Operačný program
Efektívna
verejná správa

Európska únia
Európsky sociálny fond

MINISTERSTVO
INVESTÍCIÍ, REGIONÁLNEHO ROZVOJA
A INFORMATIZÁCIE
SLOVENSKEJ REPUBLIKY

Výstup č. 4.1.2

Koncept pre zavádzanie analytického spracovania údajov do praxe

Zmluva o dielo č. 445/2022

Projekt:

Zlepšenie využívania údajov vo verejnej správe

ITMS kód projektu:

314011S979

Tento projekt je podporený z Európskeho sociálneho fondu.

Obsah

ZOZNAM SKRATIEK	
Skratka	**Význam**
ACID	Atomicita, konzistentnosť, izolácia a trvácnosť (Atomicity, Consistency, Isolation, and Durability)
API	Aplikačné programovacie rozhranie (Application Programming Interface)
CIP	Centrálna integračná platforma
CMÚ	Centrálny model údajov
CSV	comma-separated values
DB	Databáza
EDIW	Európska digitálna peňaženka s identitou (European Digital Identity Wallet)
eID	Elektronická identita
eIDAS	Nariadenie Európskej únie č. 910/2014 o elektronickej identifikácii a dôveryhodných službách pre elektronické transakcie na vnútornom európskom trhu.
ELT	Extract, load, transform
ESB	Enterprise Service Bus
ETL	Extract, transform, load
EU	Európska únia
GDPR	Všeobecné nariadenie o ochrane osobných údajov (General Data Protection Regulation)
GIS	Geografický informačný systém
HTML	HyperText Markup Language
iPaaS	Integration Platform as a Service
IS CSRÚ	Informačný systém centrálnej správy referenčných údajov
IS VS	Informačný systém verejnej správy
JSON	JavaScript Object Notation
JSON-LD	JSON pre linkované údaje (JSON for Linking Data)
KAV	Konsolidovaná analytická vrstva
MIRRI SR	Ministerstvo investícii, regionálneho rozvoja a informatizácie
MOU	Manažment osobných údajov
MV SR	Ministerstvo vnútra SR
OLAP	Online analytické spracovanie (Online analytical processing)
OVM	Orgán verejnej moci

PIMS	Systémy na správu osobných informácií (Personal Information Management System)
RDF	Resource Description Framework
SDG	Jednotná digitálna brána (Single Digital Gateway)
SQL	Structured Query Language
TSDB	Databáza pre časové rady (time-series database)
RDBMS	Systém na manažment relačných databáz (Relational database management system)
URI	Jednotný referencovateľný identifikátor
VC	Overiteľné poverenia (Verifiable Credentials)
W3C	World Wide Web Consortium
XML	Extensible Markup Language

1 Manažérske zhrnutie

Tento dokument o koncepte pre zavádzanie analytického spracovania údajov do praxe sa venuje:

- aktualizácii príručiek a návodov pre prípady použitia využívanie údajov,
- zhodnoteniu projektov zavádzania využívania údajov,
- plánovaniu zavádzania využívania údajov,
- odporúčaniam pri riešení problémov,
- vyhodnoteniu poučení.

Pri vypracovaní tohto dokumentu sme zhodnotili súčasný stav projektov (KAV a dopytové projekty pre Zlepšenie využívania údajov verejnej správy) a navrhli sme plán ďalšieho zlepšenia využívania údajov:

- vo vzťahu k dostupným údajom a nástrojom v KAV,
- vo vzťahu k potrebám dátovej transformácie pre vybrané inštitúcie.

V dokumente navrhujeme aj revíziu operačného modelu aktuálneho KAV v rozsahu prevádzky, rolí a ďalšieho rozširovania KAV. V súlade s navrhovanou dátovou a analytickou vrstvou sme vytvorili návod ako modelovať a popisovať dáta a analytické aplikácie, ktoré budú zavádzané do KAV. Tento návod je detailne popísaný v dokumente 1.3.2 Usmernenia pre zrozumiteľné zdokumentovanie dátových štruktúr, procesov tvorby dát, štatistických metodológií (ak boli použité), dátových zdrojov, kontextov a ďalšie aspekty popisu dát).

Použiteľnosť nástrojov bola zanalyzovaná pre konkrétne role a aktérov. Špecifikovali sme prístupovú politiku s prípadmi použitia pre nasledujúce role:

- administrátor,
- dátový architekt,
- biznis analytik,
- vývojár,
- koncový používateľ,
- analytické jednotky/ rezorty.

Definovali sa princípy pre KAV, aby boli inštitúcie motivované spolupracovať a aby sa dátová a analytická vrstva dala flexibilne rozvíjať, čo znamená, že:

- údaje sa zdieľajú na jedno centrálne miesto,
- údaje sa zbierajú pre analytické účely (prípady použitia získaných interných a externých údajov),
- dokumentujú sa prínosy, prečo spolupracovať a zapájať sa do dátovej transformácie,
- vytvárajú sa dostupné návody a praktické aplikácie.

2 Úvod

Konsolidovaná analytická vrstva má za cieľ podporiť nielen súčasnú, ale aj budúcu generáciu analytikov a dátových vedcov pre verejný sektor, ktorá bude využívať dostupné dáta pre lepšie rozhodnutia a hodnotenia následkov týchto rozhodnutí vo verejnom sektore. Preto musí technologický „stack" KAV byť čo najviac odolný voči budúcim zmenám, musí počítať s rôznorodosťou spracovávaných dát, analytických a vizualizačných nástrojov pre rôzne typy prípadov použitia a v neposlednom rade nesmie vytvoriť takzvaný „vendor-lockin".

Moderné organizácie už bežne využívajú multi-cloudové prostredie, aby naplnili svoje biznis požiadavky a požiadavky na lokality a načasovanie, a tiež v neposlednom rade aj na predídenie vendor-lockin". Vymeniť dodávateľa je bolestivé, pokiaľ jedno riešenie dominuje v dátovej infraštruktúre organizácie. Je spojené s migráciou obrovského množstva dát do nového systému, pričom nemožno dopredu zaručiť, že nové riešenie bude spĺňať všetky požiadavky. Preto je výmena dodávateľa spravidla označovaná ako riskantná. Monolitické riešenia často vyžadujú vysokú prvotnú investíciu z pohľadu času a financií, čo prispieva k mentalite utopených nákladov a odmietaniu implementácie nového, vhodnejšieho riešenia. Preto modulárne riešenia zásadne uľahčujú rozhodovanie o aktualizácií nástrojov na prácu s dátami a výrazne zjednodušujú proces nákupu. Preto je vhodné vyberať si takých dodávateľov, ktorý podporujú open-source technológie (hlbšie zdôvodnenie tejto stratégie vysvetľuje spoločnosť Netflix tu). Ďalšou výhodou open source riešení je, že okolo nich spravidla funguje živá a početná skupina používateľov, vďaka čomu je ľahšie nájsť odborníkov so skúsenosťami s takýmto riešením.

Komunitu analytikov a dátových vedcov treba podporovať v neustálom rozvoji ich technických, ale aj mäkkých zručností, aby vedeli čo najlepšie vykonávať svoju prácu. V nasledujúcej tabuľke sú zhrnuté technické zručnosti, ktoré sú už roky na vrchole zoznamu požiadaviek mnohých zamestnávateľov a vzdelávacích inštitúcií. K mäkkým zručnostiam patria:

- Komunikácia – väčšina analytických pozícií si vyžaduje excelentné komunikačné zručnosti. Dátový vedec musí porozumieť biznis požiadavkám alebo problému, vyžiadať si dodatočné informácie a dáta od zúčastnených strán a komunikovať zásadné zistenia.
- „Storytelling" – Štatistické výpočty sú zbytočné, pokiaľ ďalšie tímy ich nevedia využiť vo svojich procesoch a rozhodnutiach. Z tohto dôvodu treba vedieť tieto výsledky týmto tímom predať cez pútavý príbeh, ktorý je krátky, k veci a ktorý poskytuje užitočné zhrnutie a zrozumiteľné vizualizácie.
- Spolupráca- pre nájdenie riešení komplexných problémov je potrebné spolupracovať s rôznorodými tímami, pochopiť ich potreby a zapracovať ich vstupy. Pri dátovej analýze sa tiež očakáva schopnosť spolupracovať s dátovými architektmi a vývojármi.
- Učenie sa – oblasť dátovej analýzy sa neustále a dynamicky vyvíja, preto základom úspechu je neustále sa vzdelávať a mať chuť skúšať nové prístupy a technológie.

Tabuľka 1: Potrebné technické zručnosti dátových analytikov

Kategória	Zručnosť	Popis
Programovacie jazyky	Python	Python je momentálne najpopulárnejší a najprispôsobivejší programovací jazyk v oblasti dátovej vedy. Dokáže pokryť všetko od dolovania dát až po vývoj webových stránok a prevádzku embeddovaných systémov v jednom unifikovanom jazyku.

Kategória	Zručnosť	Popis
	R	R je softvérový balík s programovacím jazykom R určený pre spracovanie dát, výpočty a grafické znázornenie dát a výsledkov. Vďaka rôznym knižniciam sa dajú jednoducho implementovať algoritmy strojového učenia ako aj rôzne ďalšie štatistické a grafické prístupy k analýze dát.
Ukladanie a spracovanie dát	Hadoop platforma	Hadoop platforma predstavuje balík open-source nástrojov, ktoré umožňujú dátovým vedcom spracovávať obrovské datasety veľkých dát cez klastre počítačov s využitím jednoduchých programovacích metód. To je potrebné hlavne vtedy, keď objem dát je väčší ako kapacita systémovej pamäte.
	SQL databázy	SQL databázy a súvisiaci programovací jazyk SQL, ktorý sa využíva na manažovanie a analýzu cenných dát uložených v relačných databázach, v ktorých majú dátové body svoj jasne definovaný model cez vzájomné vzťahy.
	NoSQL databázy	NoSQL databázy – ide o ukladanie, manažovanie a analýzu dát, ktoré nemajú relačnú štruktúru. Ide napríklad o priestorové údaje alebo grafové údaje, či textové datasety.
	Algoritmy pre dolovanie dát	Ide o algoritmy napríklad strojového učenia, ktoré dokážu spracovať obrovské množstva dát. Cieľom tohto spracovania môže byť napríklad analýza klastrov, odľahlých hodnôt, vzorcov či klasifikácia. Mnohé z týchto algoritmov možno použiť aj na predpoveď výstupnej premennej u nového, nepozorovaného dátového bodu. Strojové učenie môže byť s „učiteľom“ alebo bez („supervised“ alebo „unsupervised“), podľa toho, či sú dostupné označené dáta na trénovanie algoritmu.
Vizualizácia dát		Ide o grafické zobrazenie dát a výsledkov analýz cez vizuálne komponenty ako grafy, mapy, grafiky, infografiky a podobne. Táto zručnosť spadá medzi technickú analýzu a schopnosť vizuálne komunikovať s koncovými užívateľmi. Dátová vizualizácia je kľúčová hlavne v oblasti veľkých dát na to, aby sa im dalo lepšie porozumieť a nájsť v nich užitočné informácie a znalosti. Na dátovú vizualizáciu existujú okrem knižníc priamo pre Python alebo R aj mnohé ďalšie nástroje s intuitívnym používateľským rozhraním.

V praxi dátovej vedy a analytiky spravidla **získavanie surových dát** z rôznych zdrojov **a ich príprava** na to, aby sa dali využívať v dátových modeloch, **predstavuje až 80 percent práce**. Práve túto prácu má uľahčiť moderná a flexibilná konsolidovaná analytická vrstva. Tie najužitočnejšie vrstvy - „stacky“ umožňujú organizáciám užívať výhody rýchlo sa vyvíjajúcich technológií a pridávať najlepšie riešenia na trhu pre zber, transformáciu a analýzu dát bez toho, aby na to bolo nevyhnutné alokovať veľa času a financií. Budovanie takéhoto „stacku“ však v mnohých organizáciách narazí na bariéry vytvorené monolitickými, zastaranými aplikáciami a nemodernou IT architektúrou.

Moderný „stack“ pre analýzu údajov je vytvorený desiatkami nástrojov a aplikácií, ktoré pracujú v harmónii a pomáhajú tímom získať rýchlejšie znalosti obsiahnuté v historických údajoch. Každý komponent takejto IT architektúry musí byť **dostatočne flexibilný**, aby ho bolo možné odstrániť nezávisle od ostatných komponentov, a zároveň **dostatočne robustný**, aby dokázal prispieť k odpovediam na komplexné otázky. Týmto princípom možno dosiahnuť neustále zlepšovanie za dostupné náklady pri vysokej spokojnosti zamestnancov.

V modernom, flexibilnom dátovom „stacku“ musia byť dáta prístupné celej škále rôznych nástrojov, preto musia byť centrálne spravované v databázach a výmena jedného nástroja za iný by nemala spôsobiť žiadne komplikácie. Dáta sú preto transformované podľa definovaného dátového modelu ešte predtým, ako sa nahrajú do akéhokoľvek konkrétneho nástroja na ich analýzu. Jedným zo spôsobov, ako to dosiahnuť, je používať nástroje ako dbt[1], ktoré umožňujú transformáciu dát (teda len „T“ časť bežných ETL nástrojov), priamo v dátovom sklade alebo v databázach, a to len písaním skriptov. Cieľom je zachovať dátovú kvalitu bez ohľadu na to, aký nástroj bude zo „stacku“ odstránený.

Prostredie KAV však musí byť prívetivé a užitočné aj pre ďalšie role:

- administrátor, ktorý musí mať k dispozícii efektívne nástroje na administráciu riešenia,
- dátový architekt, ktorý zaznamenáva dátové modely a schémy v dátovom katalógu, definuje transformácie s pomocou dátových analytikov a dátové toky,
- biznis analytik, ktorý spravidla potrebuje analytické nástroje bez potreby skriptovania a programovania.
- vývojár, ktorý zabezpečuje predovšetkým integrácie a transformácie dát a dátové toky podľa požiadaviek dátového architekta. V ideálnom prípade sa všetky tieto oblasti dajú robiť aj v používateľskom rozhraní bez potreby programátorských zručností, a tým pádom ich môže robiť aj dátový analytik a/alebo dátový architekt.
- koncový používateľ, ktorý musí mať k dispozícii interaktívne a zrozumiteľné vizualizácie výsledkov dátových analýz.

[1] Zdroj: https://www.getdbt.com/ , Dátum referencie: 28.04.2021

3 Analýza a dizajn pre KAV 2.0 v oblasti integrácie analytických nástrojov

Prvá fáza KAV sa zameriava predovšetkým na tradičné zdroje dát, ktorými sú administratívne zdroje. Spravidla ide o relačné dáta, ktoré nie sú považované za veľké („big data“). Rozvoj KAV 2.0 by však mal zabezpečiť aj možnosť spracovávať nové zdroje veľkých dát, ktorými sú napríklad sociálne siete, senzorové či telekomunikačné siete, a to aj v takmer reálnom čase. Len tak možno dosiahnuť dátami riadený štát, ktorý sa dokáže rozhodovať na základe objektívnych dôkazov ako moderné podniky.

3.1 Prehľad typov dát spracovávaných v KAV 2.0

Rôznorodosť dátových typov potrebných pre pokročilé analýzy dát zapríčiňuje, že prostredie KAV 2.0 musí byť multi-databázové, aby rôzne dátové typy boli efektívne uložené vo vhodnej databáze a pripravené pre ďalšie spracovanie a analýzu. V nasledujúcom texte sa budeme venovať dátovým typom z pohľadu dolovania dát a dátovej vedy. Popisované typy dát sa vyskytujú aj vo forme veľkých dát. Zo spomínaného pohľadu existujú dva typy dát rôznej zložitosti:

1. **Dáta bez závislostí *(Nondependency-oriented Data)***: Zvyčajne sa to týka jednoduchých typov dát, ako sú viacrozmerné dáta alebo textové dáta. Tieto typy dát sú najjednoduchšie a najčastejšie sa s nimi stretávame. V týchto prípadoch záznamy údajov nemajú žiadne špecifikované závislosti medzi údajovými položkami alebo atribútmi (premennými). Príkladom je súbor záznamov o jednotlivcoch, ktoré obsahujú ich vek, pohlavie a poštové smerové číslo.
2. **Dáta so závislosťami *(Dependency-oriented Data)***: V týchto prípadoch môžu medzi údajovými položkami existovať implicitné alebo explicitné vzťahy. Napríklad dataset sociálnej siete obsahuje množinu uzlov, ktoré sú navzájom spojené množinou hrán (vzťahov). Na druhej strane časové rady obsahujú implicitné závislosti.

Vo všeobecnosti údaje so závislosťami sú náročnejšie z dôvodu zložitosti spôsobenej už existujúcimi vzťahmi medzi údajovými položkami.

3.1.1 Dáta bez závislostí („Nondependency-oriented data“)

Táto forma dát je najjednoduchšia a zvyčajne sa vzťahuje na viacrozmerné dáta. Tieto dáta zvyčajne obsahujú množinu záznamov. Záznam sa tiež označuje ako dátový bod, inštancia, príklad, transakcia, entita, objekt alebo vektor vstupných premenných v závislosti od daného prípadu použitia. Každý záznam obsahuje množinu polí, ktoré sa tiež označujú ako vstupné premenné, dimenzie charakteristické znaky. Tieto polia popisujú rôzne vlastnosti daného záznamu. Ako príklad uvádzame súbor demografických údajov znázornený v tabuľke nižšie. Tu sú ilustrované demografické vlastnosti jednotlivca, ako je vek, pohlavie a poštové smerové číslo (PSČ).

Tabuľka 2: Ukážka viacrozmerného datasetu

Meno	Vek	Pohlavie	PSČ
John S.	45	M	811 09
Mayona L.	31	F	040 13
Sayani L.	11	F	841 17
Jack M.	56	M	974 01
Wei L.	63	M	011 09

Nasledujúca tabuľka sumarizuje konkrétnejšie typy dát v tejto skupine.

Tabuľka 3: Prehľad konkrétnych typov dát bez závislostí

Typ daru bez závislostí	Popis
Kvantitatívne viacrozmerné dáta	Ide o viacrozmerné dáta, v ktorých sú čisto číselné premenné. Atribúty v tabuľke (Tabuľka 2) sú dvoch rôznych typov. Vekové pole má hodnoty, ktoré sú číselné v tom zmysle, že majú prirodzené usporiadanie. Takéto atribúty sa označujú ako spojité, číselné alebo kvantitatívne. Tento podtyp je obzvlášť vhodný na analytické spracovanie, pretože zo štatistického hľadiska je oveľa jednoduchšie pracovať s kvantitatívnymi údajmi.
Kategorické dáta a dáta o zmiešaných atribútoch	Mnoho dátových súborov v reálnych aplikáciách môže obsahovať kategorické atribúty, ktoré preberajú diskrétne neusporiadané hodnoty. Napríklad v tabuľke (Tabuľka 2) majú atribúty ako pohlavie a PSČ diskrétne hodnoty bez prirodzeného usporiadania medzi nimi. V prípade dát o zmiešaných atribútoch existuje kombinácia kategorických a číselných atribútov. Úplné dáta v tabuľke (Tabuľka 2) sa považujú za dáta so zmiešanými atribútmi, pretože obsahujú číselné aj kategorické atribúty. Atribút zodpovedajúci pohlaviu je špeciálny, pretože je kategorický, ale iba s dvoma možnými hodnotami. V takýchto prípadoch je možné uložiť umelé usporiadanie medzi týmito hodnotami a použiť algoritmy určené pre číselné dáta pre tento typ. Toto sa označuje ako binárne dáta a možno to považovať za špeciálny prípad číselných alebo kategorických dát. Binárne dáta tvoria "most" na transformáciu číselných alebo kategorických atribútov do spoločného formátu, ktorý je vhodný na spracovanie v mnohých prípadoch použitia.
Binárne a množinové dáta	Binárne dáta možno považovať za špeciálny prípad buď viacrozmerných kategorických dát, alebo viacrozmerných kvantitatívnych dát. Ide o špeciálny prípad viacrozmerných kategorických dát, v ktorých každý kategorický atribút môže mať jednu z najviac dvoch samostatných hodnôt. Je to tiež špeciálny prípad viacrozmerných kvantitatívnych dát, pretože medzi týmito dvoma hodnotami existuje poradie. Hodnota 1 znamená, že prvok by mal byť zahrnutý do množiny.

Textové dáta	Textové dáta je možné zobraziť buď ako reťazec, alebo ako viacrozmerné dáta v závislosti od toho, ako sú reprezentované. Vo svojej surovej forme textový dokument zodpovedá reťazcu, čo je typ dát so závislosťou. Každý reťazec je postupnosť znakov (alebo slov) zodpovedajúcich dokumentu. Textové dokumenty sú však zriedka reprezentované ako reťazce, pretože je ťažké priamo využívať poradie medzi slovami efektívnym spôsobom pre rozsiahle aplikácie. V praxi sa používa vektorovo-priestorová reprezentácia, kde sa na analýzu používajú frekvencie slov v dokumente. V tomto zobrazení sa teda stráca presné usporiadanie slov. Tieto frekvencie slov sa zvyčajne normalizujú štatistikami, ako je dĺžka dokumentu alebo frekvencie jednotlivých slov v zbierke. Slová sa niekedy označujú aj ako termíny. Zodpovedajúca dátová matica $n \times d$ pre zbierku textu s n dokumentmi a termínmi d sa označuje ako matica termínov dokumentu („document-term matrix“). Ak sú textové dáta reprezentované vo forme vektorového priestoru, môžu sa považovať za viacrozmerné kvantitatívne dáta, ak atribúty zodpovedajú slovám a hodnoty zodpovedajú frekvenciám týchto atribútov. Tento druh kvantitatívnych dát je však špeciálny, pretože väčšina atribútov má nulové hodnoty a len niekoľko atribútov má nenulové hodnoty. Tento jav sa označuje ako riedkosť dát („sparsity“).

3.1.2 Dáta so závislosťami („Dependency-oriented data“)

Môže existovať niekoľko typov závislostí, ktoré môžu byť implicitné alebo explicitné:

1. **Implicitné závislosti**: V tomto prípade závislosti medzi údajovými položkami nie sú výslovne špecifikované, ale je známe, že "typicky" existujú v danej doméne. Napríklad po sebe idúce hodnoty teploty zhromaždené senzorom budú pravdepodobne navzájom veľmi podobné. Preto, ak sa hodnota teploty zaznamenaná senzorom v určitom čase výrazne líši od hodnoty zaznamenanej v nasledujúcom okamihu, potom je to mimoriadne nezvyčajné a treba to zanalyzovať. Tam je odlišnosť oproti viacrozmerným datasetom, v ktorých sa s každým dátovým záznamom zaobchádza ako s nezávislou entitou.
2. **Explicitné závislosti**: Zvyčajne sa to týka grafu alebo sieťových údajov, v ktorých sa hrany používajú na určenie explicitných vzťahov. Grafy sú veľmi silnou abstrakciou, ktorá sa často používa ako prechodná reprezentácia na riešenie problémov pri analýze dát v kontexte iných typov dát.

Nasledujúca tabuľka sumarizuje konkrétnejšie typy dát v tejto skupine.

Tabuľka 4: Prehľad konkrétnych typov dát so závislosťami

Typ dát so závislosťami	Popis
Dáta časových radov	Dáta časových radov obsahujú hodnoty, ktoré sa zvyčajne generujú kontinuálnym meraním v čase. Napríklad environmentálny senzor bude nepretržite merať teplotu. Takéto dáta majú zvyčajne implicitné závislosti zabudované do hodnôt prijatých v priebehu času. Napríklad susedné hodnoty zaznamenané snímačom teploty sa budú zvyčajne v priebehu času plynulo meniť.

