoval a vyúčtoval 544,54 GB dat. BigQuery účtuje dotazy podle množství zpracovaných dat. Pokud by byl tento dotaz prováděn denně, bylo by každý den účtováno

544,54 GB zpracovaných dat. Celková hodnota výpočetních prostředků použitých dotazem se měří v milisekundách slotu. Dotaz spotřeboval 2 582 083 milisekund, což představuje zhruba 43 minut výpočetního času. Úloha je spuštěna v umístění EU a používá standardní dialekt SQL (nikoliv starší SQL). Běží také s interaktivní prioritou, což znamená, že se spustí, jakmile jsou k dispozici zdroje, a obvykle se rychle dokončí.

Vzhledem k těmto podrobnostem je možné pro odhad nákladů na denní spuštění tohoto dotazu použít cenovou strukturu služby BigQuery. V době psaní této analýzy byla použita dostupná cenový struktura z prosince 2023, kdy si platforma BigQuery účtovala 5 USD za data zpracovávaná při každém 1 TiB (oficiální dokumentace Google BigQuery, 2023*)*. Při této ceně, která je minimální by firmu ABC stálo denní spouštění tohoto dotazu:
544.54 GB/den × $6,25 TiB × 30 dnů/měsíc = 3,32 USD/měsíc

Tyto náklady zohledňují pouze poplatky za zpracování dat a neberou v úvahu náklady na ukládání nebo streamovací vložení.

Při takovém množství dat jako zpracovává vybraná firma je výhodnější mít paušální sazby, které se pohybují v minimální ceně 10 000 USD. Při tomto řešení je pak cena každého dotazu nepodstatná, protože firma platí za dostupnost slotů, nikoli za množství dat, která skript zpracuje. Při fixní ceně za sloty je hlavní výhodou předvídatelnost nákladů, nezávisle na zpracovaných datech. Celkové měsíční náklady by byly pevně stanoveny na 16 000 USD při 800 slotech bez ohledu na to, kolikrát by skript byl spuštěn. To však neznamená, že performance by se nezměnila, při pravidelném spouštění skriptu vícekrát denně. Naopak, čím častější sběhnutí, tím více zabraných slotů, tím pomalejší performance databáze.

Nyní zvažme výkonnostní hledisko. Nejdelší fáze dotazu (S07: Join+) trvala 7 sekund. To naznačuje, že operace spojení je částí dotazu, která je nejnáročnější na zdroje. Různé fáze dotazu se načítají z více zdrojů dat s různým počtem záznamů, od 5 do 97,83 tisíc záznamů. Čím větší je datový zdroj, tím více prostředků je potřeba k jeho zpracování. Denní provádění takto náročného dotazu bez optimalizace datového modelu může vést ke zbytečným nákladům a neefektivnímu využívání zdrojů. Dobře navržený datový model by mohl snížit množství zpracovávaných dat tím, že je strukturuje způsobem, který umožňuje efektivnější dotazování. Například použití rozdělených tabulek, clusterových tabulek nebo

materializovaných pohledů by mohlo pomoci snížit množství dat, která je třeba prohledat pro každý dotaz, a tím snížit náklady a zvýšit výkon. Navíc pokud jsou dotazovaná data statická nebo se často nemění, může zhmotnění výsledků do tabulky a dotazování této tabulky namísto každého spuštění celého skriptu vést k výrazným úsporám nákladů. Závěrem analýza ukazuje, že bez optimalizovaného datového modelu by každodenní spouštění tohoto skriptu bylo nákladné a nebylo by optimální z hlediska výkonu. Doporučuje se prozkoumat postupy datového modelování, aby byl proces dotazování efektivnější a nákladově úspornější.

Při analýze, kolik analytiků denně používá data z databáze, pomocí Job History funkce bylo zjištěno průměrně 9 analytiků za poslední měsíc, kteří selektují data na denní bázi, a dalších 14 na týdenní, což dělá průměrně 10,86 spuštění za den.

Níže je vypracována tabulka č. 1, kde jsou uvedeny momentální stavy před implementací řešení a budoucí odhad, který je standardem pro zpracování dat ve vybrané společnosti.

Tabulka 1: Metriky současného řešení vs standard datového modelu (vlastní zpracování)

	Momentální řešení při spuštění skriptu 1x denně	Momentální řešení při průměrném počtu spuštění denně 10,86x	Průměrný Standard datového modelu ve společnosti.
Čas proběhnutí	8 s	-	120 s
Čas ve slotu	43 min 2 s	467 min 20 s	70 min
Procesovaná data	544,54 GB	5 913,7 GB	1095 GB
USD/měsíc za spuštění 1x denně skript bez paušální sazby	$3,32	$36,10	$6,67

Firma sice používá paušální sazby, tudíž reálně nezaplatí částku 887 055 dolarů, ale zabere nynější řešení 6,6x větší množství slotů, což zasekává databázi. Vzhledem k tomu, že analytici vyvolávají skripty v nejvytíženější hodiny dopoledne, je tahle neefektivnost zásadní pro zavedení datového modelu, který je levnější, zabere méně slotů, které hlavně budou potřeba v noci, kdy běží datové modely a nechá prostor pro customizované skripty přes den.

4.2.4 Propojení s platformou pro překlad

Je nutné také komentáře překládat, aby bylo možné je interpretovat. Pro tuto potřebu nyní vybraná firma spolupracuje s platformou Sentisquare, která nejen, že komentáře přeloží, ale také udělá kategorizaci těchto komentářů a přiřadí ji kategorii a štítek. Pro export dat je nyní vytvořen skript, který selektuje raw data a následně posílá do této platformy. Zde nebude zacházeno do detailů, jak přesně tato operace funguje, ale je nutné si alespoň malinko přiblížit její průběh. Po selekci, následné transformaci jsou data posílána do tabulky v BigQuery, kde jsou uložena.

Pro každého analytika je tedy nutností u této tabulce vědět, mít k ní přístup a vyexportovat tyto data a napojit je zpět ke svým analýzám a reportingu. To je možné pouze manuálně v Google Sheets, nebo také ve skriptu, který by bylo nutné spouštět s jednodenním zpožděním, protože kategorizace a překlad trvá několik hodin, stejně tak odeslaní dat zpět do databáze.

4.2.5 Integrace dat se systémem pro lepší sdílení výsledných analýz napříč společností.

Po extrakci dat, spojení transformovaných dat ze Sentisquaru a vytvoření analýzy je na řadě distribuce týdenních analýz napříč týmy produktových manažerů. To se uskutečňuje pomocí standardizovaného procesu, který zajišťuje pravidelnost a konzistenci informací. Každé pondělí ráno analytici připraví reporty obsahující aktualizovaná data a hlavní poznatky z předchozího týdne. Tyto reporty jsou poté zaslány buď pomocí příspěvku na komunikační platformě Slack nebo e-mailem přímo relevantním manažerům, přičemž jsou dodržena všechna pravidla ohledně ochrany dat a soukromí. Součástí příspěvku bývá většinou připojen i link na analýzu v Excelu. Na tomto úložištích mají manažeři možnost získat přístup k reportům kdykoliv v průběhu týdne, což jim umožňuje flexibilitu v přístupu k informacím a data mohou být využita pro ad-hoc analýzy nebo při tvorbě prezentací pro své týmy.

Jiné týmy využívají externí platformu Looker, která slouží k vizualizaci výstupů, jak již bylo zmíněno v teoretickém východisku a použitých metodách. Čili některé týmy integrují data do platformy Looker. Tento postup umožňuje efektivní sdílení informací s produktovými manažery, ale zároveň představuje určité omezení. Hlavní nevýhodou je neefektivnost připojení Google Sheet tabulek, jelikož je zde potom závislost na analyticích, že manuálně tyto data doplní a napojí sentisquare. Analytická část je dále možná přeskočit, jelikož v Lookeru si analytici udělají dashboard, který při zachování stejného pojmenování sloupců, bude fungovat stále stejně, analyzovat stejně, a tudíž je zde ušetřen jeden krok. Avšak bez konzistentního a integrovaného datového modelu může docházet k nekonzistenci ve způsobu, jakým jsou data interpretována a prezentována různým týmům. Výsledkem je potenciální nesoulad v rozhodovacích procesech a strategiích napříč odděleními.

4.2.1 SWOT analýza

Tato diplomová práce se zaměřuje na zlepšení informační strategie vybrané firmy v oblasti zákaznického zážitku. Informační strategie vybrané firmy má své odvozené cíle v oblasti analýzy dat, kde se firma snaží o efektivnost analyzování dat a procesů a s tím související maximalizace rychlosti dat a transparentnost analýz a výsledků. Vycházejíc z této strategie bude definována SWOT analýza, aby byly odhaleny silné a slabé stránky nynějšího stavu a příležitosti a hrozby.

Silné stránky

Mezi hlavní silné stránky společnosti v tuto chvíli patří vysoká kvalita dat. Společnost má k dispozici kvalitních dat z různých zdrojů, která jsou relativně čistá a dobře strukturovaná, což poskytuje solidní základ pro budování datového modelu. Tým pro zpracování dat má také hluboké technologické znalosti v oblasti analýzy dat a tím jsou schopni rychle reagovat na změny a požadavky trhu, což umožňuje pružně přizpůsobovat se novým požadavkům a implementovat nové technologie a postupy. Velkou silnou stránkou jsou právě tyto technologie jako například BigQuery a Looker, které má společnost k dispozici, což umožňuje efektivní zpracování a využití datových zdrojů.

Slabé stránky

Absence definovaného datového modelu vede k neefektivnímu zpracování a interpretaci dat, což omezuje schopnost společnosti efektivně využívat své datové zdroje. V současném stavu jsou k vytvoření analýzy vyžadovány schopnosti, které nejsou v popisu práce analytiků (například znalost jazyka SQL), tudíž kapacity analytiků jsou ještě více zahlceny z důvodu požadování práce, která je mimo jejich oblast působnosti. Dále také některé procesy, jako je sběr dat nebo ruční překlad komentářů, jsou závislé na manuální práci, což znovu může zpomalit tempo a zvýšit chybovost.

Příležitosti

Společnost má možnost vytvořit a implementovat robustní datový model, který bude sloužit jako základ pro efektivní zpracování a analýzu dat, což může vést k lepšímu porozumění zákazníkům a trhu. Také je velká příležitost v rozšíření informací v datovém modelu o informace z jiných modelů pomocí vybudování datového modelu, kde je to možné. Také automatizace opakujících se procesů může zvýšit efektivitu a snížit náklady, například pomocí DAGu.

Hrozby

Konkurenční společnosti mohou investovat do pokročilých analytických nástrojů a technologií, což může zvýšit konkurenční tlak a snížit podíl na trhu společnosti. S narůstajícím počtem kybernetických hrozeb je důležité zajistit bezpečnost dat a ochranu soukromí zákazníků, aby se zabránilo možným únikům dat nebo kybernetickým útokům. Největší hrozbou v kontextu s návrhem řešení může být nesplnění požadavků při budování datového modelu a minutí se účinku, když z tohoto důvodu analytici nebudou používat tento model.

4.2.2 Hodnocení celkové efektivity procesu a definice požadavků na nové řešení

V návaznosti na celou analýzu současného stavu a vyhotovené SWOT analýzy lze říct, že v současném procesu distribuci dat napříč týmy je vyžadována vysoká míra manuální práce a nedostatek automatizace. Selektovaná data jsou exportována z databáze do Excelu, což představuje první úroveň zpracování. Toto řešení, ač flexibilní a umožňující personalizaci pro jednotlivé týmy, přináší značný potenciál pro lidské chyby při manipulaci s daty. Další komplikace nastává při integraci těchto dat s platformou Looker, kde je nutné data nejen importovat, ale také udržovat aktuální a synchronizovaná. Bez centrálního datového modelu je každá aktualizace závislá na manuální intervenci analytiků, což výrazně zpomaluje celkový proces a omezuje jeho škálovatelnost, nemluvě o veliké znalosti, které nejsou typické pro datové analytiky, například znalost SQL. Avšak nejužším místem je počet selekcí dat v databázi BigQeury, který dělají celý proces velmi nákladný ať už na počet slotů, nebo velikost zprocesovaných dat.

Navíc, výsledná analýza není okamžitě k dispozici produktovým manažerům, ale vyžaduje další kroky ke sdílení a interpretaci. Zpoždění v distribuci může mít za následek zastaralost dat, což je v rychle se měnícím podnikovém prostředí nežádoucí. Také se zvyšuje riziko nesrovnalostí, kdy různé týmy mohou interpretovat stejná data rozdílně, což vede k nekonzistentním rozhodnutím a akcím napříč společností.

