Operačný program
Efektívna
verejná správa

Európska únia
Európsky sociálny fond

MINISTERSTVO
INVESTÍCIÍ, REGIONÁLNEHO ROZVOJA
A INFORMATIZÁCIE
SLOVENSKEJ REPUBLIKY

Výstup 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

Realizačná zmluva o poskytnutí služieb a o dielo č. 445/2022

Projekt:

Zlepšenie využívania údajov vo verejnej správe

ITMS kód projektu:

314011S979

Document review and approval

Revision history

Version	Author	Date	Revision
1.0	Ceľuchová Bošanská Bárdy Janík	31.07.2023	

This document has been reviewed by

	Reviewer	Date reviewed
1		
2		
3		
4		
5		

This document has been approved by

	Subject matter experts		
	Name	Signature	Date reviewed
1			
2			
3			
4			
5			

ZOZNAM SKRATIEK	
Skratka	**Význam**
BI	Business Intelligence (Analytické nástroje v rámci podnikovej inteligencie)
CIP	Centrálna integračná platforma
CMÚ	Centrálny model údajov
CSV	Súbor hodnôt oddelených čiarkou (Comma Separated Value)
DAG	Smerový acyklický graf (Directed Acyclic Graph)
ETL	Extrakcia, prenos a načítanie (nástroje pre „Extract, Load, Transfer“)
GDPR	Všeobecné nariadenie o ochrane osobných údajov (General Data Protection Regulation)
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)
KPI	Kľúčový ukazovateľ výkonnosti (Key Performance Indicator)
MIRRI SR	Ministerstvo investícii, regionálneho rozvoja a informatizácie
ML	Strojové učenie (Machine Learning)
MOU	Manažmentu osobných údajov
MV SR	Ministerstvo vnútra SR
OVM	Orgán verejnej moci
PII	Osobne identifikovateľné údaje (Personally Identifiable Information)
RDF	Resource Description Framework
SAM	Model predmetnej oblasti (Subject Area Model)
SLA	Dohodnutá úroveň poskytovanej služby (Service Level Agreement)
URI	Jednotný referencovateľný identifikátor
VC	Overiteľné poverenia (Verifiable Credentials)
W3C	World Wide Web Consortium
XML	Extensible Markup Language

Obsah

1 Úvod a zhrnutie

1.1 Kontext

Dokument bol pripravený v rámci projektu „Zlepšenie využívania údajov vo verejnej správe“. Tento projekt má ambíciu transformovať fungovanie inštitúcií verejnej správy tak, aby dokázali maximálne efektívne spravovať a zdieľať údaje, využívať údaje pre lepšie rozhodovanie na základe faktov a dôkazov, pre zlepšenie efektivity a adresnosti služieb na základe lepšieho využívania dát.

Projekt Zlepšenie využívania údajov vo verejnej správe realizuje Dátová kancelária verejnej správy ako špeciálna jednotka Ministerstva investícii, regionálneho rozvoja a informatizácie (ďalej aj MIRRI SR).

Výstup reflektuje požiadavky Dátovej kancelárie a nadväzuje na agendu Dátovej kancelárie, pričom dokument je rozšírený o zoznam konkrétnych odporúčaní pre úpravu rezortnej legislatívy (vrátane zodpovednosti za realizáciu v praxi). Výstup je zosúladený s pripravovaným zákonom o údajoch a zároveň prepája realizáciu princípu jedenkrát Dokument ďalej obsahuje:

- návod ako popisovať dáta a analytické aplikácie,
- vysvetlenie nástrojov na popisovanie,
- previazanie popisov a dokumentácie s Centrálnym modelom údajov a meta-informačným systémom,
- príklady správneho popisu.

Výstup vznikol ako realizácia aktivity číslo 1 Manažment kvality údajov a činnosti 1.3 — Metodika pre manažment údajov vo verejnej správe, ktorej zámerom je vytvoriť koncepciu moderného využívania údajov a na jej základe interaktívnu príručku a usmernenia pre zrozumiteľné zdokumentovanie dátových štruktúr, procesov tvorby dát, štatistických metodológií (ak boli použité), dátových zdrojov, kontextov a ďalšie aspekty popisu dát. Ďalej je zámerom vytvoriť aj koncept publikovania a archivácie údajov verejnej správy vo forme otvorených údajov.

Zámerom výstupu je pomôcť inštitúciám verejnej správy orientovať sa vo svojich povinnostiach vo vzťahu k popisovaniu údajov, ktoré spravujú, ako i analytických výstupov, ktoré produkujú. Cieľom je zjednotiť spôsob popisovania a modelovania smerom k jednotnému Centrálnemu modelu údajov. Zároveň je potrebné odbúrať legislatívne bariéry pri manažmente údajov vo vzťahu k prepájaniu jednotlivých dátových modelov do jednotného CMÚ.

1.2 Metodika realizácie výstupu:

Realizácia dokumentu pozostávala z nasledovných krokov:

- Vytvorenia konceptu popisu analytických údajov,
- Vytvorenia konceptu popisu analytických aplikácií,
- Výberu nástrojov na popisovanie.

Tento projekt je podporený z Európskeho sociálneho fondu.

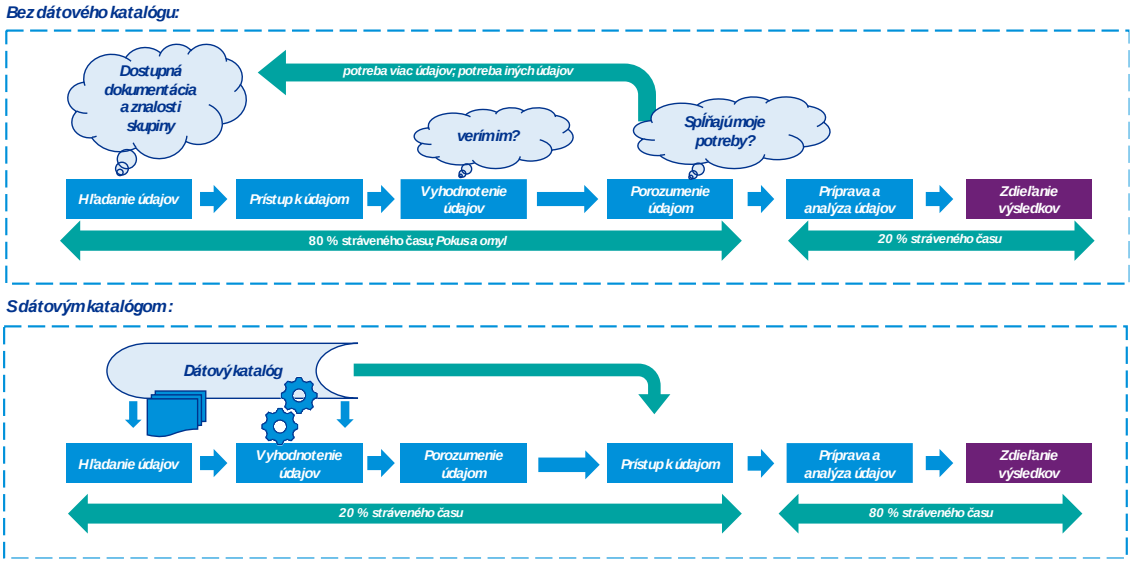
- Vytvorenia príkladu správneho popisu s prepojením na Centrálny model údajov a meta-informačný systém.

Zrealizovala sa aj právna analýza osobitných právnych predpisov v gescii iných ústredných orgánov štátnej správy mimo MIRRI, ktorá predpokladá aktívnu spoluprácu príslušných dotknutých orgánov štátnej správy, ktoré tvoria bariéru v referencovaní (Referencovanie (definícia, § 49, zákon o e-Governmente). Právna analýza zahŕňala aj právne predpisy mimo gescie MIRRI s potenciálnym alebo reálnym rozporom pri aplikácii zákonov (zákon č. 305/2013 Z. z. o e-Governmente; zákon č. 177/2018 Z. z. proti byrokracii a ich vzťah k rezortným právnym úpravám iných ústredných orgánov štátnej správy).

1.3 Správa dátových aktív

Základom pre popisovanie dátových aktív (datasetov pre analýzu a dátových produktov ako napríklad analytických aplikácií) je **dátový katalóg**. Dátový katalóg umožňuje používateľom, dátovým vedcom, dátovým analytikom, správcom údajov a ďalším odborníkom skúmať dostupné datasety, porozumieť ich obsahu a vzájomne spolupracovať a zdieľať znalosti o dátových zdrojoch. V ideálnom prípade im pomáha k väčšej sebestačnosti pri vyhľadávaní a získavaní relevantných údajov, ktoré môžu využívať v analytických aplikáciách a v iných dátových produktoch.

Organizácie, ktoré nemajú katalóg dátových aktív a ich zdrojov, sú brzdené vo svojom úsilí vykonávať analýzu a prieskum dát v rámci organizácie (Obrázok 1).. Vďaka dátovému katalógu môžu konzumenti údajov stráviť menej času snahou o pochopenie údajov a namiesto toho môžu väčšinu času venovať analýze alebo použitiu údajov na dosiahnutie zamýšľaného cieľa



Obrázok 1: Výzvy, ktoré rieši dátový katalóg

Ako ukazuje Obrázok 1, dátový katalóg poskytuje centrálne miesto, kde môžu konzumenti údajov preskúmať dostupné údaje organizácie a poskytovatelia údajov zdieľať a vymieňať si poznatky a informácie o svojich dátových zdrojoch. Bez katalógu zostávajú cenné poznatky o údajoch a dátových aktívach roztrieštené; konzumenti sú

Tento projekt je podporený z Európskeho sociálneho fondu.

odkázaní na „ad hoc“ získavanie znalostí od kompetentných ľudí a ich vlastné schopnosti, aby pochopili obsah a účel dátového aktíva. Vďaka funkčnému dátovému katalógu a procesu popisu dátových aktív konzumenti údajov môžu ľahšie objaviť nové dátové aktíva a pochopiť ich účel a poskytovatelia údajov môžu profitovať zo zníženia záťaže pri odpovediach na otázky a dopyty týkajúce sa ich údajov, ktoré by mohli byť ľahko zodpovedané pomocou popisných metadát v dátovom katalógu.