Typ dát so závislosťami	Popis
	Povaha časovej závislosti sa môže v závislosti od prípadu použitia výrazne líšiť. Napríklad niektoré formy dát zo snímačov môžu vykazovať periodické vzorce meraného atribútu v priebehu času. Dôležitým aspektom dolovania časových radov je extrakcia takýchto závislostí v dátach. Na formalizáciu otázky závislostí spôsobených časovou koreláciou sú atribúty rozdelené do dvoch typov: 1. **Kontextové atribúty**: Toto sú atribúty, ktoré definujú kontext, na základe ktorého sa implicitné závislosti vyskytujú v dátach. Napríklad v prípade dát zo snímačov sa za kontextový atribút môže považovať časová pečiatka, pri ktorej sa hodnota merania odčítala. Iné typy dát môžu mať viac ako jeden kontextový atribút. 2. **Atribúty správania**: Atribúty správania predstavujú hodnoty, ktoré sa merajú v konkrétnom kontexte. Pri príklade senzora je teplota hodnotou atribútu správania. Je možné mať viac ako jeden atribút správania. Ak napríklad viaceré senzory zaznamenávajú dáta pri synchronizovaných časových pečiatkach, výsledkom je viacrozmerný dataset časových radov. Kontextové atribúty majú zvyčajne silný vplyv na závislosti medzi hodnotami atribútov správania v dátach. Dáta časových radov sú relatívne bežné v mnohých senzorových aplikáciách, prognózach a analýzach finančného trhu.
Diskrétne sekvencie a reťazce	Diskrétne sekvencie možno považovať za kategorickú analógiu dát časových radov. Rovnako ako v prípade dát časových radov je kontextovým atribútom časová pečiatka alebo index pozície v poradí. Atribút správania je kategorická hodnota, preto sú diskrétne sekvenčné dáta definované podobným spôsobom ako dáta časových radov. Napríklad postupnosť webových prístupov, v ktorých sa pre 100 rôznych prístupov zhromažďuje adresa webovej stránky a pôvodná IP adresa žiadosti. To predstavuje diskrétnu postupnosť dĺžky $n = 100$ a dimenzionality $d = 2$. Obzvlášť častým prípadom v sekvenčných dátach je jednorozmerný scenár, v ktorom hodnota d je 1. Takéto sekvenčné dáta sa tiež označujú ako reťazce. Jednou z dôležitých obmien je prípad, keď sekvencia neobsahuje kategorické atribúty, ale množinu ľubovoľného počtu neusporiadaných kategorických hodnôt. Napríklad transakcie v supermarketoch môžu obsahovať postupnosť súborov položiek. Každý súbor môže obsahovať ľubovoľný počet položiek. Takéto sekvencie nie sú v skutočnosti viacrozmerné sekvencie, ale sú to jednorozmerné sekvencie, v ktorých je každý prvok sekvencie množinou na rozdiel od jednotkového prvku. Diskrétne sekvencie sú pre algoritmy dolovania často náročnejšie, pretože nemajú plynulú hodnotovú kontinuitu ako dáta časových radov.
Priestorové dáta	V priestorových dátach sa meria mnoho nepriestorových atribútov (napríklad teplota, tlak). Napríklad meteorológovia často zhromažďujú teploty morskej hladiny, aby predpovedali výskyt hurikánov. V takýchto prípadoch priestorové súradnice zodpovedajú kontextovým atribútom, zatiaľ čo atribúty, ako je teplota, zodpovedajú atribútom správania. Zvyčajne existujú dva priestorové atribúty. Rovnako ako v prípade dát časových radov je tiež možné mať viacero atribútov

Typ dát so závislosťami	Popis
	správania. Napríklad pri aplikácii teploty morskej hladiny je možné merať aj iné atribúty správania, ako je tlak. Dolovanie priestorových dát úzko súvisí s dolovaním dát v časových radoch, pretože atribúty správania v najčastejšie študovaných priestorových prípadoch použitia sú kontinuálne, hoci niektoré prípady použitia môžu používať aj kategorické atribúty. Preto sa kontinuita hodnôt pozoruje naprieč súvislými priestorovými lokalitami, rovnako ako sa pozoruje kontinuita hodnôt v súvislých časových pečiatkach v dátach časových radov.
Časopriestorové dáta	Osobitnou formou priestorových dát sú časopriestorové dáta, ktoré obsahujú priestorové aj časové atribúty. Presná povaha dát závisí aj od toho, ktoré z atribútov sú kontextové a ktoré sú behaviorálne. Dva druhy časopriestorových dát sú najbežnejšie: 1. Priestorové aj časové atribúty sú kontextové: Tento druh údajov možno považovať za priame zovšeobecnenie priestorových aj časových dát. Tento druh dát je obzvlášť užitočný, keď sa súčasne meria priestorová a časová dynamika konkrétnych atribútov správania. Napríklad keď je potrebné merať zmeny teploty povrchu mora v priebehu času. V takýchto prípadoch je teplota atribútom správania, zatiaľ čo priestorové a časové atribúty sú kontextové. 2. Časový atribút je kontextový, zatiaľ čo priestorové atribúty sú behaviorálne: Tento druh dát možno tiež považovať za dáta časových radov. Priestorová povaha atribútov správania však v mnohých scenároch poskytuje aj lepšiu interpretovateľnosť a cielenejšiu analýzu. Najbežnejšia forma týchto údajov vzniká v kontexte analýzy trajektórie. Treba zdôrazniť, že akékoľvek 2- alebo 3-dimenzionálne dáta časových radov možno mapovať na trajektórie. Je to užitočná transformácia, pretože to znamená, že algoritmy dolovania trajektórie sa môžu použiť aj pre 2- alebo 3-dimenzionálne dáta časových radov.
Sieťové a grafové dáta	Pri sieťových a grafových dátach môžu dátové hodnoty zodpovedať uzlom v sieti („nodes“), zatiaľ čo vzťahy medzi dátovými hodnotami môžu zodpovedať hranám („edges“) v sieti. V niektorých prípadoch môžu byť atribúty priradené k uzlom v sieti. Aj keď je tiež možné priradiť atribúty k hranám v sieti, je to oveľa menej bežné. Hrana môže byť nasmerovaná alebo nesmerovaná, v závislosti od daného prípadu použitia. Napríklad webový graf môže obsahovať nasmerované hrany zodpovedajúce smerom hypertextových odkazov medzi stránkami, zatiaľ čo priateľstvá v sociálnej sieti Facebook sú nesmerované. Niektoré príklady dát, ktoré sú znázornené ako grafy, sú tieto: – Webový graf: Uzly zodpovedajú webovým stránkam a hrany zodpovedajú hypertextovým odkazom. Uzly majú textové atribúty zodpovedajúce obsahu na stránke. – Sociálne siete: V tomto prípade uzly zodpovedajú aktérom sociálnych sietí, zatiaľ čo hrany zodpovedajú priateľským väzbám. Uzly môžu mať atribúty

Typ dát so závislosťami	Popis
	zodpovedajúce obsahu sociálnej stránky. V niektorých špecializovaných formách sociálnych sietí, ako sú e-mailové siete alebo siete chatovacích aplikácií, môžu mať hrany obsah, ktorý je s nimi spojený. Tento obsah zodpovedá komunikácii medzi rôznymi uzlami. Sieťové dáta sú veľmi všeobecnou reprezentáciou a môžu byť použité na riešenie mnohých prípadov použitia založených na podobnosti s inými typmi dát. Napríklad viacrozmerné dáta môžu byť konvertované na sieťové dáta vytvorením uzla pre každý záznam v databáze, ktorý predstavuje podobnosti medzi uzlami po hranách. Takáto reprezentácia sa pomerne často používa v mnohých prípadoch použitia na dolovanie dát založených na podobnosti, ako je napríklad analýza klastrov. Je možné použiť algoritmy na detekciu komunít na určenie klastrov v sieťových dátach a potom ich namapovať späť na viacrozmerné dáta.

3.2 Prehľadu najvhodnejších databáz podľa dátových typov pre KAV 2.0

3.2.1 Relačné databázy pre štruktúrované údaje

Tieto databázy vedia ukladať hlavne nasledovné typy dát:

- Kvantitatívne viacrozmerné dáta,
- Kategorické dáta a dáta o zmiešaných atribútoch,
- Binárne a množinové dáta,
- Diskrétne sekvencie a reťazce,
- Dáta časových radov (ako vhodná databáza sa javí PostgreSQL, hoci existujú aj špecializované systémy pre časové rady(„time-series database (TSDB)“)[2]),
- Priestorové dáta (Funkcia geodatabázy je databázová vrstva, ktorá je vytvorená nad niekoľkými rôznymi systémami správy podnikových databáz. Patria sem Oracle, Microsoft SQL Server, IBM DB2 a PostgreSQL. PostGIS[3] je jednou z najznámejších a najkomplexnejších priestorových databáz. Je rozšírením open-source databázy PostgreSQL.)[4].

Systémy na správu relačných databáz („Relational Database Management Systems (RDBMS)“) vznikli v 70. rokoch 20. storočia na ukladanie údajov vo forme tabuliek s riadkami a stĺpcami, pričom na vyhľadávanie a údržbu databázy sa používajú príkazy jazyka SQL (Structured Query Language). Relačná databáza je v podstate súbor tabuliek, z ktorých každá má schému, ktorá pevne definuje atribúty a typy údajov, ktoré sa v nich uchovávajú, ako aj kľúče, ktoré identifikujú konkrétne stĺpce alebo riadky na

[2] Zdroj: https://medium.com/@neslinesli93/how-to-efficiently-store-and-query-time-series-data-90313ff0ec20, Dátum referencie: 29.03.2023

[3] Zdroj: https://postgis.net/, Dátum referencie: 29.05.2023

[4] Zdroj: https://engage.safe.com/blog/2021/11/7-spatial-databases-enterprise/, Dátum referencie: 29.05.2023

uľahčenie prístupu. Na poli RDBMS kedysi vládli spoločnosti Oracle[5] a IBM[6], ale dnes sú rovnako populárne mnohé možnosti s otvoreným zdrojovým kódom, ako napríklad MySQL[7], SQLite[8] a PostgreSQL[9].

Relačné databázy sa ustálili vo svete podnikania vďaka niektorým veľmi atraktívnym vlastnostiam. Integrita údajov je v relačných databázach absolútne najdôležitejšia. RDBMS spĺňajú požiadavky atómovosti, konzistentnosti, izolácie a trvanlivosti (alebo ACID-kompatibility[10]) tým, že zavádzajú množstvo obmedzení, ktoré zabezpečujú spoľahlivosť a presnosť uložených údajov, vďaka čomu sú ideálne na sledovanie a ukladanie položiek, ako sú čísla účtov, objednávky a platby. Tieto obmedzenia sú však spojené s nákladnými kompromismi. Kvôli obmedzeniam schémy a typu sú systémy RDBMS nevhodné na ukladanie neštruktúrovaných alebo pološtruktúrovaných údajov. Rigidná schéma tiež predražuje nastavenie, údržbu a rast RDBMS. Nastavenie RDBMS vyžaduje, aby používatelia mali vopred pripravené konkrétne prípady použitia; akékoľvek zmeny schémy sú zvyčajne náročné a zdĺhavé. Okrem toho tradičné RDBMS boli navrhnuté na prevádzku na jednom uzle počítača, čo znamená, že ich rýchlosť je pri spracovaní veľkých objemov údajov výrazne nižšia. Všetky tieto vlastnosti spôsobujú, že tradičné RDBMS nie sú vhodné na spracovanie moderných veľkých objemov údajov.

3.2.2 Nerelačné (NoSQL) databázy pre neštruktúrované alebo semi-štruktúrované údaje

Tieto databázy sú vhodné predovšetkým na ukladanie nasledujúcich typov údajov:

- Textové dáta (viď kapitolu 3.2.2.2, ale textové súbory sa dajú uložiť aj v relačných databázach alebo v cloudovej službe úložiska objektov ako Amazon S3[11], Azure Blob Storage[12] alebo Google Cloud Storage[13]),
- Priestorové dáta (Napríklad cez Apache Couch DB[14] - databázový systém založený na dokumentoch, ktorý možno rozšíriť o podporu priestorových dát pomocou zásuvného modulu s názvom Geocouch, alebo Elasticsearch[15] - databázový systém založený na dokumentoch, ktorý podporuje dva typy geografických údajov: polia geo_point, ktoré podporujú dvojice zemepisných

[5] Zdroj: https://db-engines.com/en/system/Oracle, Dátum referencie: 29.03.2023

[6] Zdroj: https://db-engines.com/en/system/IBM+Db2, Dátum referencie: 29.03.2023

[7] Zdroj: https://db-engines.com/en/system/MySQL, Dátum referencie: 29.03.2023

[8] Zdroj: https://db-engines.com/en/system/SQLite, Dátum referencie: 29.03.2023

[9] Zdroj: https://db-engines.com/en/system/PostgreSQL, Dátum referencie: 29.03.2023

[10] Zdroj: https://www.yugabyte.com/acid/, Dátum referencie: 29.03.2023

[11] Zdroj: https://aws.amazon.com/s3/, Dátum referencie: 29.05.2023

[12] Zdroj: https://azure.microsoft.com/en-in/products/storage/blobs/, Dátum referencie: 29.05.2023

[13] Zdroj: https://cloud.google.com/storage, Dátum referencie: 29.05.2023

[14] Zdroj: https://couchdb.apache.org/, Dátum referencie: 29.05.2023

[15] Zdroj: https://www.elastic.co/, Dátum referencie: 29.05.2023

šírok a dĺžok, a polia geo_shape, ktoré podporujú body, čiary, kružnice, polygóny, multipolygóny atď.),

- Časopriestorové dáta (na ukladanie sa používa priestorovo-časová databáza, ktorá spravuje informácie o priestore aj čase, ide o rozšírenie priestorových databáz a časových databáz, najpoužívanejším príkladom je GeoMesa[16] - cloudová časopriestorová databáza postavená na Apache Accumulo a Apache Hadoop (podporuje aj Apache HBase, Google Bigtable, Apache Cassandra a Apache Kafka). GeoMesa podporuje všetky jednoduché funkcie OGC a zásuvný modul GeoServer.)),
- Sieťové a grafové dáta (viď kapitolu 3.2.2.3).

V polovici nultých rokov už existujúce RDBMS nedokázali zvládnuť meniace sa potreby a exponenciálny rast niekoľkých veľmi úspešných online podnikov a v dôsledku toho vzniklo mnoho nerelačných (alebo NoSQL) databáz. Bez vtedajších známych riešení tieto online podniky vymysleli nové prístupy a nástroje na spracovanie obrovského množstva neštruktúrovaných údajov, ktoré zhromaždili: Spoločnosť Google vytvorila MapReduce[17] a BigTable[18]; Amazon vytvoril DynamoDB[19]; Yahoo vytvoril Hadoop[20]; Facebook vytvoril Cassandru[21] a Hive[22] (ide o distribuované úložisko); LinkedIn vytvoril Kafku[23]. Niektoré z týchto firiem otvorili zdrojové kódy svojej práce; niektoré publikovali výskumné práce s podrobnými informáciami o svojich návrhoch, čo viedlo k rozšíreniu databáz s novými technológiami a databázy NoSQL sa stali hlavným hráčom v tomto odvetví.

Databázy NoSQL sú agnostické voči schéme a poskytujú flexibilitu potrebnú na ukladanie a manipuláciu s veľkými objemami neštruktúrovaných a pološtruktúrovaných údajov. Používatelia nemusia počas nastavovania vedieť, aké typy údajov budú uložené, a systém sa dokáže prispôsobiť zmenám v typoch údajov a schémach. Databázy NoSQL sú navrhnuté na distribúciu údajov v rôznych uzloch, sú vo všeobecnosti horizontálne škálovateľné a odolnejšie voči chybám. Tieto výkonnostné výhody však majú aj svoju cenu - databázy NoSQL nie sú kompatibilné s ACID a konzistencia údajov nie je zaručená. Namiesto toho poskytujú „prípadnú konzistenciu“ („eventual consistency“): keď sa staré údaje prepisujú, vrátia výsledky, ktoré sú dočasne trochu nesprávne. V súčasnosti existuje niekoľko rôznych kategórií NoSQL, z ktorých každá slúži na určité špecifické účely.

[16] Zdroj: https://www.geomesa.org/, Dátum referencie: 29.05.2023

[17] Zdroj: https://research.google/pubs/pub62/, Dátum referencie: 29.03.2023

[18] Zdroj: https://cloud.google.com/bigtable/docs/, Dátum referencie: 29.03.2023

[19] Zdroj: https://cloudacademy.com/blog/amazon-dynamodb-ten-things/, Dátum referencie: 29.03.2023

[20] Zdroj: https://www.sas.com/nl_nl/insights/big-data/hadoop.html#hadoopworld, Dátum referencie: 29.03.2023

[21] Zdroj: https://cassandra.apache.org/_/index.html, Dátum referencie: 29.03.2023

[22] Zdroj: https://hive.apache.org, Dátum referencie: 29.03.2023

[23] Zdroj: https://engineering.linkedin.com/blog/2016/04/kafka-ecosystem-at-linkedin, Dátum referencie: 29.03.2023

3.2.2.1 Úložiská typu kľúč-hodnota („Key-Value Stores“)

Úložiská typu kľúč-hodnota, ako napríklad Redis[24], DynamoDB[25] a Cosmos DB[26], ukladajú iba páry kľúč-hodnota a poskytujú základné funkcie na načítanie hodnoty spojenej so známym kľúčom. Najlepšie fungujú s jednoduchou databázovou schémou a vtedy, keď je dôležitá rýchlosť. Široké stĺpcové úložiská („Wide Column Stores“), ako napríklad spomínaná Cassandra, ScyllaDB[27] a HBase[28], ukladajú údaje do rodín stĺpcov alebo tabuliek a sú vytvorené na správu petabajtov údajov v obrovskom distribuovanom systéme.

3.2.2.2 Úložiská dokumentov („Document Stores“)

Úložiská dokumentov, ako napríklad MongoDB[29] a Couchbase[30], ukladajú údaje vo formáte XML alebo JSON, pričom názov dokumentu je kľúč a obsah dokumentu je hodnota. Dokumenty môžu obsahovať mnoho rôznych typov hodnôt a môžu byť vnorené, vďaka čomu sú mimoriadne vhodné na správu pološtruktúrovaných údajov v distribuovaných systémoch.

MongoDB je v súčasnosti najpopulárnejšia databáza NoSQL a priniesla merateľné prínosy niektorým podnikom, ktoré mali problémy so spracovaním svojich neštruktúrovaných údajov pomocou tradičného prístupu RDBMS. Tu sú dva príklady z odvetvia[31]: Po tom, čo sa spoločnosť MetLife roky snažila vytvoriť centralizovanú databázu zákazníkov v systéme RDBMS, ktorá by zvládla všetky jej poistné produkty, niekto na internom hackathone v priebehu niekoľkých hodín vytvoril takúto databázu s MongoDB, ktorá sa dostala do produkcie za 90 dní. YouGov, firma zaoberajúca sa prieskumom trhu, ktorá zbiera 5 gigabitov údajov za hodinu, ušetrila 70 % úložnej kapacity, ktorú predtým používala, prechodom z RDBMS na MongoDB.

3.2.2.3 Grafové databázy

Grafové databázy, ako napríklad Neo4J[32], JanusGraph[33] a Amazon Neptune[34], predstavujú údaje ako sieť súvisiacich uzlov alebo objektov s cieľom uľahčiť vizualizáciu údajov a analýzu grafov. Grafové databázy

[24] Zdroj: https://redis.io/, Dátum referencie: 29.05.2023

[25] Zdroj:, https://aws.amazon.com/dynamodb/, Dátum referencie: 29.05.2023

[26] Zdroj: https://azure.microsoft.com/en-us/products/cosmos-db, Dátum referencie: 29.05.2023

[27] Zdroj: https://www.scylladb.com/, Dátum referencie: 26.05.2023

[28] Zdroj: https://hbase.apache.org/, Dátum referencie: 26.05.2023

[29] Zdroj: https://www.mongodb.com/, Dátum referencie: 26.05.2023

[30] Zdroj: https://www.couchbase.com/, Dátum referencie: 26.05.2023

[31] Zdroj: https://arstechnica.com/information-technology/2016/03/to-sql-or-nosql-thats-the-database-question/, Dátum referencie: 26.05.2023

[32] Zdroj: https://neo4j.com/, Dátum referencie: 26.05.2023

[33] Zdroj: https://janusgraph.org/, Dátum referencie: 26.05.2023

[34] Zdroj: https://aws.amazon.com/neptune/, Dátum referencie: 26.05.2023

sú obzvlášť užitočné na analýzu vzťahov medzi heterogénnymi dátovými bodmi, napríklad pri prevencii podvodov alebo v grafe priateľov na Facebooku.

3.3 Dátové modelovanie

Údaje sú vo svojej podstate chaotické. Neboli tak vytvorené zámerne, avšak stávajú sa takými, keď sa zlievajú z veľmi veľa zdrojov do „staging" úložiska alebo do dátového jazera, kde následne môže prebehnúť ich transformácia (viac o procedúrach ETL alebo ELT sa nachádza v dokumente 1.1.6 Štandardizácia dátovej transformácie). Takéto údaje pred transformáciou pripomínajú hromadu stavebníc Lego, ktoré boli kedysi krásnymi ucelenými stavbami, ale úplne sa rozpadli a niekto zahodil návody. Preto hľadanie zmyslu údajov v tomto chaotickom stave je komplikované a časovo náročné. Ide o zdĺhavý, únavný a „frustrujúci proces hľadania potrebných dátových prvkov, snahy o ich prepojenie a vykonanie profilovania a čistenia údajov. Dôvera v údaje sa zvyčajne stráca už počas tejto cesty. Strojové učenie potrebuje surové údaje transformované na použitie v modeli strojového učenia prostredníctvom procesu nazývaného „feature engineering"[35]. Ide o proces výberu, manipulácie a transformácie surových údajov na premenné („features"), ktoré možno použiť pri učení. Aby strojové učenie dobre fungovalo na nových úlohách, môže byť potrebné navrhnúť a natrénovať lepšie premenné („features"). Premenná („feature") je akýkoľvek merateľný vstup, ktorý možno použiť v prediktívnom modeli - môže to byť napríklad vek a finančný príjem. Je to dôležitý proces pri vytváraní modelov strojového učenia, ktorý zabezpečuje dataset vo vhodnej forme a dostatočnej kvalite pre úspešné fungovanie modelu a pre jeho výkon po nasadení. Spravidla tento dataset má tabuľkový charakter, kde v stĺpcoch sa nachádzajú jednotlivé premenné („features") a v riadkoch jednotlivé pozorovania alebo objekty záujmu, ktorých sa tieto premenné týkajú (napríklad jednotlivé právnické osoby).

Dátové modelovanie sa teda stáva kľúčovým, keď máme údaje z veľa zdrojov a vzťahy medzi nimi, ktoré sa často menia. **Ak sú dáta statické alebo ak sú už dostupné v súlade s Centrálnym modelom údajov , je na zváženie, či je potrebné takéto modelovanie a či nestačí len dobre manažovať dátové „pipelines" vrátane dátovej transformácie.** Dátové modelovanie vie tiež pomôcť s rýchlejším nahrávaním údajov z viacerých systémov ako aj s potrebou jednoducho sledovať a auditovať údaje.

Existujú dva bežné prístupy k dátovému modelovaniu pri navrhovaní dátových skladov, ktoré majú odlišné ciele, princípy a výhody:

1. Dimenzionálne modelovanie[36],
2. Modelovanie „Data Vault"[37]

Čo je dimenzionálne modelovanie?

Dimenzionálne modelovanie je metóda návrhu dátového skladu, ktorá sa zameriava na vytvorenie jednoduchej a intuitívnej štruktúry pre používateľov a analytikov. Organizuje údaje do faktov a dimenzií,

[35] Zdroj: https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10, Dátum referencie: 26.05.2023

[36] Zdroj: https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/dimensional-modeling-techniques/, Dátum referencie: 23.05.2023

[37] Zdroj: https://datavaultalliance.com/about/what-is-datavault/data-vault-2-0-model/, Dátum referencie: 23.05.2023

kde fakty sú číselné miery biznis udalostí a dimenzie sú opisné atribúty, ktoré poskytujú kontext a podrobnosti pre fakty. Napríklad tabuľka faktov o poskytovaní služieb môže obsahovať miery, ako sú počet a poplatky, a dimenzie, ako sú typ služby, zákazník, dátum a miesto. Pri dimenzionálnom modelovaní sa používa schéma hviezdy alebo snehovej vločky, kedy je tabuľka faktov obklopená tabuľkami dimenzií, ktoré vytvárajú tvar hviezdy alebo snehovej vločky.

Čo je modelovanie „Data Vault“?

Modelovanie „Data Vault“ je metóda návrhu dátového skladu, ktorá sa zameriava na vytvorenie flexibilnej a škálovateľnej štruktúry na integráciu a uchovávanie historických údajov. Organizuje údaje do hubov, prepojení a satelitov, kde huby sú tabuľky, ktoré uchovávajú jedinečné kľúče subjektov, prepojenia sú tabuľky, ktoré uchovávajú vzťahy medzi subjektmi, a satelity sú tabuľky, ktoré uchovávajú atribúty a históriu subjektov alebo vzťahov. Napríklad hub zákazníka môže obsahovať ID zákazníka a dátum načítania, prepojenie zákazník - služba môže obsahovať ID zákazníka a ID služby a satelit zákazníka môže obsahovať meno, adresu, telefón a e-mail zákazníka. Pri modelovaní „Data Vault“ sa používa normalizovaná schéma, v ktorej sú tabuľky prepojené cudzími kľúčmi a vytvárajú tvar podobný sieti.

V čom sa líšia?

Dimenzionálne a „Data Vault“ modelovanie údajov sa líši svojím účelom, štruktúrou, výkonom a flexibilitou. Cieľom dimenzionálneho modelovania je poskytnúť používateľsky prívetivý a na dotazy optimalizovaný dátový sklad, ktorý podporuje business intelligence a analytiku. Používa denormalizovanú schému na zníženie počtu tabuliek a prepojení, ale môže zaviesť redundanciu a nekonzistentnosť údajov. Tento model zvyčajne ponúka rýchlejší výkon dotazov a jednoduchšie procesy ETL, ale môže vyžadovať viac úložného priestoru a častejšie aktualizácie. Je tiež menej flexibilný a viac závislý od biznis požiadaviek a zdrojov údajov, čo znamená, že ho možno bude potrebné prepracovať alebo prestavať, keď sa zmenia požiadavky alebo zdroje. Naopak, modelovanie „Data Vault“ sa snaží poskytnúť dátovo orientovaný a voči zmenám odolný dátový sklad, ktorý podporuje integráciu a správu údajov. Používa normalizovanú schému, ktorá zvyšuje počet tabuliek a prepojení, ale zabraňuje redundancii a nekonzistentnosti údajov. Tento model ponúka pomalší výkon dotazov a zložitejšie procesy ETL, ale môže vyžadovať menej úložného priestoru a menej časté aktualizácie. Je tiež flexibilnejší a nezávislejší od biznis požiadaviek a zdrojov údajov, čo mu umožňuje prispôsobiť sa novým alebo meniacim sa požiadavkám alebo zdrojom bez toho, aby to ovplyvnilo existujúci dátový sklad.

V čom sú si podobné?

Dimenzionálne modelovanie a modelovanie „Data Vault“ majú niektoré podobnosti, napríklad ich modulárne komponenty, ktoré možno opakovane používať a rozširovať. Pri dimenzionálnom modelovaní sa používajú fakty a dimenzie, zatiaľ čo pri modelovaní „Data Vault“ sa používajú huby, prepojenia a satelity. Obe metódy dokážu zachovať historické údaje a sledovať zmeny v čase; dimenzionálne modelovanie to robí prostredníctvom pomaly sa meniacich dimenzií alebo dimenzií typu 2, zatiaľ čo modelovanie „Data Vault“ to robí prostredníctvom satelitov a dátumov načítania. Okrem toho sa tieto dve metódy môžu kombinovať s inými metódami na vytvorenie hybridnej architektúry dátového skladu; dimenzionálne modelovanie sa môže použiť ako prezentačná vrstva alebo vrstva „data mart“, zatiaľ čo modelovanie „Data Vault“ sa môže použiť ako „staging“ vrstva alebo integračná vrstva.

Ako si vybrať tú najlepšiu?

Pokiaľ ide o určenie, ktorá technika modelovania je najlepšia, neexistuje jednoznačná odpoveď, pretože závisí od kvality, objemu, zložitosti a použitia údajov. Napríklad, ak sú zdroje údajov spoľahlivé a konzistentné, potom môže byť vhodnejšie dimenzionálne modelovanie, zatiaľ čo ak sú zdroje údajov nespoľahlivé a nestále, môže byť vhodnejšie modelovanie „Data Vault“. Podobne, ak je objem údajov malý alebo stredne veľký, môže byť dimenzionálne modelovanie efektívnejšie, zatiaľ čo modelovanie „Data Vault“ môže byť škálovateľnejšie pri práci s veľkými alebo rastúcimi objemami údajov. Okrem toho dimenzionálne modelovanie môže byť lepšie pre nízku až strednú zložitosť údajov, zatiaľ čo modelovanie „Data Vault“ môže byť prispôsobivejšie pre vysokú alebo rastúcu zložitosť. Napokon, dimenzionálne modelovanie môže byť vhodnejšie, keď sa údaje používajú najmä na reportovanie, analýzu alebo vizualizáciu, zatiaľ čo modelovanie „Data Vault“ môže byť robustnejšie, keď sa používa na integráciu, správu alebo audit.