Z tohoto hodnocení jsou vymezeny požadavky na nové řešení, které jsou potvrzeny datovými analytiky z vybrané společnosti. Nové řešení by mělo:

1. Snížit chybovost dat, která je tvořena nedostatečnou informovaností analytiků při tvorbě skriptů.
2. Zautomatizovat proces ukládání transformovaných dat.
3. Mít transparentní a lehce komunikovatelné výsledky analýzy pomocí analytické platformy.
4. Být levnější a efektivnější z hlediska výkonu databáze.

Řešení těchto problémů leží ve vývoji a nasazení komplexního datového modelu. Tento model by poskytoval jednotný základ pro všechny analytické a reportovací procesy, což by zajišťovalo konzistenci a zjednodušovalo by přístup k datům. S centralizovaným datovým

modelem by mohly data selektována každý den automaticky a jenom jednou, což by řešilo naše nejužší místo, a dále pak budou tyto data distribuovány v reálném čase do tabulek, které by byly clustrované, což by zlepšilo rychlost Lookeru a celkové zanepráznění BigQuery. Díky sjednocenosti a následné selekci požadovaných metrik každého týmu přímo v Lookeru by měla za následek větší přesnost rozhodování. Kromě toho by integrace dat s nástroji třetích stran, jako je Looker, byla usnadněna, čímž by se eliminovala potřeba manuálního zpracování a uvolnil se tak analytický tým pro složitější úlohy a hlubší analýzy.

4.2.3 Shrnutí analytické části a definování požadavků na nové řešení

Závěrem, současný proces je charakterizován nedostatečnou efektivitou a přináší s sebou řadu výzev, které by byly řešitelné implementací robustního datového modelu. Implementace komplexního datového modelu a automatizace distribuce datových analýz přispějí k podpoře rozhodování ve firmě snížením chyb a zvýšením přesnosti dat. Přechod na automatizovaný systém přináší revoluční zlepšení v přesnosti a spolehlivosti dat. Podnik už nebude muset spoléhat na náročné ruční procesy, které jsou náchylné k chybám. Místo toho budou stále k dispozici aktuální a přesná data, což je základ pro strategické rozhodování. Dále model přispěje k lepší konzistenci a jednotnosti dat. Centralizovaný datový model znamená, že všechny týmy a oddělení budou pracovat s daty, která jsou konzistentní a sjednocená a vedená pod jedním týmem, který má na starosti jejich správu. Toto odstraní jakékoli rozpory v interpretaci, což vede k jasnějším a koordinovanějším rozhodnutím. A hlavním aspektem je zvýšením efektivity a produktivity. Automatizace rutinních úloh umožní uvolnit kapacitu databáze a také analytiky od časově náročných úkolů, jako je sběr a příprava dat. Místo toho se budou moci soustředit na komplexnější analýzy a strategické úlohy. To zlepší nejen produktivitu analytického týmu, ale také celkovou efektivitu podnikání. V poslední řadě budou data budou nyní snadno integrovatelná s nástroji jako je Looker, což umožní využít pokročilé funkce těchto platforem pro hlubší analýzu a lepší vizualizace.

5 Vlastní návrhy řešení

V této kapitole bude vytvořena analýza rizik spojených s navrhovaným řešením. Dále bude rozveden návrh řešení do praktických kroků, a poté bude vyhodnoceno zdali některá rizika byla minimalizována, jaká opatření byla přijata.

5.1 Analýza rizik

Analýza rizik je zásadní při tvoření nového řešení, pomáhá identifikovat možné negativní události, které by zapříčinily ztráty ať už časové, peněžní, datové nebo další. V prvním kroku bude vytvořena identifikace možných rizik, dále bude vytvořena analýza s pravděpodobností výskytu a možným dopadem.

5.1.1 Identifikace rizika

V této kapitole bude identifikováno patnáct potenciálních rizik, ze kterých budou poté vybrány jen nejvíc aktuálních 7 rizik.

1. **Riziko obtížnosti správy:** Čím složitější je datový model, tím obtížnější je spravovat a udržovat data. Správa zahrnuje úkoly jako zálohování, obnovování, monitorování a zabezpečení dat. Komplexní datové modely mohou vyžadovat specifické nástroje a postupy pro efektivní správu, což může zvýšit náklady a nároky na datové analytiky, či zdroje. (Database Engineering, 2023)

2. Riziko obtíží při údržbě: Pokud není datový model dostatečně dobře zdokumentován, může to vést k obtížím při údržbě. Změny v datovém modelu, jako jsou přidání nových tabulek, sloupců nebo vztahů, mohou být složité a riskantní, pokud není jasné, jak tyto změny ovlivní stávající data a aplikace. Nedostatečná dokumentace může také ztížit práci novým členům týmu, kteří se snaží porozumět struktuře dat a provádět úpravy.

3. Riziko chyb: Chybějící správné propojení mezi jednotlivými částmi datového modelu může vést k nesouladům a chybám v analýzách. Může být také obtížnější porozumět všem vazbám a závislostem mezi daty. To může vést k nekonzistencím v datech, špatně definovaných metrikách a chybným analýzám nebo špatným rozhodnutím založeným na těchto datech.

4. **Riziko zpomalení databáze:** Riziko zpomalení databáze v důsledku implementace datového modelu může nastat z několika důvodů. Nesprávný návrh datového modelu, zvýšená složitost dotazů, nedostatečná indexace, růst objemu dat a nedostatečné zdroje mohou vést k tomu, že databáze není schopna efektivně zpracovávat požadavky aplikace. To může zpomalit odezvu databáze a způsobit výrazné zhoršení výkonu aplikace, což může mít negativní dopad na uživatelskou zkušenost a provozní efektivitu. Je nezbytné pečlivě plánovat a testovat datový model a průběžně provádět optimalizace a údržbu databáze, aby se minimalizovalo toto riziko.

5. Riziko snížené flexibility a adaptability: Příliš komplexní datové modely mohou být méně flexibilní a obtížněji přizpůsobitelné změnám v požadavcích ať už od vývojářů, či datových analytiků. To může být problematické v prostředí, kde se rychle mění požadavky a je nutné pružně reagovat na nové podněty.

6. **Problémy s přenosností dat:** Pokud není integrace mezi nástroji provedena úspěšně, může dojít k problémům s přenosností dat mezi nimi. To znamená, že data vytvořená nebo upravovaná v jednom nástroji nemusí být snadno přenositelná do druhého nástroje nebo platformy. To může vést k redundanci dat, nekonzistencím nebo ztrátě důležitých informací.

7. Riziko komunikačních problémů: Neúspěšná integrace může také vést k problémům s komunikací mezi nástroji a platformami. Například, pokud Terraform nedokáže správně komunikovat s Lookerem, může dojít k chybám při přenosu dat, synchronizaci nebo spuštění automatizovaných procesů mezi těmito nástroji. To může omezit schopnost týmu efektivně analyzovat Looker data .

8. Zvýšené náklady a ztráta efektivity: Neúspěšná integrace může také zvýšit náklady na vývoj a správu systému. Pokud tým musí investovat více času a zdrojů do řešení problémů spojených s nekompatibilitou mezi nástroji, může to mít negativní dopad na celkovou efektivitu a rentabilitu projektu.

9. **Nesprávný návrh datového modelu**: Pokud není datový model správně navržený a optimalizovaný pro konkrétní potřeby aplikace, může to vést k zbytečnému zatížení

databáze. Nepřiměřeně složité dotazy, nadměrné indexování nebo neefektivní struktura tabulek mohou způsobit zpomalení databáze.

10. Zvýšená složitost dotazů: Komplexní datový model může způsobit složitější dotazy, které vyžadují více času a výpočetních zdrojů k vykonání. Pokud je datový model příliš komplexní, může to vést k vytváření dotazů, které vyžadují složitější spojování tabulek, filtraci dat nebo agregaci výsledků, což může zpomalit výkon databáze. Pokud také není datový model správně indexován, mohou dotazy trvat déle než je nutné. Nedostatečná indexace může vést k plánování dotazů, které vyžaduje prohledání celého objemu dat, což může zpomalit výkon databáze.

11. **Riziko špatného školení:** Nedostatečné školení týmu na používání Lookeru a dalších nástrojů může vést k nízké produktivitě a nesprávným rozhodnutím z důvodu nedostatečné znalosti funkcí a možností nástrojů, ztrátě příležitostí pro rozvoj nových analytických schopností, zvýšeným nákladům na podporu a potřebě častější pomoci IT týmu nebo externím poskytovatelům, a konečně k nedostatečné adopci nových nástrojů a postupů, což může bránit plnému využití jejich potenciálu a přinášet odpor vůči změnám. Školení týmu je proto klíčové pro úspěšné nasazení nových technologií a minimalizaci rizik spojených s neefektivním využitím těchto platforem.

12. Riziko špatné komunikace: Nedostatečná komunikace mezi týmy odpovědnými za různé části procesu, jako je vytváření datového modelu, správa Terraformu a konfigurace Lookeru, může vést k nesrovnalostem a zpožděním v implementaci a údržbě systému. Nedostatečná komunikace může také způsobit situace, kdy jedna část systému není správně přizpůsobena změnám provedeným v jiné části, což v konečném důsledku může vyžadovat další práci na opravách a aktualizacích. Zajištění pravidelné a efektivní komunikace mezi týmy je nezbytné pro úspěšnou implementaci a udržení integrovaného a spolehlivého systému.

13. **Riziko vysokých nákladů:** Vysoká cena platforem, jako je Looker, BigQuery, GitLab a další, nebo závislost na nich, představuje značné riziko pro organizaci. Finanční zátěž spojená s jejich nákupem a licencováním může zatížit rozpočet. Závislost na dodavateli může omezit flexibilitu organizace při hledání alternativních řešení a zvýšit riziko neefektivního využití zdrojů. Nepříznivé změny v cenách, podmínkách nebo dokonce zrušení podpory mohou mít značný dopad na provoz a strategii organizace.

Nedostatečná schopnost udržet platby nebo přerušení služeb může znamenat ztrátu přístupu k důležitým funkcím a datům, což může mít vážné dopady na provoz a rozhodování. Proto je důležité, aby organizace prováděla důkladnou analýzu nákladů a přínosů před nákupem těchto platforem.

14. **Riziko špatné datové kvality:** Toto riziko představuje možnost vzniku chyb, nekonzistencí a nesprávných informací v databázi, což může mít vážné důsledky pro podnikání. Tato rizika mohou vzniknout z několika důvodů, včetně chybného zadávání dat, nedostatečné validace dat, duplikace záznamů, neúplných nebo zastaralých dat a nekonzistentních formátů. Důsledky špatné datové kvality mohou zahrnovat nesprávné rozhodnutí na základě chybných dat, ztrátu důvěryhodnosti uživatelů, narušení procesů a zvýšené náklady na opravy a údržbu. Řízení tohoto rizika zahrnuje implementaci procesů a technologií pro kontrolu a zlepšování kvality dat, jako je automatizovaná validace dat, pravidelná kontrola a čištění dat, a důkladná školení pro uživatele, aby se minimalizovalo riziko vzniku špatné datové kvality. (Database Engineering, 2023)

15. Riziko nedostatku dat: Pokud zákazníci nebudou motivováni odpovídat na dotazníky, může to vést k nedostatku relevantních dat a zpětné vazby. Nedostatek dat a zpětné vazby může znemožnit organizaci plně porozumět potřebám a preferencím zákazníků a vést k neefektivním strategiím a rozhodnutím. Je důležité investovat do marketingu a komunikace, aby se zvýšil povědomí a zájem zákazníků o poskytované služby a vytvořila se aktivní komunita, která poskytuje hodnotnou zpětnou vazbu a data pro další rozvoj a vylepšování služeb.

5.1.2 Analýza rizik

V tabulce č. 2 lze vidět výběr nejvíce aktuálních rizik, pro které je vypracovaná matice rizik. Nejvíce rizikové je zpomalení databáze, čímž se rapidně zvednou náklady společnosti, všechny firemní analýzy budou mít zpoždění, několik modelů může selhat pokud nedoběhnou v určený čas. Pravděpodobnost zpomalení databáze je na stupnici od jedné do pěti hodnocena číslem tři, což je středně pravděpodobná varianta při tvoření nového modelu, zvláště pokud se jedná o model zpracovávající frontendová data, uložená v samostatné tabulce.