Používateľmi dátového katalógu sú:

- Správcovia údajov a dátoví kurátori, ktorí vidia, ako ich údaje zapadajú do verejnej správy ako celku, a môžu túto vyššiu perspektívu využiť pri plánovaní efektívneho manažmentu údajov a zabezpečenia kvality údajov.
- Konzumenti údajov:
  - Dátoví analytici profitujú z popisných metadát, ktoré poskytujú kontext dátových aktív, a tým pádom vedia správne interpretovať výsledky dátových analýz a reportov v analytických nástrojoch.
  - Dátoví inžinieri a dátoví vedci budú môcť objavovať, chápať a využívať existujúce údaje, pričom sa vyhnú vytváraniu duplicitných údajov alebo nesprávnemu využitiu údajov v procese dátovej vedy.
  - IT architekti, ktorí navrhujú rôzne dátové „pipelines“ pre prípravu datasetov analytických údajov a pre dátové integrácie medzi IS VS.
  - Akíkoľvek ďalší konzumenti údajov na analytické spracovanie budú používať dátový katalóg údajov na širokú škálu dopytov po údajoch.
- Vedúci zamestnanci zameraní na rozvoj dátového programu, dátovej integrácie a analytického využívania údajov získajú lepšie informácie o dátovom prostredí verejnej správy a môžu prijímať lepšie informované strategické rozhodnutia pre organizáciu.

Tento projekt je podporený z Európskeho sociálneho fondu.

2 Koncept popisu analytických údajov

Analytické údaje sa spravidla popisujú v dátovom katalógu (viď. Kapitola 1.3), ktorý slúži na inventarizáciu dátových aktív a metadáta s nimi spojené. Analytické údaje sú spravidla vo forme datasetov, ktoré môžu alebo nemusia byť vytvorené z rôznych objektov evidencie. Datasety sú súbory údajov a tabuľky, ku ktorým majú prístup zamestnanci verejnej správy, prípadne ďalšie autorizované osoby ako napríklad vedci z akademického sektora. Môžu sa nachádzať v dátovom jazere alebo dátovom sklade, či už rámci KAV alebo mimo nej, v databázach IS VS alebo v akomkoľvek inom zdieľanom dátovom zdroji. Môžu alebo nemusia byť synchronizované so základnými registrami. V dátovom katalógu teda chceme popísať takéto datasety analytických údajov, aby sa dali jednoducho dohľadať a porozumieť ich obsahu – ich dátovému modelu a jednotlivým dátovým prvkom. Tiež treba popísať dôležité a užitočné metadáta k týmto datasetom z nasledovných kategorií metadát:

- **Metadáta o ľuďoch**: opisujú role a konkrétnych ľudí, ktorí pracujú s údajmi, vrátane spotrebiteľov, dátových kurátorov, správcov, odborníkov na danú oblasť atď.
- **Metadáta vyhľadávania:** Tieto metaúdaje podporujú označovanie a kľúčové slová, ktoré pomáhajú ľuďom nájsť údaje.
- **Metadáta spracovania údajov:** Táto kategória rozpracúva rôzne transformácie a derivácie, ktoré sa uplatňujú pri manažmente údajov a dátovej integrácii počas ich životného cyklu.
- **Metadáta dodávateľov údajov:** Tieto metadáta zahŕňajú údaje získané z externých zdrojov, pretože informujú o zdrojoch a predplatnom alebo licenčných obmedzeniach spojených s údajmi.

Každý analytický dataset musí mať okrem svojho základného popisu toho, o čom je, zdokumentovanú svoju schému. Ak sa jedná o tabuľku alebo relačnú databázu, popis schémy zachytáva Tabuľka 1. Dátové prvky však môžu byť aj zložené, potom by v type bolo napísané „zložený dátový typ“ a v popise by bol odkaz na ďalšiu tabuľku, v ktorej by už boli vysvetlené jednotlivé dátové prvky.

Tabuľka 1: Popis schémy analytického datasetu – jeho jednotlivých dátových prvkov

Názov dátového prvku	Typ	Popis dátového prvku	Značka (tag)	Kardinalita	Link na termín v podnikovom slovníku
Názov alebo skratka identifikujúca dátový prvok	*Text, číslo, dátum, hodnota z číselníka, Boolean,*	*Definícia dátového prvku*	*PII.dôverné, PII.vyhradené, zdrojový ISVS, poskytovateľ, agenda a podobne*	*0..N, 1..N, 0..1, 1..1*	*Hyperlink*
DN	dátum	Dátum narodenia osoby.	PII.vyhradené isvs_191 MV SR	0..1	ontology/physical-person/dateOfBirth

Tento projekt je podporený z Európskeho sociálneho fondu.

Okrem schémy by sa pre každý dataset analytických údajov malo v dátovom katalógu nachádzať aj:

- **Ukážka vzorových údajov**, s maskovaním, ak sa jedná o PII,
- Informácia o **dátovej „lineage“** (viď kapitola 2.1),
- Ideálne aj **profilovanie údajov**, pričom rôzne dátové katalógy poskytujú rôznu úroveň sofistikovanosti profilovania údajov. Profilovanie údajov sa venuje skúmaniu, analyzovaniu a vytváraniu užitočných súhrnov údajov. Výsledkom tohto procesu je vysokoúrovňový prehľad, ktorý pomáha odhaliť problémy s kvalitou údajov, riziká a celkové trendy. Analytické algoritmy zisťujú charakteristiky analytických datasetov, ako je napríklad priemer, minimum, maximum, percentil a frekvencia.
- Kontaktné údaje vlastníka,
- Miesto pre diskusiu a spoluprácu poskytovateľov a konzumentov údajov,
- Ak sa jedná o tabuľkové údaje, tak aj „raw“ SQL dotaz používateľa na načítanie vzorových údajov a profilovanie tabuľky.

Popisovanie analytických údajov ako datasetov pre strojové učenie

Ak sa datasety používajú na strojové učenie, je kľúčom k úspechu ich správne zdokumentovať, aby boli správne využívané, neboli dezinterpretované a aby nedochádzalo k skresleniam v modeloch strojového učenia. Takéto popisovanie je však nad rámec bežného dátového katalógu a vyžaduje si žiaľ veľa manuálnej práce a spoluprácu viacerých zúčastnených strán. Preto možno tento typ popisovania považovať za nadstavbu nad dátovým katalógom a postupne ho rozvíjať podľa potrieb tímov dátových analytikov a dátových vedcov. Veľmi prepracovaným konceptom sú Dátové karty („Data Cards“) so svojím „Playbookom“[1], navrhnuté spoločnosťou Google ako ich výskumný projekt. Dátové karty sú štruktúrované súhrny základných faktov o rôznych aspektoch datasetov pre strojové učenie, ktoré zainteresované strany potrebujú v celom životnom cykle projektu na zodpovedný vývoj modelov strojového učenia a umelej inteligencie.

Šablóna dátovej karty zachytáva 13 tém, ktoré sa často diskutujú pri rozhodovaní. Tieto témy nie sú tradične zachytené v technickej dokumentácii datasetu, ktorá sa nachádza v dátovom katalógu a v dátovom slovníku. Ide o nasledujúce témy, ktorým sa podrobnejšie venujeme v kapitole 5.1:

1. Sumár datasetu – úvodné prehľadné informácie,
2. Autorstvo datasetu,
3. Prehľad o datasete a jeho charakteristikách (táto časť sa do istej miery prekrýva s tabuľkou (Tabuľka 1) a profilovaním údajov v dátových katalógoch. Avšak ak sa jedná napríklad o datasety veľkých údajov alebo iné špecifické datasety, môže byť pre bežný dátový katalóg problematické ho správne zdokumentovať),
4. Príklad dátových bodov – opäť ide o istú duplicitu s ukážkou vzorových údajov v dátovom katalógu, avšak pre strojové učenie a komplexnejšie údaje môže byť dôležité vypichnúť typický a atypický dátový bod či odľahlú hodnotu, čo nie je možné v bežnom dátovom katalógu.