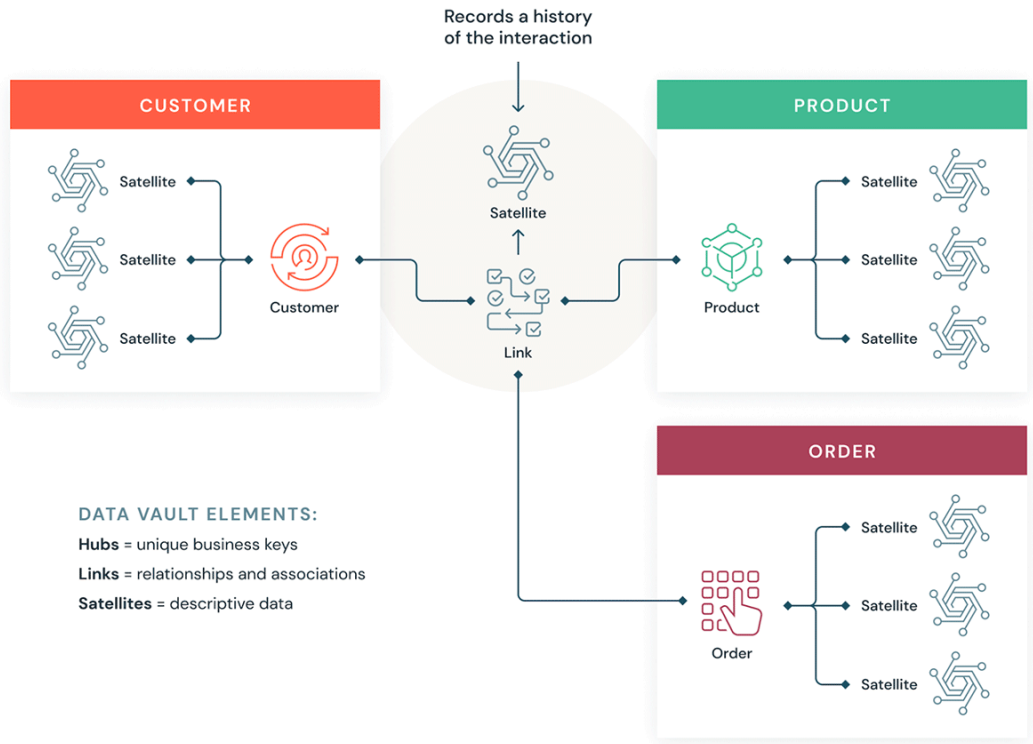
3.3.1 Odporúčanie pre „Data Vault“ modelovanie

Ak sú teda splnené podmienky pre potrebu dátového modelovania, ako dostupnosť surových údajov z veľkého počtu zdrojov, ktoré nedodržiavajú Centrálny model údajov a majú dynamické vzťahy, odporúčame modelovanie „Data Vault“ z nasledujúcich dôvodov:

- Podporujú ho aj moderné platformy ako Snowflake[38] a Databricks[39],
- Je flexibilnejší a nezávislejší od biznis požiadaviek a zdrojov údajov, teda sa dokáže prispôsobiť novým alebo meniacim sa požiadavkám alebo zdrojom bez toho, aby to ovplyvnilo existujúci dátový sklad,
- Lepšie podporuje rastúcu zložitosť údajov,
- Používa normalizovanú schému, ktorá síce zvyšuje počet tabuliek a prepojení, ale zabraňuje redundancii a nekonzistentnosti údajov,
- V koncepte KAV je potrebné pre „staging“ vrstvu a dátovú integráciu,
- Procedúry ETL potrebujú podstatne menej refaktorovania, keď sa zmení dátový model,
- Data Vault je štýl modelovania optimalizovaný pre zápis, podporuje agilné prístupy k vývoju a výborne sa hodí pre dátové jazerá a prístupy typu lakehouse,
- Pri modelovaní dátového trezora sa ako primárne kľúče odporúča používať haš identifikátorov, čo je v súlade s dokumentom 1.1.5 Štandardizácia anonymizácie údajov.

[38] Zdroj: https://www.snowflake.com/blog/feature-engineering-business-vault/, Dátum referencie: 26.05.2023

[39] Zdroj: https://www.databricks.com/blog/2022/06/24/prescriptive-guidance-for-implementing-a-data-vault-model-on-the-databricks-lakehouse-platform.html, Dátum referencie: 26.05.2023



Obrázok 1: Ako funguje modelovanie „Data Vault“ so vzájomne prepojenými hubmi, spojmi a satelitmi[40]

Viaceré spoločnosti majú pozitívne skúsenosti s týmto modelovaním už pri viac ako 100 rôznych zdrojov údajov[41]. Niektoré z týchto zdrojov môžu obsahovať podobné údaje, zatiaľ čo iné sú jedinečné. Typicky pred prijatím metodiky modelovania „Data Vault 2.0“ je každý tím zodpovedný za pochopenie zložitostí každého zdroja údajov, ich čistenie a spájanie s ďalšími údajmi, aby ich mohol formovať podľa svojich potrieb. Existuje veľa zložitostí, redundancie a nekonzistentnosti. Modelovanie „Data Vault 2.0“ dokáže vyriešiť veľa z týchto problémov.

Modelovanie metodiky „Data Vault“ je založené na prístupe zhora nadol, ktorý sa zameriava na pochopenie a modelovanie biznis konceptov nezávisle od systémov. Súčasťou implementácie je spoznanie agend, ako aj zdrojov údajov. Pri vykonávaní základného profilovania a písaní transformácií sa objavujú viaceré zvláštnosti údajov. Je kľúčové dbať na to, aby sa všetky objavy zdokumentovali na centrálnom mieste (napr. Confluence, viac v dokumente č. 1.3.2: Usmernenia pre zrozumiteľné zdokumentovanie dátových štruktúr, procesov tvorby dát, štatistických metodológií (ak boli použité), dátových zdrojov, kontextov a ďalšie aspekty popisu dát).

[40] Zdroj: https://www.databricks.com/blog/2022/06/24/prescriptive-guidance-for-implementing-a-data-vault-model-on-the-databricks-lakehouse-platform.html, Dátum referencie: 26.05.2023

[41] Zdroj: https://medium.com/indigoag-eng/using-a-data-vault-to-support-data-science-1fa4835b6db3, Dátum referencie: 26.05.2023

Jedným zo zásadných postupov metodiky je vkladať do dátového trezora surových údajov vždy len všetky údaje tak, ako sú. Dátoví vedci sa nemusia obávať, že by sa niečo stratilo pri predchádzajúcich transformáciách. Historické zmeny sú zachované a možno ich analyzovať pomocou satelitov, ktoré umožňujú cestovanie v čase. Všetky atribúty zo všetkých zdrojov sa do dátového trezora načítajú pri prvom behu bez potreby čakania, pričom požiadavky na dodatočné údaje realizuje dátové inžinierstvo.

V prípadoch, kedy je potrebné použiť kľúče zdrojových systémov ako identifikátory a údaje sa načítavajú z viacerých zdrojov, je „Same as Link" (Rovnaké ako prepojenie) spôsob, ako deduplikovať údaje a priradiť k sebe údaje z rôznych systémov. Ide o veľmi užitočný vzorec tohto modelovania, ktorý poskytuje flexibilitu pri vykonávaní zhody medzi údajmi z rôznych zdrojov.

Každý, kto sa venuje dátovej analytike alebo dátovej vede, sa na dáta pozerá trochu inou optikou a s iným cieľom ako IT odborník, ktorý buduje dátový sklad alebo lakehouse podľa modelovania „Data Vault“. Spolupráca medzi týmito rolami je neuveriteľne cenná, a to tak počas modelovania a budovania dátového trezora v rámci dátového skladu alebo lakehouse, ako aj po ňom, keď sa údaje profilujú ešte hlbšie a objavujú sa ďalšie potenciálne problémy. Úzka spolupráca umožňuje obom stranám zdieľať odborné znalosti v danej oblasti, identifikovať medzery, ktoré si vyžadujú ďalšie zdroje údajov, identifikovať a zdokumentovať neodhalené problémy s kvalitou údajov a vytvoriť ešte lepšiu dokumentáciu.

4 Najžiadanejšie nástroje pre KAV 2.0

Najžiadanejšie nástroje sa zisťovali dotazníkovým prieskumom v rámci projektu KAV v roku 2022, vykonaným medzi nasledujúcimi analytickými jednotkami:

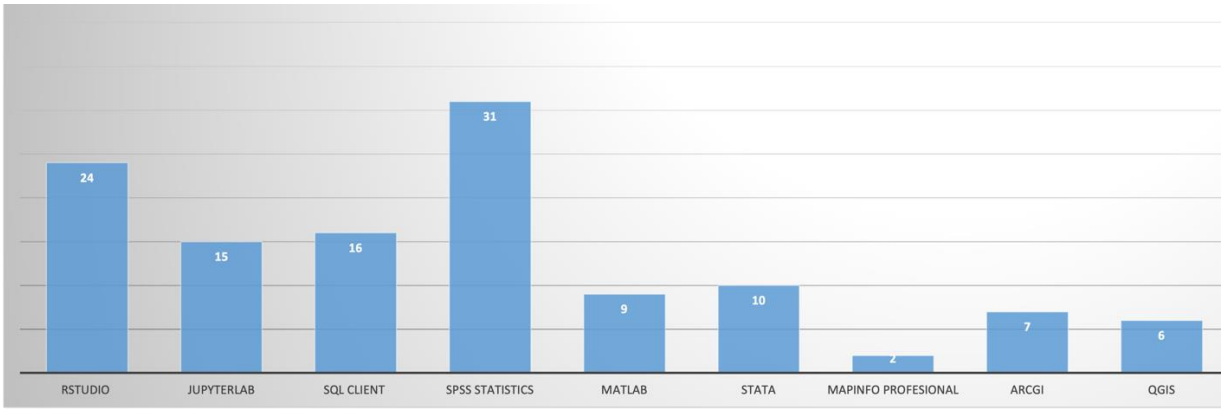
1. Inštitút kultúrnej politiky,
2. Analytický útvar Ministerstva obrany SR,
3. Kancelária Rady pre rozpočtovú zodpovednosť,
4. Inštitút environmentálnej politiky Ministerstva životného prostredia SR
5. Centrum pre hospodárske otázky Ministerstva hospodárstva SR
6. Inštitút digitálnych a rozvojových politík Ministerstva investícií, regionálneho rozvoja a informatizácie SR,
7. Inštitút dopravnej politiky Ministerstva dopravy SR,
8. Inštitút pôdohospodárskej politiky Ministerstva pôdohospodárstva a rozvoja vidieka SR,
9. Analytické centrum Ministerstva spravodlivosti SR,
10. Inštitút vzdelávacej politiky Ministerstva školstva, vedy, výskumu a športu SR,
11. Inštitút finančnej politiky Ministerstva financií SR,
12. Národný kontrolný úrad - odbor stratégie a analýz,
13. Inštitút správnych a bezpečnostných analýz Ministerstva vnútra SR.

Prieskum sa žiaľ nerealizoval medzi inými skupinami koncových používateľov, ktorí nemusia byť natoľko analyticky a programátorsky zruční, ako zamestnanci analytických jednotiek. Pre túto skupinu koncových používateľov sme zanalyzovali a odporučili reportovace nástroje / analytické nástroje bez potreby programovania alebo skriptovania v kapitole 4.2.

4.1 Najžiadanejšie analytické nástroje pre KAV 2.0

Pri prioritizácii nákupu analytických nástrojov treba vyhodnotiť nasledujúce aspekty:

- Možnosť nasadenia do prostredia KAV,
- Využitie pre prioritné prípady použitia (viď kapitola 6.3.2),
- Aktuálny záujem a schopnosť ich využívania analytickými jednotkami (Obrázok 2) alebo inými používateľmi,
- Rozvoj potrebných technických zručností dátových vedcov a analytikov (Tabuľka 1),
- Cenový model a cenová dostupnosť.



Obrázok 2: Prehľad záujmu o analytické nástroje medzi analytickými jednotkami[42]

4.1.1 Podpora jazykov R, SQL a Python

S podporou jazyka R, SQL a Python ako aj Jupyter Notebookmi sa počíta v každej verzii dátovej a analytickej vrstvy. Ide o najrozšírenejšie jazyky vo svete analytického spracovania údajov a dátovej vedy, predovšetkým vďaka tomu, že ide o open source a sú k dispozícii zadarmo. Sú kľúčové pre rozvoj technických zručností dátových vedcov a analytikov (Tabuľka 1). Možno ich používať na vytváranie dátových „pipelines" ako aj na samotné analytické spracovanie údajov a prípravu modelov založených aj na strojovom učení. Knižnice v jazyku Python a R podporujú aj rozličné vizualizácie údajov. RStudio má aj svoju cloudovú verziu[43] a produkty pre RStudio ponúka aj Amazon AWS Marketplace[44], Microsoft Azure Marketplace[45] a Google Cloud Platform[46].

Keďže ide o písanie zdrojového kódu alebo skriptu, je dôležité podporovať aj verzionovanie cez systém git a rozličné nástroje, ako napríklad GitHub[47]. Tiež je dôležité vedieť zdieľať svoje vytvorené modely aj s metrikami ich hodnotenia (viac v dokumente 1.3.2 Usmernenia pre zrozumiteľné zdokumentovanie dátových štruktúr, procesov tvorby dát, štatistických metodológií (ak boli použité), dátových zdrojov, kontextov a ďalšie aspekty popisu dát), použitú verziu datasetov na vytvorenie a overenie modelu, prípadne na jeho využitie, a v prípade strojového učenia aj spravovať verzie natrénovaného modelu, ktorý možno použiť na ďalšie predikcie.

[42] Zdroj: Prieskum záujmu o softvérové nástroje, vykonaný v rámci projektu KAV v roku 2022 (popísaný v úvode kapitoly 4)

[43] Zdroj: https://posit.cloud, Dátum referencie: 30.05.2023

[44] Zdroj: https://aws.amazon.com/marketplace/seller-profile?id=6185573f-e9d3-4df1-a8da-2cd4996a3561, Dátum referencie: 30.05.2023

[45] Zdroj: https://azuremarketplace.microsoft.com/en-us/marketplace/apps?search=rstudio, Dátum referencie: 30.05.2023

[46] Zdroj: https://console.cloud.google.com/marketplace/browse?q=RStudio, Dátum referencie: 30.05.2023

[47] Zdroj: https://github.com, Dátum referencie: 30.05.2023

4.1.2 SPSS Statistics verzus STATA

Balík **IBM SPSS Statistics** je integrovaný súbor produktov zameraný na profesionálnych dátových analytikov. Zaoberá sa celým analytickým procesom, od plánovania až po zber údajov, analýzu, reportovanie a nasadenie. Umožňuje prístupu k viac ako 100 rozšíreniam, vďaka ktorým sa dajú využívať bezplatné knižnice napísané v jazykoch R, Python a v syntaxi SPSS. IBM SPSS Tento nástroj podporuje pokročilé štatistické postupy, ktoré zahŕňajú lineárne a nelineárne štatistické modely, ako aj prediktívne simulačné modelovanie, ktoré zohľadňuje neisté vstupy, geopriestorové analýzy a vlastné tabuľky. Ponúka aj nástroje na prípravu údajov, riešenie pre chýbajúce hodnoty a platnosť údajov, rozhodovacie stromy a prognózovanie. Dajú sa v ňom robiť grafy aj diagramy.

SPSS Statistics je teraz k dispozícii ako možnosť predplatného alebo trvalej licencie či licencie na dobu určitú[48]. Predplatné začína od 99 dolárov na používateľa, k čomu sa prikupujú ďalšie moduly za predplatné. Licencie na dobu určitú a trvalé licencie sú tradičným spôsobom nákupu programu SPSS Statistics prostredníctvom predajcu IBM alebo obchodného partnera IBM.

Hlavnou nevýhodou tohto nástroja je to, že nedokáže spracovávať veľké údaje a neponúka škálovateľnosť cloudových riešení. Tiež najlepšie funguje s databázou Db2 od IBM[49], ktorá je súčasťou proprietárneho dátového „stacku“.

Stata je štatistický softvérový nástroj, ktorý sa dá použiť na rôzne účely. Je to populárny nástroj na štatistickú analýzu. Vytvorila ho spoločnosť Stata Corp v roku 1985. Je to licencovaný produkt, ročná licencia pre jedného používateľa začína na 840 dolárov na rok[50]. Pracuje aj s rôznymi operačnými systémami vrátane Linuxu, Windows a Mac OS. Neponúka cloudovú verziu, možno ju len nainštalovať na server vo verejnom cloude[51].

V programe Stata je získavanie údajov a manipulácia s nimi pomerne jednoduchá. Aj keď používateľ nepozná syntax, vie sa ju rýchlo naučiť. Integruje aj jazyk Python cez PyStata. Ponúka vytváranie užitočných reportov pomocou vizualizácie dátových modelov. Je to jeden z najvýkonnejších softvérových nástrojov na správu údajov, analýzu údajov a grafiku, ktoré sú k dispozícii. Dá sa ľahko rozširovať a online je dostupných veľa zdrojov zdarma na učenie sa. Do Stata nástroja sa však nedajú doprogramovať nové funkcionality, podporuje len isté typy dát a príprava grafov trvá veľmi dlho.

Záver

Stata a SPSS Statistics sú veľmi podobné nástroje. Zásadný rozdiel je to, že SPSS má užívateľsky prívetivé grafické používateľské rozhranie s rozbaľovacími ponukami, ktoré väčšina používateľov desktopového softvéru intuitívne pochopí. STATA používa inštrukcie v syntaxi príkazového riadka. SPSS má pri používaní prostredníctvom grafického používateľského rozhrania obmedzenejšie funkcie. Preto „power users“ často uprednostňujú STATA. Oba nástroje majú nevýhodu, že nevedia byť súčasťou

[48] Zdroj: https://www.ibm.com/products/spss-statistics/pricing, Dátum referencie: 30.05.2023

[49] Zdroj: https://medium.com/@SimonLightstone/4-reasons-why-db2-on-cloud-is-the-best-database-for-spss-d922a9a2e9e4 , Dátum referencie: 30.05.2023

[50] Zdroj: https://www.stata.com/order/new/gov/single-user-licenses/dl/,

[51] Zdroj: https://blog.stata.com/2019/11/05/stata-in-the-cloud/,

modulárneho dátového a analytického „stacku“ s otvoreným zdrojovým kódom, ktorý v plnej miere využíva prostredie verejného cloudu a vie spracovávať veľké údaje.

4.1.3 Matlab

Platforma Matlab a Simulink predstavuje analytický svet sám o sebe s veľmi špecifickými nástrojmi na modelovanie a simuláciu, ktoré sú aj veľmi úzko špecializované pre rôzne odvetvia ako neurovedy, fyziku, biotechnológie, telekomunikačné systémy, elektroniku a mnoho ďalších. Srdcom MATLABu je jazyk MATLAB, ktorý je založený na maticiach a umožňuje najprirodzenejšie vyjadrenie výpočtovej matematiky.

Vstavané funkcie MATLABu ponúkajú špičkové prostriedky na vykonávanie výpočtov vrátane optimalizácie, lineárnej algebry, numerického riešenia obyčajných diferenciálnych rovníc , analýzy údajov, spracovania signálov a mnohých ďalších vedeckých úloh. Pre väčšinu týchto funkcií sa používajú moderné algoritmy. Existuje ich veľa pre animácie aj 2-D a 3-D grafiku. MATLAB podporuje aj externé rozhranie. Dá sa integrovať aj s inými programovacími jazykmi ako Python, Java a C/C++. Používateľ môže vytvárať vlastné funkcie v jazyku MATLAB. Nie je teda obmedzený na používanie len vstavaných funkcií.

MATLAB poskytuje ďalšie balíky nástrojov napríklad pre neurónové siete, symbolické výpočty, spracovanie obrazu, návrh riadiacich systémov a štatistiku.

Rôzne spôsoby využitia MATLABu sú:

- Vývoj algoritmov,
- Riešenie lineárnej algebry,
- Vykresľovanie grafov pre datasety veľkých údajov,
- Vizualizácia a analýza údajov,
- Numerické výpočty matíc.

Výhodou je, že dokáže spracovávať aj veľké údaje, pričom najlepšiu podporu má pre verejný cloud Amazon AWS, na ktorom sa dá vytvoriť aj klaster grafických kariet NVIDIA pre hĺbkové učenie[52]. Mnohé produkty sa dajú využívať aj v iných cloudových prostrediach ako Microsoft Azure.

Záver

Vzhľadom na uzavretý ekosystém, špecifickosť jazyka Matlab, ale predovšetkým vzhľadom na jeho vysokú cenu ho odporúčame len na tie prípady použitia, na ktoré nemožno použiť žiaden iný nástroj uvedený v kapitole 4.1.

[52] Zdroj: https://www.mathworks.com/help/cloudcenter/ug/matlab-deep-learning-container-on-aws.html, Dátum referencie: 30.05.2023

4.1.4 ARCGI verzus QGIS verzus Mapinfo Professional (premenované na Precisely)

Ide o špecializované geografické informačné systémy (GIS), ktoré slúžia na spracovanie a analýzu priestorových údajov.

ArcGIS od spoločnosti Esri je integrovanou kolekciou softvérových produktov geografického informačného systému pre rôzne platformy. ArcGIS možno nasadiť do cloudového prostredia, privátneho aj do Amazon AWS alebo Microsoft Azure, alebo ho možno využívať ako manažovanú službu SaaS v cloude[53]. Umožňuje spracovanie priestorových analýz, správu údajov a mapovanie 3D údajov na základe polohy v súlade so štandardmi. Zároveň sa pomocou neho dá spolupracovať s ostatnými zdieľaním poznatkov nielen prostredníctvom máp, ale aj aplikácií a reportov.

Nástroj ArcGIS ponúka prístup s jednotným prihlásením (SSO), správu používateľov na základe skupín, prideľovanie oprávnení a licencií a monitorovanie interných ukazovateľov činnosti v rámci organizácie, ako aj ukazovateľov pre verejne zdieľané mapy/aplikácie. Vytváranie máp je veľmi jednoduché. Údaje sa dajú nahrať či už pretiahnutím tabuľky, nahratím lokálnych dátových súborov alebo pripojením obsahu uloženého v cloude. Následne inteligentné mapovanie vygeneruje návrhy na vykreslenie vizualizácií založených na údajoch a zároveň ponúkne vlastné možnosti štýlovania. K dispozícii je zbierka základných máp vo vektorovom formáte pre vernosť zobrazenia vo vysokom rozlíšení s flexibilným štýlovaním, ktorá pozostáva zo satelitných snímok, máp ulíc, krajiniek, oceánov a ďalších.

Na portáli sa dá vykonávať niekoľko typov analýz. Tieto typy analýz možno rozdeliť do kategórií na základe údajov použitých v analýze. Ak sa plánuje vykonať analýza údajov, ktoré obsahujú premenné („features“) alebo tabuľkové údaje, vykonáva sa analýza premenných. Nástroje priestorovej analýzy sa dajú použiť na vykonávanie bežných analytických funkcií, ako je vyhľadávanie „hot spots“, lokalizácia ulíc a adries, vyhľadávanie miesta, smerovanie alebo prístup ku geodatabáze. Nástroje geoanalytiky možno použiť na vykonávanie analytických funkcií, ktoré analyzujú vzory a agregujú údaje v kontexte priestoru aj času. Ak sú údaje v rastrovom formáte, k dispozícii rastrovú analýzu. Nástroje na analýzu rastrov sa používajú na analýzu a spracovanie snímok a rastrových datasetov, ako sú multispektrálne satelitné a letecké snímky, výškové, vedecké a klasifikované datasety. Nástroje môžu analyzovať snímky na základe algoritmov, extrahovať priestorové vzory, spracovávať terény na odvodenie povrchov a sumarizovať a spravovať údaje. Dajú sa tiež tiež vykonávať pracovné postupy hĺbkového učenia („deep learning“) a analýzy časových radov. K dispozícii sú aj nástroje, ktoré sa dajú prispôsobiť na mieru. Výhodou je aj dátová vrstva s rôznorodými databázami – relačnou, časopriestorovou aj s podporou pre veľké dáta, grafovou pre podporu znalostných grafov a objektovou.

Čo sa týka licencovania, ArcGIS ponúka ročné licencie v závislosti od počtu používateľov, pričom ročná licencia pre jednotlivca začína na 100 dolároch ročne. K dispozícii sú rôzne balíky produktov a služieb, pre podnikové licencie treba kontaktovať obchodné oddelenie.

QGIS ponúka mnoho bežných funkcií GIS poskytovaných základnými funkciami a zásuvnými modulmi. Ide o nástroj s otvoreným zdrojovým kódom, ktorá je k dispozícii aj v online verzii (zadarmo aj v pro platenej verzii)[54]. Vektorové a rastrové údaje sa dajú zobrazovať a prekrývať v rôznych formátoch a projekciách bez konverzie do interného alebo spoločného formátu. Medzi podporované formáty patria

[53] Zdroj: https://enterprise.arcgis.com/en/cloud/, Dátum referencie: 30.05.2023

[54] Zdroj: https://qgiscloud.com/pages/plans?locale=en, Dátum referencie: 30.05.2023

napríklad: tabuľky a pohľady s priestorovou podporou pomocou PostGIS, SpatiaLite a MS SQL Spatial, Oracle Spatial, vektorové formáty podporované nainštalovanou knižnicou OGR vrátane ESRI shapefiles, MapInfo, GML a mnohé ďalšie. Rastrové a obrazové formáty podporované nainštalovanou knižnicou GDAL (Geospatial Data Abstraction Library) sú napríklad GeoTIFF, ArcInfo ASCII GRID, JPEG, PNG a mnohé ďalšie. Podporuje aj online priestorové údaje poskytované ako webové služby.

Pomocou prívetivého grafického rozhrania sa dajú vytvárať mapy a interaktívne skúmať priestorové údaje. Vektorové a rastrové vrstvy sa dajú vytvárať, upravovať, spravovať a exportovať vo viacerých formátoch. Podporuje vizualizácie a úpravy údajov OpenStreetMap.

Možno vykonávať analýzu priestorových údajov v priestorových databázach a iných formátoch podporovaných OGR. QGIS v súčasnosti ponúka nástroje na vektorovú analýzu, vzorkovanie, geoprocesing, geometriu a správu databáz. Vytvorené mapy a analýzy sa dajú aj publikovať na internete.

QGIS možno používať ako klienta WMS, WMTS, WMS-C alebo WFS a WFS-T a ako server WMS, WCS alebo WFS. (Pozri časť Práca s údajmi OGC.) Okrem toho môžete svoje údaje publikovať na internete pomocou webového servera s nainštalovaným UMN MapServer alebo GeoServer.

QGIS sa dá prispôsobiť špeciálnym potrebám vďaka rozšíriteľnej architektúre zásuvných modulov a knižniciam, ktoré možno použiť na vytvorenie zásuvných modulov. Dokonca sa dajú vytvárať nové pomocou jazyka C++ alebo Python.

MapInfo Pro (Precisely) je **desktopové** mapové riešenie pre analytikov geografických informačných systémov (GIS) na vizualizáciu, analýzu, úpravu, interpretáciu a výstup údajov - odhaľuje vzťahy, vzory a trendy. Údaje sa dajú vizualizovať pomocou symbolov, tém a značiek na mape. Na jednej mape sa dá prekryť viacero datasetov a rozlíšiť tak vzory, ktoré by inak neboli viditeľné. Integrovať s mapami sa dajú aj podnikové údaje s demografickými údajmi. Nad nimi sa dajú použiť výkonné nástroje na priestorové dopytovanie a modelovanie. Dajú sa spúšťať rôzne scenáre na presné a aktuálne zobrazenie možností lokality. Vytvorené modely môžu byť ľahko zrozumiteľné aj pre netechnických zamestnancov vďaka interaktívnej a intuitívnej vizualizácii. Tento nástroj ponúka ročné až trojročné licencie, pričom cena začína na 785 librách na rok na používateľa[55].

Záver

Nástroje GIS sa v skutočnosti obmedzujú na 4 jednoduché myšlienky:

- Vytváranie geografických údajov.
- Ich spravovanie,
- Analyzovanie,
- A ich zobrazenie na mape.

Toto sú základné funkcie, ktoré dobre plnia oba softvéry GIS - QGIS alebo ArcGIS. Ani s jedným zo softvérov na tvorbu máp GIS - QGIS alebo ArcGIS - nemôžete urobiť chybu.