Druhým nejrizikovějším scénářem jsou špatná datová kvalita, které mohou zapříčinit špatné výsledky analýz, nepřesnou interpretaci a neporozumění zákaznickému prožitku, což může ovlivnit implementaci nových nepřesných webových funkcí. Třetí nejrizikovější je riziko zvýšení nákladů. Tato možnost může nastat pokud se změní ceny platforem, které společnost využívá, jako například Looker, či BigQuery na úroveň, která pro podnik bude riziková. Pravděpodobnost výskytu je nízká, avšak dopad velmi vysoký pro celé Net Revenue.

Dalšími riziky jsou obtížnost správy a nesprávný návrh datového modelu. Oba scénáře mají střední až vyšší pravděpodobnost a střední dopad. Pokud data nejdou jednoduše spravovat z důvodů komplexity modelu či chybějící dokumentace, je nemožné pro zaměstnance hledat případné chyby, či implementovat nové funkce. Nesprávný návrh datového modelu může způsobovat delší procesy, pomalejší procesování dat a nešikovnou definici v Lookeru, která může zpomalovat tuto platformu. Dalším je riziko špatného školení, které má nižší pravděpodobnost výskytu, však vyšší dopad, jelikož při špatném školení, analytici nejsou schopni plně využívat potenciál Lookeru a dělat méně stabilní a vhodné metriky a vizualizace.

V poslední řadě je zde nedostatek dat a problém s přesností dat. Přesnost dat způsobená špatnou kooperací mezi použitými nástroji může způsobit chybné analýzy, pravděpodobnost, že to nastane je však velmi nízká. Nedostatek dat je také velmi málo pravděpodobné, avšak dopad by byl vysoký jelikož by společnost neměla podklady k strategickým rozhodnutí zakládajících se na zákazníkově preferenci.

Tabulka 2: Klasifikace rizik (vlastní zpracování)

	Riziko	**Pravděpodobnost výskytu (1-5)**	**Dopad (1-5)**	**Hodnota rizika**
1.	1. Obtížnost správy	4	3	**SR 12**
2.	4. Zpomalení databáze	3	5	**VR 15**
3.	6. Problémy s přenosností dat	1	3	NR 3
4.	9. Nesprávný návrh datového modelu	3	3	**SR 9**
5.	11. Riziko špatného školení	2	4	SR 8
6.	13. Riziko zvýšení nákladů	2	5	**VR 10**
7.	14. Špatná datová kvalita	3	4	**VR 12**
8.	15. Nedostatek dat	1	4	SR 4

Tabulka 3: Matice rizik (vlastní zpracování)

Pst/Dopad	Velikost rizika				
5 = velmi vysoká	SR	SR	**VR =2**	VR	VR
4 = Vysoká	NR	SR	**SR = 1**	VR	VR
3 = Střední	NR	SR	**SR = 4**	**VR = 7**	VR
2 = Nízká	NR	NR	SR	**SR = 5**	**VR = 6**
1 = Velmi nízká	NR	NR	**NR = 3**	**SR = 8**	SR
	1 = Velmi nízká	2 = Nízká	3 = Střední	4 = Vysoká	5 = velmi vysoká

Zelená sekce označuje nízkou velikost rizika, žlutá střední a červená vysokou velikost rizika. Samotná opatření, která budou zaručovat minimalizaci těchto rizik budou vypracována na konci šesté kapitoly po zhodnocení navrhnutého řešení

5.2 Postupné kroky pro vytvoření datového modelu

V této kapitole budou detailněji popsány praktické kroky, které jsou nutné pro vytvoření datového modelu a automatizaci selekce dat. Tato kapitola bude obsahovat zkrácené příklady skriptů.

5.2.1 Vytvoření skriptu pro tabulky

Při stavění datového modelu a tabulek lze pojmout dva protichůdné přístupy a to normalizace a denormalizace. Normalizace je proces snižování redundance dat a zlepšování integrity dat rozdělením dat do menších a jednodušších tabulek. Denormalizace je proces zvyšování redundance dat a zlepšování přístupu k datům spojováním dat do větších a složitějších tabulek. Oba přístupy mají své výhody a nevýhody v závislosti na kontextu a cílech modelování dat. Normalizace může například snížit anomálie dat a náklady na jejich ukládání, ale může zvýšit počet spojů a dotazů. Denormalizace může zlepšit výkonnost a jednoduchost dotazů, ale může zvýšit duplicitu dat a náklady na jejich údržbu. Proto je třeba normalizaci a denormalizaci vyvážit, abyste dosáhli optimálního kompromisu mezi kvalitou a efektivitou dat. V modelu, který bude v návrhu postaven se bude vycházet z normalizace, jelikož surová data jsou v jedné velké tabulce, jejíž zprocesování je velmi nákladné (Database Engineering, 2023). Proto bude vystavěno více tabulek, které budou selektovat jen potřebná data a budou dále zpracovávat informace i jiných modelů.

S touto myšlenkou budou vytvořeny dvě tabulky, přípravná tabulka tzv. prep tabulka a flat tabulka. Flat tabulka je jednoduchá databázová struktura, která ukládá data v jedné tabulce. Na rozdíl od relačních databází, které mohou obsahovat více tabulek s vazbami a relacemi mezi nimi tzv. dimenze, flat tabulka obsahuje všechny potřebné informace v jediné tabulce bez nutnosti joinu na dimenzionální klíč.

Pro optimalizaci selekce frontendových dat, které jsou ve formátu JSON, bude vytvořena prep tabulka, kde budou vybrány všechny atributy nehledě na preference analytiků. Tato tabulka se bude plnit každý den, nebude zde vytvořena žádná transformace, kromě selekce a přiřazení správného datového typu z JSONu.

Druhou bude flat tabulka, kde budou vytvořeny metriky, transformace, napojení na další modely a extrakce dat z těchto externích modelů. Důvodem pro vytvoření prep tabulky je ten, že pokud budou nároky na novou implementaci, či změnu ve flat tabulce, nebude potřeba zpětně plnit tabulku z frontendových logů, neb ty budou selektovány a ukládány v prep tabulce, a tím pádem se bude tato flat tabulka obnovovat přímo z prep tabulky. Tím se ušetří hodně slotů a processingu, protože už nebude nutné procesovat JSON, který je nákladný. Struktura prep tabulky bude následovná:
V prvním kroku bude vytvořena tabulka prep_frontend_logs, kde jsou definovány všechny datatypy daných sloupců.

```
CREATE TABLE IF NOT EXISTS `xxx.project.prep_frontend_logs`(
    event_id STRING,
    visitor_id STRING,
    session_id STRING,
    bid INTEGER,
    language STRING,
    email STRING,
    score NUMERIC,
   ...);
```

V druhém kroku proběhne selekce dat, která byla již ukázána v kapitole 5.3.3, zde jsou přidány JOINy na tabulky, které například sjednocují konvence pojmenovávání typu zařízení, co zákazník užívá, nebo jazyk, tak aby byl pro celou společnost, pro všechny data sjednocený. Dále se zde filtrují reakce botů nebo záznamy, které by kvůli nějaké chybě neměly visitor_id nebo session_id. Eventy, které chceme, aby byly vybrány jsou implementovány v pomovné tabulce zvané config_event_name_type. V případě, že by byl přidán nový event, který zaznamenává zákaznický zážitek, pak by musel být přidán do config tabulky, do té doby nebude zařazen do každodenního výběru dat.

```
CREATE TEMPORARY TABLE tmp_prep_events AS(
    SELECT
        LOWER(nested.eventId) AS event_id,
        LOWER(nested.visitorId) AS visitor_id,
        LOWER(nested.sessionId) AS session_id,
        NULLIF(nested.bid, 0) AS bid,
        COALESCE(lang.converted, nested.langId) AS language,
        JSON_EXTRACT_SCALAR(nested.props, '$.email') AS email,
        ...
        CURRENT_TIMESTAMP() AS inserted_at
    FROM `frontend.logs` AS nested
    LEFT JOIN `config_device_type_conversion` AS conv -sjednocení zařízení
```

```
        ON nested.userAgentInfo.deviceType = conv.original
    LEFT JOIN `config_language_code_conversion` AS lang -sjednocení jazyku
        ON nested.langId = lang.original
    INNER JOIN `config_event_name_type` es -selekce pouze chtěných eventů
        ON nested.eventName = es.event_name
    WHERE nested.serverTimestamp >= '{{ ds }}' AND nested.serverTimestamp <
'{{ next_ds }}'
    AND nested.props IS NOT NULL
    AND nested.trafficCategory IS NULL  --výsledky bez záznamy botů
    AND nested.visitorId <> ''
    AND nested.sessionId <> ''
    );
```

V poslední řadě, budou nově vybraná data implementována pomocí Merge statementu do námi vytvořené prep tabulky. V prvním kroku vkládáme hodnoty, které se v tabulce ještě nevyskytují čili všechny nové hodnoty. V druhém kroku aktualizujeme ty, které se změnily. Tento krok se hodí, pokud přijde změna a je některé hodnoty byly pozměněny. Například se změnilo Booking ID (bid).

```
MERGE ` xxx.project.prep_frontend_logs ` fs
USING tmp_prep_events rwd
    ON fs.event_id = rwd.event_id
    AND fs.visitor_id = rwd.visitor_id
    AND fs.session_id = rwd.session_id
    AND fs.bid = rwd.bid
        ...
WHEN NOT MATCHED THEN
    INSERT (
        event_id,
        visitor_id,
        session_id,
        bid,
        ...)
    VALUES(
        event_id,
        visitor_id,
        session_id,
        bid,
        ...)
WHEN MATCHED
    AND fs.event_id <> rwd.event_id
    OR fs.visitor_id <> rwd.visitor_id
    OR fs.session_id <> rwd.session_id
    OR fs.bid <> rwd.bid
        ...
THEN UPDATE SET
    event_id = rwd.event_id,
    visitor_id = rwd.visitor_id,
```

```
    session_id = rwd.session_id,
    bid = rwd.bid,
    ...
    inserted_at = CURRENT_TIMESTAMP();
```

Po vytvoření prep tabulky, kde máme naimplementovaná raw data, přetransformovaná do správného formátu, je možné vytvořit flat tabulku. Tato tabulka bude mít také v prvním kroku vytvoření samotné tabulky, kde budou vypsány všechny sloupce se správným datovým typem. V druhém kroku budou vyselektovány data na daný den, aby byly připravené pro další transformaci.

```
CREATE TEMP TABLE tmp_prep_fe_data AS (
    SELECT
        event_id,
        visitor_id,
        session_id,
        NULLIF(bid, 0) AS bid,
        NULLIF(email, '') AS email,
        'unknown' AS survey_status,
        score,
        comment,
        survey_type,
        device_platform,
        device_model,
        device_type,
        os_name,
        os_version,
        app_version,
        CASE
            WHEN module IN ('account', 'mmb', 'mmb-platform')
                THEN 'mmb'
            WHEN module = 'booking' AND action = 'show'
                THEN 'booking'
            WHEN user_step = 'homepage'
                THEN 'homepage'
            WHEN module = 'search'
                THEN 'search'
            WHEN module = 'helpcenter'
                THEN 'helpcenter'
            ELSE user_step
        END AS user_step,
        is_logged_in,
        CASE
        ...
        END AS refund_method,
        is_partial_offer,
        module,
        action,
```

```
        event_name,
        event_name_type,
        is_answered,
        server_timestamp AS survey_initiated_at
    FROM `prep_frontend_logs`
  WHERE server_timestamp >= '{{ ds }}' AND server_timestamp < {{ next_ds }}
  AND session_id NOT IN (SELECT session_id FROM `xxx.prep_frontend_logs`
WHERE survey_type = 'test')
```

Už při prvním selektu dat dochází k transformaci, například se zde přiřazuje user_step, kde lze odvodit, v jakém kroku kupování letenek zákazník vyplnil dotazník, zdali to bylo při hledání letenek, nebo v sekci spravování rezervace etc. V dalším kroku budou provedeny složitější transformace jako například napojení na tabulky, díky kterým bude možno získat více informací o chování zákazníka. Například metrika customer_journey_stage, která udává, v jakém stádiu konkrétního bookingu zákazník vyplnil dotazník. Nebo zda daný zákazník měl přerušený let, zda si zařádal o vrácení peněz, zda nás kontaktovat přes zákaznickou linku a další. Zde je příklad jedné z těchto napojení, kde lze vidět, že se vybírají jenom ty booking_id, které byly vytvořeny maximalně před devíti měsíci. Tyto booking_id se později napojí pomocí joinu zpátky a vytvoří se metrika.

```
CREATE TEMP TABLE tmp_refund_requests AS (
    SELECT
        f.booking_id,
        MIN(f.timestamp) AS refund_requested_timestamp
    FROM `flat_vraceni_penez` AS f
    INNER JOIN tmp_prep_fe_data AS b
        ON f.booking_id = b.booking_id
    WHERE f.is_valid = TRUE AND f.is_deleted = FALSE
    AND CAST(f.timestamp AS DATE) >= CAST(b.survey_initiated_at AS DATE) - 274 -
-taking into consideration last 9 months
    GROUP BY f.booking_id
);
```