[1] Zdroj: https://sites.research.google/datacardsplaybook/, Dátum referencie: 18.07.2023

5. Motivácia vytvorenia datasetu a zámer jeho využívania,
6. Prístup, doba uchovávania a vymazanie datasetu - ide o dôležitú tému obzvlášť pre datasety veľkých údajov alebo pre citlivé datasety, na ktoré sa viažu špecifické licenčné a zmluvné podmienky,
7. Pôvod a uchovávanie datasetu („Provenance“) – týka sa zberu, kritérií zberu, vzťahu k zdroju, verzii a údržby,
8. Citlivé atribúty, týkajúce sa napríklad PII,
9. Rozšírené používanie – pojednáva o tom, s akými ďalšími datasetmi možno daný dataset prepojiť a v akých modeloch strojového učenia ho možno použiť, alebo v akých modeloch už bol použitý – tu je dôležité prelinkovanie s popisom analytickej aplikácie – modelu strojového učenia, čomu sa venujeme v kapitole 3.1.2 a 5.2,
10. Aké transformácie boli vykonané na datasete – tu je opäť prienik s dátovou „lineage“ dátových katalógov a s katalogizovaním „pipelines“, tak je vhodné informácie z týchto nástrojov prepojiť s väčšou mierou detailu,
11. Aké anotácie a označovanie bolo vykonané, ak sa jedná o dataset pre strojové učenie s učiteľom,
12. Aké typy validácie boli použité,
13. Aké typy vzorkovania boli použité,

Na zabezpečenie jednoznačnosti interpretácie dátovej karty je vhodné mať prepojenie na podnikový slovník, ktorý definuje použité koncepty v danom datasete, všade tam, kde je to vhodné a užitočné. Na postupné – iteratívne a paralelné vypĺňanie týchto kariet sú definované aj súbory rôznych aktivít so zúčastnenými stranami[2]. V kapitole 5.1 sa venujeme vybranej podmnožine takéhoto popisovania analytických údajov, ktoré sa používajú ako datasety pre modely strojového učenia. Jednoduché analytické datasety vo forme tabuliek či databáz sa popisujú priamočiaro podľa tabuľky vyššie (Tabuľka 1) a automatizovane nástrojmi na popisovanie (viď kapitola 4). Analytické datasety môžu tiež vychádzať z objektov evidencie a odkazovať sa na ich dokumentáciu v MetaIS (Confluence),[3] ktorú je do budúcna potrebné previesť aj do dátového katalógu. Tiež aj vzhľadom na isté prieniky s bežným záznamom v dátovom katalógu je dobré mať aj pre datasety pre strojové učenie nejakú formu katalógového spisu v dátovom katalógu s odkazom na podrobnejší popis na základe Dátovej karty, uložený napríklad v nástroji MetaIS (Confluence).

Prepojenie popisu s podnikovým slovníkom

Správa taxonómií a ontológií sa vykonáva v nástroji VocBench[4], z ktorého sa publikuje na portál znalosti.gov.sk (spoločne tak predstavujú obdobu podnikového slovníka). Podnikový slovník pre analytické údaje musí z neho vychádzať a byť s ním automatizovane synchronizovaný. Žiaľ každý dátový katalóg má svoj vlastný dátový model metadát aj pre nejakú formu podnikového slovníka, takže treba tam tieto informácie dostať z nástroja VocBench.

[2] Zdroj: https://sites.research.google/datacardsplaybook/activities/ a https://github.com/PAIR-code/datacardsplaybook/ , Dátum referencie: 18.07.2023

[3] Zdroj: https://wiki.vicepremier.gov.sk/display/IN/OBJEKTY+EVIDENCIE, Dátum referencie: 31.07.2023

[4] Zdroj: https://znalosti.gov.sk/vocbench3/#/Home, Dátum referencie: 18.07.2023

Tento projekt je podporený z Európskeho sociálneho fondu.

Podnikový slovník v dátových katalógoch je riadený slovník, ktorý pomáha stanoviť konzistentný význam pojmov, vytvoriť spoločné chápanie a budovať znalostnú bázu. Glosárové termíny môžu tiež pomôcť pri organizácii alebo objavovaní dátových aktív a dátových prvkov. Podnikové slovníky spravidla organizujú termíny pomocou hierarchických, ekvivalentných a asociatívnych vzťahov. Glosáre sú teda zbierkou hierarchie termínov, ktoré patria do určitej domény, pričom:

- Termín glosára je špecifikovaný preferovaným termínom pre pojem alebo terminológiu (príklad – fyzická osoba, občan).
- Termín glosára musí mať jedinečnú a jasnú definíciu, aby sa zaviedlo konzistentné používanie a chápanie termínu.
- Termín môže obsahovať synonymá, iné termíny používané pre ten istý pojem (príklad - občan, klient, poberateľ dávky atď.).
- Termín môže mať podradené termíny, ktoré termín ďalej špecifikujú. Príklad: termín slovníka pobyt môže mať podradené termíny – trvalý pobyt, prechodný pobyt atď.
- Termín môže mať aj príbuzné termíny na zachytenie súvisiacich pojmov.
- Termín by mal byť prelinkovaný so zoznamom dátových aktív, prostredníctvom ktorého možno zistiť všetky dátové aktíva súvisiace s daným termínom.
- Každý termín má mať stav životného cyklu (napríklad koncept, v schvaľovaní, schválený, zneplatnený, vymazaný).
- Termín má tiež skupinu recenzentov, ktorí kontrolujú a prijímajú zmeny v slovníku v rámci životného cyklu terminológií a ontológií (CMÚ).

Termíny z glosára sa môžu používať na označovanie alebo tagovanie ako dodatočné metadáta dátových aktív na opis a kategorizáciu.

2.1 Zdokumentovanie dátovej „lineage“

Dátová „lineage“ dokumentuje cestu údajov cez IT systémy organizácií a ukazuje, ako údaje medzi nimi prúdia a ako sa cestou transformujú na rôzne účely. Využíva údaje o údajoch – metadáta, ktoré umožňujú koncovým používateľom aj odborníkom na správu údajov sledovať históriu dátových aktív a získať informácie o ich biznis význame alebo technických atribútoch. **Preto je dátová „lineage“ kľúčovou súčasťou popisu analytických údajov v dátovom katalógu.**

Záznamy o dátovej „lineage“ môžu pomôcť dátovým vedcom, dátovým analytikom a iným používateľom pochopiť údaje, s ktorými pracujú, a zabezpečiť, aby boli relevantné pre ich potreby. Dátová „lineage“ zohráva dôležitú úlohu aj v programe manažmentu údajov, v master data manažmente a pri dodržiavaní predpisov. Okrem iných aspektov týchto iniciatív zjednodušuje dva kritické postupy manažmentu údajov: analýzu základných príčin problémov s kvalitou údajov a vplyv zmien na datasety.

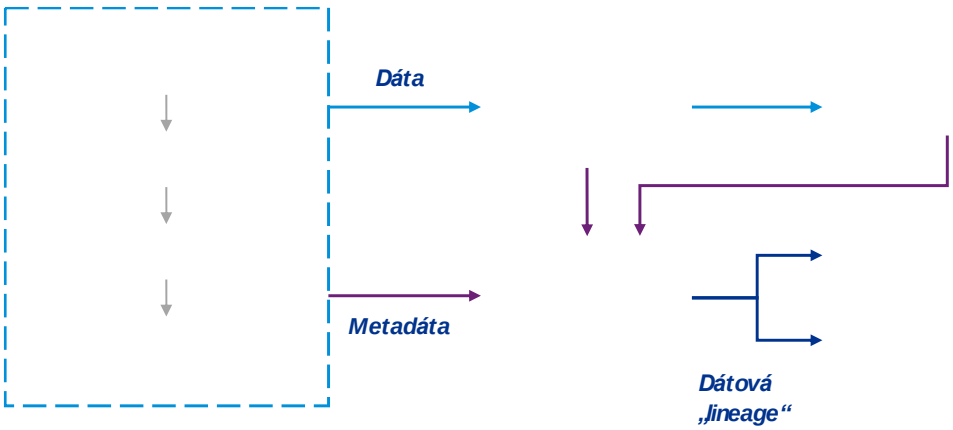
Informácie o dátovej „lineage“ sa zhromažďujú z prevádzkových systémov pri spracovaní údajov a z dátových skladov a dátových jazier, ktoré uchovávajú datasety pre aplikácie BI a analytické aplikácie. Okrem podrobnej dokumentácie sa môžu vytvárať mapy a diagramy toku údajov, ktoré graficky vizualizujú dátovú „lineage“ a mapujú ju na biznis procesy.

Prečo je dôležité venovať sa dátovej „lineage“?

Ako už bolo spomenuté, informácie o dátovej „lineage“ sú kľúčové pre manažment údajov a analytické spracovanie údajov. Bez prístupu k týmto informáciám je oveľa ťažšie naplno využiť potenciálnu hodnotu údajov. Nižšie sú uvedené niektoré ďalšie výhody:

- **Presnejšie a užitočnejšie analýzy**: Vďaka tomu, že analytické tímy a ďalší používatelia vedia, odkiaľ údaje pochádzajú a čo znamenajú, dátová „lineage“ zlepšuje ich schopnosť nájsť údaje, ktoré potrebujú na účely BI, analýz a dátovej vedy. To vedie k lepším výsledkom analýzy a zvyšuje pravdepodobnosť, že práca na analýze údajov prinesie zmysluplné informácie, ktoré budú podnetom pre dátami riadené rozhodovanie.
- **Efektívnejší manažment údajov**: Dátová „lineage“ pomáha aj pri sledovaní toku údajov a vykonávaní ďalších kľúčových oblastí procesu manažmentu údajov. Pomáha manažérom a členom tímu pre manažment údajov zabezpečiť, aby boli údaje platné, čisté a konzistentné a aby boli zabezpečené, spravované a správne používané. Pomáha aj pri riadení migrácie údajov, odbúravaní dátových síl a zisťovaní a riešení nedostatkov v analytických datasetoch a v objektoch evidencie.
- **Prísnejšie zabezpečenie údajov a ochrana súkromia**: Organizácie môžu využívať informácie o dátovej „lineage“ na identifikáciu citlivých údajov, ktoré si vyžadujú obzvlášť prísne zabezpečenie. Môžu sa použiť aj na nastavenie rôznych úrovní prístupových oprávnení používateľov na základe zásad bezpečnosti a ochrany osobných údajov a na posúdenie potenciálnych rizík pre údaje v rámci stratégie riadenia rizík organizácie.
- **Lepšie dodržiavanie predpisov**: Lepšia ochrana bezpečnosti údajov informovaná o dátovej „lineage“ môže organizáciám pomôcť zabezpečiť, aby dodržiavali zákony o ochrane osobných údajov a iné predpisy. Dobre zdokumentovaná dátová „lineage“ tiež uľahčuje vykonávanie interných auditov dodržiavania predpisov a podávanie správ o úrovni dodržiavania predpisov.