[55] Zdroj: https://www.precisely.com/product/precisely-mapinfo/mapinfo-pro, Dátum referencie: 30.05.2023

QGIS je zadarmo. Má podporu viacerých jazykov a spolieha sa na úsilie dobrovoľníkov, čo je naozaj dobré. Má obrovskú podporu na „stack exchange[56]“. Čím viac pracujete v QGIS, tým viac skrytých skvostov nájdete, ako napríklad:

- Interaktívne pivotové tabuľky s GroupStats,
- Jednoduché pridávanie CSV,
- Ohromujúca kartografická symobológia a možnosti označovania.

ArcGIS je jednou z najlepších investícií do GIS. Je rozšíriteľný a má najväčšiu komunitu používateľov. Taktiež poskytuje výukové programy so vzorovými údajmi na získanie praktických skúseností. ModelBuilder a automatizácia sú špičkové.

V najhrubších rysoch je hodnotenie takéto: MapInfo Pro < ArcView < ArcEditor < QGIS < ArcInfo. Víťazom je ArcInfo.

4.2 Prehľad vhodných reportovacích / analytických nástrojov bez potreby programovania alebo skriptovania

V týchto nástrojoch sa tiež analyzujú údaje a vytvárajú sa dashboardy, ktoré používatelia môžu skúmať a používať napríklad na lepšie rozhodovanie. Na trhu je nespočetne veľa takýchto nástrojov. Tieto moderné nástroje boli navrhnuté aj s ohľadom na netechnických používateľov, preto ich uvádzame, aj keď neboli výsledkom prieskumu medzi analytickými jednotkami. Tieto nástroje umožňujú expertom v danej oblasti odpovedať na otázky v rámci agendy bez závislosti od vývojárov a dátových analytikov. V nasledujúcej tabuľke (Tabuľka 5) vyberáme len tie populárne nástroje, ktoré sú cenovo dostupné, podporujú aj skriptovanie a dátovú analýzu s pomocou jazykov R a Python, a dajú sa ľahko zakomponovať do dátového „stacku“ vo verejnom cloude.

Tabuľka 5: Prehľad vlastností populárnych reportovacích / analytických nástrojov

Parameter	Power BI	Tableau	Qlik Sense	Looker
Verzia zadarmo s plnou funkcionalitou	Áno	V samostatnom nástroji	V samostatnom nástroji	Nie
Vývojárske prostredie	Desktop	Desktop	Webový prehliadač	Cloud
Integrácia s dátovým „stackom“ v cloude	Áno	Áno	Áno	Áno

[56] Zdroj: https://stackexchange.com, Dátum referencie: 30.05.2023

Parameter	Power BI	Tableau	Qlik Sense	Looker
Podpora R a Pythonu	Áno	Áno	Áno	Áno
Analytika s podporou umelej inteligencie	Áno	Áno	Áno	Nie
Vyhľadávanie s podporou strojového spracovania prirodzeného jazyka	Áno	Nie	Áno	Nie
Nástroj na prípravu dát	Áno	V samostatnom nástroji	V samostatnom nástroji	Nie
Nástroje na dátové modelovanie	Áno	V samostatnom nástroji	Áno	Áno
Preferovaný dátový model	Star-schéma	Plochý	Snowflake	Plochý
Nezávislý od databázy	Áno	Áno	Áno	Nie
Vstavaná bezpečnosť na úrovni riadkov	Áno	Áno	Áno	Nie
Zmiešané typy modelov	Áno	Nie	Nie	Nie
Prístup k modelu tretej strany	Áno	Nie	Nie	Nie
Komentovanie a spolupráca	Áno	Áno	Áno	Nie
Vizualizácia na mieru a open-source	Áno	Nie	Áno	Áno

Záver

Power BI je veľmi pokročilým nástrojom, ktorý spĺňa všetky parametre a vďaka tomu, že je ľahšie dostupný ako súčasť rozšíreného balíčka Microsoft Office produktov, existuje aj dostatok ľudí, ktorí s ním vedia pracovať. V súkromnej sfére sú ale veľmi populárne nástroje Tableau a Qlik Sense, vďaka rozšíreným funkcionalitám na analýzu a vizualizáciu údajov. Jedným z faktorov, ktoré musí organizácia zohľadniť pri výbere, sú celkové náklady, ktoré jej vzniknú. Náklady úplne závisia od počtu ľudí, ktorí potrebujú prístup, od verzie, ktorú sa spoločnosť rozhodne používať, a od licenčných poplatkov, ktoré vzniknú.

Tableau Desktop stojí 70 dolárov na mesiac pre jednotlivcov, zatiaľ čo pre tímy a organizácie sa pohybuje od 12 do 70 dolárov na mesiac v závislosti od ich požiadaviek. Qlik Sense Analytics sa dodáva v 2 rôznych variantoch:

- For Business - riešenie SaaS na operacionalizáciu analytiky v skupinách a tímoch za 30 dolárov na mesiac
- Qlik Sense® Enterprise - multicloudové riešenie na škálovanie a rozšírenie analytiky naprieč oddeleniami a organizáciami. Stojí 70 dolárov mesačne pre profesionálov a 40 dolárov mesačne pre analytikov.

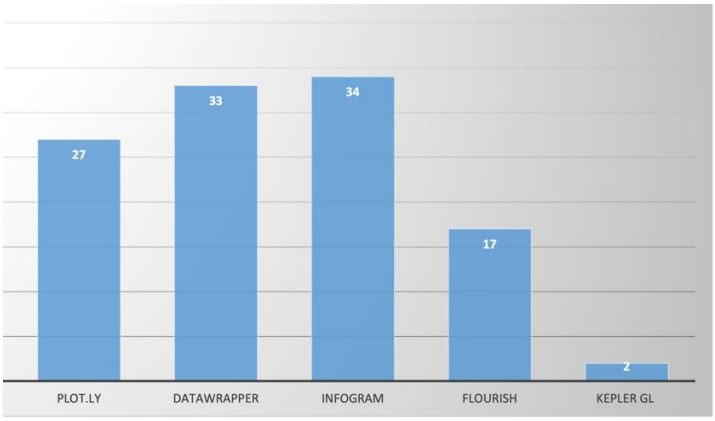
V tomto parametri ale Tableau vyniká - organizácia spustila nástroj s názvom Tableau Public, ktorý je bezplatný. Ponúka takmer celý rozsah funkcionalít, ktoré má Tableau Desktop (okrem jednej nevýhody - svoju prácu treba ukladať do galérie Tableau Public, nie do lokálneho počítača). Stojí za to sa naň pozrieť.

Ďalším určujúcim faktorom, ktorý ovplyvní výber, sú zručnosti koncového používateľa. Ak sú používatelia technicky zdatní, vedia skriptovať alebo programovať a hľadajú aplikáciu, ktorá im umožní vykonávať komplexné transformácie a vizualizácie údajov na jednej platforme, budú sa skôr prikláňať k aplikácii Qlik Sense. Na druhej strane, ak sú menej technicky zdatní a vyžadujú len prácu s transformovanými údajmi, lepšia voľba Tableau. Takou strednou cestou v tomto smere je Power BI.

4.3 Prehľad vhodných vizualizačných nástrojov pre KAV

Vizualizačné nástroje sa vyberali podľa nasledujúcich kritérií:

- Vhodnosť nasadenia do prostredia KAV,
- Aktuálny záujem a schopnosť ich využívania analytickými jednotkami,
- Rozvoj potrebných technických zručností dátových vedcov a analytikov (Tabuľka 1),
- Cenový model a cenová dostupnosť.



Obrázok 3: Prehľad záujmu o vizualizačné nástroje medzi analytickými jednotkami[57]

Vizualizačné nástroje sú poslednou nadstavbou nad dátovým a analytickým „stackom“. Využívajú sa na tú formu vizualizácie a prezentovanie výsledkov analytického spracovania a analytických modelov, ktorú nie je možné spraviť v nástrojoch popísaných v kapitole 4.1 a 4.2 (aj programovacie jazyky R a Python majú k dispozícii mnoho knižníc na vizualizáciu údajov, modelov a ich výsledkov). Spravidla tieto nástroje na vizualizáciu sú napojené na už analyticky spracované údaje alebo na údaje, ktoré sú výstupom analytických modelov alebo strojového učenia. Môže ísť teda o:

- Nahratie tabuľkového súbora ako CSV do nástroja,
- Napojenie na „živý“ zdroj údajov ako databázu,
- Integrovanie s API.

Spravidla tieto vizualizačné nástroje poskytujú aj vnorenie výslednej vizualizácie na ľubovoľnú webovú stránku, alebo export do správ, prezentácií či ako obrázky, napríklad aj pre vytlačenie. Poskytujú aj krásne, interaktívne šablóny na to, aby sa dal vytvoriť okolo údajov a ich analýz príbeh (na to sa sústreďuje hlavne nástroj Flourish a Infogram). Datawrapper je obzvlášť populárny pre vytváranie grafík, založených na dátach, ktoré sú súčasťou správ, novín alebo online spravodajstva. Pomocou Plot.ly sa dajú robiť krásne interaktívne aplikácie ako dashboardy a reportovacie nástroje založené na údajoch s minimom zdrojového kódu. Jednoducho sa dá napojiť na mnoho databáz v cloudovom „stacku“. Kepler.gl je výkonný open source nástroj na geopriestorovú analýzu rozsiahlych súborov údajov.

Záver

Všetky nástroje uvedené na obrázku (Obrázok 3) majú svoje unikátne vlastnosti a sú dostupné. Flourish, Infogram a Datawrapper má aj verziu zadarmo. Kepler.gl je bezplatný open source. Plot.ly má podnikovú licenciu na používateľa zhruba 950 dolárov na rok. Závisí teda len na konkrétnom prípade použitia, v ktorom sa dá spraviť najefektívnejšie a najatraktívnejšie.

[57] Zdroj: Prieskum záujmu o softvérové nástroje, vykonaný v rámci projektu KAV v roku 2022 (popísaný v úvode kapitoly 4)

4.4 Využitie analytických nástrojov v multi-databázovom prostredí

Keďže zdroje údajov neustále rastú, vykonávanie analytického spracovania údajov s viacerými databázami sa stalo neefektívnym a nákladným. Preto vznikli tri prístupy na centralizáciu dát: dátový sklad, dátové jazero a dátový „lakehouse“.

Dátový sklad („Data Warehouse“)

Dátový sklad podporuje tok údajov z prevádzkových systémov do analytických/rozhodovacích systémov vytvorením jediného úložiska údajov z rôznych zdrojov (interných aj externých). Vo väčšine prípadov je dátový sklad relačná databáza, ktorá uchováva spracované údaje optimalizované na zhromažďovanie poznatkov. Zhromažďuje údaje s vopred stanovenou štruktúrou a schémou pochádzajúce z transakčných systémov a podnikových aplikácií a údaje sa zvyčajne používajú na prevádzkové výkazníctvo, analýzy, dashboardy a vizualizácie. Dátový sklad ponúka vyššiu kvalitu údajov a rýchlejšie výsledky dotazov. Cloudové ponuky pre dátové sklady sú Google Big Query[58], Amazon Redshift[59], Azure Synapse Analytics[60] a Snowflake[61].

Dátový sklad nedokáže spracovať surové alebo neštruktúrované údaje a údržba skladu s neustále rastúcim množstvom údajov je nákladná. Takisto nie je najlepším riešením na pokročilé analytické spracovanie údajov, ako je strojové učenie alebo prediktívna analýza.

Dátové jazerá („Data Lakes“)

Keďže údaje, ktoré sa majú dostať do dátových skladov, sa musia pred uložením spracovať - môže to vzhľadom na obrovské množstvo neštruktúrovaných údajov zabrať značný čas a zdroje. V reakcii na to začali podniky v roku 2010 budovať dátové jazerá, ktoré uchovávajú všetky štruktúrované a neštruktúrované údaje v akomkoľvek rozsahu. Dátové jazero má oddelenú vrstvu ukladania a spracovania údajov v porovnaní so starším dátovým skladom, kde je za ukladanie aj spracovanie zodpovedný jeden nástroj. Dátové jazero ukladá údaje do spomínaného objektového úložiska, ako je Amazon S3, Google Cloud Storage alebo Azure Data Lake Storage. Azure Data Lake[62] je takouto ponukou od spoločnosti Microsoft. Okrem toho na tomto trhu pôsobí napríklad Snowflake[63], Cloudera[64] Dátové jazerá ukladajú surové, pološtruktúrované aj štruktúrované údaje a môžu sa zriadiť bez toho, aby bolo potrebné najprv definovať štruktúru a schému údajov. Dátové jazero vhodné pre dátového vedca, ktorý dokáže spracovať surové údaje. Naproti tomu dátový sklad je viac prispôsobený koncovým používateľom. Dátový sklad je ideálny na strojové učenie, prediktívnu analýzu, profilovanie používateľov

[58] Zdroj: https://cloud.google.com/bigquery/, Dátum referencie: 30.05.2023

[59] Zdroj: https://aws.amazon.com/redshift/, Dátum referencie: 30.05.2023

[60] Zdroj: https://azure.microsoft.com/en-us/products/synapse-analytics/, Dátum referencie: 30.05.2023

[61] Zdroj: https://www.snowflake.com/en/data-cloud/workloads/data-warehouse/, Dátum referencie: 30.05.2023

[62] Zdroj: https://azure.microsoft.com/en-us/solutions/data-lake,

[63] Zdroj: https://www.snowflake.com/en/data-cloud/workloads/data-lake/,

[64] Zdroj: https://www.cloudera.com/products/sdx/data-lake-service.html,

atď. Dátové jazerá umožňujú používateľom spúšťať analýzy bez toho, aby museli údaje presúvať do samostatného analytického systému, čo podnikom umožňuje získavať poznatky z nových zdrojov údajov, ktoré predtým neboli k dispozícii na analýzu, napríklad vytváraním modelov strojového učenia pomocou údajov z logovacích súborov, sociálnych sietí a zariadení internetu vecí. Vďaka tomu, že všetky údaje sú ľahko dostupné na analýzu, môžu dátoví vedci odpovedať na nové typy otázok alebo riešiť staré otázky pomocou nových údajov.

Dátové jazerá riešia mnohé problémy dátových skladov, ale majú nízku kvalitu údajov a výkonnosť dotazov nie je dostatočne efektívna. Okrem toho si vyžaduje ďalšie nástroje na spúšťanie dotazov SQL. Častou výzvou architektúry dátového jazera je, že bez zavedenia vhodného rámca kvality a správy údajov, kedy do dátového jazera prúdia terabajty štruktúrovaných a neštruktúrovaných údajov, je často veľmi ťažké triediť ich obsah. Ak dátové jazero nie je dobre organizované, môže to viesť k problému stagnácie, čo znamená, že údaje sú síce uložené na jednom mieste, ale sa v plnej miere nevyužívajú. Dátové jazerá sa tak môžu zmeniť na „dátové bažiny“, pretože uložené údaje sa stanú príliš chaotickými na to, aby sa dali použiť.

Dátové lakehouses

Dátový lakehouse je novou architektúrou, ktorá spája to najlepšie z oboch svetov - dátových skladov a dátových jazier. Slúži ako jednotná platforma pre dátové sklady a dátové jazerá. Má funkcie manažmentu údajov, ako je napríklad ACID transakcia vychádzajúca z konceptu dátového skladu a nízkonákladové ukladanie ako dátové jazero. Poskytuje priamy prístup k zdrojovým údajom, umožňuje súbežné operácie čítania a zápisu údajov a podporu schém pre manažment údajov.

Všetky údaje - štruktúrované, pološtruktúrované, neštruktúrované, sa ukladajú v dátovom jazere bez toho, aby sa vykonávalo ich spracovanie. Neskôr sa na vytvorenie konkrétneho prípadu použitia údajov použijú rôzne nástroje na spracovanie. Okrem toho výkonnostné optimalizácie, ako je indexovanie, skompaktnenie údajov, pomáhajú dosiahnuť rýchlejšie výsledky dopytov podobne ako v dátovom sklade. Podporuje aj toky údajov („streaming data“), takže dokáže aktualizovať dashboardy v reálnom čase.

Architektúra dátového lakehouse ponúka zvýšenú spoľahlivosť údajov tým, že znižuje objem údajov prenesených cez ETL procedúry cez ukladanie nespracovaných údajov. Údaje sa tak nebudú duplikovať vo viacerých systémoch, kam by ich nahrali ETL procedúry. Vďaka zníženiu počtu procedúr ETL a odstráneniu duplikácie sa znížia aj náklady. Okrem toho ponúka aj lepšiu správu údajov a otvára údaje pre viaceré prípady použitia.

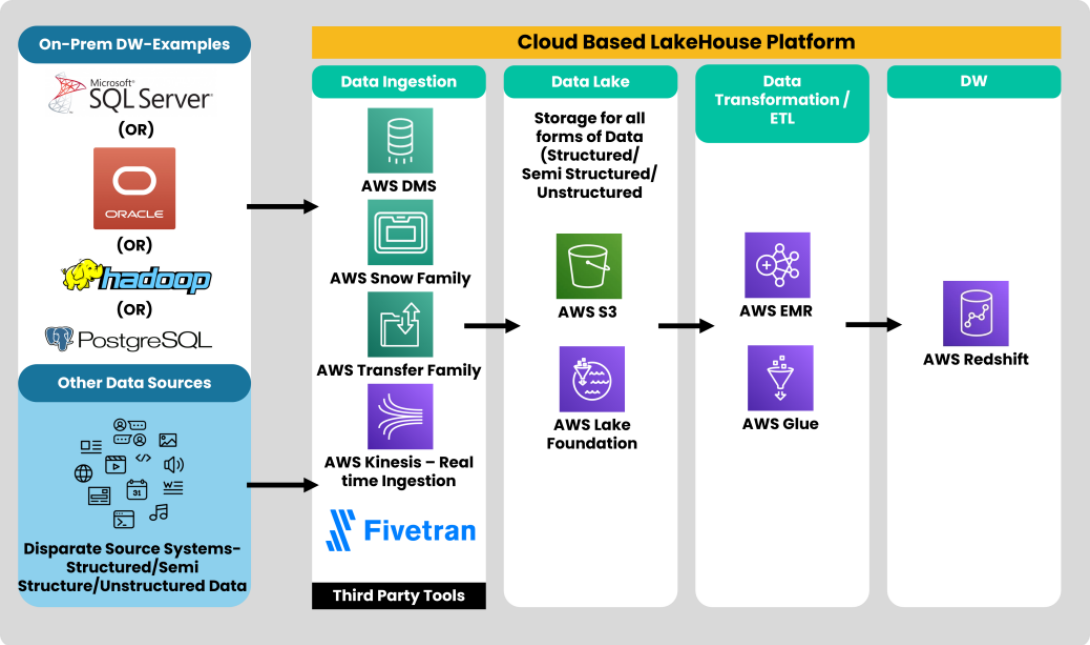
Niektoré z kľúčových funkcionalít dátového lakehouse sú nasledovné:

- **Podpora transakcií ACID** - ACID zabezpečuje konzistenciu transakcií a integritu údajov. Pomáha zachovať integritu údajov, keď rôzne komponenty vykonávajú súbežné operácie alebo v prípade zlyhania. Je to základná vlastnosť dátového skladu a implementovala sa aj do dátového lakehouse.
- **Surový alebo neštruktúrovaný formát údajov** - Dátový sklad podporoval len štruktúrované údaje, ale v tomto prípade máme k dispozícii podporu pre neštruktúrované typy údajov vrátane zvuku, videa atď.
- **Podpora streamovania** – Moderné informačné systémy generujú údaje ako neobmedzený tok, ktorý treba analyzovať v reálnom čase. Dátový lakehouse má podporu pre streamovanie údajov a generovanie poznatkov v reálnom čase.

- **Oddelené ukladanie a výpočet** - Ukladanie a výpočet sú oddelené, vďaka čomu sú nezávisle škálovateľné podľa potrieb prípadu použitia. Umožňuje tiež spúšťať dotazy pomocou rôznych výpočtových uzlov.

Riešenie pre dátový lakehouse ponúka napríklad Google v riešení BigLake[65], Cloudera[66], Microsoft Fabric[67] alebo Databricks[68]. Dátový lakehouse možno vyskladať aj v cloude Amazon AWS nasledovným spôsobom (Obrázok 4):

- Python APIs pre načítanie údajov,
- Úložisko S3 pre dátové jazero,
- Amazon Glue pre procedúry ETL,
- Amazon Redshift pre dátový sklad.



Obrázok 4: Implementácia architektúry dátového lakehouse v cloude Amazon AWS[69]

V nasledujúcej tabuľke (Tabuľka 6) sú zhrnuté rozdiely medzi dátovým skladom, dátovým jazerom a dátovým lakehouse.

[65] Zdroj: https://cloud.google.com/biglake, Dátum referencie: 29.05.2023

[66] Zdroj: https://www.cloudera.com/products/open-data-lakehouse.html, Dátum referencie: 29.05.2023

[67] Zdroj: https://www.microsoft.com/en-us/microsoft-fabric, Dátum referencie: 29.05.2023

[68] Zdroj: https://www.databricks.com, Dátum referencie: 29.05.2023

[69] Zdroj: https://www.tigeranalytics.com/blog/practical-guide-data-lakehouse-across-aws-azure-gcp-snowflake/, Dátum referencie: 29.05.2023

Tabuľka 6: Porovnanie dátového skladu, dátového jazera a dátového lakehouse (zelenou sú označené výhody, z ktorých vie profitovať aj KAV)

Aspekt porovnania	Dátový sklad	Dátové jazero	Dátový lakehouse
Typ dátového úložiska	Dobre pracuje so štruktúrovanými údajmi	Dobre pracuje s pološtruktúrovanými a neštruktúrovanými údajmi	Dokáže spracovať so štruktúrovanými, pološtruktúrovanými a neštruktúrovanými údajmi
Účel	Optimálne pre prípady použitia v oblasti analýzy údajov a business intelligence (BI)	Vhodné pre úlohy v oblasti strojového učenia (ML) a umelej inteligencie (AI)	Vhodné pre úlohy v oblasti analýzy údajov aj strojového učenia
Náklady	Úložisko je nákladné a časovo náročné vybudovať	Úložisko je nákladovo efektívne, rýchle a flexibilné	Úložisko je nákladovo efektívne, rýchle a flexibilné
Súlad s ACID	Zaznamenáva údaje spôsobom, ktorý je v súlade s ACID, aby sa zabezpečila najvyššia úroveň integrity	Nesúlad s ACID: aktualizácie a vymazania sú zložité operácie	Súlad s ACID, aby sa zabezpečila konzistencia, keď viacero strán súčasne číta alebo zapisuje údaje

Ako sa teda medzi týmito architektúrami pre KAV rozhodnúť?

Dátový sklad je dobrou voľbou pre organizácie, ktoré hľadajú vyspelé, štruktúrované dátové riešenie, ktoré sa zameriava na prípady použitia v oblasti business intelligence a dátovej analýzy. Dátové jazerá sú zas vhodné pre organizácie, ktoré hľadajú flexibilné, nízkonákladové riešenie na spracovanie veľkých objemov údajov, ktoré by umožnilo implementovať strojové učenie a dátovú vedu na neštruktúrovaných údajoch.

Ak však prístupy dátového skladu a dátového jazera nespĺňajú požiadavky na údaje, alebo ak treba nájsť spôsob, ako na údajoch implementovať pokročilé analytické spracovanie údajov aj úlohy strojového učenia, v takom prípade je dátový lakehouse rozumnou voľbou. **Keďže ambíciou KAV je implementovať aj prípady použitia založené na strojovom učení, avšak zároveň pre niektoré skupiny používateľov je požiadavkou mať k dispozícii aj používateľsky intuitívnejšie nástroje na analytické spracovanie údajov, odporúčame ísť cestou dátového lakehouse.** Práve platformy pre dátové lakehouse sa snažia maximalizovať výhody dátových skladov a dátových jazier, a zároveň potlačiť nevýhody. Dátové lakehouses sa na začiatku môžu zložito budovať. Uľahčí to však využívanie platformy ako napríklad spomínaný Databricks či Cloudera, vytvorenej na podporu otvorenej architektúry datových lakehouses. Navyše mnohé takéto platformy sa platia podľa využívania a možno ich nasadiť na lídrov na trhu s verejnými cloudmi ako Amazon AWS, Google Cloud a Microsoft Azure, takže počiatočná investícia do rozbehnutia dátovej a analytickej infraštruktúry je nízka a bez rizika. Ďalším podporovaným trendom je takzvaný „multicloud“, kedy nasadené platformy a cloudové služby počítajú s tým, že údaje a aplikácie sa nachádzajú u viacerých poskytovateľov verejných, prípadne aj privátnych cloudov.

Od výberu poskytovateľa cloudu, architektúry a platformy pre podporu dátového lakehouse následne závisí aj dostupnosť a jednoduchosť nasadenia rozličných analytických nástrojov. Najvyššiu mieru flexibility má využívanie programovacích jazykov R a Python, ktoré majú otvorený zdrojový kód a sú podporované všetkými veľkými poskytovateľmi cloudových služieb a platforiem pre dátový lakehouse. Rôzne analytické a reportovacie nástroje podporujú priame predpripravené pripojenia na rôzne databázy.

Využívanie oddelených analytických nástrojov popísaných v kapitole 4.1 (okrem používania priamo R a Python) a 4.2 komplikuje vrstvu spoločného dátového katalógu a jednotného prístupu k dátam. To sa dá vyriešiť vytvorením si takejto vrstvy na mieru, čo nie je úplne triviálna úloha. Tento prístup môže byť doplnený o ekosystémy založené na platformách Snowflake, Databricks alebo Cloudera, ktoré vytvárajú tú najmodernejšiu verziu dátového skladu s podporou dátovej analýzy a business intelligence. Tieto ekosystémy sú **natívne pre prostredia verejných cloudov** a môžu v praxi rýchlo škálovať a priniesť okamžité hmatateľné výsledky pre analytické jednotky. Ďalšou obrovskou výhodou platforiem ako Snowflake a Databricks, ktoré sú momentálne titánmi na trhu[70], je, že dokážu podporovať analýzu dát pre podporu rozhodovania v takmer reálnom čase, čo otvára úplne nové príležitosti pre rozvoj KAV v budúcnosti. Na rozdiel od riešení ako Google Big Query, Amazon Redshit, Amazon EMR, ktoré sú v prvom rade úložiskami dát v cloude pripravenými pre analýzu, sa Snowflake a Databricks snažia vytvoriť celý dátový „stack" v cloude pre moderné organizácie. Snowflake sa pozicionuje ako „dátový cloud", kým Databricks ako „dátový lakehouse". Snowflake bol spočiatku známy ako dátový sklad pre biznis analytikov a dátových inžinierov. Databricks sa zas preslávil ako „data lake" pre dátových vedcov a inžinierov strojového učenia. Avšak tieto rozdiely sa začínajú zahmlievať. Snowflake rozšíril svoju ponuku o Snowpark pre dátových vedcov, podporu pre Python, Snowflake pre Apache Iceberg a podporu transakčných databáz, aby si získal náklonnosť vývojárov v open-source a dátových vedcov. Medzitým zas Databricks nedávno uviedol Databricks SQL, funkcionality Delta Lake a Unity katalóg metadát a dát, ktoré sú zamerané na zákazníkov orientovaných na tradičnejšie dátové úložiská, bezpečnosť a správu dát, čo je každodenným chlebíčkom Snowflake.