V dalším kroku budou všechny temporary tabulky, které byly vytvořeny pro získání informací z cizích modelů, dány dohromady pomocí JOINů. Bude uměle vygenerováno unikátní ID survey_fwid pro každý vytvořený záznam. Dále pro sloupce typu *device*, které obsahují více hodnot, budou vytvořeny datové typy STRUCT, z důvodu lepší organizace tabulky.

```
CREATE TEMP TABLE tmp_results AS (
    SELECT
```

```
        ROW_NUMBER () OVER () + mf.max_survey_fwid AS survey_fwid,
        rslt.event_id,
        rslt.visitor_id,
        rslt.session_id,
        rslt.assessment_id,
        rslt.bid,
        rslt.transaction_id,
        rslt.email,
        rslt.survey_status,
        rslt.survey_type,
        STRUCT(dd.code, dd.vendor, dd.version, dd.type) AS device,
        rslt.device_platform,
        STRUCT(rslt.os_name AS code, rslt.os_version AS version) AS
operating_system,
        refund_compensation_type,
        ...
        CASE WHEN tcc.bid IS NOT NULL THEN TRUE ELSE FALSE END AS
has_carrier_cancellation,
        CASE WHEN ref.bid IS NOT NULL THEN TRUE
             ELSE FALSE END AS has_refund_request,
        CASE WHEN chc.bid IS NOT NULL THEN TRUE
             ELSE FALSE END AS has_chargeback,
        CASE WHEN fch.bid IS NOT NULL THEN TRUE
             ELSE FALSE END AS has_declined_fraud_check,
        CASE WHEN tkt.bid IS NOT NULL THEN TRUE
             ELSE FALSE END AS has_ticket_activity,
        ...
        ROW_NUMBER() OVER(PARTITION BY rslt.visitor_id, rslt.session_id,
rslt.survey_type, rslt.bid ORDER BY rslt.is_answered DESC, rslt.is_main_event
DESC, survey_initiated_at DESC) AS rn
    FROM tmp_intermediate_results AS rslt
        LEFT JOIN tmp_carrier_cancellations AS tcc
            ON rslt.bid = tcc.bid
        LEFT JOIN tmp_result_with_associated_refunds AS assr
            ON rslt.bid = assr.bid
        LEFT JOIN tmp_check_in_status AS cia
            ON rslt.bid = cia.bid
        LEFT JOIN tmp_bids_flagged_FR AS tfr
        ON rslt.bid = tfr.bid
        LEFT JOIN tmp_refund_requests AS ref
            ON rslt.bid = ref.bid
        LEFT JOIN tmp_chargebacks AS chc
            ON rslt.bid = chc.bid
        LEFT JOIN tmp_declined_fraud_checks AS fch
            ON rslt.bid = fch.bid
        ...
        CROSS JOIN max_fwid mf
        QUALIFY rn = 1
);
```

Dále následuje znovu merge statement jako u předchozí tabulky. Když je skript pro obě tabulky vytvořen a finální transformace je úspěšná a otestovaná, je nutná kontrola dat. Na obrázku č. 6 je demonstrovaná ukázka dat v tabulce. Byl vybrán pouze malý fragment sloupců, kterých je v tabulce 45.

Row	score	comment	survey_status.code	survey_type.code	device.code	device.vendor
1	1	Way too hard t...	completed	ease_of_use	*null*	*null*
2	4	*null*	completed	ease_of_use	oppo cph2211	oppo
3	2	*null*	completed	mmb_experience	oppo cph2477	oppo
4	2	*null*	completed	ease_of_use	oppo cph2477	oppo
5	1	*null*	completed	ease_of_use	oppo cph2127	oppo
6	1	*null*	completed	mmb_experience	oppo cph2127	oppo
7	3	Sredno	completed	ease_of_use	oppo cph2195	oppo
8	5	*null*	completed	ease_of_use	oppo rmx3241	oppo
9	5	*null*	completed	mmb_experience	oppo rmx3241	oppo
10	1	Impossible to a...	completed	ease_of_use	oppo cph2251	oppo
11	1	*null*	completed	mmb_experience	oppo cph2251	oppo
12	1	Cannot print or ...	completed	ease_of_use	oppo cph2251	oppo
13	1	*null*	completed	mmb_experience	oppo cph2251	oppo
14	2	Nulla da dichiar...	completed	ease_of_use	oppo cph2211	oppo
15	2	*null*	completed	mmb_experience	oppo cph2211	oppo
16	5	*null*	completed	ease_of_use	iphone	apple
17	4	navigating dest...	completed	ease_of_use	ios-device	apple
18	1	Are you fucking...	completed	ease_of_use	iphone	apple
19	5	good	completed	ease_of_use	iphone	apple
20	5	*null*	completed	ease_of_use	ipad	apple

Obrázek 7: Ukázka dat z výsledné tabulky (vlastní zpracování)

V poslední řadě je nutné vytvořit prostý select z již vytvořené flat_survey tabulky, který poslouží pro tvorbu flat_sentiquare. Tato tabulka bude později použita pro účely týmu, který vytvoří pomocí API adresy napojí data do platformy na kategorizaci a překlad. Důvodem k vytvoření třetí tabulky je žádost analytiků propojit informace z dalších modelů přímo jen do sentisquare, ne do celého modelu, jejich přítomnost v Lookeru není nutná.

```
CREATE TABLE IF NOT EXISTS `commercial-sandbox-69f69f41.cx.flat_sentisquare`(
    event_id STRING,
    visitor_id STRING,
    session_id STRING,
    assessment_id INTEGER,
    answer_id INTEGER,
    bid INTEGER,
    score NUMERIC,
    comment STRING,
    ... );

CREATE TEMP TABLE tmp_slot AS(
  SELECT
    bid,
    ROW_NUMBER()OVER(PARTITION BY bid ORDER BY requested_timestamp DESC) AS rn
  FROM `xxx.project.flat_garance`
```

```
  WHERE typ_garance = 'slot'
  AND CAST(requested_timestamp AS DATE) >= '2024-12-15' - 274  --jsou vybrány
data za posledních 9 mesíců.
  QUALIFY rn = 1);

CREATE TEMP TABLE tmp_survey AS(
SELECT DISTINCT
        prep.event_id,
        prep.visitor_id,
        prep.session_id,
        prep.assessment_id,
        prep.answer_id,
        prep.bid,
        ...
    FROM survey_tp_prep prep
        ...
    LEFT JOIN tmp_ slot AS shot
      ON prep.bid = shot.bid);
```

5.2.1 Tvorba DAGu

Dalším krokem je vytvoření DAGu, který byl představen v teoretické části. Tento python soubor bude zajišťovat, že se skripty pro obě tabulky spustí v danou dobu. A vzhledem k tomu, že bylo v tabulkách užito i jiných zdrojových tabulek z jiných modelů, budou vytvořeny sensory, zaručující správnou posloupnost.

Prvním krokem při tvorbě DAGu je import knihoven, které budou využity.

```
import jinja2
from datetime import datetime, timedelta

from dags.config import var
from airflow.models import DAG

from airflow.providers.google.cloud.operators.bigquery import
BigQueryInsertJobOperator
from astronomer.providers.core.sensors.external_task import (
    ExternalTaskSensorAsync as ExternalTaskSensor,
)
from astro_plugins.sensors.astro_http_sensor import AstroHttpSensor

from astro_plugins.utils import retrieve_sql_path
from astro_plugins.utils.sensor import get_execution_times
```

Následně bude nadefinován název, vlastník modelu, email pro odesílání eror upozornění, čas spuštění, sla, která určují maximální dobu, po kterou má DAG vykonat spuštěnou úlohu, kolikrát opakuje danou úlohu, dokud neselže, po kolika minutách bude úloha opakována. Tento konkrétní DAG bude poběží od 20.12.2023 a bude se spouštět jednou denně, od půlnoci.

```
dag = DAG(
    dag_id="survey_model_1d",
    default_args={
        "owner": "CX",
        "email": "alerts-tym@slack.com",
        "email_on_failure": True,
        "email_on_retry": False,
        "depends_on_past": False,
        "start_date": datetime(2023, 12, 20, 0, 0, 0),
        "sla": timedelta(hours=6),
        "retries": 10,
        "retry_delay": timedelta(minutes=15),
    },
    schedule_interval=timedelta(days=1),
    max_active_runs=1,
    template_searchpath=retrieve_sql_path(__file__),
    template_undefined=jinja2.Undefined,
    params={
        "src_gcp_project_id": var("src_gcp_project_id"),
        "dst_gcp_project_id": var("dst_gcp_project_id"),
    },
)
```

Níže je představena definice sensoru, na který později bude odkázáno při definici závislosti na cizích modelech. Pro tuto definici je nutné znát DAG id, jméno úlohy, api údaje, v jakém intervalu daná úloha běží, a povolené statusy.

```
def get_astro_http_sensor(dag_id, task_id, delta, api_credentials,
allowed_states):
    return AstroHttpSensor(
        dag=dag,
        task_id=f"{dag_id}_{task_id}_sensor" if task_id else f"{dag_id}_sensor",
        external_dag_id=dag_id,
        external_task_id=task_id,
        api_credentials=api_credentials,
        poke_interval=60,
        timeout=8 * 60 * 60,
        allowed_states=allowed_states,  # skipped status indicates that no new
data was present for processing
        execution_date_fn=lambda execution_date: get_execution_times(
```

```
            execution_date=execution_date,
            delta_minutes=delta,
            end_delta=dag.schedule_interval,
        ),
    )
astro_http_sensors = [
    get_astro_http_sensor(dag_id, task_id, delta, api_credentials,
allowed_states)
    for dag_id, task_id, delta, api_credentials, allowed_states in [
        (
            "nazev_dagu_1d",
            "nazev_tasku_123",
            6 * 60,
            "api_credentials",
            ["success", "skipped"],
        ),
        (
            "nazev_dagu_6h",
            "nazev_tasku_374",
            6 * 60,
            "api_credentials",
            ["success", "skipped"],
        )
```

V dalším kroku bude definována úloha pro všechny tabulky. Je nutné vytvořit název úlohy a implementovat konfiguraci, kde bude nadefinována cesta k souboru obsahující skript. V posledním kroku jsou nastaveny závislosti. Jako první poběží všechny úlohy v sekci astro_http_sensors, poté prep tabulka, následně flat_survey a v poslední řadě flat_sentisquare.

```
flat_survey = BigQueryInsertJobOperator(
    task_id="flat_survey",
    configuration={
        "query": {
            "query": "{% include 'sql/flat_survey.sql' %}",
            "useLegacySql": False,
        },
        "labels": query_labels,
    },
    **common_args_tasks,
)

flat_sentisquare = BigQueryInsertJobOperator(
    task_id="flat_sentisquare",
    configuration={
        "query": {
            "query": "{% include 'sql/flat_sentisquare.sql' %}",
            "useLegacySql": False,
```

```
        },
        "labels": query_labels,
    },
    **common_args_tasks,
)

prep_frontend_logs = BigQueryInsertJobOperator(
    task_id="prep_frontend_logs",
    configuration={
        "query": {
            "query": "{% include 'sql/prep_frontend_logs.sql' %}",
            "useLegacySql": False,
        },
        "labels": query_labels,
    },
    **common_args_tasks,
)
(
    astro_http_sensors
    >> prep_frontend_logs
    >> flat_survey
    >> flat_sentisquare
)
```

5.2.2 Vytvoření terraformu

Pro všechny tabulky je nyní nutné vytvořit terraform, který bude zaručovat, že tabulka má správnou strukturu. Pro jakékoliv budoucí změny je nutné nejdříve vytvořit Merge Request (MR) s nově vytvořeným sloupcem, či změnou. Tento MR bude vytvořen pomocí GitLabu, kde je nutné mít potvrzení dalšího spolupracovníka pro provedení změny. V terraformu budou nastaveny také partitioning a clustering pro dané tabulky. Tento krok tudíž zaručuje, že je zaznamenávána jakákoliv činnost, která se s tabulkami děje, čímž se omezuje chybovou a díky clusterování a partitioningu se zlepší performance databáze při zpracovávání frontendových dat.