Ako ukazuje Obrázok 2, nástroje na sledovanie dátovej „lineage“, popísané v kapitole 4.2, zhromažďujú metadáta a mapujú dátové toky, ako aj vzťahy a závislosti medzi rôznymi dátovými prvkami. Metadáda sa zbierajú, ako údaje tečú rôznymi systémami, a využívajú sa na vytvorenie informácií o dátovej „lineage“.



Obrázok 2: Príklad procesu dátovej „lineage“

Dátová „lineage“ verzus klasifikácia údajov a pôvod údajov („data provenance“)

Tento projekt je podporený z Európskeho sociálneho fondu.

Dátová „lineage“ je tiež úzko prepojená s klasifikáciou údajov a pôvodom údajov („data provenance“), dvoma ďalšími procesmi manažmentu údajov.

Klasifikácia údajov: Ide o zaradenie údajov do rôznych kategórií na základe ich vlastností, predovšetkým na účely zabezpečenia a dodržiavania predpisov. Klasifikácia sa používa na kategorizáciu údajov na základe toho, aké sú citlivé (tomu sa venujeme v dokumente „*1.1.5 Štandardizácia anonymizácie údajov*“). Týmto spôsobom sa oddeľujú objekty evidencie a datasety, ktoré potrebujú vyššiu úroveň zabezpečenia a prísnejšie kontroly prístupu, od tých, ktoré ju nepotrebujú. Dátová „lineage“ poskytuje informácie o datasetoch, ktoré môžu pomôcť pri ich klasifikácii.

Pôvod údajov („data provenance“): Niekedy sa považuje za synonymum dátovej „lineage“, ale alternatívne sa považuje za užšie zameranú na pôvod údajov vrátane ich zdrojového systému a spôsobu ich generovania. V tomto kontexte môžu dátová „lineage“ a pôvod údajov fungovať ruka v ruke, pričom pôvod údajov poskytuje dokumentáciu na vysokej úrovni o tom, odkiaľ údaje pochádzajú a čo s ich pôvodom súvisí.

Dátová „lineage“ a manažment údajov

Veľkou výzvou je preklenúť medzeru medzi definovaním politík manažmentu údajov a ich implementáciou. Práve tu prichádza na rad dátová „lineage“. Tým, že dokumentuje zdroje a toky údajov, umožňuje tímom pre manažment údajov monitorovať, ako sa údaje pohybujú v systémoch, ako sa upravujú a používajú. Informácie o dátovej „lineage“ im pomáhajú zaistiť, aby bola zavedená správna bezpečnosť údajov a kontrola prístupu a aby sa údaje uchovávali, udržiavali a používali v súlade so zásadami ich manažmentu.

Dátová „lineage“ môže tiež uľahčiť špecifické úlohy súvisiace s manažmentom údajov, ako už bolo spomenuté. Pri analýze príčin chýb údajov poskytujú záznamy o dátovej „lineage“ prehľad o postupnosti fáz spracovania, ktorými dataset prechádza. Úrovne kvality možno skúmať v každej fáze, aby sa zistilo, odkiaľ pochádzajú chyby údajov. Pri práci smerom dozadu od miesta, kde bola chyba identifikovaná prvýkrát, môže správca údajov skontrolovať, či údaje v predchádzajúcich bodoch spĺňali očakávania alebo či vtedy už obsahovali chybu. Určením fázy, v ktorej boli údaje pri vstupe v súlade s požiadavkami, ale pri výstupe boli chybné, môžu zamestnanci zapojení do programu manažmentu údajov odstrániť hlavnú príčinu chyby namiesto toho, aby len opravovali zlé údaje.

Dátová „lineage“ je užitočná aj pri vykonávaní analýzy vplyvu, aby sa udržal prehľad o problémoch spôsobených zmenami formátov a štruktúr zdrojových údajov, čo je bežný problém v dnešných čoraz dynamickejších dátových prostrediach. Keď sa údaje zmenia, môže to mať nezamýšľané následky v ďalšom priebehu. Tým, že správca údajov pracuje od bodu vytvorenia alebo zberu údajov smerom vpred, môže sa spoľahnúť na dokumentáciu o dátovej „lineage“, ktorá mu pomôže sledovať závislosti údajov a identifikovať fázy spracovania, ktoré sú ovplyvnené zmenami. Tieto fázy sa potom môžu prepracovať tak, aby sa prispôsobili zmenám a zabezpečili konzistentnosť údajov v rôznych systémoch.

Kľúčové techniky dátovej „lineage“

Na zhromažďovanie a dokumentovanie informácií o dátovej „lineage“ možno použiť rôzne techniky. Nemusia sa nevyhnutne vylučovať - organizácia môže používať viac ako jednu techniku v závislosti od potrieb a povahy dátového prostredia. Medzi dostupné techniky patria tieto:

Tento projekt je podporený z Európskeho sociálneho fondu.

- **Označovanie údajov** („**Data tagging**“): Skúmaním metadát možno na datasety použiť značky, ktoré ich pomáhajú opísať a charakterizovať na účely dátovej „lineage“. Označovanie môžu vykonávať správcovia údajov, iní členovia tímu pre manažment údajov a koncoví používatelia manuálne alebo automaticky pomocou softvéru. Napríklad nástroje na určovanie dátovej „lineage“ a funkcionality na určovanie dátovej „lineage“ zabudované do softvéru na manažment údajov často obsahujú automatizované algoritmy, ktoré môžu používatelia spustiť na označovanie datasetov.
- **„Lineage“ založená na vzoroch**: Tento prístup hľadá vzory vo viacerých datasetoch, napríklad podobné dátové prvky, riadky a stĺpce. Ich prítomnosť naznačuje, že datasety spolu súvisia a môžu byť súčasťou toku údajov, zatiaľ čo rozdiely v hodnotách alebo atribútoch údajov sú znakom toho, že údaje boli transformované pri prechode z jedného systému do druhého. Transformácie údajov a dátové toky sa potom môžu zdokumentovať ako súčasť záznamov o dátovej „lineage“.
- **„Lineage“ založená na parsovaní**: V tomto prípade nástroje na analýzu dátovej „lineage“ analyzujú logiku transformácie údajov, „runtime“ logy, toky integrácie údajov a iný kód na spracovanie údajov s cieľom identifikovať a extrahovať informácie o dátovej „lineage“. Parsovanie ponúka komplexný prístup k sledovaniu dátovej „lineage“ v rôznych systémoch a môže byť presnejšie ako „lineage“ založená na vzoroch, ale je aj zložitejšie.

Ďalší prístup je plne manuálny: rozhovory s biznis používateľmi, analytikmi BI, dátovými vedcami, správcami údajov, vývojármi dátovej integrácie a ďalšími zamestnancami a dodávateľmi o tom, ako sa údaje pohybujú v systémoch a ako sa používajú a upravujú. Získané informácie sa môžu použiť na zmapovanie tokov a transformácií údajov, možno ako východiskový bod pre iniciatívu dátovej „lineage“ pred zavedením automatizovanejších techník.

Osvedčené postupy v oblasti dátovej „lineage“

Tu je niekoľko osvedčených postupov, ktoré pomôžu udržať proces dátovej „lineage“ na správnej ceste a zabezpečiť, aby poskytoval presné a užitočné informácie o datasetoch:

- **Od začiatku zapojte vedúcich zamestnancov a používateľov**: Programy manažmentu údajov potrebujú podporu a účasť vedúcich a výkonných zamestnancov, aby boli úspešné, a to isté platí aj pre dátovú „lineage“. Podpora zo strany vrcholového manažmentu je nevyhnutná na získanie súhlasu a financovania. Mali by sa zapojiť aj relevantní riaditelia a používatelia, aby sa zabezpečilo, že tímy manažmentu údajov plne chápu, ako sa údaje používajú v analytických, prípadne v agendových procesoch, a aby sa overilo, že informácie o dátovej „lineage“ sú relevantné a platné.
- **Zdokumentujte biznis aj technickú dátovú „lineage“ biznis aj technických údajov**: Biznis dátová „lineage“ sa zameriava na to, odkiaľ údaje pochádzajú, ako prúdia a aký je ich agendový kontext. Technická dátová „lineage“ poskytuje podrobnosti o transformáciách údajov, integráciách a potrubiach („pipelines“), ako aj kombináciu pohľadov na „lineage“ na úrovni tabuliek, stĺpcov a dotazov. Zdokumentovanie oboch poskytuje užitočné informácie používateľom a analytickým tímom na jednej strane a dátovým architektom, odborníkom zaoberajúcim sa modelovaním údajov a CMÚ, analytikom kvality údajov a ďalším IT odborníkom na strane druhej.

Tento projekt je podporený z Európskeho sociálneho fondu.