4.4.1 Analytické spracovanie veľkých údajov („big data") cez distribuované a paralelné spracovanie

Zatiaľ čo potreby v oblasti ukladania dát a výpočtovej techniky v posledných desaťročiach rástli míľovými krokmi, tradičný hardvér nedokázal držať krok. Podnikové údaje sa už nezmestia do štandardného úložiska a výpočtový výkon potrebný na zvládnutie väčšiny úloh analytického spracovania veľkých objemov údajov môže trvať týždne, mesiace alebo ho jednoducho nie je možné vykonať na štandardnom počítači. Na prekonanie tohto nedostatku sa vyvinuli mnohé nové technológie, ktoré zahŕňajú spoluprácu viacerých počítačov a distribuujú databázu na tisíce komoditných serverov. Keď je sieť počítačov prepojená a spolupracuje na plnení rovnakej úlohy, počítače tvoria klaster. Klaster si možno predstaviť ako jeden počítač, ktorý ale vďaka použitiu zosieťovaného komoditného hardvéru môže výrazne zlepšiť výkon, dostupnosť a škálovateľnosť v porovnaní s jedným výkonnejším strojom, a to pri nižších nákladoch. Apache Hadoop[71] je príkladom distribuovanej dátovej infraštruktúry, ktorá

[70] Zdroj: https://research.contrary.com/reports/databrick-vs-snowflake, Dátum referencie: 12.12.2022

[71] Zdroj: https://hadoop.apache.org, Dátum referencie: 30.05.2023

využíva klastre na ukladanie a spracovanie obrovského množstva údajov a ktorá umožňuje architektúru dátového jazera.

Hadoop sa skladá z troch hlavných komponentov:

1. Hadoop Distributed File System (HDFS) - ide spôsob ukladania a sledovania údajov na viacerých (distribuovaných) fyzických pevných diskoch;
2. MapReduce - je rámec na spracovanie údajov na distribuovaných procesoroch;
3. Yet Another Resource Negotiator (YARN) - je rámcom na správu klastra, ktorý riadi distribúciu takých vecí, ako je využitie procesora, pamäte a pridelenie šírky pásma siete na distribuovaných počítačoch.

Obzvlášť pozoruhodnou inováciou je vrstva spracovania Hadoopu: MapReduce je dvojstupňový výpočtový prístup na spracovanie veľkých (viacterabajtových alebo väčších) datasetov distribuovaných vo veľkých klastroch komoditného hardvéru spoľahlivým spôsobom odolným voči chybám. Prvým krokom je rozdelenie údajov medzi viacero počítačov („Map“), pričom každý z nich paralelne vykonáva výpočty na svojej časti údajov. Ďalším krokom je spojenie týchto výsledkov párovým spôsobom („Reduce“). Existuje mnoho open source implementácií tejto myšlienky. Hadoop je vytvorený na iteratívne výpočty, skenovanie obrovského množstva údajov v rámci jednej operácie z disku, distribúciu spracovania na viacero uzlov a ukladanie výsledkov späť na disk. Vyhľadávanie zettabajtov indexovaných údajov, ktoré by v tradičnom prostredí dátového skladu trvalo 4 hodiny, sa dá vykonať pomocou Hadoopu a HBase[72] za 10 až 12 sekúnd. Hadoop sa zvyčajne používa na vytváranie komplexných analytických modelov alebo aplikácií na ukladanie veľkého objemu údajov, ako sú retrospektívna a prediktívna analýza, strojové učenie a porovnávanie vzorov či segmentácia zákazníkov.

MapReduce však spracováva údaje v dávkach, a preto nie je vhodný na spracovanie údajov v reálnom čase. Apache Spark[73] bol vytvorený v roku 2012 s cieľom vyplniť túto medzeru. Spark je nástroj na paralelné spracovanie údajov, ktorý je optimalizovaný na rýchlosť a efektívnosť tým, že spracováva údaje v pamäti. Funguje na rovnakom princípe ako MapReduce, ale pracuje oveľa rýchlejšie, pretože väčšinu výpočtov vykonáva v pamäti a na disk zapisuje len vtedy, keď je pamäť plná alebo keď je výpočet dokončený. Tento výpočet v pamäti umožňuje Sparku spúšťať programy až 100x rýchlejšie ako Hadoop MapReduce v pamäti alebo 10x rýchlejšie na disku. Keď je však súbor údajov taký veľký, že sa nedostatok pamäte RAM stáva problémom (zvyčajne stovky gigabajtov alebo viac), Hadoop MapReduce môže prekonať Spark. Spark podporuje tieto programovacie jazyky na písanie algoritmov a programov: Python, Scala, Java, R alebo SQL. Spark má tiež rozsiahlu sadu knižníc na analýzu údajov, ktoré pokrývajú širokú škálu funkcií: Spark SQL na SQL a štruktúrované údaje; MLib na strojové učenie, Spark Streaming na spracovanie tokov údajov a GraphX na analýzu grafov. Keďže Spark sa zameriava na výpočty, nie je vybavený vlastným systémom ukladania dát a namiesto toho beží na rôznych systémoch ukladania dát, ako sú Amazon S3, Azure Storage a Hadoop HDFS.

Hadoop a Spark nie sú jediné technológie, ktoré využívajú klastre na spracovanie veľkých objemov údajov. Ďalší populárny výpočtový prístup k distribuovanému spracovaniu dopytov sa nazýva masívne paralelné spracovanie („Massively Parallel Processing (MPP)“). Podobne ako MapReduce, MPP

[72] Zdroj: https://hbase.apache.org, Dátum referencie: 30.05.2023

[73] Zdroj: https://spark.apache.org, Dátum referencie: 30.05.2023

rozdeľuje spracovanie údajov medzi viacero uzlov a uzly spracúvajú údaje paralelne, aby sa zvýšila ich rýchlosť. Na rozdiel od Hadoopu sa však MPP používa v RDBMS a využíva architektúru "zdieľania ničoho" - každý uzol spracúva svoj vlastný výsek údajov pomocou viacjadrových procesorov, vďaka čomu sú mnohonásobne rýchlejšie ako tradičné RDBMS. Niektoré databázy MPP, ako napríklad Pivotal Greenplum[74], majú vyspelé knižnice strojového učenia, ktoré umožňujú analýzu priamo v databáze. Podobne ako v prípade tradičných RDBMS však väčšina databáz MPP nepodporuje neštruktúrované údaje a aj štruktúrované údaje si budú vyžadovať určité spracovanie, aby vyhovovali infraštruktúre MPP; preto si nastavenie dátovej „pipeline“ pre databázu MPP vyžaduje ďalší čas a zdroje. Keďže databázy MPP sú kompatibilné s ACID a poskytujú oveľa vyššiu rýchlosť ako tradičné RDBMS, zvyčajne sa používajú v špičkových podnikových riešeniach na ukladanie údajov, ako sú Amazon Redshift, Pivotal Greenplum a Snowflake.

Výber medzi MPP a Hadoop už nie je pre KAV až také jednoznačné, ako ukazuje Tabuľka 7. Vzhľadom na pár vlastností Hadoop, ktoré sú v súlade s princípmi ďalšieho rozvoja dátovej a analytickej vrstvy (kapitola 5), ho odporúčame, keďže sa dá jednoduchšie využívať v praxi výberom nadstavbového riešenia od dodávateľa ako: Amazon Elastic MapReduce[75], Cloudera CDH Hadoop Distribution[76], Hortonworks Data Platform (HDP)[77], IBM Open Platform[78] a Microsoft Azure's HDInsight[79]. Samozrejme predpokladom je, že niektorý z ďalších aspektov Hadoop neblokujú úspešnú implementáciu daného prípadu použitia.

Tabuľka 7:

Aspekt porovnania	Masívne paralelné spracovanie	Hadoop
Otvorenosť platformy	Uzavretá a proprietárna. Pri niektorých technológiách nie je možné ani stiahnutie dokumentácie pre tých, čo nie sú zákazníci.	Úplne otvorený zdrojový kód so zdrojmi od dodávateľa aj komunity, ktoré sú voľne dostupné na internete.
Hardvérové požiadavky	Mnohé riešenia sú len pre zariadenia, softvér nemôžete nasadiť na vlastný klaster. Všetky riešenia vyžadujú špecifický hardvér podnikovej triedy, ako sú rýchle disky, servery s veľkým	Funguje akýkoľvek hardvér, niektoré usmernenia týkajúce sa konfigurácií poskytujú dodávatelia. Väčšinou sa odporúča používať lacný komoditný hardvér s priamo pripojeným

[74] Zdroj: https://datasense.be/technology/pivotal-greenplum/, Dátum referencie: 30.05.2023

[75] Zdroj: https://aws.amazon.com/emr/, Dátum referencie: 30.05.2023

[76] Zdroj: https://www.cloudera.com/products/open-source/apache-hadoop/key-cdh-components.html, Dátum referencie: 30.05.2023

[77] Zdroj: https://www.cloudera.com/content/dam/www/marketing/resources/datasheets/hdp-datasheet.pdf.landing.html, Dátum referencie: 30.05.2023

[78] Zdroj: https://www.ibm.com/docs/en/spectrum-scale-bda?topic=STXKQY_BDA_SHR/bl1adv_openplatform.html, Dátum referencie: 30.05.2023

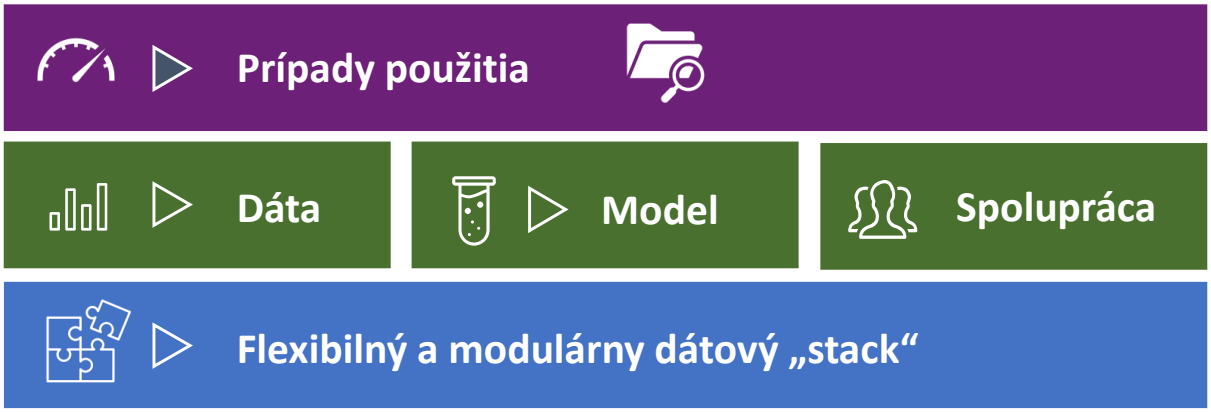
[79] Zdroj: https://azure.microsoft.com/en-us/products/hdinsight, Dátum referencie: 30.05.2023

Aspekt porovnania	Masívne paralelné spracovanie	Hadoop
	množstvom pamäte ECC RAM, 10GbE/Infiniband atď.	úložiskom („Direct-attached storage (DAS)“).
Škálovateľnosť (na uzly)	V priemere desiatky uzlov, 100-200 je maximum.	100 uzlov v priemere, počet tisíc je maximum.
Škálovateľnosť (používateľské údaje)	Desiatky terabajtov v priemere, petabajt je maximum.	Stovky terabajtov v priemere, desiatky petabajtov je maximum.
Oneskorenie dotazu	10-20 milisekúnd	10-20 sekúnd
Priemerný čas spustenia dotazu	5-7 sekúnd	10-15 minút
Maximálny čas vykonávania dotazu	1-2 hodiny	1-2 týždne
Optimalizácia dotazov	Komplexné podnikové „enginy“ na optimalizáciu dotazov uchovávané ako jedno z najcennejších podnikových tajomstiev.	Žiadna optimalizácia alebo len s naozaj obmedzenou funkčnosťou, niekedy ani nie na základe nákladov.
Cena technológie	Desiatky až stovky tisíc dolárov za uzol	Zadarmo alebo až tisíce dolárov za uzol
Prístupnosť pre koncových používateľov	Jednoduché prívetivé rozhranie SQL a jednoduché interpretovateľné funkcie v databáze	SQL nie je úplne v súlade s ANSI, používateľ by sa mal starať o logiku vykonávania, základné rozloženie údajov. Funkcie sa zvyčajne musia napísať v jazyku Java, skompilovať a umiestniť na klaster.
Cieľová skupina koncových používateľov	Biznis analytici a databázoví administrátori s bežnými skúsenosťami	Vývojári v jazyku Java a skúsení databázoví administrátori
Redundancia jednej úlohy	Nízka, úloha zlyhá pri zlyhaní uzla MPP, ktorý ju vykonáva.	Vysoká, úloha zlyhá len vtedy, ak zlyhá uzol, ktorý riadi vykonávanie úlohy.
Cieľové systémy	Všeobecné dátové sklady a analytické systémy.	Účelové „enginy“ na spracovanie údajov.
Vendor Lock-in	Typický prípad.	Zriedkavý prípad zvyčajne spôsobený nesprávnym použitím technológie.

Aspekt porovnania	Masívne paralelné spracovanie	Hadoop
Maximálna súbežnosť	Desiatky až stovky dopytov.	Až 10-20 úloh.
Zložitosť implementácie riešení	Stredná	Vysoká

5 Princípy rozvoja KAV a spolupráce s OVM a tretími stranami

Modernizácia dátovej a analytickej vrstvy je odpoveďou na splnenie dnešnej potreby organizácií konať rýchlo, pracovať na flexibilnosti a prinášať inovácie. Modernizácia dátovej vrstvy musí mať za cieľ zvýšiť agilitu pri implementovaní prípadov použitia pre analytické spracovanie údajov, zlepšiť používateľskú prívetivosť, ponúkať rôzne možnosti a nástroje pre rôzne role používateľov a rýchlo škálovať. Pritom sa musia dodržiavať osvedčené postupy, ktoré sú rozhodujúce pre maximalizáciu prínosov modernizácie dátovej vrstvy. Táto modernizácia KAV stojí na piatich stavebných kameňoch, ako ukazuje obrázok.

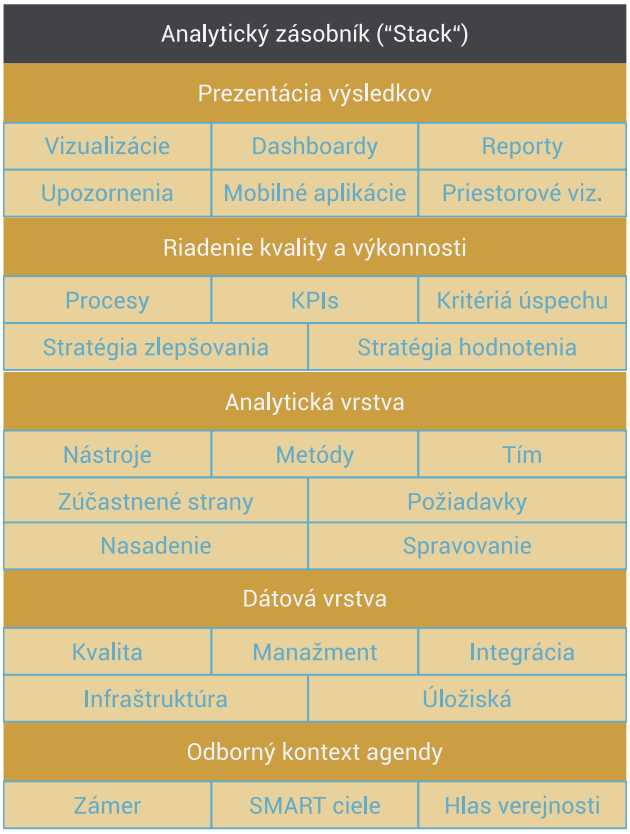


Obrázok 5: Kamene, na ktorých stojí rozvoj KAV

Modernizácia dátovej a analytickej vrstvy umožňuje verejnej správe naplno využiť hodnotu svojich jedinečných dátových aktív, rýchlejšie vytvárať poznatky prostredníctvom dátového inžinierstva založeného na umelej inteligencii a dokonca vyťažiť hodnotu aj zo starších údajov. Moderná dátová architektúra umožňuje, aby sa údaje verejnej správy stali škálovateľnými, dostupnými, spravovateľnými a analyzovateľnými pomocou cloudových služieb. Okrem toho zabezpečuje súlad s legislatívou o bezpečnosti a ochrane osobných údajov a zároveň umožňuje prístup k údajom v rámci celej verejnej správy. Pomocou moderného prístupu k údajom môžu organizácie verejnej správy poskytovať lepšie zákaznícke skúsenosti, lepšie sa rozhodovať, znižovať náklady a zefektívňovať svoj chod.

Základné vrstvy celkového systému analytického spracovania údajov sú nasledovné, ako znázorňuje aj Obrázok 6:

- Prezentácia výsledkov,
- Riadenie kvality a výkonnosti,
- **Analytická vrstva,**
- **Dátová vrstva,**
- **Odborný kontext agendy.**



Obrázok 6: Vrstvy systému analytického spracovania údajov

Dátová vrstva predpokladá, že sú v organizácii zavedené postupy efektívneho manažmentu dát a implementované patričné technológie, ktoré dokážu sprostredkovať údaje v požadovanom formáte a kvalite analytickej vrstve. V tomto dokumente sa venujeme predovšetkým analytickej vrstve, na ktorej sa vytvárajú analytické modely a modely rozhodovania, ale aj dátovej vrstve, na ktorej prebieha meranie a zber dát, a odbornému kontextu dát, ktorý napomáha porozumieť dátam a správne ich interpretovať. Prístup je založený na dátovej vede a postupoch pre vyvíjanie algoritmov umelej inteligencie. Základné tri princípy, ktoré treba mať na pamäti pri budovaní analytickej a dátovej vrstvy na základe ponaučení z implementácie a dobrej praxe, sú nasledovné:

1. Budovať dôveryhodný a modulárny dátový „stack", založený predovšetkým na otvorenom zdrojovom kóde a s minimalizáciou takzvaného „vendor-lockinu",
2. Platiť len za skutočné využívanie,
3. Posúvať sa smerom vpred s uchopiteľnými prípadmi použitia pri využití existujúcich zručností v tímoch.

Ďalej v kapitole uvádzame deväť osvedčených postupov na základe uvedených princípov, ktoré je potrebné dodržiavať pri modernizácii dátovej a analytickej vrstvy pre KAV.

5.1 Vytváranie flexibilných, rozšíriteľných dátových schém

Organizácie získavajú silnú konkurenčnú výhodu tým, že zvyšujú svoju schopnosť skúmať údaje a využívať pokročilú analytiku. Na dosiahnutie tohto cieľa prechádzajú na denormalizované, meniteľné

dátové schémy s menším počtom fyzických tabuliek na organizáciu údajov s cieľom maximalizovať výkon. Používanie flexibilných a rozšíriteľných dátových modelov namiesto rigidných umožňuje rýchlejší prieskum štruktúrovaných a neštruktúrovaných údajov. Znižuje tiež zložitosť, pretože dátoví architekti a vývojári nemusia na vyhľadávanie relačných údajov vkladať abstrakčné vrstvy, ako sú dodatočné prepojenia („joins“) medzi vysoko normalizovanými tabuľkami.

Dátové modely sa môžu stať rozšíriteľnými pomocou techniky dátového modelovania „Data Vault 2.0“ (kapitola 3.3), čo je normatívna metóda transformácie surových údajov na inteligentné, využiteľné poznatky. Aj grafové databázy NoSQL využívajú neštruktúrované údaje a umožňujú aplikácie vyžadujúce masívnu škálovateľnosť, možnosti práce v reálnom čase a prístup k dátovým vrstvám v systémoch umelej inteligencie. Organizácie môžu ukladať údaje aj pomocou JavaScript Object Notation (JSON), čo umožňuje zmenu štruktúry databázy bez toho, aby to ovplyvnilo Centrálny model údajov.

5.2 Zameranie sa na architektúru založenú na doméne, ktorá je v súlade s biznis požiadavkami

Dátoví architekti sa odkláňajú od klastrov centralizovaných podnikových dátových jazier k architektúram založeným na doméne (takzvané dátové siete („data mesh“), viac v dokumente 1.1.6 Štandardizácia dátovej transformácie). V rámci nich sa v podnikoch používajú techniky virtualizácie údajov na organizáciu a integráciu distribuovaných dátových aktív. Prístup založený na doméne bol nápomocný pri plnení špecifických biznis požiadaviek s cieľom urýchliť čas uvedenia nových dátových produktov a služieb na trh. Pre každú doménu môže vlastník produktu a produktový tím udržiavať dátový katalóg s možnosťou vyhľadávania spolu s poskytovaním dokumentácie (definície, koncové body API, schémy a ďalšie) a iných metadát konzumentom. Doména ako ohraničený kontext tiež umožňuje používateľom získať dátový plán, ktorý zahŕňa údaje, integráciu, ukladanie a architektonické zmeny.

Tento prístup výrazne skracuje čas strávený budovaním nových dátových modelov v dátovom jazere, zvyčajne z mesiacov na dni. Namiesto vytvorenia centralizovanej dátovej platformy sa môžu nasadiť logické platformy pre jednotlivé agendy alebo skupiny prípadov použitia, ktoré sú spravované v rámci rôznych oddelení alebo rôznych OVM. V prípade architektúry orientovanej na doménu sa využíva prístup „dátovej infraštruktúry ako platformy“ („data infrastructure as a platform“) so štandardizovanými nástrojmi na údržbu dátových aktív, ktoré urýchľujú implementáciu. Z „data mesh“ sa údaje dajú presúvať (streamovať) do dátových skladov, dátových jazier aj dátových lakehouses[80].

5.3 Odstránenie dátových síl v rámci verejnej správy

Organizačné silá sťažujú verejnej správe riadenie procesov a rozhodovanie na základe presných informácií. Odstránenie organizačných síl umožní prijímať informovanejšie rozhodnutia a efektívnejšie využívať údaje. Je zrejmé, že spoľahlivá architektúra dátovej a analytickej vrstvy musí odstrániť silá vykonaním auditu informačných systémov, legislatívy, kultúry a cieľov.

Dôležitou súčasťou modernizácie dátovej vrstvy je sprístupnenie interných údajov ľuďom, ktorí ich potrebujú, keď ich potrebujú. Keď sa v rôznych úložiskách nachádzajú rovnaké údaje, vytvorené

[80] Pozri tu: https://www.confluent.io/events/current-2022/streaming-data-into-your-lakehouse/ alebo tu: https://www.kai-waehner.de/blog/2022/07/05/data-streaming-for-data-ingestion-into-data-warehouse-and-data-lake/ , Dátum referencie: 30.05.2023

duplicity údajov takmer znemožňujú určiť, ktoré údaje sú relevantné. V modernej dátovej a analytickej vrstve sa odbúravajú dátové silá a informácie sa čistia a overujú, aby sa zabezpečila ich presnosť a úplnosť. OVM musia v podstate prijať kompletný a centralizovaný „master data manažment" (v prostredí slovenského eGovernmentu ide o dôslednú implementáciu referenčných registrov, viac v dokumente 2.1.1 Popis štandardizácie dátových prvkov a entít) a Centrálnu integračnú platformu, aby automatizovali správu všetkých informácií z rôznych kanálov na jednom mieste a umožnili dlhodobé odstránenie dátových síl.

5.4 Oddelenie prístupových bodov k údajom

Údaje dnes už nie sú obmedzené na štruktúrované údaje, ktoré možno analyzovať pomocou tradičných nástrojov. V dôsledku veľkých objemov údajov a cloud computingu je k dispozícii obrovské množstvo štruktúrovaných a neštruktúrovaných údajov, ktoré obsahujú dôležité informácie pre organizácie, často z rôznych dôvodov ťažko prístupné. Z toho vyplýva, že dátová vrstva by mala byť schopná spracovať údaje zo štruktúrovaných aj neštruktúrovaných zdrojov, a to v štruktúrovanej aj neštruktúrovanej podobe. Ak to OVM neurobia, prídu o dôležité informácie potrebné na prijímanie informovaných rozhodnutí.

Údaje možno vystaviť prostredníctvom rozhraní API, aby sa obmedzil a ochránil priamy prístup k zobrazovaniu a úprave údajov a zároveň sa umožnil rýchlejší a aktuálnejší prístup k štandardným datasetom. Údaje sa dajú ľahko opakovane používať medzi tímami, čo urýchľuje prístup k nim a umožňuje bezproblémovú spoluprácu medzi analytickými tímami. Týmto spôsobom možno efektívnejšie implementovať prípady použitia analytického spracovania údajov.

5.5 Zváženie verejných cloudových platforiem nielen pre budovanie KAV

Cloud computing je pravdepodobne najvýznamnejšou hnacou silou nového revolučného prístupu k dátovej vrstve na rýchle škálovanie možností a nástrojov pre analytické spracovanie údajov. Klesajúce náklady na cloud computing a nárast dátových nástrojov priamo v pamäti umožňujú organizáciám využívať najsofistikovanejšiu pokročilú analytiku. Poskytovatelia cloudových služieb prinášajú revolúciu v tom, ako organizácie všetkých veľkostí získavajú, nasadzujú a prevádzkujú dátovú infraštruktúru, platformy a aplikácie vo veľkej škále. **Nasadenie KAV do vybraného prostredia verejného cloudu je preto nevyhnutnou podmienkou ďalšieho rozvoja**. S cloudovou integračnou platformou alebo master data manažmentom môžu organizácie využívať výhody hotových a nakonfigurovaných riešení, v rámci ktorých môžu bezproblémovo nahrávať svoje údaje, automatizovať tvorbu katalógov a obohacovať ich.

Cloudové platformy a služby vo verejných cloudoch eliminujú potrebu údržby hardvéru, hostingu aplikácií, aktualizácií verzií a bezpečnostných záplat. Náklady možno vo veľkej miere optimalizovať vypínaním nevyužitých klastrov a platením len za skutočne využívané služby na mesačnej báze. Okrem toho dátové platformy založené na cloude prinášajú aj vyššiu úroveň kontroly nad dátovými produktmi a bezpečnosťou.

5.6 Integrácia modulárnych, najlepších a ideálne otvorených platforiem

Mnohé organizácie prechádzajú na **modulárne dátové architektúry, ktoré využívajú komponenty s najlepšou dostupnosťou a často aj s otvoreným zdrojovým kódom („open source"), ktoré možno podľa potreby vymeniť za nové technológie bez toho, aby to ovplyvnilo ostatné časti architektúry**. Organizácia využívajúca tento prístup môže rýchlo poskytovať nové, dátovo náročné digitálne služby

miliónom zákazníkov a pripojiť sa ku cloudovým aplikáciám na veľkej škále. Organizácie môžu tiež vytvoriť nezávislú dátovú vrstvu, ktorá zahŕňa komerčné databázy a komponenty s otvoreným zdrojovým kódom.

Tie najužitočnejšie „stacky" umožňujú organizáciám užívať výhody rýchlo sa vyvíjajúcich technológií a pridávať najlepšie riešenia na trhu pre zber, transformáciu a analýzu dát bez toho, aby na to bolo nevyhnutné alokovať veľa času a financií. Budovanie takéhoto „stacku" však v mnohých organizáciách narazí na bariéry vytvorené monolitickými, zastaranými aplikáciami a nemodernou IT architektúrou.