Nejdříve bude nadefinovaná struktura a datový typ sloupců pro každou tabulku. Standardem je také definovat popis, co daný sloupec představuje, aby se zamezilo jakémukoliv zmatení při práci s daty. Níže je příklad pro flat_survey.

```
[{
    "name": "survey_fwid",
    "mode": "REQUIRED",
    "type": "INTEGER",
    "description": "Unique identifier of the flat table row"
  },
  {
```

```
    "name": "event_id",
    "mode": "NULLABLE",
    "type": "STRING",
    "description": "The unique identifier of the log in event_logs_nested."
  },
  ...
  {
    "name": "device",
    "mode": "NULLABLE",
    "type": "RECORD",
    "description": "Denotes distinction of device model",
    "fields": [
        {
          "name": "code",
          "mode": "NULLABLE",
          "type": "STRING",
          "description": "Device model customer used."
        },
        {
        "name": "vendor",
        "mode": "NULLABLE",
        "type": "STRING",
        "description": "Name of the brand or vendor of the device"
        },
        ...
        }]}]
```

Takto vytvořené JSON soubory jsou dále nadefinovány do terraform souboru, kde je nastaven jejich clustering a partitioning. Nejdříve se nastaví module, který už je pro daný projekt vytvořen, díky dalším modelům, čili stačí přidat pouze nově vytvořené tabulky. Každý servisní účet, kterému nejsou přidělena práva, nebude mít přístup k daným tabulkám. Flat_sentisquare i flat_survey tabulka bude mít nastavený partitioning po dni, tudíž při výběru dat za daný den, skript nebude procesovat všechna data a hledat vybrané datum, ale bude hledat podle indexu přiřazeného pro daný den, čímž se zlepší běh databáze. Stejně tak clustering je nastaveno pro základní ID v každé tabulce. Clustering tyto ID seřadí k sobě, aby v případě výběru konkrétního ID procesovat pouze dané sekce, ve kterých data ukládá. Ve flat_survey se nejvíce používají v JOINech například event_id, visitor_id a session_id, proto jsou nastavena pro clustering.

```
module "bq_access_project" {
  source = "xus/infra/bigquery-data-access/prod"
  version = "~> 0.1.1"

  dataset_id = "project"
  data_viewers = [
    "group:673@spolecnost.com",
    "serviceAccount:astro-tym@0829.iam.gserviceaccount.com",]
  ...
```

```
  tables = [
    {
      table_id = "flat_sentisquare"
      schema   = file("bq_schemas/project/flat_sentisquare.json")
      description = "Flat table preparing survey data for a transfer to
sentisquare"
      time_partitioning = {
        type = "DAY"
        field = "survey_answered_at",
      }
      range_partitioning = null
      clustering = ["event_id", "visitor_id", "session_id"]
      expiration_time    = null
      labels             = {}
    },
    {
      table_id = "flat_survey"
      schema   = file("bq_schemas/project/flat_survey.json")
      description = "Fact table where each record represents a survey answer.“
      time_partitioning = {
        type = "DAY"
        field = "survey_initiated_at",
        expiration_ms = null
        require_partition_filter = false
      }
      range_partitioning = null
      clustering = [
        "session_id", "visitor_id", "event_id"]},
  ...
```

Po vyhotovení kompletního DAGu, skriptů, terraformu a otestování v lokálním prostředí, je možné vytvořený MR poslat na schválení a poslat tyto změny na produkci. Po aplikaci MR se díky terraformu vytvoří tabulky a bude manuálně zapnut DAG, aby se tyto tabulky plnily každý den. Je možné manuálně nastavit zpětné doplnění modelu, aby byly dostupná data z minulosti.

5.2.3 Napojení na Looker a vytvoření Exploru

Momentálně jsou upravená data dostupná v databázi. Dalším krokem je propojení dat do platformy pro vizualizaci a tvoření dashboardu. Výsledná tabulka flat_survey je tedy nutná nadefinovat v developerském prostředí v jazyku LookML. Nejdříve zde bude vytvořeno propojení s BigQuery a poté definice sloupců, jejich datového typy a metrik, které budou klíčové pro analytiky. Jednou z nejvíce používaných metrik v celé společnosti je Net Promoter Score, která určuje jak jsou zákazníci obecně spokojení s nabízeným produktem,

například se vyhodnocují A/B testy pomocí této metriky. Dalšími jsou například počet zobrazených dotazníků, počet zodpovězených dotazníků, což udává poměr v jakém je úspěšnost odpovídání. Dále pak průměrné skóre. Tyto metriky jsou možné kombinovat s dimenzemi, čili je možné například vyhodnotit jaké je skóre pro každý typ dotazníku, etc.

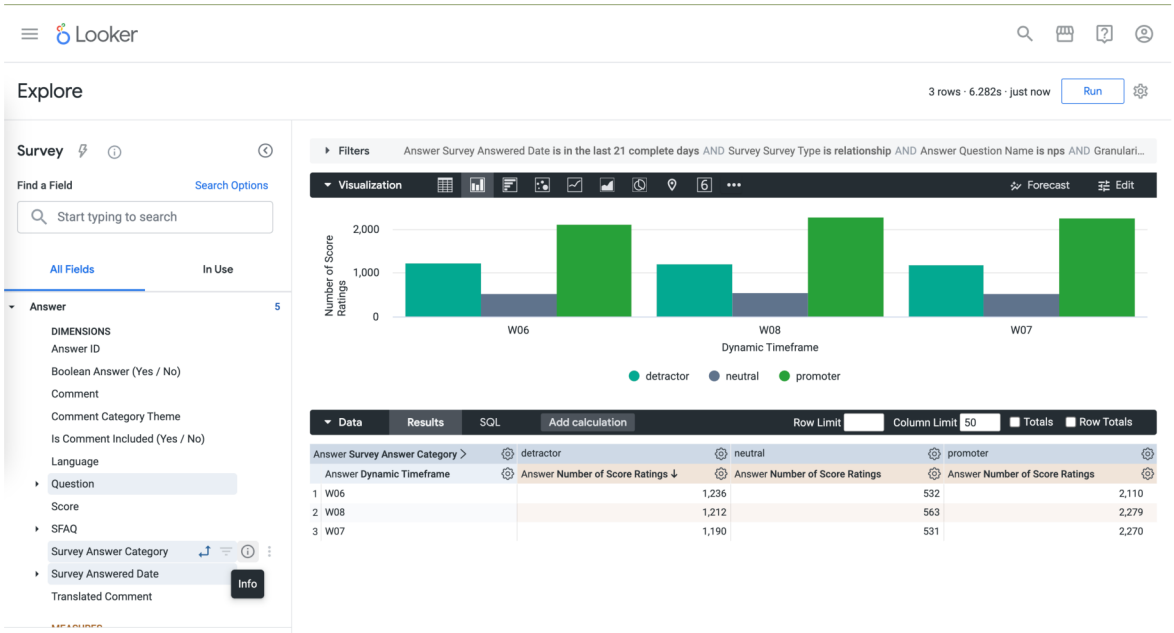
```
view: flat_survey {
  view_label: "Survey"
  sql_table_name: `xxx.project.flat_survey` ;;

    dimension: event_id {
      type: string
      label: "Event ID"
      description: "The unique identifier of the Frontend Log"
      value_format_name: id
      sql: ${TABLE}.event_id ;;
    } ...

## MEASURES:

measure: displayed_surveys_count {
      label: "Number of Only Displayed Surveys"
      description: "Number of unanswered surveys"
      type: count_distinct
      filters: [survey_status__code: "-completed"]
      sql: CONCAT(${visitor_id}, ${session_id}, ${survey_type__code});;
    }
measure: average_score {
      type: average
      description: "Average of survey answers score."
      value_format: "0.00"
      sql: ${score} ;;
    }
measure: net_promoter_score {
      label: "Net Promoter Score"
      description: "Net Promoter Score"
      type: number
      sql: (${promoter_count}-
            ${detractor_count})*1.0/NULLIF(${nps_count}, 0) *100;;
      value_format: "0.00"
      drill_fields: [bid, max_score_nps, max_value_text]
    } ...
```

Následně se definuje celý model, kde se napojí již definovaná tabulka. Lze připojit i jiné modely a vytvořit tak možnost pro analytiky vybírat informace z jiných modelů, pokud sdílejí stejné ID.

```
## Survey model
```

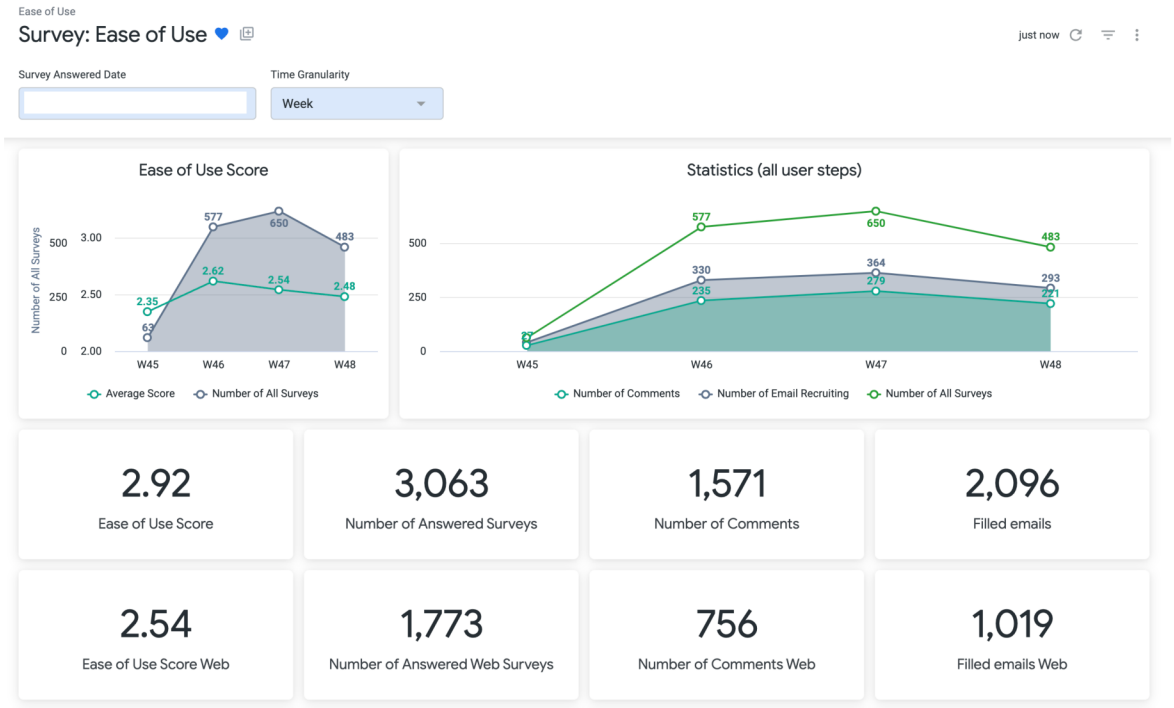
```
explore: survey {
  description: "The Survey Model includes data from all of the company's surveys
such as Ease of Use, NPS, Refund Survey, SFAQ"
  group_label: "Customer Experience"
  hidden: no
  label: "Survey"
  view_name: flat_survey
  extends: [survey]
  fields: [ALL_FIELDS*]

  join: flat_customer_info {
    view_label: "Customer Info"
    relationship: many_to_one
    sql_on: ${flat_survey.bid} = ${flat_customer_info.bid};;
  }

  join: flat_agent {
    view_label: "Agent"
    relationship: one_to_one
    type: left_outer
    sql_on: ${flat_survey.agent__email} = ${flat_agent.agent_person__email};;

  join: parameter {
    view_label: "Granularity (applicable to dynamic dimensions)"
  }
```

Následně se znovu vytvoří MR pro schválení, změny se aplikují a model je připravený pro analytiky na vytvoření dashboardu. Explore pro daný model je představen na obrázku č.7, kde lze vidět příkladná analýza počtu zákazníků, kteří patří do skupin detractor, neutral nebo promoter, což vyhodnocuje, kolik zákazníků je spokojeno se vzhledem stránek. Již v samotném exploru je možné vytvořit vizualizaci s filtry dle vlastního výběru. Vizualizace tohoto typu je pak možná uložit do dashboardu, který se sdílí s produktovými manažery. Konkrétně v této analýze byly použity filtry pro typ dotazníku, typ otázky a filtrace časového úseku 21 dní. V položce *Visualization* je možné vybrat další typy vizualizace, ku příkladu čárkový graf, korelační diagram, koláčový graf a další. Je možné zde vytvořit taky personalizované metriky, což je obzvláště přínosné pro analytiky, kteří nejsou zběhlí v LookML jazyku, nebo nemají developerská práva.