- **Prepojenie dátovej „lineage“ so skutočnými agendovými, analytickými a IT potrebami**: Dátová „lineage“ by nemala byť akademickým cvičením. Aby priniesla očakávané prínosy, musí pomôcť umožniť lepšie rozhodnutia a stratégie, ako aj efektívnejší manažment údajov, lepšiu kvalitu údajov a ďalšie prínosy v oblasti manažmentu údajov. V opačnom prípade pôjde pravdepodobne o zbytočnú investíciu.
- **Uplatnite prístup k dátovej „lineage“ na úrovni celej verejnej správy**: Proces dátovej „lineage“, ktorý sa zameriava len na niektoré datasety, tiež nebude taký prínosný, ako by mohol byť. Aby sa to naozaj oplatilo, malo by ísť o komplexné úsilie, ktoré zahŕňa všetky údaje verejnej správy, ktoré sa zdieľajú a analyzujú, pričom základom práce na dátovej „lineage“ by mal byť jeden archív metadát.
- **Vytvorte dátový katalóg s vloženými informáciami o dátovej „lineage“**: Nájsť a pochopiť relevantné údaje je pre používateľov BI a analytiky často veľkou výzvou. Vytvorením dátového katalógu im môže tím pre manažment údajov poskytnúť súpis dostupných dátových aktív, ktorý obsahuje aj informácie o dátovej „lineage“.

2.2 Proces popisu analytických údajov v dátovom katalógu

Pre dôslednú implementáciu dátového katalógu analytických údajov (s využitím nástrojov popísaných v kapitole 4), ktorý bude zachytávať popis analytických údajov podľa konceptu v tejto kapitole, je potrebné vykonať týchto desať krokov v oblasti plánovania a budovania dátového katalógu.

1. Identifikujte potrebné metadáta a zdokumentuje ich hodnotu v rámci manažmentu údajov

Všetky efektívne programy manažmentu údajov sú podporované správou biznis aj technických metadát. Metadáta dávajú obsahu datasetov kontext a poskytujú informácie, vďaka ktorým sú údaje použiteľné a zrozumiteľné v celej verejnej správe. Správna správa metadát pomáha organizáciám riadiť svoje údaje s cieľom zlepšiť kvalitu údajov a zvýšiť prevádzkovú efektívnosť prostredníctvom implementácie politík, metodík a štandardov v oblasti údajov. Zdokumentovanie týchto očakávaných prínosov jednotlivých typov metadát môže byť súčasťou odôvodnenia nákladov a prínosov pre dátový katalóg.

Na zabezpečenie zberu správnych údajov je potrebné odpovedať na základnú otázku: ktoré metadáta dokumentovať?

Ktoré metadáta zdokumentovať?

Naplnenie dátového katalógu tvarom, štruktúrou a sémantikou údajov je prvým krokom pri jeho vytváraní. Väčšina používateľov údajov, ako sú dátoví vedci, dátoví inžinieri, analytici a ďalší, sa na údaje odvoláva v zmysle schémy alebo tabuľky, v ktorej sa údaje nachádzajú. Dátový katalóg musí poskytovať odpovede na nasledujúce otázky:

- Kde nájdem pseudonymizovanú evidenciu občanov, ktorí poberajú dávky v hmotnej núdzi? – V dátovom katalógu bude popísaný konkrétny dataset.
- Ako sa generujú faktúry? - Faktúra obsahuje jednu alebo viacero objednávok. Skontrolujte, či sa údaje nachádzajú v tabuľkách "invoices" a "orders". V prípade, že bola faktúra zaplatená, platbu nájdete v tabuľke "payments".

Tento projekt je podporený z Európskeho sociálneho fondu.

2. Identifikujte spôsoby využitia rôznych nástrojov na správu metaúdajov

Hoci sa pojmy dátový katalóg, podnikový slovník a dátový slovník niekedy používajú zameniteľne, nie je to to isté. Podnikový slovník definuje biznis pojmy používané v celej verejnej správe a poskytuje autoritatívny zdroj na ich pochopenie. Dátový slovník poskytuje technické informácie o údajoch, ktoré môžu zahŕňať vlastnosti takých atribútov, ako sú typ údajov, dĺžka, platné hodnoty, predvolené hodnoty, vzťahy s inými dátovými poľami, pravidlá transformácie údajov, biznis pravidlá a obmedzenia. Dátové slovníky podporujú používanie fyzických metadát, ktoré obsahujú podrobnosti o tom, kde sa údaje nachádzajú a ako sú uložené. Podnikové slovníky sú zamerané na biznis aspekty manažmentu údajov a na ich sémantické porozumenie a interoperabilitu, zatiaľ čo dátové slovníky sú doménou technických správcov údajov. Dátový katalóg môžu používať doménoví aj technickí správcovia, pretože zahŕňa aspekty ostatných dvoch nástrojov.

3. Navrhnite model predmetnej oblasti pre svoje údaje

Efektívny dátový katalóg sleduje využitie údajov na biznis vrstve, nielen technickú implementáciu systémov. Model predmetnej oblasti („Subject Area Model (SAM)“)[5], ktorý definuje rôzne predmetné oblasti pre údaje organizácie a biznis koncepty jednotlivých agend, ktoré sú v nich obsiahnuté, ukazuje biznis používateľom umiestnenie údajov bez obmedzenia aplikáciami, súbormi alebo databázami. SAM slúži ako základ návrhu a modelovania dátovej architektúry a mal by z neho vychádzať dátový katalóg aj podnikový slovník.

4. Vytvorte podnikový slovník

Tento slovník je pokrytý procesom a metodikami tvorby Centrálneho modelu údajov, ktorý je popísaný v dokumente „*1.1.2 Štandardizácia pre modelovanie údajov*“.

5. Vytvorte dátový slovník

Dátový slovník by mal obsahovať opisy a mapovania každej dátovej tabuľky alebo datasetu a všetkých ich metadátových entít. Potom sa stane základom pre import metadát do dátového katalógu. Aj v tomto prípade sú dôležití dátoví kurátori a správcovia údajov, pretože tí poskytnú usmernenie o biznis metadátach, ktoré sa majú použiť v dátovom katalógu - podľa zdroja, konceptu a tematickej oblasti. Momentálne sa dátový slovník buduje v nástroji Talend, ktorý je základom platformy CIP. Venujeme sa mu tiež v dokumente „*1.1.2 Štandardizácia pre modelovanie údajov*“.

6. Získajte metadáta z databáz a iných zdrojov údajov

Dátové katalógy používajú metadáta na identifikáciu tabuliek a datasetov pre používateľov. Katalóg prehľadáva databázy organizácie a iné dátové úložiská a načítava súvisiace metadáta do svojho súpisu dátových aktív. Predtým, ako organizácia začne budovať dátový katalóg, je potrebné identifikovať a zaznamenať zdroje metadát. Je to dôležitý krok a podobne ako predchádzajúce dva si vyžaduje, aby organizácia mala spoľahlivý program manažmentu údajov. V tomto prípade sú potrební správcovia údajov

[5] Zdroj: https://opengovernance.odpi.org/coco-pharmaceuticals/scenarios/defining-subject-areas/, Dátum referencie: 19.07.2023

Tento projekt je podporený z Európskeho sociálneho fondu.

a IS VS, ktorí poskytnú prehľad o správnych zdrojoch údajov, ktoré sa majú použiť, ako aj o správnych zdrojov metaúdajov. Moderné dátové katalógy poskytujú možnosti priamo sa napojiť na väčšinu používaných databáz alebo na webové aplikácie či systémy ako SAP. Jediná cesta udržateľnosti získavania aktuálnych metadát je cez automatizované postupy, ideálne bez potreby ľudského zásahu.

Všetky hlavné databázy a dátové úložiská (napr. AWS S3) majú k dispozícii rozhrania API, ktoré umožňujú extrahovať metadáta, ktoré predstavujú tvar a sémantiku údajov. Existujú scenáre, v ktorých sa nedá pripojiť priamo k databáze. Napríklad sa nemôžu zverejniť citlivé údaje alebo treba zmapovať spravovanú databázu, ktorá nie je verejne dostupná. V takýchto scenároch by mala existovať možnosť použiť vzorové datasety a exporty z dátového skladu ako alternatívu k priamemu pripojeniu k databáze.

V najhorších scenároch, keď všetko zlyhá, by mala byť vybudovaná kapacita rýchlo zachytiť údaje manuálne bez automatizácie. Vzhľadom na frekvenciu zmien všetkých klientskych knižníc rozdielnych databáz nemožno zaručiť dokonalý proces alebo nástroj. Preto mať možnosť opraviť problémy manuálne je rozhodujúce pre vybudovanie robustného dátového katalógu.

7. Priraďte kontaktné miesta

Po vytvorení dátového katalógu je dôležité určiť, kto sú dôležité osoby pre jednotlivé dátové aktíva. Preto je dôležité priradiť dátovým aktívam vlastníkov. To umožní konzumentom údajov s ďalšími otázkami alebo požiadavkami osloviť správnu osobu.

Otázky rôznych konzumentov údajov možno rozdeliť do dvoch kategórií:

1. Biznis kontext tohto dátového aktíva – ukážka otázky:
   - „Čo znamená hodnota z číselníka „nezaradený“ pre tento dátový prvok?“
2. Technické atribúty pre dátové aktívum – ukážka otázky:
   - „Kto môže pridať tento nový dátový prvok do ontológie a dátového modelu?“

Dátový katalóg môže mať mnoho typov vlastníkov (napríklad správca údajov, dátový kurátor, technický vlastník, biznis vlastník a podobne). Dôležitú úlohu však zohrávajú správca údajov, prípadne dátový kurátor a technický vlastník. Správca údajov alebo dátový kurátor dáva používateľom vedieť, na koho sa majú obrátiť so všetkými informáciami súvisiacimi s biznis kontextom údajov a s ich správnou sémantickou interpretáciou. Technický vlastník má medzitým k dispozícii odpovede na technicky orientované otázky, ktoré môžu mať konzumenti údajov pri ich analýze (alebo pri téme dátovej integrácie).