V raných štádiách nastavovania analytického spracovania údajov možno takýto dátový „stack" vybudovať vyskladaním open source a freemium nástrojov. Ako sa však začne analyzovať stále viac historických údajov v rôznych agendách, bude potrebné riešiť viac špecifické problémy a nuansy pri tvorbe podkladov pre rozhodovanie, ako aj spracovávať stále väčšie objemy dát v rôznorodých formátoch. Aby dátami riadené rozhodovanie ostalo konkurencieschopné a prinášalo hodnotu za peniaze, musia sa dátové tímy naučiť „stack" modernizovať tak, aby sa nemíňali veľké objemy financií na náročné implementácie a aby sa nestávali obeťou takzvaného „vendor lock-inu" (ide o situáciu, kedy dokáže IT architektúru upravovať a aktualizovať len jeden dodávateľ pôvodného riešenia). Vymeniť dodávateľa je bolestivé, pokiaľ jedno riešenie dominuje v dátovej infraštruktúre organizácie. Je spojené s migráciou obrovského množstva dát do nového systému, pričom nemožno dopredu zaručiť, že nové riešenie bude spĺňať všetky požiadavky. Preto je výmena dodávateľa spravidla označovaná ako riskantná. Monolitické riešenia často vyžadujú vysokú prvotnú investíciu z pohľadu času a financií, čo prispieva k mentalite utopených nákladov a odmietaniu implementácie nového, vhodnejšieho riešenia. Preto modulárne riešenia zásadne uľahčujú rozhodovanie o aktualizácií nástrojov na prácu s dátami a výrazne zjednodušujú proces nákupu.

Monolitické riešenia tiež vyžadujú dôsledné procesy správy a dohľadu, ktoré štandardizujú dátové formáty, preverujú duplikované alebo nepresné údaje, manažujú dodržiavanie súladu s metodikami a požiadavkami na bezpečnosť a poskytujú rámec pre celkovú dátovú politiku organizácie. To predstavuje problém pri výmene dodávateľa, ak procesy správy a dohľadu sú implementované v rámci jedného nástroja business intelligence (BI), kvôli čomu by bolo potrebné s novým dodávateľom začať v bode nula.

Údaje sa v dátovom „stacku" budú synchronizovať s informačnými systémami prostredníctvom centrálnej integračnej platformy a biznis logiku budú spracovávať mikroslužby, ktoré sa budu nachádzať v kontajneroch. Okrem zjednodušenia integrácie medzi rôznorodými nástrojmi a platformami rozhrania založené na API znižujú riziko zavádzania nových problémov do existujúcich aplikácií a urýchľujú čas uvedenia do produkcie. Uľahčujú aj výmenu jednotlivých komponentov.

5.7 Kolaboratívne vysoko výkonné nástroje pre analytikov a doménových expertov

Takéto nástroje umožňujú analytikom rýchlo si medzi sebou, ale aj medzi ostatnými členmi tímu, zdieľať získané poznatky a reporty. Samotné historické údaje a ich analytické spracovanie nedokážu mať vplyv na zlepšenie chodu organizácie, pokiaľ sa nestanú integrálnou súčasťou odbornej práce a rozhodovacích procesov zamestnancov.

V priemere 60 až 73 percent všetkých údajov organizácie sa nevyužíva na žiadnu analýzu. Dôvodov tohto zlyhania vo využívaní údajov je mnoho, ale najzásadnejším je ten, že tímy v rámci organizácie si ani nie sú isté, ktoré dáta sú relevantné pre riešenie danej otázky alebo problému. Ak postupy a nástroje podporujú spoluprácu medzi zamestnancami na rôznych oddeleniach, ktorí rozumejú kontextu údajov

a informáciám v nich obsiahnutých, začne sa využívať v praxi vyššie percento objemu údajov. Odborníci na danú agendu potrebujú znalosti, aby mohli viesť svoje tímy, a spravidla potrebujú odpovede založené na hĺbkovej analýze daného problému. Ich každodenná skúsenosť s danou agendou im umožňuje pýtať sa veľmi cenné doplňujúce otázky, ktoré by sa ľudia bez takéhoto úzkeho kontaktu s problematikou nevedeli pýtať. Ak dátový analytický tím a odborníci na danú agendu spolu úzko spolupracujú, dokážu ďalej stavať na svojej expertíze a posúvať vpred znalosti o tom, ako čo najlepšie využiť analytické spracovanie historických údajov na zlepšenie danej agendy. V praxi to znamená, že odborníci na danú agendu bez technologických znalostí by mali byť schopní využívať kolaboratívne analytické nástroje tak, aby sa mohli:

- podieľať na procese dátového modelovania,
- využívať výhody prednastavených šablón na dátovú analýzu,
- robiť dopyty nad dátovým lakehousom,
- vytvárať dátové vizualizácie,

a to bez znalosti programovania alebo nutnosti spoliehať sa na dátový tím, aby im vytvorili prispôsobené dashboardy alebo report. Dátoví analytici analyzujú dáta, interpretujú ich a robia na ich základe predikcie, avšak nie sú blízko situáciám v praxi, v ktorých dáta vznikajú. Spolupráca dátových analytikov a odborníkov na danú agendu na otázkach „prečo?“ vie predísť nákladným pochybeniam, kedy dátoví analytici bez tejto spolupráce môžu urobiť nesprávne predpoklady.

Získavanie nových poznatkov z dát je „nákazlivé“. Ak sa člen tímu dozvie niečo z dát, čo ovplyvní jeho pracovný úspech zásadným spôsobom, bude okamžite zvedavý, aké ďalšie poznatky sú v dátach, ktoré čakajú na to, aby boli objavené. Ak členovia tímu naprieč organizáciou alebo rezortom dostanú nástroje na to, aby participovali na analytickom procese, sú motivovaní zdieľať svoje vstupy a hľadať v dátach spoločne ďalšie poznatky. Zvedavosť podporuje kultúru inovácie a nových objavov, ktoré sú na nezaplatenie pre každú organizáciu.

5.8 Zhodnotenie zručností

Pri implementovaní dátového programu je dôležité mať po ruke správne zručnosti v tímoch. Kľúčovou výhodou architektúry založenej na open source je, že uľahčuje prístup k zručnostiam, keďže organizácie môžu efektívne načrieť do širšieho okruhu ľudí s potrebnými zručnosťami. Je tiež dôležité, aby sa organizácie snažili rozvíjať správne zručnosti u svojich interných zamestnancov. Mnohé organizácie, ako bolo zrejmé aj z prieskumu medzi analytickými jednotkami v kapitole 4.1, majú dedičstvo analytických a štatistických zručností, ktoré je potrebné premostiť do moderných analytických a dátových vedeckých pozícií. V tejto oblasti je kľúčová úloha vzdelávania a rozvoja a vedúci zamestnanci sa musia zapojiť do spoločného úsilia, aby zabezpečili, že jednotlivci budú mať k dispozícii správne školenia, ktoré im umožnia skutočne využívať dostupné údaje. Nakoniec sa nesmie zabúdať na silu komunity, ktorá podporuje vzájomné vzdelávanie - práve vďaka tomu sa open source stal takou hybnou silou v oblasti technológií. Na internete je voľne dostupné (alebo za malý poplatok) veľké množstvo školiacich materiáloch – od dokumentácie, návodov, cez blogy, tipy, diskusné fóra, ukážky, projekty, nadstavby až po online prednášky a vzdelávacie kurzy. Konkrétne sú dané nasledovné očakávania pre definované role:

- Administrátori databáz a analytických nástrojov sa musia naučiť spravovať dátovú a analytickú vrstvu v cloude v dátovom lakehouse,

- Dátoví architekti sa budú musieť naučiť fungovať s dynamickými schémami údajov v „data mesh“ a dátovým modelovaním „Data Vault 2.0“,
- Biznis analytici budú musieť porozumieť novým možnostiam, ktoré prinášajú nové analytické nástroje s veľkým objemom údajov ako aj metodike dátového modelovania „Data Vault 2.0“,,
- Vývojári, či už interní alebo externí, sa budú musieť naučiť vytvárať dátové „pipelines“ na mieru v novom prostredí dátového lakehouse,
- Koncovi používatelia budú musieť vedieť využiť pridanú hodnotu interaktívnych vizualizačných a analytických nástrojov (ktoré nepotrebujú znalosť programovacieho jazyka), napríklad vo forme interaktívnych dashboardov pre zlepšovanie prevádzky alebo lepšie strategické rozhodovanie,
- Analytické jednotky/ rezorty budú musieť vyškoliť svojich kľúčových zamestnancov pre analytické spracovanie údajov na dátových vedcov, aby vedeli prioritne používať otvorené jazyky ako R a Python a rozumeli procesu dátovej vedy (viac o tomto procese sa nachádza v dokumente 4.1.1 Príručka regulácie 2.0), prípadne aj strojovému učeniu a MLOps.

Správne zručnosti umožnia vedúcim zamestnancom v oblasti dátového programu presadzovať silnú kultúru údajov v rámci celej verejnej správy - vytvárať kolektívne porozumenie o hodnote dátového programu, ktorú môže priniesť, a o tom, čo treba na získanie tejto hodnoty spraviť. Hlavným cieľom dátového programu musí byť, aby OVM mali svojich lídrov v tejto oblasti a aby pod týmto vedením spolupracovali, pokiaľ ide o údaje.

5.9 Budovanie dátovej a analytickej vrstvy s ohľadom na ochranu súkromia („privacy by design“) a bezpečnosť údajov

Ďalšou výhodou architektúry založenej na dátovom lakehouse je to, že ponúka platformy pre implementovanie jednotného modelu spravovania („governance“) a bezpečnosti pre všetky dáta. Čím sa dá jednoduchšie a spoľahlivejšie s dátami pracovať, tým viac budú organizácie otvorené zdieľať dáta a implementovať nové prípady použitia pre analytické spracovanie údajov.

Z pohľadu ochrany súkromia a bezpečnosti údajov treba dodržiavať nasledujúcich 6 princípov, pričom viac o tejto téme sa nachádza v dokumentoch 1.1.3 Štandardizácia pre bezpečnosť a ochranu údajov a 1.1.5 Štandardizácia anonymizácie údajov. Konkrétne pretavenie týchto princípov do praxe závisí od výberu konkrétneho poskytovateľa cloudu a platformy pre dátový lakehouse, ako aj od ďalších služieb, nástrojov a platforiem použitých pre dátovú a analytickú vrstvu.

1. Správa identít a prístupu na základe najnižších oprávnení

Správa identít a prístupu („Identity and Access Management(IAM)“) je základom bezpečnosti údajov. Stará sa o to, aby správni ľudia mali prístup k správnym zdrojom. IAM sa zaoberá týmito aspektmi autentifikácie a autorizácie: správa účtov, správa identít, autentifikácia, riadenie prístupu (autorizácia) a federácia identít. Najvyššie oprávnenia prístupu do databáz má len zopár preverených a poverených administrátorov. Ideálne je tiež implementovať logické platformy po doménach, ako je uvedené v kapitole 5.2, čo umožňuje automaticky obmedzovať prístup k dátovej a analytickej vrstve pre vývojárov, biznis analytikov, dátových vedcov a koncových používateľov v danej doméne, keďže neexistuje jednotný prístup k centrálnej platforme. V každom prípade by žiadna z týchto rolí nemala mať oprávnenie meniť údaje v databázach, alebo ich z nich v celom rozsahu prenášať do iného cieľa. Tvorba dátových „pipelines“ s procedúrami pre ETL musí byť pod dohľadom viacerých očí, aby bola zachovaná dôveryhodnosť a integrita údajov.

2. **Ochrana údajov pri prenose a v pokoji („at rest“)**

Údaje treba katalogizovať a priradiť im úroveň citlivosti s mechanizmom riadenia prístupu podľa princípu 1. Údaje pri prenose musia byť šifrované.

3. **Zabezpečenie siete a ochrana koncových bodov**

Zabezpečte svoju sieť a monitorujte a chráňte integritu siete interných a externých koncových bodov prostredníctvom bezpečnostných zariadení alebo cloudových služieb, ako sú firewally.

4. **Nastavenie modelu zdieľanej zodpovednosti**

Bezpečnosť a dodržiavanie predpisov sú spoločnou zodpovednosťou poskytovateľa platforiem a služieb, verejnej správy ako zákazníka a poskytovateľa cloudu. Je dôležité pochopiť, ktorá strana je zodpovedná za ktorú časť, a ako

5. **Splnenie požiadaviek na súlad a ochranu osobných údajov**

Na splnenie regulačných požiadaviek je potrebné zakryť alebo upraviť informácie umožňujúce identifikáciu osôb. Údaje do KAV už prúdia pseudonymizované tak, aby sa dali dopĺňať informácie z ďalších zdrojov o danej dotknutej osobe alebo subjekte. Ďalšie citlivé údaje na úrovni „Vyhradené“ alebo „Dôverné“ musia byť anonymizované, napríklad vek sa musí generalizovať do vekových skupín, napríklad po 5 rokoch, adresa trvalého pobytu sa musí odseknúť na poštové smerovacie číslo (ak tento postup anonymizácie nevyhovuje danému prípadu použitia, musí sa splniť požadovaná hodnota k-anonymity, viac v dokumente .1.5 Štandardizácia anonymizácie údajov). Ak je to možné, úsilie o dodržiavanie predpisov treba automatizovať.

6. **Monitorovanie zabezpečenia systému**

Na monitorovanie aplikácií, platformy a infraštruktúry treba používať automatizované nástroje. Na skenovanie infraštruktúry na zraniteľnosti a odhaľovanie bezpečnostných incidentov sa využíva automatizované skenovanie v rámci „pipelines“ nepretržitej integrácie a nepretržitého nasadzovania („Continuous Integration (CI)/Continuous Deployment (CD)“).

Všetky zmeny v databázach ako aj zdieľania údajov musia byť tiež monitorované a logované s časovou pečiatkou a identifikátorom osoby, ktorá k databáze pristupovala.

6 Plánovanie zavádzania využívania údajov pre analytické spracovanie

Ďalšie plánovanie v oblasti využívania údajov pre analytické spracovanie vychádza zo zhodnotenia prebiehajúcich a ukončených projektov z dopytových výziev pre lepšie využívanie údajov, ako aj z vyhodnotenia ponaučení všetkých doterajších aktivít v tomto smere.

6.1 Zhodnotenie projektov z výzvy pre zlepšenie využívania údajov verejnej správy

Z celkového počtu 88 dopytových projektov, podaných v rámci dopytových výziev dátového programu, bolo len 11 dopytových projektoch zameraných na lepšie využívanie údajov v rámci výzvy pre zlepšenie využívania údajov verejnej správy:

1. **Projekt_588**: Riadenie IT aktív vo verejnej správe, MIRRI SR.
2. **Projekt_654**: Lepšie využívanie údajov pre kontrolu výskytu a šírenia baktérií - Antimikrobiálna rezistencia v SR (AMR), Národné centrum zdravotníckych informácií,
3. **Projekt_680:** Rozvoj rizikovej analýzy a implementácia antifraudového systému v agendových IS PPA, PPA
4. **Projekt_683:** Dynamický cenový model, Štatistický úrad,
5. **Projekt_684:** Socioekonomické aspekty Big Data**,** Štatistický úrad,
6. **Projekt_687:** Zvýšenie kvality služieb Finančnej správy pre občanov a podnikateľov,
7. **Projekt_691**: Analytické centrum dopravných informácií, Žilinská univerzita v Žiline,
8. **Projekt_697**: Zvýšenie efektívnosti využívania energetických údajov za účelom úspory financií v budovách verejnej správy, Technická univerzita v Košiciach,
9. **Projekt_722**: Dynamický architektonický model governmentu, MIRRI SR
10. **Projekt_745**: Prediktívne modely preventívnych programov: Úrad verejného zdravotníctva SR,
11. **Projekt_746**: Zavedenie analytickej vrstvy pre lepšie využívanie údajov VšZP, Všeobecná zdravotná poisťovňa

Tabuľka 8: Zhrnutie dopytových projektoch zameraných na lepšie využívanie údajov

ID projektu	Zámer	Potrebné údaje	Výsledok
Projekt_588	**Optimalizácia prevádzky organizácií verejnej správy:** Zefektívňovanie riadenia celého životného cyklu IT aktív, optimalizácia nákladov na hardvér a softvér	Údaje poskytnuté infraštruktúrou pripojených OVM, údaje z obdržaných z licenčných zmlúv, nákupných objednávok a pod.	Výsledky realizácie projektu poskytnú kľúčové informácie k určeniu čo najoptimálnejšej východiskovej pozície, definovaniu skutočných potrieb OVM týkajúcich sa IT aktív a poskytnú podporné informácie pre

ID projektu	Zámer	Potrebné údaje	Výsledok
			rozhodovanie o obstarávaní nových IT aktív.
Projekt_654	**Optimalizácia prevádzky organizácií verejnej správy:** Nové procesy podporené kvalitnými a aktuálnymi údajmi, ktoré budú viesť k optimalizácií liečebných a epidemiologických postupov - dopad na efektívne predpisovanie antibiotickej liečby.	Údaje z laboratórnych vyšetrení, ktoré sú upravené ustanovením § 5 a 6 zákona č. 153/2013 Z. z. o národnom zdravotníckom informačnom systéme, ďalšie údaje identifikované v rámci projektu. Vybuduje sa 7 registrov (Register baktérií, Medzinárodná klasifikácia chorôb, Register subjektov AMR, Register epidemiologických a liečebných postupov, Register šírenia infekcií, Register hlásení, Register laboratórnych vyšetrení LOINC).	Systém bude na základe analýzy dát zachytávať pozitívne výsledky a bude určovať podmienky pre generovanie alertu. Bude možné vyhodnotiť jednotlivé liečebné a epidemiologické postupy a na základe skúsenosti ich vylepšovať.
Projekt_680	**Spojenie úradníka a stroja: inovácie procesov:** Existujúca riziková analýza nepokrýva všetky typy podpôr. Zámerom je vytvoriť modely a postupy pre jednotlivé procesy kontrol a využiť dostupné údaje pre automatizáciu procesu a vyššiu efektivitu pri odhaľovaní podvodov a plánovaní kontrol.	Údaje v elektronickej podobe zo samotných žiadostí, dáta o žiadateľoch ako aj dáta verejnej správy pre overovanie dát v súlade s princípom „jedenkrát a dosť“.	Využitie týchto dát ako aj rozšírenie funkcionality zvýši presnosť detekčných mechanizmov podvodov a rozšírenie existujúcich procesov rizikovej analýzy aj na ďalšie typy podpôr. Vytvorenie registra rizikových subjektov.
Projekt_683	**Zvýšenie výkonnosti vnútorných procesov:** Využívanie alternatívnych metód zberu údajov pre včasnejšiu štatistickú produkciu v oblasti	Potrebné údaje, ich formáty a možnosti získavania, ako je napr. webscraping, scaner data, nákup údajov a pod., budú zanalyzované. Predbežne sa javia užitočné údaje od	Hlavnými výsledkami bude: – Aktuálnejšie údaje v čase, – Detailné trendy,

ID projektu	Zámer	Potrebné údaje	Výsledok
	Spotrebiteľských cien a cien produkčných štatistík.	Heureka[81], či už nákupom alebo cez webscraping.	– Zníženie náročnosti na zber štandardným spôsobom, – Presnejšie výpočty inflácie. **Výsledky projektu sa podarí realizovať.**
Projekt_684	**Lepší návrh politík a regulácií:** Využitie moderných nástrojov na spracovanie veľkého objemu údajov (Big Data) a definovanie korelácií medzi správaním sa obyvateľstva inými socioekonomickými aspektami. Závery projektu môžu byť využité na zlepšenie rozhodovanie ako na národnej tak aj na regionálnej úrovni.	Údaje o pohybe obyvateľstva od telekomunikačných operátorov, údaje o príspevkoch na sociálnych sieťach.	Možnosť posudzovať dochádzanie obyvateľstva za prácou a zaťaženosti jednotlivých miest. Včasnejší odhad vývoja HDP s nízkou administratívnou záťažou. Sledovanie sociálneho napätia v spoločnosti s možnosťou identifikovať pálčivé témy, ktoré treba adresovať zmenou politiky a regulácie. **Výsledky projektu sa podarí realizovať.**
Projekt_687	**Zvýšenie kvality služieb:** Využitie analytických metód, analytických údajov a aplikovanie algoritmov strojového učenia na dosiahnutie zníženia administratívnej záťaže Call centra.	Zanalyzujú sa možnosti zabezpečenia zdrojov dát z vlastných zdrojov, voľne dostupných zdrojov ako aj prípadne ďalších platených no zdrojov. Základom sú údaje o využívaní portálu Finančnej správy a z kontaktov Call centra (Navštívené informačné obsahy v čase, čas zotrvania na informačnom obsahu, počet použití, čo používateľ hľadal, aké	Možnosť identifikovať. v akých prípadoch je používateľ nútený využiť služby Call centra a priniesť tak používateľom cielené informačné služby a vyhľadávanie informácií, aby tieto informácie našiel alebo ich našiel rýchlejšie a zbytočne nepoužíval služby Call centra.

[81] Zdroj: https://heureka.group/sk-sk/datove-prehlady/, Dátum referencie: 30.05.2023

ID projektu	Zámer	Potrebné údaje	Výsledok
		výsledky mu zobrazilo, na aký výsledok klikol, ako dlho zotrval na zvolenej stránke, aké ďalšie výsledky zvolil, aké ďalšie vyhľadávania uskutočnil.)	
Projekt_691	**Lepšie riadenie zdrojov a plánovanie:** Podporovanie rozhodovania v oblasti dopravy za účelom optimalizácie riešení podporených dátovou analýzou. Primárne využitie sa predpokladá v oblasti regulácií a formulovania národných politík.	Údaje z vlastných dopravných prieskumov, z komunikačných systémov, kamerových systémov, dispečerských monitoringov, logov zo zariadení plus využitie voľne prístupnej dátovej základne (Dopravné reporty a štatistiky Národného dopravného informačného centra, Model cestnej siete Slovenskej správy ciest, Priestorové geografické údaje, Nehodovosť v cestnej doprave MV SR, Odvetvové a prierezové štatistické databázy zo Štatistického úradu SR, Dopravné rozvrhy verejnej dopravy, Datasety z meteorologických systémov, Špecifické údaje o lokalitách z Integrovaného záchranného systému, Dáta o zimnej údržbe ciest) a historických databáz.	Výsledkom môže byť napríklad "Simulačný interaktívny model budúceho stavu zvolených lokalít cestnej siete SR na základe vybraných parametrov – nárazová zmena intenzity dopravy, odstavenie vybraných objektov dopravnej infraštruktúry, zvyšovanie objemu vozidiel na území a pod.", "Modely prepojených dopravných dát – rozvrhy integrovanej dopravy ako podklad pre optimalizovaný plán dopravnej cesty pre cestujúcich, a pod.".
Projekt_697	**Optimalizácia prevádzky organizácií verejnej správy:** Podpora lepšieho rozhodovania sa inštitúcií vlastniacich budovy v oblasti šetrenia elektrickou energiou.	Údaje z meračov elektrickej energie, ktoré budú v reálnom čase posielať údaje cez internet do Centra analýzy energetických údajov, údaje, ktoré poskytne prevádzkovať energetickej infraštruktúry –	Výsledkom sú ušetrené prostriedky, ktoré sú teraz vynakladané na odpočet údajov z elektromerov, okamžitý prehľad o aktuálnej spotrebe a sankciách za prekročenie dohodnutých odberov.

ID projektu	Zámer	Potrebné údaje	Výsledok
	Systém ponúkne prehľady a predikcie, ktoré zjednodušia plánovanie rozpočtových položiek na energie. Zároveň umožní zabrániť škodám na elektrickej infraštruktúre, ktoré je možne detekovať vhodnými technológiami merania a prvkov umelej inteligencie nad nameranými údajmi.	distribučná spoločnosť, ktorá zabezpečuje pripojenie monitorovaného objektu do energetickej sústavy.	
Projekt_722	**Lepší návrh politík a regulácií:** Zefektívnenie centrálneho prístupu k riadeniu architektúry eGovernmentu.	Príprava vstupných zdrojov dát v požadovanom formáte, ako napr. exporty z existujúcich informačných systémov, tabuľkové štruktúry, či iné podporované otvorené štandardy. Súčasťou je tiež príprava údajov evidovaných v papierovej podobe, príp. iných kontextových informácií vyhodnotených ako relevantných z pohľadu ich digitalizácie a prínosu k definovaným prípadom použitia. Pôjde o architektonické údaje, strategické údaje zo strategických materiálov, údaje z portfólia, údaje o službách, procesné údaje, projektové údaje.	Vďaka vytvorenému DAMoG modelu znalostí o fungovaní MIRRI SR je možné vykonávať lepšie, na dátach postavené, informované rozhodnutia o strategických prioritách a ďalšom smerovaní architektúry. Takýto model bude prospešný pri vykonávaní dopadových ("what- if") analýz a návrhu riešení pre efektívne fungujúcu verejnú správu.
Projekt_745	**Prediktívne verejné zdravotníctvo:** Určiť a predpovedať, akým spôsobom sa bude vyvíjať zdravotný stav populácie na základe individuálnych príspevkov,	Údaje o výskyte ochorení, Klinické štúdie týkajúce sa ochorenia, Anonymizované zdravotné záznamy pacientov prostredníctvom integrácie na NCZI alebo alternatívnymi spôsobmi, Pacientami generované	Vyššia efektivita a účinnosť preventívnych programov vďaka ich adresnému nastaveniu. **Projekt bol zrušený.**

ID projektu	Zámer	Potrebné údaje	Výsledok
	a pôsobenia determinantov zdravia. Identifikácia pacientov, ktorí sú najviac ohrození chorobami, ako sú srdcové choroby alebo demencia, čo umožňuje skoršiu diagnostiku a lacnejšiu, cielenejšiu, prispôsobenú prevenciu. Vybrať a cieliť vhodné intervencie.	dáta zo senzorov a mobilných zariadení, Analýza a návrh zberu údajov v oblasti „antropometrického a fyziologického merania (tlak krvi, telesná výška, hmotnosť, telesné obvody) a odbery krvi)	
Projekt_746	**Lepší dozor a dohľad nad regulovaním prostredím:** Účelne, hospodárne a efektívne využívať prostriedky z verejného zdravotného poistenia. **Spojenie úradníka a stroja: inovácie procesov:** Zefektívnenie procesov systému vnútornej a externej kontroly a systému revízii. **Plánovanie budúcich kapacít:** Sofistikované analýzy, ktoré by vedeli predikovať trendy vývoja zdravotnej starostlivosti, segmentovať poskytovateľov a ich výkony. **Prediktívne kontroly:** Využitie umelej inteligencie a strojového učenia pri predikcii chybného a podvodného	Údaje o Poistenec, Kategórie platiteľa poistenca, Platiteľ, Predpísané poistné, Zaplatené poistné, Doklad vymáhania/platobný výmer, Exekúcia, Poskytovateľ ZS, Faktúra od PZS, Zdravotná starostlivosť, Fyzická kontrola PZS, Regres, Komunikácia s EÚ, Návrh na zdravotnú starostlivosť, Centrálne nakupovaný produkt, Zdravotná pomôcka (sklad zdravotných pomôcok), Čakacia listina, Dohoda o poskytovaní ZS medzi PZS a poistencom, a informácie z historických databáz. Dávky od poskytovateľov zdravotnej starostlivosti o vykázanej zdravotnej starostlivosti, schválená zdravotná starostlivosť VšZP., Dávky od UDZS - Informácie o mŕtvych pacientoch., Reklamácie zdravotnej starostlivosti – Spätná väzba od	Zanalyzovať predpokladanú stratu na neúčelnom preplácaní zdravotnej starostlivosti v desiatkach až stovkách miliónov EUR. Podpora zamestnancov v podobe výsledkov heuristických a prediktívnych algoritmov ako napríklad v podobe predspracovania dávky výkonov na preplácanie s odporúčaniami, kde mohlo dôjsť k chybe a odporúčaniami pre poskytovateľov zdravotnej starostlivosti ako chybu opraviť. Sofistikované analýzy, ktoré by vedeli predikovať trendy vývoja zdravotnej starostlivosti, segmentovať poskytovateľov a ich výkony. Viac vykonaných kontrol bez navýšenia počtu revíznych lekárov.