Obrázek 8: Ukázka dat z Looker Exploru (vlastní zpracování)

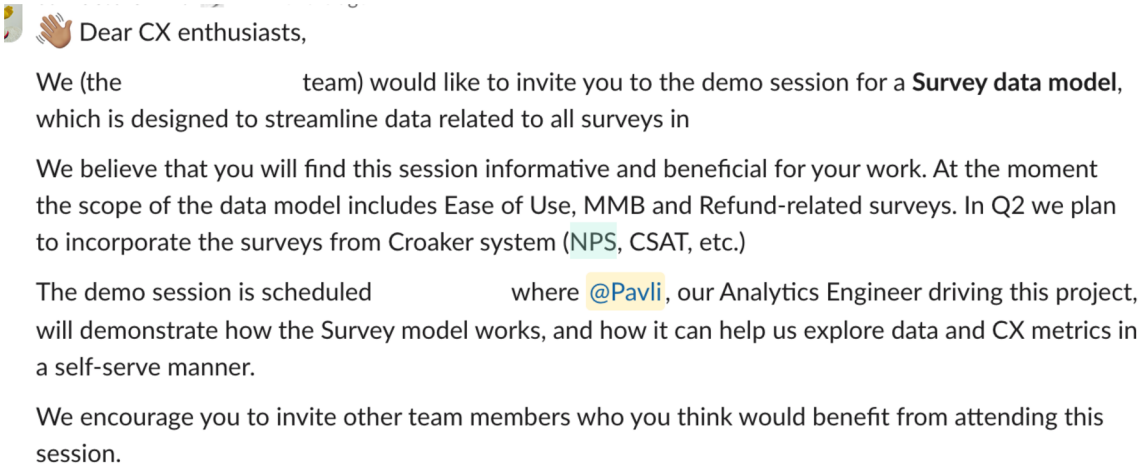
Na obrázku č.8 je příklad dashboardu, kde jsou naznačeny příkladné vizualizace, které by analytici sdíleli s vyšším managementem. Takto vytvořený dashboard si je možné uložit do výběru, pro lepší dostupnost a obnovuje se každý den.



Obrázek 9: Ukázka dat z Looker Exploru (vlastní zpracování)

5.2.4 Vytvoření dokumentace

Finálním krokem je vytvoření dokumentace s popisem pro všechny metriky a všechny sloupce. Také by měl být součástí odkaz na DAG, tabulky v BigQeury, Explore a Dashboard v Lookeru. Pro maximální informovat bylo oznámení tohoto modelu zveřejněno na komunikační platformě a byla vytvořena prezentace, pro všechny analytiky zabývající se zákaznickým prožitkem, kde bude demonstrováno, jak s daným modelem zacházet.



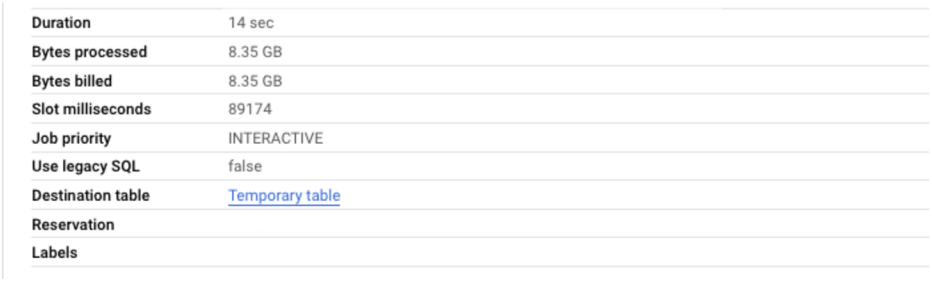
Obrázek 10: Oznámení první verze demonstrace modelu (vlastní zpracování)

5.3 Zhodnocení navrhovaného řešení

V rámci návrhu řešení bylo zaměřeno na komplexní návrh a implementaci datového modelu, přičemž byl kladen důraz na zvýšení efektivity databáze BigQuery. K tomu je potřeba sdílet výsledky, které obě tabulky z hlediska výkonnosti performují.

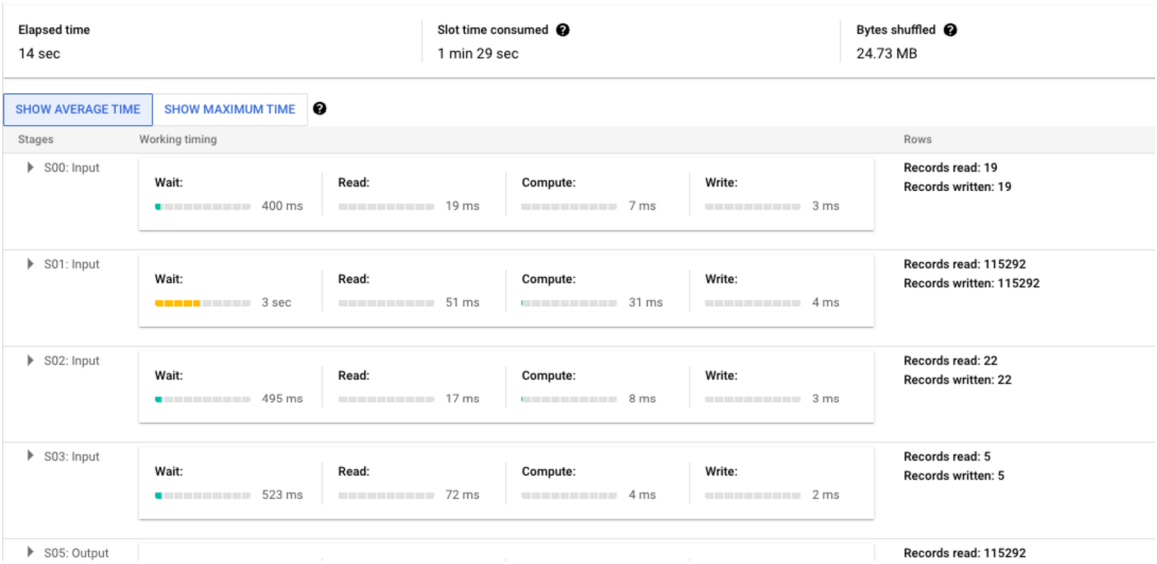
5.3.1 Výkonnost databáze

Prep tabulka a její výkonnost je zaznamenána na obrázku č.9 a 10, tato úloha byla dokončena za 14 sekund, což svědčí o rychlém provedení, zejména s ohledem na objem dat. Úloha zpracovala 8,35 gigabajtů (GB) dat. Dotaz využil 89 174 milisekund slotů, což odráží množství výpočetních prostředků použitých v průběhu času (oficiální dokumentace Google BigQuery, 2023). Cena by zde byla odhadnuta na 8,35 GB/den × $6,25 TiB × 30 dnů/měsíc = 0,05 USD/měsíc



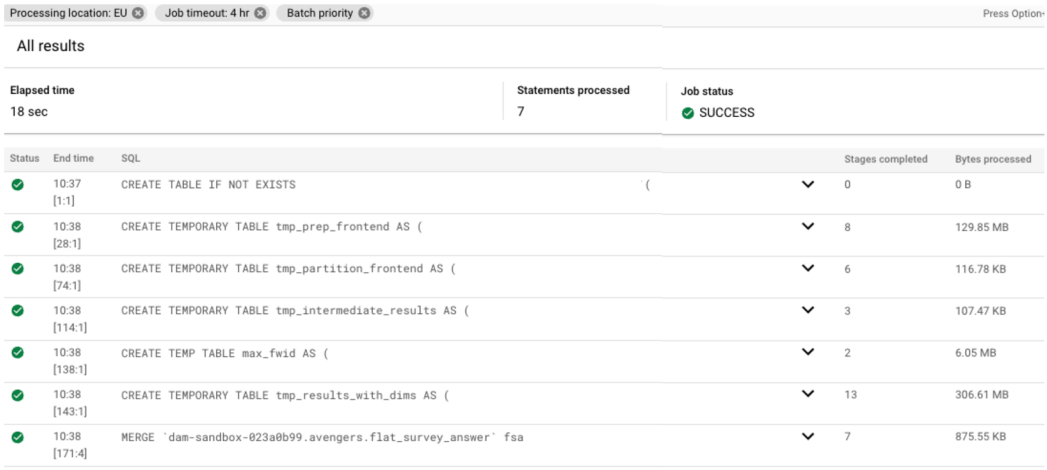
Duration	14 sec
Bytes processed	8.35 GB
Bytes billed	8.35 GB
Slot milliseconds	89174
Job priority	INTERACTIVE
Use legacy SQL	false
Destination table	Temporary table
Reservation	
Labels	

Obrázek 11: Zobrazení vytíženosti příkazů prep_frontend (vlastní zpracování)



Obrázek 12: Detailnější zobrazení vytíženosti příkazů prep_frontend (vlastní zpracování)

Tabulka flat_survey procesovala celkem 306, 61 MB za 14 sekund a je složena z několika příkazů, což lze viděl na obrázku č. 12. Pro detailní zobrazení jako u prep tabulky bychom museli zanalyzovat každý jednotlivý příkaz, proto pro celkové zhodnocení zprocesovaných dat a rychlosti bude použito jen celkového souhrnu.



Obrázek 13: Celkové zobrazení vytíženosti příkazů pro flat_survey (vlastní zpracování)

Tabulka flat_sentisquare procesovala 19.58 GB během 1 min 51 s. Celkem tedy všechny tři tabulky zprocesovaly 28,24 GB dat a doba trvání byla 2 minuty a 19 sekund. Z poskytnutých informací vyplývá, že spuštěné skripty v rámci databáze vykazují vysokou úroveň efektivity. Prvním důležitým faktorem je rychlost provedení, například úloha zpracovávající 8,35 GB dat byla dokončena za pouhých 14 sekund, což naznačuje schopnost databáze efektivně zpracovávat velké objemy dat v krátkém časovém úseku.

Dále, analýza využití výpočetních zdrojů ukazuje, že dotazy využily výpočetní prostředky efektivně. Skutečnost, že úloha využila 38 minut a 14 sekund slotů, naznačuje, že databáze dokázala optimalizovat využití výpočetních zdrojů a minimalizovat ztráty času. Tato efektivní správa výpočetních zdrojů přispívá k celkové rychlosti a výkonu databáze.

Nízké náklady bez paušální sazby jsou dalším důležitým aspektem. Cena zpracování celkových 28,24 GB dat byla odhadnuta na 0,17 USD/měsíc. S ohledem na objem zpracovávaných dat a rychlost provedení lze tuto cenu považovat za přiměřenou a efektivní v kontextu poskytovaných služeb. Tento faktor je klíčový zejména pro organizace s omezeným rozpočtem, které hledají efektivní a ekonomická řešení. Pro vybranou společnost je však hlavní množství obsazených slotů, které jsou velmi nízké.

Tabulka 4: Metriky navrženého datového modelu (vlastní zpracování)

	Momentální řešení při průměrném počtu spuštění denně 10,86x	Průměrný Standard datového modelu ve společnosti.	Navržený datový model
Čas proběhnutí	-	120 s	139 s
Čas ve slotu	467 min 20 s	70 min	38 min 14 s
Procesovaná data	5 913,7 GB	1095 GB	28,24 GB
USD/měsíc za spuštění 1x denně skript bez paušální sazby	$36,10	$6,67	$0,17

Z porovnání stavu před implementací navrhovaného řešení, datového standardu ve vybrané společnosti a navrženým řešení lze vidět, že nově vybudovaný model splňuje standard. Procesovaný čas je delší, ale podstatné jsou spotřebované sloty, které jsou naceněné. Cena bez paušální sazby je zde čistě orientační, jelikož společnost platí za počet slotů. Datový model a jeho schopnost rychle zpracovat velké objemy dat, efektivně využívá výpočetní zdroje a tím minimalizuje náklady a přispívá k celkové úspěšnosti a efektivitě databázového prostředí. Čili cíl zefektivnit běh databáze byl splněn.