Pri vytváraní dátového katalógu možno vlastníkom prideliť úlohy. Tieto úlohy majú zabezpečiť, aby bol dátový katalóg dobre zdokumentovaný a užitočný pre ostatných členov tímu a konzumentov.

8. Profilujte údaje s cieľom poskytnúť používateľom štatistiky

Tieto profily sú informačné súhrny, ktoré používateľom dátového katalógu vysvetľujú metadáta. Napríklad profil databázy často obsahuje počet tabuliek, súborov a počty riadkov. V podnikovom slovníku by sa profilovanie údajov zameralo na biznis metadáta

Tento projekt je podporený z Európskeho sociálneho fondu.

a ich používanie v rámci organizácie správcami a používateľmi biznis údajov. Súčasťou profilovania majú byť aj prehľady o tom, ktoré údaje sú citlivé, v súlade s kategorizáciou a klasifikáciou uvedenou v dokumente „*1.1.5 Štandardizácia anonymizácie údajov*“.

9. Identifikujte vzťahy medzi zdrojmi údajov

Objavte súvisiace údaje vo viacerých dátových úložiskách a zapracujte tieto informácie do dátového katalógu, aby používatelia mohli pochopiť vzťahy. Napríklad dátový analytik môže potrebovať konsolidované údaje o poberateľoch sociálnych dávok pre analytickú aplikáciu. Prostredníctvom dátového katalógu a dátového slovníka môže analytik zistiť, že päť datasetov v piatich rôznych systémoch obsahuje relevantné údaje.

10. Dokumentujte každú interakciu

Keď začnete dokumentovať svoje údaje v dátovom katalógu, môže sa vám množstvo informácií, ktoré chcete zachytiť, zdať spočiatku priveľké. Ak predpokladáme, že existujú v dátovej vrstve len dve databázy a každá z nich má niekoľko desiatok tabuliek. Každá tabuľka má ďalej niekoľko polí. V tomto okamihu sa zdá, že sa už pozeráte na niekoľko tisíc dátových aktív. Preto treba začať výberom jednej metodiky a časom pomaly pridávať dokumentáciu. Tým sa zabezpečí, že sa v priebehu niekoľkých mesiacov dosiahne určité percento pokrytia, možno 90 % alebo menej. Medzi bežné metodiky patria napríklad:

- **„Vždy, keď sa o tom dozviete, zdokumentujte to**“: Každý by mal prevziať zodpovednosť za aktualizáciu dátového katalógu, keď sa dozvie niečo nové, čo ešte nebolo zdokumentované.
- **„Keď dôjde k zmene zdrojového kódu, zmeňte dokumentáciu“:** Keď tímy zverejnia nové funkcionality systémov alebo dátových „pipelines“, príslušní členovia tímu by mali aktualizovať aj dokumentáciu údajov.
- **„Vyhraďte členom tímu čas“:** Požiadajte každého člena tímu, aby venoval jednu hodinu týždenne alebo možno 15 minút každé ráno dátovému katalógu. To im umožní pridať novú dokumentáciu pre dátové aktíva, ktoré dobre poznajú, alebo preskúmať tie, ktoré nepoznajú.

Všetky dátové aktíva by mali mať v rámci dátového katalógu bohatú textovú dokumentáciu, aby mali používatelia možnosť zvýrazniť kľúčové body. Dátové katalógy by mali používateľom poskytovať aj možnosť zoskupovať aktíva do spoločných celkov. To sa môže uskutočniť prostredníctvom označovania údajov. Ak napríklad chcete mať možnosť zobraziť správu o všetkých osobne identifikovateľných údajoch („personally identifiable information (PII)“), môžete označiť všetky tabuľky a polia, ktoré obsahujú takéto údaje, symbolom „PII“.

Ak dátový katalóg umožní používateľom viesť konverzáciu o údajoch, môže to viesť k ďalším prínosom. Keď má konzument údajov alebo používateľ katalógu nejakú otázku týkajúcu sa údajov a táto otázka je zodpovedaná - potom by mala byť otázka, odpoveď a konverzácia, ktorá viedla k odpovedi, zdokumentovaná v rámci katalógu. To umožňuje ďalšiemu používateľovi katalógu alebo konzumentovi údajov s podobnou otázkou, aby si mohol pozrieť predchádzajúcu konverzáciu a pochopiť kontext okolo odpovede. Ušetrí sa tým čas, aby sa nemuseli opätovne hľadať odpovede na tie isté otázky.

11. Zachyťte informácie o „lineage“ údajov

Tento projekt je podporený z Európskeho sociálneho fondu.

Nástroje na extrakciu, prenos a načítanie (ETL) sa používajú na extrakciu údajov zo zdrojových systémov, ich transformáciu a čistenie a načítanie do cieľového dátového úložiska (viac v dokumente „*1.1.6 Štandardizácia dátovej transformácie*“). Pri budovaní dátového katalógu metadáta zozbierané počas procesu ETL obsahujú dokumentáciu o „lineage“ údajov (kapitola 2.1), ktorá sleduje, odkiaľ údaje pochádzajú, ako prúdia cez systémy, a ďalšie informácie. Dátová „lineage“ pomáha používateľom pochopiť dátové aktíva v katalógu a umožňuje správcom údajov a analytikom vysledovať chyby údajov až k ich prvotnej príčine v zdrojových systémoch alebo v ich transformácii preskúmaním toku údajov.

12. Zaručte aktuálnosť dátového katalógu

Jednou z hlavných výziev, ktorým organizácie čelia, je udržiavať dátový katalóg aktuálny. Vývojári zvyčajne raz za čas zmenia štruktúru databáz a často vytvárajú nové dátové „pipelines“. Dátoví vedci a analytici spravidla rovnako často vytvárajú dátové kocky alebo presúvajú údaje medzi analytickými prostrediami a vytvárajú nové analytické aplikácie. S odvolaním sa na tieto vzory by mal dátový katalóg tieto zmeny podľa možnosti automaticky identifikovať a podľa toho sa aktualizovať.

Na zabezpečenie aktuálnosti dátového katalógu je dôležitá určitá interakcia používateľov katalógu, aby sa dvakrát skontrolovala kvalita a pravdivosť informácií. Dátový katalóg údajov môže využívať funkcionalitu správy, ktoré používateľov navedú na základe domnienky, že upraviť zastaranú dokumentáciu údajov.

13. Usporiadajte dátový katalóg na použitie konzumentmi údajov

Každá organizácia používa dátový katalóg podľa svojich požiadaviek a potrieb. Preto je potrebné stanoviť štandardy a normy pre spôsob, akým sa má dátový katalóg využívať naprieč verejnou správou. Tu je dôležité poznamenať, že spôsob, akým sa plánuje používať dátový katalóg, do veľkej miery ovplyvní spôsob popisu dátových aktív a vytvárania dokumentácie. Preto ak neviete, ako budú používatelia a konzumenti údajov používať dátový katalóg, je vysoko pravdepodobné, že čas, ktorý sa strávi dokumentovaním údajov, povedie k neadekvátnym výsledkom.

Väčšina databáz a systémov je navrhnutá na používanie IT expertmi. Dátové katalógy a podnikové slovníky by mali byť navrhnuté pre rôznych konzumentov údajov – od používateľov cez dátových analytikov až po technológov. Aj v tomto prípade by štruktúra katalógov a slovníkov mala vychádzať z modelu predmetnej oblasti (SAM), ktorý sa navrhol skôr v procese. Okrem toho by tieto nástroje mali byť prístupné nielen prostredníctvom počítačov, ale aj tabletov a smartfónov. Dátový slovník môže byť na rozdiel od dátového katalógu a podnikového slovníku organizovaný podľa funkčných oblastí a aplikácií vzhľadom na technickú povahu jeho obsahu.

Niektoré bežné postupy, ktoré možno podniknúť na optimalizáciu interakcie s dátovým katalógom, sú nasledovné:

- Nastavte štandardizované formáty dokumentácie a jej používanie v rámci databáz, schém, dátových prvkov a dátovej „lineage“.
- Určite kľúčové moduly, z ktorých sa dá naučiť používanie dátového katalógu v praxi, a označte aktíva zahrnuté v každom module spoločnou témou.

Tento projekt je podporený z Európskeho sociálneho fondu.

- Vytvorte metodické pokyny na základe štandardov týkajúce sa používania katalógu údajov. Tým sa hlboko zakorení kultúra údajov medzi členmi tímu.

2.2.1 Osvedčené postupy na vytvorenie katalógu údajov

Budovanie dátového katalógu, ako aj podnikového slovníka a dátového slovníka a ich následné využívanie na zhromažďovanie, organizovanie a kurátorovanie metadát sú úlohy, na ktorých by sa mali podieľať tímy z IT aj priamo z agendy organizácie. Tým sa zabezpečí, že metadáta budú zamerané na potreby biznis používateľov a umožnia ich konzistentnú správu v rámci celej verejnej správy.