ID projektu	Zámer	Potrebné údaje	Výsledok
	správania poskytovateľov zdravotnej starostlivosti.	poistenca, Údaje o zdravotnom sektore – Dáta od NCZI, MZ SR a iných subjektov, Číselníky VšZP, Vzdialenosť a rozloženie poskytovateľov zdravotnej starostlivosti.	

Zámery dopytových projektov boli veľmi zaujímavé, relevantné, v mnohých prípadoch aj veľmi ambiciózne, keďže zahŕňali aj prediktívne analýzy a rozsiahly dopad na svoju oblasť (Projekt_680, Projekt_691, Projekt_697, Projekt_745, Projekt_746). Niektoré projekty (Projekt_683, Projekt_684, Projekt_691, Projekt_697) dokonca plánovali využiť pri analytickom spracovaní aj veľké údaje („big data“). Tiež pokrývali rôzne ponúkané oblasti analytického spracovania údajov, čo potvrdilo užitočnosť a aktuálnosť pripravených príručiek a návodov na prípravu projektov.

Náročnejšia sa však ukázala fáza realizácie, v ktorej neúmerne veľa času zaberá získanie údajov v dostatočnej kvalite, obzvlášť z externých zdrojov (trvanie minimálne polroka, bežne aj rok a pol). Keďže nepomerne viac dopytových projektov bolo predložených na manažment údajov, uvedomujú si potrebu zvyšovať kvalitu a dostupnosť svojich údajov a lepšie ich spravovať aj samotné organizácie.

Pri realizácii týchto dopytových projektov by určite pomohla dostupnosť centrálnej dátovej a analytickej vrstvy KAV, pretože by to ušetrilo čas, zdroje a náklady pri implementácii, kedy sa často v rámci projektu muselo „vynaliezať koleso“ (ako napríklad nastaviť dočasne dátový a analytický „stack“ v Amazon EMR v AWS, čo ale aj zároveň slúži ako ďalší dôkaz, že sa takýto „stack“ pre daný prípad použitia dá vytvoriť rýchlo a lacno, za nízky mesačný poplatok).

6.2 Zhodnotenie ponaučení

Pre úspešné realizovanie vízie Konsolidovanej analytickej vrstvy momentálne nie sú dostatočne pevné štyri základné kamene:

1. **Infraštruktúra** s modulárnym a flexibilným dátovým „stackom“, ktorá by v maximálnej možnej miere podporovala riešenia s otvoreným zdrojovým kódom a platby len za skutočne využité služby každý mesiac.
2. **Údaje** – centrálna integračná platforma momentálne nemá vybudované dátové „pipelines“, ktoré by efektívne premiestňovali veľký objem dát z danej agendy zo zdrojových databáz do dátovej vrstvy Konsolidovanej analytickej vrstvy. Ak daný prípad použitia vyžaduje údaje z externého zdroja mimo štátnej správy, ide tiež spravidla o časovo veľmi náročnú a nepredvídateľnú etapu z celej implementácie.
3. **Prípady použitia** – aktuálne neexistuje odsúhlasený zásobník efektívne realizovateľných prípadov použitia pre analýzu údajov, ktoré by prinášali hmatateľné výsledky v podobe lepšieho rozhodovania, napríklad v oblasti finančných úspor alebo návratnejších investícií, lepšej regulácie, efektívnejšieho plnenia legislatívou daných povinností či pružnejšieho poskytovania služieb občanom a podnikateľom. Väčšina projektov z dopytovej výzvy bola príliš ambiciózna.
4. **Spoločný cieľ s merateľnými výsledkami** – momentálne nie sú nastavené dostatočné motivačné faktory, aby všetky zúčastnené strany – poskytovatelia údajov, prevádzkovatelia infraštruktúry, používatelia údajov, zodpovedné OVM – ťahali za jeden koniec a snažili sa čo najefektívnejšie

dostať do cieľa. Veľkou výzvou je to, že väčšina teoreticky zaujímavých prípadov použitia si vyžaduje intenzívnu medzirezortnú spoluprácu.

Keďže takáto plošná zmena v uvažovaní a práci s údajmi ako s cenným zdrojom je náročná a často neúspešná na prvý pokus aj v podnikoch v súkromnom sektore, odporúčame pokračovať flexibilne a v malom a striktne sa držať troch princípov uvedených v kapitole 5. Predovšetkým je zásadné vytvoriť infraštruktúru pre KAV, za ktorú sa bude platiť len v takej miere, v akej sa bude využívať. To sa dá dosiahnuť len vo verejných cloudoch, ktoré navyše už v dnešnej dobe spĺňajú nielen tri princípy uvedené v kapitole 5, ale aj umožňujú rýchlo a efektívne pripraviť moderné prostredie pre analytikov a neustále ho inovovať o nové nástroje (viď kapitoly 4, 4.3 a 4.4). Cloudové platformy a služby umožňujú organizáciám presunúť výdavky z kapitálových do prevádzkových výdavkov, takže to prináša určitý účtovný zisk. Správa spoločnosti Forrester s názvom Navigácia v období poklesu v roku 2023: Technology Executive, poukazuje na "silnú cloudovú stratégiu" ako na jeden z kľúčových prvkov, ako čeliť recesii. Prípad zníženia nákladov môže byť obmedzený. A ako upozorňuje Maisto, senior analytik vo Forrester: „sen o platení podľa toho, ako cloud používate“ sa môže rýchlo zmeniť na „nočnú moru o platení podľa toho, ako naň zabudnete“ - trochu ako keď si v januári zaregistrujete členstvo v posilňovni a v auguste zistíte, že pravidelne platíte za niečo, čo ste pravidelne nevyužívali. To je však problém, keď organizácia premigruje do verejného cloudu svoje obrovské portfólio informačných systémov. V prípade postupného zavádzania KAV do praxe tento problém nehrozí, pretože klastre v cloude, ktoré sa aktuálne nevyužívajú na výpočet analytických algoritmov, sa vedia automaticky vypínať. Tiež portfólio cloudových služieb (SaaS) nebude také rozsiahle, aby sa nedalo sledovať ich využívanie.

Projekt KAV plánoval zaobstarať do dátovej vrstvy nasledovné datsety, ktoré považujeme za užitočné pre mnohé prípady použitia:

- Finstat,
- ITMS,
- Environmentálne záťaže.

Obzvlášť pri údajoch z ITMS vidíme „nízko visiace ovocie“, keďže ide o rozsiahle údaje z jedného informačného systému, ktoré by sa mohli dostať do dátovej vrstvy KAV priamo z jeho databáz. Následne by sa nad týmito údajmi dalo robiť viacero zaujímavých analýz, napríklad o investovaní a stave projektov, aj v podobe dashboardov, ako aj prediktívne analýzy o čerpaní a úspešnosti projektov.

6.3 Vysokoúrovňový plán ďalšieho rozvoja

Nasledujúci vysokoúrovňový plán pomôže vyhnúť sa bežným problémom dátovej stratégie, z ktorých viaceré boli identifikované aj v kapitole 6.2:

- Nedostatok prehľadu alebo pochopenia toho, ako sa údaje v organizácii používajú.
- Nedostatočné zosúladenie medzi OVM pri zdieľaní údajov a poznatkov.
- Ťažkosti pri zabezpečovaní súladu so zákonmi o ochrane osobných údajov a regulačnými požiadavkami.
- Roztrieštenosť dátových nástrojov, ktorá môže spôsobiť neefektívnosť, zbytočné náklady, zmätok, chyby a oneskorenia.
- Nedostatok opatrení na zabezpečenie údajov a politík správy údajov na ochranu integrity údajov.
- Ťažkosti pri integrácii štruktúrovaných a neštruktúrovaných údajov z rôznych zdrojov.

Vysokoúrovňový plán ďalšieho rozvoja KAV a zavádzania analytického spracovania údajov do praxe navrhujeme rozdeliť do troch fáz:

1. Fáza 1 – vytvorenie dátovej a analytickej vrstvy vo verejnom cloude a tímu, ktorý s ňou vie pracovať,
2. Fáza 2 – akčné cestovné mapy pre realizovateľné prípady použitia, pomocou ktorých sa dá rýchlo dostať do cieľa s merateľnými výsledkami,
3. Fáza 3 – strategické plánovanie ďalšieho rozvoja na základe výsledkov a ponaučení Fázy 2, napríklad môže ísť o zvyšovanie úrovne dátovej a analytickej vyspelosti (Tabuľka 9).

6.3.1 Fáza 1 – Dátová a analytická vrstva vo verejnom cloude a tím

1. Zosúlaďte svoju víziu

Spôsob, akým sa využívajú alebo plánujú využívať údaje, musí byť v súlade s cieľmi OVM ako aj dátového programu. Na základe 2-3 ročnej vízie treba určiť, kam údaje zapadajú do celkového obrazu.

V tejto fáze treba vedieť jasne formulovať „prečo“ – prečo majú OVM zdieľať svoje údaje do KAV? Čo tým získajú? Plán dátového programu a stratégie digitálnej transformácie v oblasti údajov môže vychádzať z potreby:

- Zvýšiť prehľad o danej agende – čo sa v nej darí, aké sú výzvy, aké kapacity sú nedostatočné a podobne, inými slovami – podporiť dátami riadené rozhodovanie.
- Vytvoriť nové služby pre občanov a podnikateľov, alebo vylepšiť tie existujúce.
- Vytvoriť efektívnejšie spôsoby využívania údajov vo verejnej správe – napríklad pre Reguláciu 2.0 a pre nastavovanie politík na základe dôkazov (viac v dokumente 4.1.1 Príručka Regulácie 2.0).

To všetko môžu byť dôstojné motívy, ale je dôležité ujasniť si problém, ktorý chcete vyriešiť, skôr, ako budete pokračovať. Z toho by mali potom aj vyplynúť viaceré prípady použitia pre fázu 2.

2. Vyberte si správnu architektúru a nástroje

Ďalej treba na základe vízie a cieľov vybrať správne nástroje a platformy:

1. Vybrať jedného alebo viacerých poskytovateľov verejného cloudu.
2. Potvrdiť si architektúru dátového lakehouse (kapitola 4.4) – bude sa využívať „data mesh“ a tvoriť logické platformy? (kapitola 5.2) Pre väčšinu organizácií je výhodné používať jednu platformu, ale niektoré môžu potrebovať viac. Uistite sa, že vyberáte správnu platformu (platformy) zo správneho dôvodu.
3. Aké typy údajov (kapitola 3.1) sa budú spracovávať v akých konkrétnych databázach (kapitola 3.2)? Výber databáz môže byť obmedzený na základe rozhodnutí v krokoch 1 a 2. V prípade problémov ich treba prehodnotiť.
4. Konektory a integračné nástroje, aj s ohľadom na rozvoj Centrálnej integračnej platformy: Konektory a integračné nástroje zvyšujú efektívnosť získavania a transformácie údajov (viac v dokumente 1.1.6 Štandardizácia dátovej transformácie). Dávajte si však pozor, aby ste sa nedostali do situácie, keď budete musieť podporovať viacero jednorazových nástrojov.
5. Vybrať vhodné analytické (kapitola 4.1), reportovacie (kapitola 4.2) a vizualizačné (kapitola 4.3) nástroje. Výber nástrojov môže byť obmedzený na základe rozhodnutí v krokoch 1 a 2. V prípade

problémov ich treba prehodnotiť. Bez ohľadu na to, aké nástroje si vyberiete, treba poznať kompromisy a dôsledky. Napríklad Power BI využíva hviezdicovú schému, zatiaľ čo Tableau používa ploché štruktúry. To má vplyv na zručnosti, ktoré sú potrebné v tíme. Takisto treba preskúmať a vybrať vhodnú licenčnú štruktúru pre dané potreby.

Po vykonaní týchto krokov ste pripravení vytvoriť vizuálnu architektúru, aby sa overilo, ako nástroje, platformy a databázy navzájom súvisia a podporujú sa.

3. Vykonajte inventarizáciu existujúcich dátových aktív

V ideálnom prípade ste už postupovali podľa 12 krokov na vytvorenie rámca na hodnotenie údajov[82]. Po posúdení ste schopní spísať dátové aktíva do dátového katalógu (viac v dokumentoch 1.1.2 Štandardizácia pre modelovanie údajov a 1.3.2 Usmernenia pre zrozumiteľné zdokumentovanie dátových štruktúr, procesov tvorby dát, štatistických metodológií (ak boli použité), dátových zdrojov, kontextov a ďalšie aspekty popisu dát). Okrem toho treba vytvoriť prioritizáciu datasetov a zdrojových systémov podľa agendových domén, aby ste získali prehľad z „vtáčej perspektívy“.

4. Osvojte si manažment údajov

Rámec manažmentu údajov je kľúčovou zložkou plánu dátového programu. Prístup by mal poskytovať vysokoúrovňový pohľad na:

- Počiatočné nastavenia a konfigurácie (t. j. riadenie prístupu).
- Či sa vyžaduje zabezpečenie na úrovni riadkov a kedy/kde.
- Ktoré datasety potrebujú certifikáciu a kto bude tento proces riadiť.
- Citlivé údaje, ktoré musia spĺňať regulačné požiadavky a musia byť pseudonymizované alebo anonymizované (viac v dokumentoch 1.1.3 Štandardizácia pre bezpečnosť a ochranu údajov a 1.1.5 Štandardizácia anonymizácie údajov),
- Vybraný dátový katalóg v dokumente .1.2 Štandardizácia pre modelovanie údajov poskytuje prehľad o línii údajov („lineage“) - napríklad o tom, kde sa citlivé údaje používajú.

5. Definujte úlohy a zodpovednosti

Na základe cieľov dátového programu a vybraných nástrojov treba definovať role a zručnosti potrebné na realizáciu programu. Existujú už tieto role v organizácii (alebo by mali existovať?).

Niektoré dôležité otázky týkajúce sa zdrojov, ktoré si treba položiť:

- Bude existovať aj centralizovaný tím pre analytické spracovanie údajov?
- Alebo budú OVM a analytické jednotky riešiť svoje vlastné potreby?
- Alebo sa nastaví hybridný model, v ktorom bude dátový tím pôsobiť ako interný konzultant pre OVM a analytické jednotky?

[82] Zdroj: https://skypointcloud.com/blog/set-up-a-data-assessment-framework/, Dátum referencie: 31.05.2023

- Súčasťou plánu môže byť pridanie alebo zníženie počtu zamestnancov. Ak nie sú možnosti zamestnať dátový tím na plný úväzok, mali by sa chýbajúce zdroje riešiť prostredníctvom školenia interných zamestnancov alebo najatím konzultantov.

6. Vytvorte plán školení

Po definovaní rolí a zodpovedností by ste mali poznať skupiny zručností potrebné na podporu implementácie dátového programu. Plán školení pomôže vybudovať interné kapacity a schopnosti na vyplnenie medzier pomocou existujúcich zdrojov (viď aj kapitola 5.8).

7. Vizualizujte svoj plán

Ak chcete získať ucelený pohľad na zavádzanie analytického spracovania údajov do praxe, vezmite výstupy vyššie uvedených krokov a rozvrhnite ich do grafu, aby ste vytvorili vizuálne znázornenie svojho plánu.

Hoci by váš plán zavádzania mal zahŕňať 2-3 ročnú víziu, musíte tiež načrtnúť ciele pre každý rok. Plán na prvý rok by mal obsahovať najviac podrobností vrátane prioritizovaných prípadov použitia (viď Fáza 2), príslušných agend, rizík a stratégií na ich zmiernenie, ako aj očakávaných výsledkov.

8. Zvalidujte plán a buďte realistickí

V neposlednom rade buďte realistickí. Nesnažte sa „uvariť oceán“. Počas implementácie si treba získať srdcia a mysle, aby viac ľudí vo verejnej správe ocenilo hodnotu údajov a ich analytického spracovania. Plánujte, čo sa dá dosiahnuť.

6.3.2 Fáza 2 – akčné cestovné mapy

Aby sa čo najrýchlejšie overila užitočnosť infraštruktúry vybudovanej vo Fáze 1, ako aj celkové nastavenie plánu, ktorý vyplynul z Fázy 1 a aby sa identifikovali ďalšie oblasti, ktoré treba riešiť strategicky, navrhujeme realizovať niekoľko akčných cestovných máp pre realizovateľné prípady použitia. Základom je ukázať v krátkej dobe merateľné výsledky, ktoré presvedčia všetky zúčastnené strany o zmysluplnosti zdieľania údajov pre analytické spracovanie. Pripraviť takúto akčnú cestovnú mapu možno v niekoľkých krokoch.

Krok 1: Identifikujte rýchle úspechy a veľmi dôležité alebo naliehavé prípady použitia.

Prvým krokom pri vytváraní akčnej cestovnej mapy je určiť, ktoré dátové a analytické prípady použitia by mali byť prioritné. To zahŕňa identifikáciu takých prípadov použitia, ktoré sa dajú ľahko implementovať, ako aj vysoko kritických alebo naliehavých prípadov použitia, ktoré sa musia riešiť, aby sa splnili ciele, definované či už v Pláne obnovy, Operačnom programe Slovensko, v programovom vyhlásení vlády alebo v akejkoľvek reforme či projekte. Ideálne by v tomto kroku mal byť dostupný zásobník takýchto prípadov použitia zoradený podľa priorít, kedy by tento krok predstavoval len výber toho najprioritnejšieho a kontrola, či všetky zdokumentované predpoklady pre jeho realizáciu sú stále platné a aktuálne.

Určenie prioritného prípadu použitia pre danú akčnú cestovnú mapu umožní:

- objasniť si ciele a zámery, ktoré chceme realizáciou prípadu použitia dosiahnuť,
- zamerať sa na najdôležitejšie aspekty cestovnej mapy a plánu implementácie,

- identifikovať potenciálne riziká a výzvy a zaviesť pohotovostné plány na zmiernenie týchto rizík,
- stanoviť realistické očakávania a jasne ich oznámiť zainteresovaným stranám.
- zosúladiť cestovnú mapu a jej plán s celkovou stratégiou dátového programu a podporiť dosiahnutie strategických cieľov.

Hoci by mal existovať aj dlhodobý plán pre dátový program, ktorým sa treba riadiť, je dôležité sústrediť sa na dosiahnutie rýchlych výsledkov implementáciou funkčných analytických riešení, ktoré prinášajú rýchlo hodnotu a vykazujú prvé pozitívne výsledky. Tým sa vybuduje potrebné momentum a podporí sa zavádzanie komplikovanejších prípadov použitia. Treba využívať rýchle prototypy riešení priamo s koncovými používateľmi, aby sa od nich mohla včas a často získavať spätná väzba a aby sa tak vyhlo riziku investovania veľkého množstva času do zbytočného riešenia.

Krok 2: Zrevidujte predpoklady a rôzne faktory, ktoré môžu ovplyvniť realizovateľnosť prípadu použitia.

Ako vyplýva aj z ponaučení, každou takouto cestovnou mapou posilňujeme tri z piatich základných kameňov úspechu. Tieto „kamene“ nám zároveň slúžia na zhodnotenie realizovateľnosti:

1. **Infraštruktúra** – máme dostupné databázy, nástroje na prácu s údajmi (napríklad na transformáciu), analytické a vizualizačné nástroje na to, aby sme zrealizovali vybraný prípad použitia? Ak nie, dá sa infraštruktúra jednoducho rozšíriť?
2. **Údaje** – máme potrebné údaje v KAV alebo ak nie, je realistické ich v dohľadnej dobe získať (rádovo mesiace, maximálne pol roka, pričom vzorka by bola dostupná skôr)?
3. **Merateľné výsledky** – dá sa úspech prípadu použitia merať, a ak áno, budú tieto merateľné výsledky viditeľné v dohľadnej dobe po realizácii (rádovo týždne, maximálne kvartál)?

Získanie údajov a ich čistenie a príprava na analýzu je najväčšou výzvou a zaberie spravidla 80 percent[83] zdrojov, či už ľudských, finančných alebo časových (v dátovej vede je to označované ako pravidlo 80 na 20[84]). Treba s tým počítať a od začiatku riadiť túto aktivitu ako rizikovú.

Výzvy pri externých poskytovateľoch údajov sú nasledovné: zmluvné podmienky, kvalita, rozsah a formát dát, spôsob ich poskytovania, cena za poskytovanie, verejné obstarávanie. Výzvy pri získavaní údajov z rôznych OVM sú neexistujúce dátové „pipelines“, neschválený Zákon o údajoch, nákladné integrácie zdrojových systémov a databáz z proprietárnych riešení, nedostatok financovania na modernizáciu informačných systémov, aby dokázali zvládať aj synchronizáciu veľkých objemov údajov alebo údajov v takmer reálnom čase pre moderné analýzy tokov.

Okrem toho je v tomto kroku dôležité identifikovať aj míľniky na vysokej úrovni na základe strategických cieľov dátového programu alebo očakávaných zmien v IT infraštruktúre či manažmente údajov, ktoré by mohli ovplyvniť realizáciu vybraného prípadu použitia. Tieto míľniky pomôžu určiť požadované tempo dokončovania prípadov použitia, ktoré podporia strategické ciele dátového programu v požadovanom časovom rámci. Treba si položiť si otázky, ako napríklad:

[83] Zdroj: https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf, Dátum referencie: 26.05.2023

[84] Zdroj: https://www.infoworld.com/article/3228245/the-80-20-data-science-dilemma.html, Dátum referencie: 26.05.2023

- Ako ambiciózne sú ciele dátového programu, ako sa ich darí napĺňať a v akom časovom rámci sú tieto ciele stanovené?
- Aké dôležité udalosti alebo externé faktory môžu ovplyvniť časový plán realizácie prípadu použitia?
- Sú známe nejaké kritické zmeny v organizácii zapojenej do prípadu použitia, ktoré môžu ovplyvniť jeho realizovateľnosť, alebo by mali byť zahrnuté do časového plánu? Môže ísť napríklad o uvedenie nového informačného systému do prevádzky, ktorým sa mení spôsob zberu potrebných údajov a ich dátový model.
- Akú úroveň bude musieť mať dátová a analytická vyspelosť (Tabuľka 9) pre úspešnosť realizácie vybraného prípadu použitia? Ak nie je aktuálne dostatočná, dokážeme sa v rámci definovaného časového rámca posunúť na vyššiu úroveň?

Tabuľka 9: Definovanie úrovní dátovej a analytickej vyspelosti[85]

Aspekt - Úroveň	Chaotická	Reaktívna	Zadefinovaná	Manažovaná	Optimalizovaná
Popis	Tímy sú preťažené, nedôvera v reporty a nedostatok kapacít na dátami riadené rozhodovanie.	Náhľad na zdravie organizácie poskytujú len limitované reporty o minulej výkonnosti s oneskorenými ukazovateľmi.	Vedúci zamestnanci majú dashboardy s definovanými kľúčovými metrikami na základe presných dát. Vedia, čo sa deje dnes a prečo.	Vedúci zamestnanci majú integrovanú podnikovú analytiku a metriky naprieč rezortmi, ktoré poskytujú komplexné pohľady na agendy a fokus na budúce výsledky.	Organizácia je hnaná podnikovými analytickými systémami, vrátane prediktívnych metrík, rozhodovania sa dejú na základe dát na všetkých úrovniach.
Štruktúra analytiky	Neexistujúca	Neexistujúca	Základná/ Minimálna	Integrovaná	Pokročilá
Proces riadenia	V silách	V silách	V silách	Integrovaný	Riadne spravovaný
KPIs	Nedefinované a ad hoc	Výkonnosť v minulosti	Aktuálne/ V reálnom čase	Merajú budúce výsledky	Prediktívne a komplexné
Dôvera v údaje	Nízka	Čiastočná	Čiastočná	Úplná	Úplná
Zručnosti	Limitované	Rôzne	Rôzne	Podporované	Vylepšené

[85] Zdroj: https://www.analytics8.com/blog/data-strategy-playbook-from-assessment-to-building-a-roadmap/, Dátum referencie: 26.05.2023

Napríklad dostať za z „Reaktívnej" úrovne do „Zadefinovanej" môže obnášať:

- Vybudovanie infraštruktúry a tímu na poskytovanie a podporu demokratizácie údajov (čo sa nedá zvládnuť v rámci jednej akčnej roadmapy, preto infraštruktúra a tím sa musia vybudovať vo Fáze 1),
- Pridanie ďalších personálnych pozícií na podporu rastu,
- Migrácia do cloudu a využívanie moderných technológií (čo sa nedá zvládnuť v rámci jednej akčnej roadmapy, preto sa toto musí nastaviť vo Fáze 1),
- Odstránenie technologickej závislosti od spoločnosti DentaQuest
- Vývoj dátového „stacku“ na vytvorenie jediného „zdroja pravdy“ s dátovým jazerom aj dátovým skladom – toto môže byť aj výstupom realizácie prípadu použitia, aby bol v budúcnosti udržateľnejší, ktorý bude následne adresovať Fáza 3.

Krok 3: Naplňte časovú os míľnikmi a aktivitami na realizáciu prípadu použitia

Po stanovení míľnikov na vysokej úrovni a dôslednej analýze realizovateľnosti prípadu použitia je ďalším krokom doplnenie časového plánu o míľniky a aktivity, ktoré budú potrebné na dobudovanie dátovej architektúry a na splnenie ďalších požiadaviek prípadu použitia v iteráciách, ktoré budú postupne zvyšovať vyspelosť údajov a analytiky a zároveň poskytovať hodnotu. V prvej iterácii si položte otázky, ako napríklad:

- Potrebujeme nastaviť nové nástroje alebo technológie? Ak áno, máme implementačného partnera alebo si ho musíme najať?
- Budeme musieť vytvoriť nové dátové „pipelines“? Máme na to interné kapacity?
- Ako vyspelý bude analytický model (Tabuľka 10)? Cieľom by malo byť mať medzi vysoko prioritnými prípadmi použitia aj taký, ktorý bude vyžadovať aspoň jeden prediktívny model. Avšak v tejto fáze sa ešte neuvažuje s plným nasadením prevádzky strojového učenia (takzvanému „ML Ops“). Táto potreba však môže byť výsledkom implementácie prípadu použitia a jeho cestovnej mapy a zohľadnená v plánovaní vo fáze 3.
- Koho musíme zapojiť?
- Aké máme očakávať výsledky v jednotlivých míľnikoch? Z výsledkov prípadu použitia môžu vzniknúť požiadavky na rozvoj dátového „stacku“ vo Fáze 3.
- Sú definície naprieč OVM konzistentné a v súlade s dátovým programom? Čo treba spraviť na zosúladenie implementácie prípadu použitia s dátovým programom, konkréte aj so správou údajov?