5.3.2 Hodnocení splnění požadavků analytiků

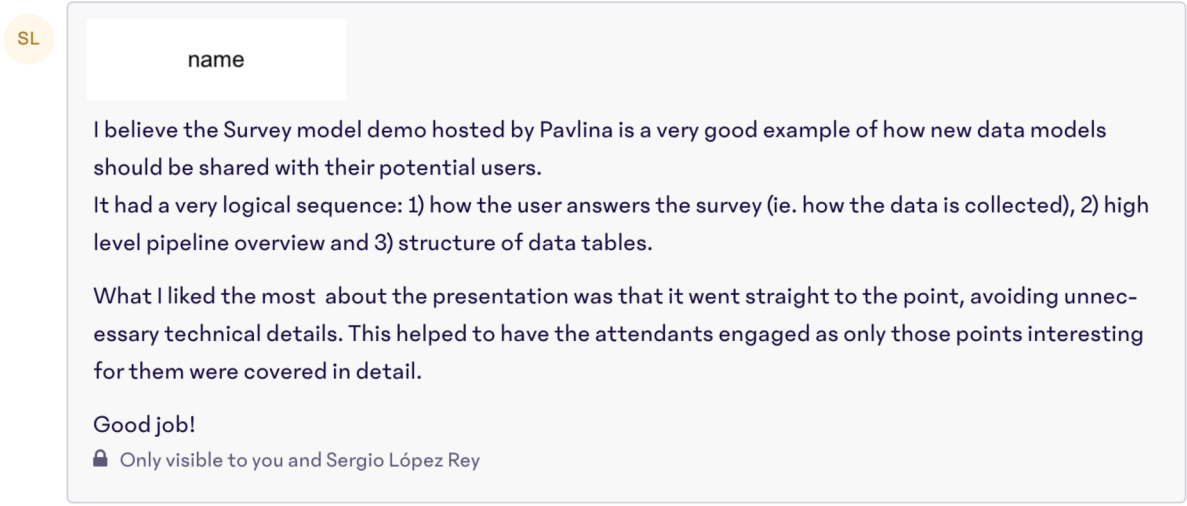
Mezi požadavky definované datovými analytiky patřilo snížit chybovost dat, zautomatizovat proces ukládání transformovaných dat, a mít transparentní a lehce komunikovatelné výsledky analýzy pomocí analytické platformy.

Navrhované řešení přináší významné vylepšení v procesu zpracování a analýzy dat s ohledem na snížení chybovosti. Tento model definuje atributy a jejich vztahy, což minimalizuje riziko nedorozumění a chybných interpretací dat. Díky vytvoření terraformu, který sleduje všechny změny ve struktuře tabulek je chybovost v změně struktury modelu značně redukována. Automatizace selekce dat také hraje klíčovou roli v redukci chyb. Eliminuje manuální zásahy, zajišťuje konzistentní aplikaci pravidel a transformací, což výrazně snižuje možnost vzniku chyb. Také rozdělení procesu na dvě fáze, "prep tabulka" a "flat tabulka", dále zlepšuje správu a ladění procesu. Toto oddělení umožňuje jasně definovat úkoly a cíle pro každou fázi, což zvyšuje transparentnost a usnadňuje správu celého procesu. Vytvořené analýzy v Lookeru jsou kontrolovány celým týmem analytiků, kteří mají přístup do kalkulací a výpočtů. Celý model je pod správnou Data Modeling týmu a za samotné analýzy je jmenován jeden člověk z týmu analytiků, který pravidelně kontroluje jejich správnost. Rozložení těchto zodpovědností maximalizuje kontrolu nad daty a chybnými změnami, které by mohly být aplikovány. Všechny tyto aspekty zajišťují minimalizaci chybovosti v celém procesu a tím je tento požadavek splněn.

Díky implementaci řady algoritmů a nástrojů umožňujících automatizované zpracování dat se podařilo minimalizovat potřebu lidské intervence. Skripty nejsou nutné na týdenní bázi spouštět. Data jsou dostupná každý den, díky DAGu a data se automaticky propojují s platformou Looker, kde se také analýzy automaticky plní do vytvořených Dashboardů. Čili pro každodenní analýzy není potřeba zásahu nikoho ze společnosti. Proces byl plně automatizován. Pouze v případě ad-hoc analýz je nutnost zásahu. Tato automatizace vede k významnému zvýšení efektivity celého procesu. Důležitým přínosem je také úspora času a zdrojů, kterou tato automatizace přináší. Efektivní využívání dostupných prostředků a dosažení lepších výsledků v kratším čase je zásadní pro úspěšné fungování procesu analýzy dat. Z těchto důvodů lze tedy konstatovat, že úkol automatizace celého procesu analýzy dat byl úspěšně splněn.

Implementace Lookeru umožnila vytvoření uživatelsky přívětivého rozhraní, které umožňuje snadnou vizualizaci a interpretaci výsledků analýzy dat. Díky interaktivním dashboardům a grafům je možné data prezentovat způsobem, který je srozumitelný a přístupný všem uživatelů, včetně těch, kteří nemají hluboké znalosti analytických nástrojů. Looker poskytuje také možnost vytváření různých reportů a sdílení dat prostřednictvím různých kanálů, což napomáhá transparentnosti výsledků analýzy. Uživatelé mohou snadno přistupovat k relevantním informacím a rychle porozumět významu dat díky popiskům, což podporuje efektivní rozhodování a spolupráci v rámci organizace. Další výhodou Lookeru je jeho flexibilita a škálovatelnost, což umožňuje přizpůsobení analytických procesů konkrétním potřebám a změnám v prostředí organizace. To vše přispívá k dosažení cíle mít transparentní a lehce komunikovatelné výsledky analýzy, čili tento cíl je splněn.

Po zavedení datového modelu a provedení informační prezentace Survey modelu analytikům, byla podána pozitivní zpětná vazba: „S modelem se jednoduše pracuje a jeho vizualizace je jednodušší, když máme všechny tabulky napojené na Looker“ (analytička vybrané společnosti, 2023), na obrázku 13 je zpětná vazba od Head of CX, vedoucího oddělení pro zákaznický prožitek.



Obrázek 14: Hodnocení vedoucího oddělení (vlastní zpracování)

5.3.3 Minimalizace rizik

Navrhovaný datový model přináší řadu vylepšení a optimalizací, které pomáhají minimalizovat rizika spojená s databázovými systémy a datovou analytikou. Použití Terraformu k správě datových modelů a tabulek je základem pro správu infrastruktury jako

kódu. Tento přístup umožňuje lepší sledování změn, zlepšuje dokumentaci a usnadňuje provádění změn ve datových modelech a strukturách tabulek. V důsledku toho je sníženo riziko neúmyslných změn a zlepšuje se spravovatelnost tabulek. Přítomnost dokumentace také minimalizuje jakékoliv obavy o složitost správy datového modelu. Bylo vytvořeno také demonstrační video, kde jsou popsány základní myšlenky modelu a vysvětlena práce s modelem. Riziko obtížné správy je minimalizováno.

Vytvořením speciální přípravné "prep" tabulky je zajištěna pružnost v případě změn ve struktuře dat bez nutnosti zpracování velkého množství surových dat. Tato tabulka efektivně funguje jako vyrovnávací paměť, která ukládá data v nepříliš zpracovaném stavu a připravuje je pro další transformace a analýzy, čímž se redukuje zátěž na systému a šetří se zdroje. Flat tabulka na druhou stranu umožňuje rychlý a efektivní přístup k datům pro analytické a reportingové účely bez potřeby složitého joinu a dotazů, což optimalizuje výkon při dotazech na data. Navíc tento přístup znamená, že analýza a reporty budou méně náročné na zdroje a rychlejší, čímž se zvyšuje celková efektivita procesů, riziko zpomalení databáze bylo minimalizováno, je však nutné při každé změně v logice a kódu otestovat, zdali se dotazy nestaly příliš náročnými.

Skriptování a automatizace selekce dat pomocí DAGu dodává procesu flexibilitu a zároveň eliminuje riziko lidské chyby při manuální selekci a transformaci dat. DAG zajistí, že se všechny úkoly vykonávají ve správném pořadí a ve správný čas, a to znamená, že data budou vždy aktualizována a připravena pro analýzu ve stanovených intervalech a tímto budou minimalizovány problémy s přesností dat.

Riziko, že datový model nebyl navržen správně, je zde minimální z několika důvodů. Předně, použití osvědčených postupů v oblasti databázového modelování ujišťuje, že datový model odpovídá konkrétním potřebám analýzy a zpracování dat. Tento přístup zajistí, že datový model bude mít stabilní základy a bude se správně vyrovnávat s různými scénáři využití dat. Pečlivé plánování a konstrukce skriptů a DAGů s jasně definovanými senzory a kroky zajišťují, že zpracování dat proběhne v přesně stanoveném pořadí, což minimalizuje možnost chyby kvůli nesprávné koordinaci datových toků. Dále, použití nástroje Terraform zajišťuje, že veškeré změny jsou zaznamenány ve verzích a jsou systematicky sledovány. To umožňuje rychlou reakci na jakékoli problémy a zaručuje průběžné vylepšování datového modelu podle aktuálních potřeb a požadavků. Tato transparentnost a možnost vrá

vrátit provedené změny zpět v případě potřeby znamená, že jakákoli potenciální chyba v designu lze rychle identifikovat a napravit. Integrace s Lookerem přináší další vrstvu ověření, protože vizualizace dat může odhalit nesrovnalosti nebo problémy, které by mohly být přehlédnuty v surových datových sadách. Analýza dat pomocí nástrojů jako Looker pomáhá rychle identifikovat potenciální slabiny v datovém modelu a napomáhá jejich korekci. Na závěr, pečlivé dokumentování celého procesu a architektury datového modelu, spolu s adekvátním sdílením informací a tréninkem uživatelů, zajišťuje, že všechny aspekty datového modelu jsou srozumitelné a transparentní. Uživatelé jsou tak schopni správně interpretovat data a používat datový model tak, jak byl navržen. Vzhledem ke všem těmto bezpečnostním opatřením a procesům je možné s důvěrou konstatovat, že datový model byl navržen správně a případná rizika spojená s jeho návrhem byla efektivně minimalizována.

Pro minimalizaci špatného školení bylo poskytnuto základní školení a materiály, které analytikům pomohou pochopit a efektivně využívat model. Nicméně, úroveň porozumění a odbornosti, kterou si analytici vyvinou, závisí převážně na jejich osobním odhodlání a ochotě se učit. V rámci této role byla představena prezentace a demonstrace s cílem seznámit analytiky s modelem a daty, avšak plná odpovědnost za další vzdělávání leží na samotných analyticích. Vedle toho, součástí předání modelu je také podrobná dokumentace, která slouží jako vzdělávací nástroj a zdroj informací pro uživatele modelu. Přestože je autor připraven poskytovat podporu, skutečné zvládnutí a uplatnění modelu vyžaduje průběžnou práci a angažovanost každého analytika zvlášť.

Riziko zvýšených nákladů bylo minimalizováno díky stanovení požadavků na nový systém. Bylo dbáno na to, aby byly zohledněny všechny nezbytné atributy a jejich vztahy, což snižuje náklady spojené s případnými pozdějšími úpravami a nekompatibilitami. Díky automatizaci je splněn cíl kompatibility s dalšími platformami, čímž se také snižuje riziko zvýšených nákladů při extenzivní správě.

Dalším z klíčových prvků je již zmíněná normalizace a denormalizace, které jsou základními stavebními kameny pro efektivní strukturaci dat. Normalizace datového modelu snižuje nadbytek nepotřebných dat a zvyšuje jejich integritu, což je zásadní pro zajištění správnosti a konzistence dat. Tímto způsobem se minimalizují rizika vzniku nekonzistentních dat, která by mohla vést k chybným analýzám a rozhodnutím založeným na nesprávných datech. Ve flat tabulce jsou provedeny transformace, díky kterým jsou data

deduplikována v případě, že se duplicity objeví ve zdrojových tabulkách. Díky terraformu, kde jsou některé sloupce vytvořeny jako nutné, tudíž nikdy nesmí být nulové je vytvořená jakási kontrola proti jakýmkoli chybám na zdroji. Tedy v případě špatné kvality dat bude tým datových analytiků okamžitě upozorněn. Tímto je dopad minimalizován, výskyt v modelu potlačen a ve zdrojových tabulkách vytvořeny protiopatření na okamžité upozornění.

Riziko nedostatku dat nebylo nijak minimalizováno, jelikož je to těžce ovlivnitelné a nebylo možné ho zcela minimalizovat. Lze však předpokládat, že organizace přijala alternativní strategie pro získání a doplnění potřebných dat. To zahrnuje rozšíření datového získávání prostřednictvím různých kanálů, jako je sledování interakce zákazníků na webových stránkách, analýzy sociálních médií, poskytnutí dalších pobídek pro zákazníky, aby poskytli zpětnou vazbu, nebo vytvoření partnerství s externími firmami pro získání relevantních tržních dat.