Okrem toho sú k dispozícii ďalšie osvedčené postupy pre tvorbu dátového katalógu, ktoré by organizácie mali mať na pamäti:

- Zahrnúť oprávnenia používateľov, monitorovanie používania, označovanie citlivých údajov a ďalšie opatrenia na ochranu bezpečnosti a súkromia údajov (ako je popísané v dokumentoch „*1.1.3 Štandardizácia pre bezpečnosť a ochranu údajov*“ a „*1.1.5 Štandardizácia anonymizácie údajov*“).
- Podporiť spoluprácu prostredníctvom funkcionalít, ako je možnosť hodnotiť a komentovať údaje a „chatovať“ s ostatnými používateľmi katalógu.
- Vypracovať program školení pre koncových používateľov, aby sa oboznámili s dátovým katalógom a mohli ho efektívne používať.
- Vytvoriť proces na aktualizáciu katalógu podľa toho, ako sa menia dátové aktíva, biznis požiadavky a procesy pri výkone agend.

Efektívne plánovanie, vývoj a implementácia dátového katalógu môže priniesť správu metadát do programu manažmentu údajov a poskytnúť trvalú hodnotu tým, že podporí lepšie pochopenie dátových aktív verejnej správy a uľahčí všetkým ich vyhľadávanie, prístup k nim a ich používanie.

Tento projekt je podporený z Európskeho sociálneho fondu.

3 Koncept popisu analytických aplikácií

Rozlišujeme dve základné analytické aplikácie:

1. Analytický report v analytickom nástroji ako PowerBI[6] alebo Qliksense alebo zostava z informačných panelov BI,
2. Modely strojového učenia.

V posledných rokoch sa analytický reporting stal jednou z najdôležitejších súčastí BI na svete a inšpiroval spoločnosti v rôznych odvetviach k strategickejšiemu mysleniu. Komplexnosť analytických informácií už nemožno jednoducho zachytiť do jednej tabuľky, napríklad v nástroji Excel. Je čoraz ťažšie vytvoriť a používať jediný report a komunikovať širokú škálu dôležitých poznatkov medzi oddeleniami, sekciami, organizáciami či inými zainteresovanými stranami. Preto organizácia potrebuje správny analytický report vo forme analytickej aplikácie, ktorá vyfiltruje dôležité údaje a poskytne interaktívnu vizualizáciu kľúčových poznatkov. Analytický report využíva kvalitatívne a kvantitatívne údaje organizácie na analýzu a hodnotenie plnenia stratégie alebo vykonávania procesov v agendách, pričom umožňuje zamestnancom prijímať rozhodnutia založené na údajoch a analýzach. Hoci analytický report vychádza zo štatistických a historických údajov a môže poskytnúť prognózu vývoja konkrétneho problému, jej použitie je rozšírené aj pri analýze aktuálnych údajov.

Tvorba takýchto reportov ako analytických aplikácií by nemala byť určená len pre špecializovaných analytikov, ktorí dokážu pracovať s dátami a interpretovať zložité informácie. Vďaka samoobslužným nástrojom BI (viac o nich v dokumente *„4.2.1 Koncept pre zavádzanie analytického spracovania údajov do praxe“*), ktoré rozširujú znalosti všetkých zamestnancov, sa analytické reporty ako analytické aplikácie môžu stať jedným z neoceniteľných nástrojov pre pokrov v oblasti riadenia a rozhodovania na základe údajov.

Analytické aplikácie založené na strojovom učení a umelej inteligencii môžu byť v súčasnosti aj súčasťou samoobslužných BI nástrojov. Pre hľadanie špecifických riešení na mieru sa však stále intenzívne využívajú analytické aplikácie v podobe modelov strojového učenia, spravidla napísaných v jazyku R alebo Python (viac o nich v dokumente *„4.2.1 Koncept pre zavádzanie analytického spracovania údajov do praxe“*).

Analytické aplikácie sa môžu tvoriť nad:

- Otvorenými údajmi,
- Údajmi verejnej správy, ktoré obsahujú citlivé údaje, avšak sú pseudonymizované a anonymizované pre ochranu súkromia (na tento typ údajov slúži KAV),
- Údajmi tretích strán, ktoré majú rôzne licenčné podmienky, avšak sú už pseudonymizované a anonymizované,
- Údaje so senzorových sietí alebo iných technológií, ktoré predstavujú plynulý tok údajov („data streaming“).

[6] Zdroj: https://www.youtube.com/watch?v=mRbeLl0TO14, Dátum referencie: 28.07.2023

Tento projekt je podporený z Európskeho sociálneho fondu.

3.1.1 Koncept popisu analytickej aplikácie v analytickom nástroji

Tradičné typy analytických reportov zvyčajne pozostávajú z:

- titulnej strany,
- obsahu,
- úvodu,
- metodiky,
- hlavnej časti,
- záverov,
- odporúčaní,
- a bibliografie.

S dynamickým, interaktívnym a samoobslužným softvérom na tvorbu reportov však môže byť štruktúra oveľa jednoduchšia a ucelenejšia. Ak je report zložitejší, možno zachovať všetky konvenčné informácie. Ale interaktívnosť reportu umožní preskúmať fakty naživo – napríklad prispôsobením časovej osi grafu, filtrovaním údajov alebo zmenou výpočtu vzorca KPI.

Jedným z najdôležitejších krokov je výber správneho typu grafu. Moderný dátový report ponúka množstvo interaktívnych grafov a vizualizácií, ktoré možno využiť vo svoj prospech a spraviť vďaka nim informácie obsiahnuté v údajoch zrozumiteľné, prehľadné a správne interpretovateľné. Ak sa vyberú správne typy grafov - také, ktoré reprezentujú informácie, ktoré chcete pomocou analytického reportu o údajoch sprostredkovať – zvýši sa tým efektivita komunikácie a produktivita pri využívaní údajov. Medzi bežné typy grafov patria interaktívne stĺpcové grafy, čiarové grafy, bublinové grafy, plošné grafy a mapy.

Okrem práce so správnymi typmi grafov je používanie dynamických údajov v reálnom čase jedným zo základných kameňov úspechu analytického reportovania. Momentálne však dátová vrstva verejnej správy nepodporuje takéto spracovanie údajov, preto to uvádzame v kapitole 6 ako oblasť dobrej praxe a ďalšieho rozvoja.

Pokiaľ ide o aspekt dizajnu analytického reportu, jasný, stručný „layout“ s vyváženou kombináciou vizuálnych prvkov je správnou cestou. Pre výkonné analytické reporty treba zabezpečiť, aby „layout“ poskytoval jasné odpovede na otázky spojené s KPI a ďalšími analytickými výstupmi, ktoré report ponúka. Mali by ste sa vyhnúť tomu, aby ste do akýchkoľvek analytických reportov nabaľovali príliš veľa grafov a „widgetov“, pretože to len odvádza pozornosť od najcennejších informácií, ktoré sú jednoznačne interpretovateľné. V záujme maximálneho úspechu rozhodovania založeného na údajoch sa treba tiež zamerať na dodržiavanie logického formátu, ktorý používateľom umožní na prvý pohľad získať užitočné informácie. Pridanie tabuliek v spodnej časti stránky umožní dosiahnuť tento logický formát, pretože zvyčajne poskytujú väčšiu hodnotu ako grafy, diagramy alebo podobné elementy.

Okrem množstva rôznych typov analytických prehľadov existuje aj mnoho typov dynamických KPI, ktoré možno použiť. Vizuálne bohatá a interaktívna povaha týchto KPI poskytuje prístup k množstvu neoceniteľných faktov, a to minulých, aktuálnych (v reálnom čase) a prognózovaných do budúcnosti. Na to, aby formát reportu v fungoval optimálne, je nevyhnutné vybrať vhodnú šablónu KPI, ktorá bude správne počítať to, čo

Tento projekt je podporený z Európskeho sociálneho fondu.

je potrebné, a používať pritom správne údaje metodicky dobrým spôsobom. Pri tvorbe analytického reportu preto treba dobre zvážiť, ktoré trendy chcete odhaliť alebo porovnať, a podľa toho vybrať KPI. Správnym výberom ukazovateľov získate možnosť odhaliť silné a slabé stránky a zároveň sprístupníte svoje informácie ostatným zainteresovaným stranám, interným aj externým. Vďaka práci s preddefinovanými šablónami budete môcť poskytovať nepretržitý prístup k najdôležitejším panelom KPI. Ideálne je, keď analytický report podáva informácie vo forme zrozumiteľného príbehu. Ak je formát analytického reportu vytvorený s ohľadom na rozprávanie príbehov, úsilie zaviesť rozhodovanie na základe údajov bude oveľa účinnejšie.