Tabuľka 10: Štyri úrovne vyspelosti analytických modelov[86]

Deskriptívna analýza	Diagnostická analýza	Prediktívna analýza	Preskriptívna analýza
Zameriava sa na analýzu historických údajov s cieľom pochopenia minulých udalostí, vzorcov, a trendov. Pomáha organizáciám odpovedať na otázku: "Čo sa stalo?" **Nástroje a techniky:** - Microsoft Excel na analýzu údajov a vizualizáciu - Tableau alebo Power BI pre vizualizáciu a dashboardy - Základné techniky čistenia údajov, ako je odstraňovanie duplicít a spracovanie chýbajúcich údajov, - Základné štatistické metriky, ako napr. priemer, medián, mód a štandardná odchýlka - Vytváranie stĺpcových a čiarových grafov na zobrazenie trendov v údajoch **Príklad**: Finančný tím môže sledovať medzimesačný a medziročný rast alebo pokles výdavkov.	Sústreďuje sa na identifikáciu príčiny minulých udalostí prostredníctvom skúmania korelácií medzi údajmi, závislostí a anomálií. Pomáha organizáciám odpovedať na otázku: „Prečo sa to stalo?“ **Nástroje a techniky:** - Štatistické nástroje a grafy na identifikáciu trendov a vzorcov v údajoch - Nástroje na dolovanie údajov na identifikáciu korelácií a vzťahov medzi údajmi - Regresná analýza na identifikáciu vzťahov medzi premennými - Analýza údajov na identifikáciu príčiny problému alebo záležitosti - Používanie korelačnej analýzy na identifikáciu vzťahov medzi premenných **Príklad**: Finančný tím môže porovnať načasovanie kľúčových iniciatív s medzimesačným a medziročným rastom výdavkov alebo poklesom a pomôcť tak určiť korelácie.	Využíva historické údaje, štatistické modely a strojové učenie na predpovedanie budúcich výsledkov a trendov. Pomáha organizáciám odpovedať na otázku: "Čo sa pravdepodobne stane v budúcnosti? **Nástroje a techniky:** - Algoritmy strojového učenia pre prediktívne modelovanie - Pokročilé štatistické techniky, ako je analýza časových radov a rozhodovacie stromy - Nástroje na dolovanie údajov na identifikáciu vzorcov a vzťahov v údajoch - Používanie analýzy časových radov na identifikáciu trendov a vzorov - Vytvorenie prediktívneho modelu na predpovedanie správania **Príklad**: Finančný tím môže vytvoriť presnejšie prognózy pre ďalší fiškálny rok.	Zameriava sa na odporúčanie najlepšieho postupu na základe rôznych scenárov a potenciálnych výsledkov. Pomáha organizáciám odpovedať na otázku: "Čo by sme s tým mali robiť?" **Nástroje a techniky:** - Optimalizačný softvér na identifikáciu optimálneho riešenia problému - Simulačné modelovanie na testovanie rôznych scenárov a výsledkov - Algoritmy strojového učenia na automatizované rozhodovanie - Platformy na spracovanie veľkých objemov údajov - Vytvorenie simulačného modelu na testovanie vplyvu rôznych stratégií **Príklad**: Finančný tím má k dispozícii spôsoby, ako optimalizovať manažment rizika privysokých výdavkov.

Nezabudnite zohľadniť všetky aktivity potrebné pre prvú iteráciu, aj keď sa zdajú byť nepodstatné. Vynechanie malej úlohy na začiatku môže mať za následok, že sa budete musieť vrátiť a znovu zapojiť niekoho, kto má teraz iné priority. Je dôležité zvážiť, ako každý nástroj, technológia alebo proces zapadá do širšej dátovej architektúry, a pristupovať k implementácii postupne.

Krok 4: Doplňte ďalšie podrobnosti a závislosti.

[86] Zdroj: https://www.domontconsulting.com/products/data-analytics-strategy-toolkit?variant=40429892337746&gad=1&gclid=CjwKCAjwscGjBhAXEiwAswQqNFmvk0mfRV0dZdwejjz1zJrTQBnsf9XnrA8-eHClyIIJsqFGULsX_xoCd_MQAvD_BwE#download, Dátum referencie: 25.05.2023

Aby sa zabezpečilo, že plán implementácie prípadu použitia je realizovateľný a dosiahnuteľný, je dôležité pridať ďalšie podrobnosti a závislosti. Patrí sem napríklad:

- Identifikácia ľudí a zdrojov, ktoré budú potrebné na dokončenie prípadu použitia, vrátane všetkých externých partnerov alebo dodávateľov, ktorí môžu byť zapojení.
- Identifikácia akýchkoľvek závislostí alebo vzájomných väzieb, ktoré môžu ovplyvniť časový plán, ako napríklad dokončenie určitých prípadov použitia, ktoré sú potrebné na to, aby sa ostatné mohli posunúť vpred.
- Vytvorenie plánu zdrojov, v ktorom sa uvedú požadované úlohy a zodpovednosti jednotlivých členov tímu, ako aj odhadovaný čas a rozpočet potrebný na každú aktivitu. Pomáha to zabezpečiť, aby mal každý jasno o svojich úlohách a povinnostiach a aby mal tím potrebné zdroje na dokončenie prípadu použitia podľa plánu.

Zohľadnenie všetkých obmedzení alebo limitov, ktoré môžu mať vplyv na časový plán alebo rozpočet, ako sú regulačné požiadavky, aspekty ochrany osobných údajov alebo technické obmedzenia. Tieto faktory si môžu vyžadovať dodatočné plánovanie a zdroje na riešenie, preto je dôležité zohľadniť ich v pláne.

Krok 5: Zahrňte plán komunikácie so zúčastnenými stranami o časovom pláne.

Komunikovať cestovnú mapu treba so všetkými zúčastnenými stranami, ktoré musia byť zapojené, aby sa podarilo vybraný prípad použitia zrealizovať. Definované míľniky sa musia dostať do časových plánov a vývojárskych šprintov zodpovedných ľudí v nasledujúcich možných zúčastnených stranách (treba opätovne preveriť, odkomunikovať a prípadne upraviť plán a podrobnosti z Kroku 4):

- Vlastníci údajov, ktorí musia byť stotožnení s tým, že ich údaje sa budú v KAV používať na analýzu, a musia k tomu existovať platné dokumenty na základe legislatívy a usmernení,
- Tím biznis analytikov, ktorý budú musieť detailnejšie zanalyzovať prípad použitia a jeho hlbší kontext vrátane porozumenia údajom,
- Tím vývojárov (externý alebo interný) zodpovedný za realizáciu dátovej integrácie (ak všetky údaje potrebné pre realizáciu prípadu použitia nie sú dostupné v KAV),
- Tím administrátorov (externý alebo interný) zodpovedný za údržbu infraštruktúry pre KAV (môže ísť o nastavenie dodatočných cloudových služieb, pomoc pri manažmente klastra či vybranej databázy),
- Tím dátových architektov, ktorý môže navrhnúť dátový model podľa vybranej metodiky, ak je to potrebné a ide o komplexnejší prípad použitia,
- Analytický tím u externého dodávateľa alebo interná analytická jednotka, ktorá bude pracovať na predprípravе dát a na tvorbe analytického modelu,
- Koncoví používatelia, ktorí budú používať analytické výstupy, aby si vyhradili čas na testovanie a overovanie výstupov a spätnú väzbu.

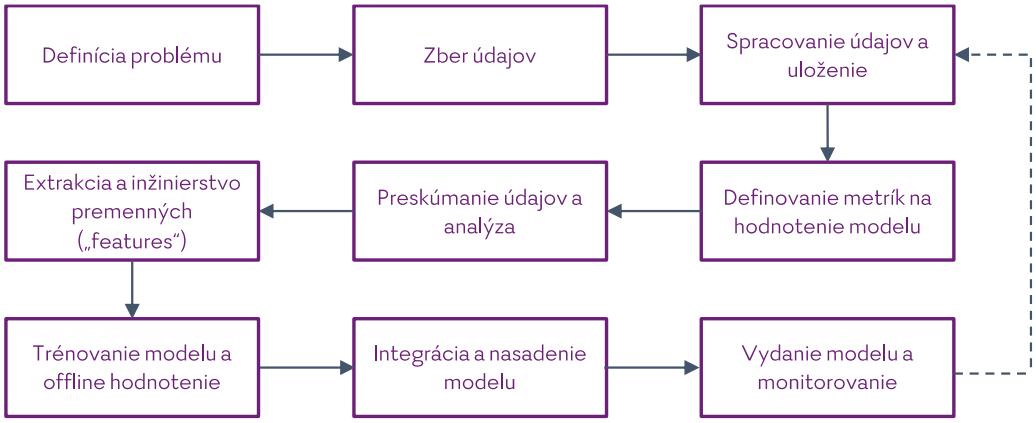
Po vypracovaní akčnej cestovnej mapy je dôležité oznámiť ho vedeniu, aby sa vytvorila podpora a zabezpečilo sa, že všetci sú v súlade a pracujú na rovnakých cieľoch. Pre efektívnu komunikáciu treba:

- Jasne a stručne komunikovať ciele cestovnej mapy, zdôrazniť, ako podporí dosiahnutie kľúčových cieľov dátového programu a zosúladí sa s celkovou stratégiou digitálnej transformácie.

- Zdieľanie časového plánu a míľnikov so zúčastnenými stranami a jasné vysvetlenie účelu a významu prípadu použitia.
- Pravidelné zapojenie zúčastnenými strán s cieľom získať spätnú väzbu a podnety k plánu a cestovnej mape v oblasti údajov, pričom sa zabezpečí, aby boli realistické a riešili potreby a záujmy všetkých relevantných strán.
- Poskytovanie aktuálnych informácií o pokroku v plnení plánu, riešenie akýchkoľvek problémov alebo výziev, ktoré sa vyskytnú, a jeho prípadné prispôsobovanie meniacim sa prioritám alebo okolnostiam.
- Zabezpečenie transparentnej a otvorenej komunikácie, ktorá podporuje spoluprácu a zapojenie všetkých príslušných strán.

6.3.3 Fáza 3 – strategické plánovanie ďalšieho rozvoja

Keďže základným vstupom pre návrh tejto fázy sú výsledky fázy 2, nie je možné v tomto dokumente túto fázu podrobne definovať. Očakáva sa ale, že v tejto fáze bude potrebné zvýšiť analytickú a dátovú vyspelosť minimálne na úroveň „Manažovaná“ (Tabuľka 9), aby aj vedúci zamestnanci OVM zapojených do dátového programu profitovali pri svojej každodennej práci a rozhodnutiach z nasadených analytických riešení. Prediktívne alebo preskriptívne analytické modely (Tabuľka 10) pravdepodobne nebudú bežnou rutinou, ale mali by byť súčasťou aspoň niektorých zabehnutých prípadov použitia, a tomu bude potrebné prispôsobiť aj infraštruktúru, procesy a zdroje. S tým súvisí ďalšia zásadná oblasť, ktorú bude treba zaviesť v tejto fáze, a tou je už spomínaná prevádzka strojového učenia („MLOps“). Ide súbor postupov, ktoré podrobne opisujú, ako zavádzať modely strojového učenia, monitorovať ich a opakovane ich trénovať štruktúrovaným a segmentovaným spôsobom. V životnom cykle MLOps je deväť fáz, ako ukazuje aj Obrázok 7:



Obrázok 7: Deväť fáz životného cyklu MLOps

1. **Definícia problému** - Pri vývoji modelov je prvým krokom identifikácia problému, ktorý sa bude riešiť pomocou umelej inteligencie.
2. **Zber údajov** - Po identifikácii problému nasleduje fáza zberu údajov. Tieto údaje sa použijú na trénovanie modelov a mali by pochádzať z vhodného zdroja. Tieto údaje sa musia ukladať aj po natrénovaní modelu, aby sa dal model neustále hodnotiť, či funguje, používať napríklad aj na predikcie a opakovane trénovať.

3. **Spracovanie alebo ukladanie údajov** - Aby ste mohli efektívne trénovať modely, budete potrebovať veľké množstvo údajov. Na ukladanie sa zvyčajne používajú dátové sklady alebo dátové jazerá. Tieto údaje sa môžu konsolidovať a čistiť v dávkach alebo ako tok, podľa toho, čo je pre daný prípad použitia výhodnejšie.
4. **Definícia metrík** - Aby sa dala merať kvalita modelov a určiť, či sú úspešné pri riešení problému identifikovaného v prvej fáze, musí existovať zhoda, ktoré metriky sa budú používať.
5. **Preskúmanie údajov** - V ďalšej fáze dátoví vedci vytvárajú hypotézy o tom, aké techniky modelovania budú na základe preskúmania údajov najužitočnejšie.
6. **Extrakcia a inžinierstvo premenných modelu („features“)** - Po vypracovaní hypotéz dátoví vedci ďalej určia, ktoré aspekty údajov sa použijú ako vstupy do modelov. Tieto aspekty sú známe ako premenné („features“). Ak je napríklad riešeným problémom algoritmus schvaľovania grantov, vlastnosti modelov môžu zahŕňať finančnú situáciu a históriu realizovania grantových projektov žiadateľov. Dátoví vedci budú musieť rozhodnúť aj o spôsobe generovania premenných (okrem toho, ktoré aspekty sa vyberú ako premenné) a musia sa zapojiť inžinieri, aby sa premenné priebežne aktualizovali, keď sa zhromaždia nové údaje. Mnoho dátových „stackov“ ponúka aj úložisko pre tieto premenné, v ktorom sa dajú ukladať, spravovať a prezerať metódy, ktorými boli vytvorené.
7. **Trénovanie modelov a offline hodnotenie** - Keď sa modely vytvoria, natrénujú a vyhodnotia, dátoví vedci vyberú prístup s najlepším výkonom. Približne 80 - 90 % zozbieraných a spracovaných údajov sa použije na trénovanie modelov, zatiaľ čo 20 - 30 % sa vyhradí na vyhodnotenie.
8. **Integrácia a nasadenie modelov** - Po natrénovaní a vyhodnotení modelov je ďalším krokom ich integrácia do produktu a následné nasadenie, zvyčajne v rámci verejného cloudu. Integrácia modelov môže zahŕňať vytvorenie nových služieb, aby výsledný produkt mohol získavať predpovede modelov, ktoré budú viesť k výsledkom pre koncových používateľov. Ak sa vrátime k príkladu so žiadosťou o grant, po spustení modelu na schvaľovanie bude agendový systém pre zamestnancov potrebovať spôsob, ako získať prístup k algoritmu pre posúdenie, kto by mal alebo nemal dostať grant, aj spolu s vysvetlením predpovede modelu.
9. **Vydanie („release“) a monitorovanie modelu** - Po nasadení je potrebné modely pozorne monitorovať, aby sa zabezpečilo, že nedôjde k problémom, ako je odchyľovanie sa údajov od pôvodných, na ktorých bol model trénovaný, čo vedie aj k závažným odchýlkam v modeli, alebo skreslenie modelu. Ďalšou výhodou monitorovania modelu je, že sa dá identifikovať, ako možno model zlepšiť pomocou opakovaného trénovania s novými údajmi.

MLOps sa istým spôsobom vyvinul z DevOps. Rovnako ako DevOps vznikol s cieľom pomôcť „operacionalizovať“ vývoj softvéru, MLOps sa stal spôsobom „operacionalizácie“ strojového učenia. Strojové učenie sa však zásadne líši od softvéru a vyžaduje si vlastný špecifický proces. Môžu mať spoločné črty, ale jednotlivé kroky životného cyklu sú pre MLOps jedinečné.

Podobne ako DevOps, aj MLOps pomáha zlepšovať kvalitu produkčných modelov, pričom zahŕňa biznisové a regulačné požiadavky a správu modelov. Niektoré z problémov, ktoré MLOps rieši, sú:

- **Neefektívne pracovné postupy** - MLOps poskytuje rámec na efektívne a účinné riadenie životného cyklu strojového učenia. Zosúladením metodických pravidiel, znalosti agendy s technickou zdatnosťou vytvára MLOps štruktúrovanejší, iteratívny pracovný postup.
- **Nedodržiavanie predpisov** - Strojové učenie je pomerne nová oblasť a regulačné orgány neustále menia svoje požiadavky a aktualizujú svoje usmernenia. MLOps preberá zodpovednosť za dodržiavanie a aktualizáciu meniacich sa predpisov.

- **Úzke miesta** - Pri zložitých, neintuitívnych algoritmoch môže často dochádzať k úzkym miestam. MLOps uľahčuje spoluprácu medzi prevádzkovými a dátovými tímami, čím pomáha znižovať frekvenciu a závažnosť týchto typov problémov. Spolupráca, ktorú MLOps podporuje, využíva odborné znalosti predtým oddelených tímov a pomáha efektívnejšie vytvárať, testovať, monitorovať a nasadzovať modely strojového učenia.

V konečnom dôsledku je kľúčovým výsledkom MLOps vysokokvalitný model, na ktorý sa dá spoľahnúť, že bude robiť to, čo má. MLOps tiež dáva možnosť pozorovať zmeny, ktoré sa dejú pri vyvíjaných modeloch, aby sa mohli zachytiť a upraviť negatívne vplyvy ešte pred spustením do produkcie.

V nedposlednom rade je v tejto fáze kľúčové vyladiť takzvanú „pozorovateľnosť údajov" („data observability")[87], vďaka ktorej sa bude dať vždy na údaje, ktoré vstupujú do analytického spracovania, spoľahnúť. Pozorovateľnosť údajov sa vzťahuje na komplexné porozumenie stavu a výkonnosti údajov v rámci informačných systémov organizácie. Nástroje na pozorovanie údajov využívajú automatizované monitorovanie, analýzu základných príčin („root cause analysis"), dátovú „lineage" a prehľad o „zdraví" údajov na proaktívne zisťovanie, riešenie a prevenciu anomálií údajov. Výsledkom tohto prístupu sú funkčnejšie dátové „pipelines", vyššia produktivita tímu, zdokonalené postupy manažmentu údajov a v konečnom dôsledku vyššia spokojnosť koncových používateľov. Pozorovateľnosť údajov je založená na nasledujúcich piatich pilieroch:

1. **Aktuálnosť**: Ide o snahu pochopiť, ako aktuálne sú tabuľky s údajmi, ako aj zistiť frekvenciu, s akou sa tabuľky aktualizujú. Aktuálnosť údajov je obzvlášť dôležitá, keď ide o rozhodovanie; koniec koncov, zastarané údaje sú v podstate synonymom premrhaného času a peňazí.
2. **Kvalita**: Dátové „pipelines" môžu byť v poriadku, ale údaje, ktoré cez ne prúdia, môžu byť „odpad". Pilier kvality sa zaoberá samotnými údajmi a aspektmi, ako je percento chýbajúcich hodnôt, percento unikátov a či sú údaje v prijateľnom rozsahu. Kvalita poskytuje prehľad o tom, či sa tabuľkám dá dôverovať na základe toho, čo možno od údajov očakávať.
3. **Objem**: Týka sa úplnosti tabuliek s údajmi a ponúka pohľad na stav zdrojov údajov. Ak sa 200 miliónov riadkov zrazu zmení na 5 miliónov, mali by ste o tom vedieť.
4. **Schéma**: Zmeny v organizácii alebo v dátovom modeli údajov, inými slovami, schéma, často naznačujú nefunkčnosť údajov. Sledovanie toho, kto a kedy vykonáva zmeny v týchto tabuľkách, je základom pre pochopenie zdravia dátového ekosystému.
5. „**Lineage**": Keď sa údaje pokazia, prvá otázka vždy znie: „Kde?". Dátová lineage poskytuje odpoveď tým, že informuje o tom, ktoré zdroje a následné procedúry boli ovplyvnené, ako aj o tom, ktoré tímy generujú údaje a kto k nim pristupuje. Dobrá dátová lineage zhromažďuje aj informácie o údajoch (označované aj ako metadáta), ktoré hovoria o riadení, biznis a technických usmerneniach súvisiacich s konkrétnymi tabuľkami údajov a slúžia ako jediný zdroj pravdy pre všetkých konzumentov.

6.3.3.1 Odporúčania pre fázu 3 a ďalší rozvoj

Aj tie najpokročilejšie dátami riadené organizácie by mali neustále skúmať nové technológie a nahrádzať komponenty ich „stacku" vhodnejšími alternatívami.

[87] Zdroj: https://www.montecarlodata.com/blog-what-is-data-observability/ , Dátum referencie: 26.05.2023

Komponenty „stacku“ sa musia dať jednoducho medzi sebou prepájať

Dátové tímy sa rýchlejšie posúvajú vpred, keď jednotlivé komponenty v „stacku“ sú ľahko vymeniteľné. Veľa populárnych nástrojov dneška ráta s touto požiadavkou, a preto poskytujú desiatky predpripravených konektorov na bežné biznis aplikácie a aplikačné programovateľné rozhrania pre integrácie na mieru. Niektoré nástroje ako Fivetran[88] alebo Stitch[89] slúžia len na to aby sa dali dáta pohodlne preniesť z jedného miesta na druhé a umožňujú tímom vytvoriť jednotnú automatizovanú dátovú „pipeline“ so štandardizovanou schémou. Vďaka tomu môžu dátoví analytici rýchlo preskúmať dáta novými spôsobmi bez toho, aby museli investovať mesiace do tvorby konektorov pre každý nový nástroj, ktorý si chcú odskúšať.

Rýchle možnosti na modelovanie a preskúmanie dát s cieľom získať poznatky

Dátové tímy musia mať flexibilitu pri členení dát ľubovoľným spôsobom, ktorý čo najlepšie pomôže adresovať daný problém a získať nové poznatky. Najlepší nástroj na túto úlohu závisí od viacerých faktorov: s akým programovacím jazykom sú dátoví analytici najviac komfortní, na akú otázku sa snažia nájsť odpoveď a aká zúčastnená strana požaduje nové poznatky. Moderný dátový „stack“ nelimituje dátových analytikov v ich prístupe. Poskytuje nástroje na prácu vo viacerých programovacích jazykoch, na vizualizáciu dát mnohými spôsobmi, a na prispôsobovanie dashboardov a reportov rôznym oddeleniam a ich najdôležitejším merateľným ukazovateľom. Na rozdiel od monolitických riešení, ktoré často predpisujú pracovný postup pri analýze dát, modulárne „stacky“ zložené z flexibilných komponentov ponechávajú používateľom možnosť diktovať si svoje najlepšie postupy pre dosiahnutie cieľa. Analytici si môžu dáta rôzne vyberať a transformovať tak, aby našli nové vzorce a trendy, kým ostatní zamestnanci môžu vidieť výstupy v jednoducho použiteľných prostrediach bez nutnosti programovania.

Infraštruktúra musí prispôsobovať svoju kapacitu aktuálnym potrebám

Akonáhle začnú byť citeľné nedostatky vo výkonnosti analytických pracovných postupov, je čas prehodnotiť celý „stack“. Hoci staršie systémy môžu poskytovať ďalšie cenové úrovne pre vyššiu kapacitu na serveroch, niekedy môžu byť problémy skôr ukryté v samotnom procese replikácie, transformácie a dopytovania dát, ako v nedostatočnej kapacite serverov. Namiesto zaťažovania vývojárov zoznamom chýb a požiadaviek na zlepšenie sa môže viac vyplatiť nájsť si také nástroje, ktoré sú špecificky vytvorené pre prácu s väčšími objemami dát a ktoré optimalizujú dátové „pipelines“ a skracujú čas na dopytovanie dát ako Databricks, Snowflake alebo BigQuery .

Analýza historických údajov nekončí pri dashboardoch

Veľa organizácii investuje nemalé prostriedky do centralizovaných BI nástrojov a očakáva, že dashboardy budú pomáhať pri ťažkých rozhodnutiach. Avšak dátový lakehouse plný dát ešte neznamená, že dátový analytický tím dokáže reagovať na každú požiadavku s včasnými a hodnotnými poznatkami. Namiesto toho, aby analytický dátový tím vytváral stále nové alebo aktualizované dashboardy, mal by mať skôr prístup k nástrojom, ktoré umožnia hĺbkovejšiu analýzu dát. Vďaka

[88] Zdroj: https://www.fivetran.com/, Dátum referencie: 31.05.2023

[89] Zdroj: https://www.stitchdata.com/, Dátum referencie: 31.05.2023

takýmto nástrojom a zavedeným postupom bude možné odhaliť nové vzorce v údajoch a dostať sa tak k podstate veci, prečo dané merateľné ukazovatele sa vyvíjajú tak, ako sa vyvíjajú, a čo možno robiť na zmenu tohto trendu. Cieľom je teda pridať do dátového „stacku" také technológie a nástroje, ktoré kompenzujú limity tradičných BI nástrojov. Napríklad v riešeniach ako Mode[90], zdieľané pracovné prostredie umožňuje analytikom preskúmať dáta rôznymi spôsobmi pomocou jazykov SQL, Python alebo R, a to v tom istom prostredí. V nástroje Sisu[91], kolaboratívne knižnice na dopyty pomáhajú analytikom zdieľať a používať dopyty efektívne naprieč všetkými dashboardmi či aplikáciami.

Treba vyhradiť priestor aj na proaktívne analýzy historických údajov

Veľa organizácií nikdy neopustí reaktívnu fázu dátovej analýzy – teda fázu, v ktorej len reaguje na vzniknuté problémy a otázky. Aj keď sú dáta dobre spravované a dátový tím sa plne venuje analýze a má zručnosti a nástroje aj na hĺbkovú analýzu, stále môže všetok svoj čas využiť len na odpovedanie na otázky prichádzajúce z jednotlivých oddelení. Často sa pri hlbšej analýze dát vynárajú ďalšie a ďalšie otázky. Analytické tímy často trávia až 80% svojho času prípravou dát, a len 20% ich samotnou analýzou. Aby sa dalo z toho cyklu vymaniť, musí tím tráviť menej času manuálnym preusporiadavaním dát alebo vytváraním nových reportov alebo dasboardov pre každú otázku, ktorá im pristane na stole, ale miesto toho musia mať čas na proaktívnu analýzu dát. Práve spomínané nástroje, ako Sisu, dokážu automatizovať tie najprácnejšie úlohy ako prípravu dát a kontinuálnu analýzu, a vedia rozšíriť danú analýzu o komplexné testovanie každého faktoru v dátach, čím môžu vyplávať na povrch nové hodnotné poznatky. Tieto inteligentné riešenia využívajú pokročilé nástroje strojového učenia na diagnostiku zmien v ukazovateľoch, akonáhle nastanú, a na následné preskúmanie každej dimenzie v dátach, čím môžu prispieť k odpovediam na otázky, o ktorých nik nevedel, že sa ich vôbec treba pýtať.

[90] Zdroj: https://mode.com/, Dátum referencie: 31.05.2023

[91] Zdroj: https://sisudata.com/, Dátum referencie: 31.05.2023