Níže v tabulce č.5 lze vidět dřívější klasifikace a nynější stav po zavedení modelu. Celkově navrhovaný datový model minimalizuje rizika související s chybami v datech, jejich integritou, efektivitou zpracování, auditovatelností, uživatelským pochopením a vizualizací, čímž podporuje vytvoření spolehlivého, bezpečného a efektivního datového eko-systému.

Tabulka 5: Klasifikace rizik po návrhu řešení(vlastní zpracování)

	Riziko	**Pravděpodobnost výskytu (1-5)**	**Dopad (1-5)**	**Hodnota rizika**
1.	1. Obtížnost správy	4 → 2	3 → 2	**SR 12**
2.	4. Zpomalení databáze	3	5	**VR 15**
3.	6. Problémy s přenosností dat	1	3 → 2	NR 3
4.	9. Nesprávný návrh datového modelu	3 → 1	3	**SR 9**
5.	11. Riziko špatného školení	2 → 1	4	SR 8
6.	13. Riziko zvýšení nákladů	2	5	**VR 10**
7.	14. Špatná datová kvalita	3 → 1	4 → 3	**VR 12**
8.	15. Nedostatek dat	1	4 → 2	SR 2

5.3.4 Přínos pro strategické řízení podniku

Promyšlený a sofistikovaný datový model přináší mnohostranné výhody, které mají značný dopad na byznysovou stránku organizace. Efektivita zpracování a preciznost analýz, dosažené formou automatizace a využitím špičkových výpočetních technik, představují klíčové prvky pro podporu informovaných rozhodnutí. Tyto rozhodnutí jsou seskupena v rámci strategického plánování, kde poskytují unikátní vhledy do dat, umožňují kvalitativnější interpretaci tržních trendů a výsledků v reálném čase. Redukce nákladů spojená s automatizací znamená, že lidské a finanční zdroje lze realokovat jiných oblastí podniku, zatímco rizika jsou díky modelu minimalizována. To se promítá do vyšší adaptability společnosti na trhu, díky implementaci nových vylepšení v zákaznickém zážitku a tudíž zachování zákazníků a zvýšení konkurenční výhody

.

V tomto ekosystému se Looker jeví jako zásadní nástroj, který s sebou nese revoluci ve vizualizaci dat a interakci uživatelů se zprávami, čímž významně usnadňuje proces rozhodování. Uživatelsky přívětivé rozhraní Lookeru, s jeho interaktivními dashboardy a možností snadného sdílení výsledků, zvyšuje transparentnost a přístupnost dat napříč celou organizací, což optimalizuje komunikaci mezi týmy. Vedle toho, flexibilní a škálovatelná povaha Lookeru umožňuje efektivní integraci se stávajícími systémy. Tímto se prostor pro další růst a inovace rozšiřuje. Aplikace Lookeru dále vytváří příležitost pro hloubkové zákaznické analýzy, což umožňuje organizacím lépe porozumět potřebám zákazníků a reagovat na ně s vysokou mírou personalizace.

Vstavěné funkce pro manipulaci s daty a jejich transformaci umožňují analytikům provádět složité analýzy bez potřeby hluboké technické znalosti, čímž se zvyšuje dostupnost datově založených postupů pro širší spektrum uživatelů v rámci organizace. To podporuje kulturu orientovanou na data a vede k lepšímu využití datových zdrojů, čímž se strategické rozhodování stává ještě více informované a zaměřené na výsledky.

V rukou týmu strategického plánování se tak dostává mocný nástroj pro analyzování trhů, optimalizaci produktového portfolia, vyhledání nových obchodních příležitostí a zlepšení zákaznického servisu. Všechny tyto aspekty, když jsou spojeny dohromady, slouží jako motor pohání inovační strategie, posilují konkurenční pozici a otevírají dveře k udržitelnému růstu a konkurenční výhodě v nepředvídatelném a neustále se měnícím

podnikovém prostředí. Strategické řízení podniku, podložené datovou analytikou na té nejvyšší úrovni, se tak stává nejen reaktivním, ale proaktivním elementem, jenž definuje nejen současné postavení podniku, ale i jeho budoucí vizi a cestu k úspěchu.

6 Závěr

V závěru této diplomové práce lze konstatovat, že implementovaný datový model a automatizované nástroje pro analýzu dat představují významný pokrok pro vybranou oblast společnosti. Zavedení terraformu a rozdělení procesu na více fází pomohlo zredukovat chybovost dat a zlepšit jejich správu. Díky Lookeru bylo možné vyvinout transparentní a intuitivní analytické rozhraní, které je přehledné a jednoduše komunikovatelné napříč celou společností. Interaktivní dashboardy a grafy zpřístupňují výsledky analýz i méně technicky zdatným uživatelům, což podporuje data-driven kulturu a rozhodovací procesy společnosti. Flexibilita a škálovatelnost Lookeru jsou klíčové pro adaptaci na budoucí požadavky a rozšiřující se analytické potřeby.

V souhrnu lze konstatovat, že provedené změny vedly k plnění požadavků datových analytiků, přispěly k zásadnímu zlepšení procesu zpracování a analýzy dat a posílily informační strategii společnosti. Zlepšení datové kvality, zefektivnění procesů a zvýšení uživatelské přívětivosti analytického rozhraní tak činí toto řešení významným přínosem pro vybranou společnost v dané oblasti. Budoucí výzkum a rozvoj v této oblasti může dále rozšířit a hlouběji rozpracovat možnosti, které tato práce otevřela, a pokračovat optimalizaci dat.

7 Seznam použité literatury

KANG Li, YI Li, LIU Dong, "Research on Construction Methods of Big Data Semantic Model", Proceedings of the World Congress on Engineering 2014 Vol – I, WCE 2014, July 2–4, 2014, London, U.K. IEEE.

MAIER, R. (1996). Benefits and quality of data modelling — Results of an empirical analysis. In: Thalheim, B. (eds) Conceptual Modeling — ER '96. ER 1996. Lecture Notes in Computer Science, vol 1157. Springer, Berlin, Heidelberg. https://doi-org.ezproxy.lib.vutbr.cz/10.1007/BFb0019927

DWIVEDI, Sunita a CHOURASIYA, Leeladhar, 2022. *Data Modeling: A Perspective In Changing Database Scenario.* Online. 19. MCRPV Bhopal. ISSN 1735188X. Dostupné z: https://www.webology.org/data-cms/articles/20220713104552amwebology%2019%20(3)%20-%20148%20pdf.pdf. [cit. 2023-12-22].

1. ANONYM. *What Is Data Mining? How It Works, Benefits, Techniques, and Examples*, 2023. Online. Ivestipedia.com. Dostupné z: https://www.investopedia.com/terms/d/datamining.asp. [cit. 2023-12-23].

2. ANONYM.*5 Benefits of using Terraform*, 2023. Online. Dev.to. Dostupné z: https://dev.to/pragyanatvade/5-benefits-of-using-terraform-4foo. [cit. 2023-12-23].

Oficiální dokumentace Google Cloud. *Cloud data warehouse to power your data-driven innovation*, 2023. Online. Cloud.google.com. Dostupné z: https://cloud.google.com/bigquery. [cit. 2023-12-25].

Oficiální dokumentace Terraform. *Google Cloud Platform Provider*, 2023. Online. Registry.terraform.io. Dostupné z: https://registry.terraform.io/providers/hashicorp/google/latest/docs/resources/bigquery_table.html. [cit. 2023-12-25].

3. ANONYM. 2023. *What is Clustering?* 2023. Online. Developers.google.com. Dostupné z: https://developers.google.com/machine-learning/clustering/overview. [cit. 2023-12-25].

Oficiální dokumentace Google BigQuery. *Introduction to partitioned tables*, 2023. Online. Cloud.google.com. Dostupné z: https://cloud.google.com/bigquery/docs/partitioned-tables. [cit. 2023-12-25].

Oficiální dokumentace Airflow. *DAGs*, 2023. Online. Airflow.org. Dostupné z: https://airflow.apache.org/docs/apache-airflow/stable/core-concepts/dags.html. [cit. 2023-12-25].

Oficiální dokumentace Looker. 2023. *Looker Studio*. Online. Looker Studio. Dostupné z: https://cloud.google.com/looker/docs. [cit. 2023-12-22].

Oficiální dokumentace Databoxu, 2023. Online. Databox documentation. Dostupné z: https://doorfortyfour.github.io/DataboxDocumentation/#/. [cit. 2023-12-22].

Oficiální dokumentace Microsoft, 2023. Power BI documentation. Online. Power BI documentation. Dostupné z: https://learn.microsoft.com/en-us/power-bi/. [cit. 2023-12-22].

Oficiální dokumentace Agency Analytics, 2022. Online. Agency Analytics Documentation. Dostupné z: https://agencyanalytics.com/help-center. [cit. 2023-12-22].

Oficiální dokumentace Geckoboardu, 2023. Online. Geckoboard.com. Dostupné z: https://www.geckoboard.com/academy/data-sources/. [cit. 2023-12-22].

SMEJKAL Vladimír. RAIS Karel. *Řízení rizik ve firmách a jiných organizacích*. 4., aktualiz. a rozš. vyd. Praha: Grada, 2013. Expert (Grada). ISBN 978-80-247-4644-9. (23)

FOTR, Jiří. 2012. *Tvorba strategie a strategické plánování: teorie a praxe*. 1. vyd. Praha: Grada, 2012. ISBN 978-80-247-3985-4

Oficiální dokumentace Google Cloud. 2023. *Single pass multi field json extraction with BigQuery and dbt (huge performance increase)*. Online. Google Cloud. Dostupné z: https://cloud.google.com/bigquery/docs/json-data. [cit. 2023-11-27].

Oficiální dokumentace Google Cloud. 2023. *BigQuery pricing*. Online. GOOGLE. Google Cloud. Dostupné z: https://cloud.google.com/bigquery/pricing. [cit. 2023-12-11].

What are some common data modeling challenges and how do you overcome them? Online. Dostupné z: https://www.linkedin.com/advice/3/what-some-common-data-modeling-challenges-how-do-you. [cit. 2024-03-02].

KNIGHT, Michelle, 2023. *Data Modeling Trends in 2024*. Online. Dostupné z: https://www.dataversity.net/data-modeling-trends-in-2024/. [cit. 2024-04-13].

AIKEN PhD, Peter, 2023. *Data-Ed Webinar: Conceptual vs. Logical vs. Physical Data Modeling*. Online. Dostupné z: https://www.dataversity.net/jul-11-data-ed-webinar-conceptual-vs-logical-vs-physical-data-modeling/. [cit. 2024-04-13].

BELLINI, P. Managing complexity of data models and performance in broker-based Internet/Web of Things architectures. Online. In: . Dostupné z: https://www.sciencedirect.com/science/article/pii/S2542660523001579?via%3Dihub. [cit. 2024-04-15].

AUAFIQ, E. a SAADANE, R., 2022. *AI-based modeling and data-driven evaluation for smart farming-oriented big data architecture using IoT with energy harvesting capabilities*. Online. Science Direct. Dostupné z: https://www.sciencedirect.com/science/article/abs/pii/S221313882200145X. [cit. 2024-04-15].

GUO, S. a TANG, J., 2021. *Study on Landscape Architecture Model Design Based on Big Data Intelligence*. Online. Science Direct. Dostupné z: https://www.sciencedirect.com/science/article/abs/pii/S2214579621000368. [cit. 2024-04-15].

SUH, S.; CHUNG, D. a LEE, B., 2006. *STEP-compliant CNC system for turning: Data model, architecture, and implementation*. Online. Science Direct. Dostupné z: https://www.sciencedirect.com/science/article/pii/S0010448506000376. [cit. 2024-04-15].

SCHERZINGER, S.; BEURSKENS, M. a LEWINSKI, K., 2024. *Data modelling as a means of power: At the legal and computer science crossroads*. Online. Science Direct. Dostupné z: https://www.sciencedirect.com/science/article/pii/S0267364923000754. [cit. 2024-04-15].

HARRISON, S. a BARKER, S., 2024. *Data–model comparisons*. Online. Science Direct. Dostupné z: https://www.sciencedirect.com/science/article/abs/pii/B9780323999311002051. [cit. 2024-04-15].

7.1 Seznam obrázků

7.2 Seznam tabulek