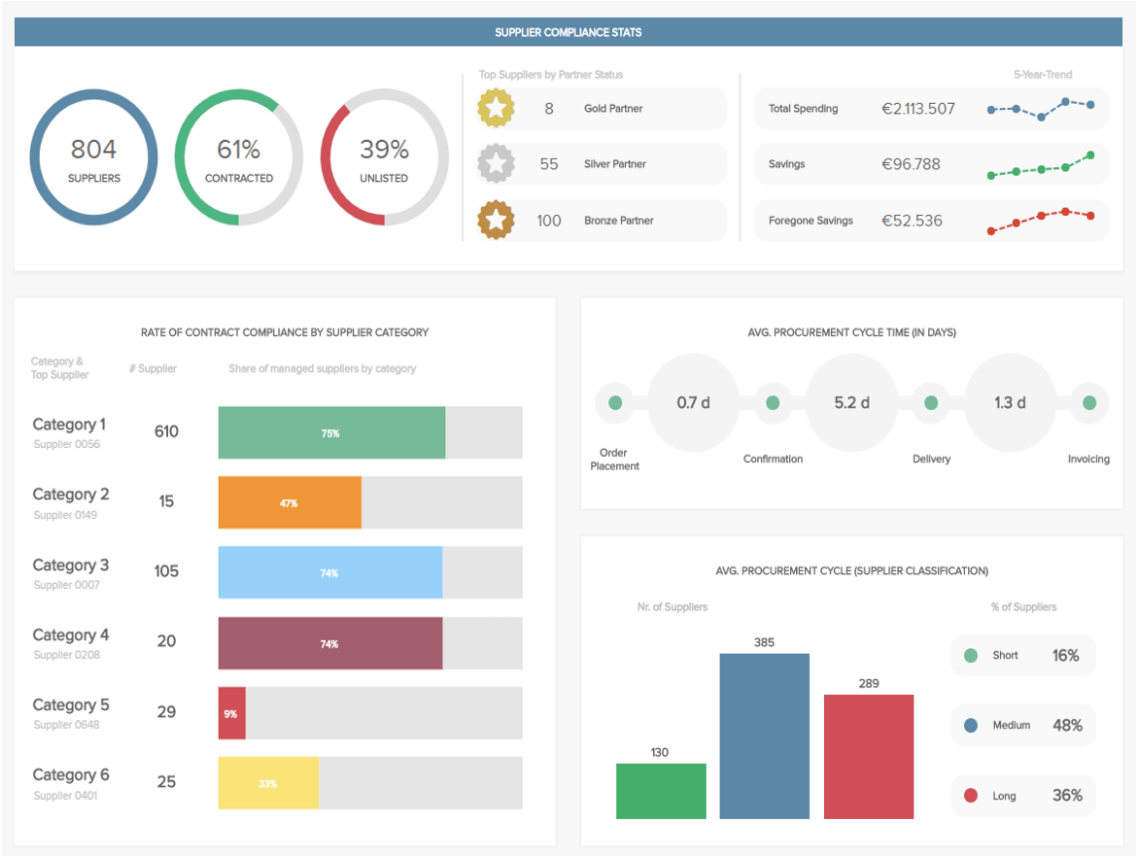
Pri dolaďovaní analytického reportu je dôležité zvážiť vlastnosti a funkcionality, ktoré údaje urobia interaktívnejšími. To aj odbúrava potrebu podrobného popisovania analytickej aplikácie formou komplexnej dokumentácie. Medzi tieto dynamické funkcionality patria napríklad:

- Klikateľné filtre grafov alebo diagramov,
- Podrobná vizualizácia údajov a rozbaľovanie ďalších detailov a metrík,
- Funkcia zväčšenia grafu,
- Dynamické textové polia a obrázky,
- Praktické informačné „toolkity“.

Ak povolíte správne funkcionality a pomôžete každému v organizácii pochopiť, ako ich používať, zabezpečíte, že vaše analytické reporty budú mať maximálnu pridanú hodnotu.

Príkladom je informačný panel verejného obstarávania (Obrázok 3), ktorý umožňuje efektívne sledovať, vyhodnocovať a optimalizovať všetky procesy obstarávania pomocou kľúčových ukazovateľov výkonnosti obstarávania, ako je miera zhody, trvanie cyklu objednávky, miera chybovosti dodávateľov a mnohé ďalšie.

Tento projekt je podporený z Európskeho sociálneho fondu.



Obrázok 3: Informačný panel verejného obstarávania, v interaktívnejšej verzii tu[7]

Optimalizácia procesov pomocou efektívneho analytického reportu o verejnom obstarávaní v konečnom dôsledku prinesie hodnotu vo všetkých agendách organizácie. Takáto analytická aplikácia vo forme informačného panelu je zrozumiteľná aj bez podrobného popisu za predpokladu, že je jasne zdokumentovaný výpočet KPI a samotný dataset (viď kapitola 2).

Ukážka informačného panelu s KPI v oblasti verejného obstarávania (Obrázok 3) poskytuje prehľad rôznych ukazovateľov, s ktorými pracuje oddelenie verejného obstarávania. Prvými zobrazenými KPI sú niektoré štatistické údaje o dodávateľoch. Spomedzi 804 dodávateľov má 61 % uzavretých rámcovú dohodu alebo dlhodobú dohodu na poskytovanie SLA a služieb, zatiaľ čo 39 % dodáva len na základe krátkodobej zmluvy o realizácii konkrétneho projektu alebo o poskytnutí služieb s konkrétnymi výstupmi. Zmluvní dodávatelia sa potom môžu klasifikovať a získať určitý status partnera, napríklad zlatý, strieborný alebo bronzový - zvyčajne sa udeľuje tým, ktorí poskytujú organizácii dobré podmienky dodania, výhodné ceny, kvalitné výstupy a ktorí sa starajú o vzťah s organizáciou. V pravej časti sú vystavené trendy nákladov a úspor za päť rokov. Vidíme, že na dodávateľov pripadajú výdavky vo výške viac ako dva milióny eur, ale oddeleniu verejného obstarávania sa podarilo ušetriť viac ako 96 000 eur.

[7] Zdroj: https://public.datapine.com/#board/NpWWdmyIS0ExxHEgLAyBw, Dátum referencie: 28.07.2023

Tento projekt je podporený z Európskeho sociálneho fondu.

Druhý KPI zobrazený na tomto informačnom paneli sa zameriava na mieru dodržiavania zmlúv podľa jednotlivých kategórií: najvyššiu mieru dodržiavania zmluvných podmienok ako napríklad dodanie podľa časového harmonogramu definovaného vo verejnom obstarávaní dosahuje prvá kategória, ktorá má aj najvyšší počet dodávateľov, pričom miera dodržiavania zmluvných podmienok u 610 z nich je 75 %. Najnižšiu mieru dosahuje štvrtá kategória s 9 % dodržiavaním podmienok u 29 dodávateľov. Posledný ukazovateľ na tomto paneli ilustruje priemerné trvanie cyklu verejného obstarávania vyjadreného v dňoch. Toto trvanie zahŕňa celý proces objednávania, od zadania objednávky až po jej potvrdenie, doručenie a nakoniec fakturáciu. Tu vidíme, že celkový priemer je 7,2 dňa, pričom obdobie medzi potvrdením a doručením objednávky na faktúru je najdlhší (5,2 dňa). Samozrejme vo verejnej správe je dôležité sledovať aj trvanie samotného procesu verejného obstarávania, ako aj míľniky implementácie niekoľkoročných projektov. Mať prehľad a monitorovať všetky procesy od začatia verejného obstarávania cez dodávanie, fakturáciu až po hodnotenie rôznych typov dodávateľov umožní optimalizovať proces verejného obstarávania a vo všeobecnosti aj efektívnosť danej organizácie. Takýchto analytických aplikácií vo forme informačných panelov môže vznikať viacero v danej problematike za predpokladu, že je zrozumiteľne popísaný ich účel, použité KPIs alebo iné premenné znázorňované v interaktívnych grafoch a vizualizáciách, ako aj samotný dataset (viď kapitola 2).

3.1.2 Koncept popisu modelov strojového učenia

Popísať analytickú aplikáciu založenú na strojovom učení je zložitejšie, ako v analytickom nástroji, keďže aj proces jej vytvorenia a využívania v praxi je oveľa zložitejší. Súčasťou takejto aplikácie sú aj takzvané „pipelines“ strojového učenia alebo dátovej vedy. Tieto „pipelines“ pozostávajú z viacerých postupných krokov, pomocou ktorých sa vykoná všetko od extrakcie a predbežného spracovania údajov až po trénovanie a nasadenie modelu. „Pipeline“ strojového učenia je „end-to-end“ konštrukt, ktorý riadi tok údajov do modelu strojového učenia (alebo súboru viacerých modelov) a výstup z neho. Zahŕňa vstupné surové údaje, tvorbu vstupných premenných („features“), výstupy, model strojového učenia a parametre modelu a výstupy predpovedí. Dobrá dokumentácia sa oplatí kvôli nasledujúcim výhodám:

- **Pomáha ľuďom dosiahnuť spoločnú úroveň porozumenia**: Dobrá dokumentácia umožňuje ľuďom z rôznych tímov a na rôznej technickej úrovni spoločne rozumieť a diskutovať o danej analytickej aplikácii.
- **Ujasňuje víziu a plány**: Správne kroky sa líšia v závislosti od plánov, vízie a obmedzení organizácie. Mať podrobnú dokumentáciu umožní každému lepšie určiť ďalšie kroky. Treba mať na pamäti, že je ťažké vidieť cieľ, ktorý sa nedá dosiahnuť. Vďaka dokumentácii sú ciele oveľa konkrétnejšie.
- **Znižuje náklady na nástup do zamestnania**: Vždy, keď na projekt alebo pozíciu príde nový človek, prvá vec, ktorú by mal urobiť, je, že si pozrie, čo už bolo urobené. To zahŕňa prejdenie si využívaných analytických metód, informácií o zozbieraných údajoch, zdôvodnenie „pipelines“ strojového učenia atď. Mať dobrú dokumentáciu výrazne skráti čas, ktorý nový zamestnanec strávi zaučením sa. A tiež šetrí čas ostatných členov tímu pri zaškoľovaní nového človeka.
- **Urýchľuje prieskumnú analýzu**: V oblastiach, ako je dátová veda a umelá inteligencia, vývojári trávia veľa času prieskumom domény a využitých metód, ktoré fungovali na daný problém. Je to veľmi dôležité na zabezpečenie toho, aby analýza

Tento projekt je podporený z Európskeho sociálneho fondu.

vykonávaná neskôr v rámci pipeline nevynechala kľúčové nuansy jedinečné pre riešený problém. Skvelá dokumentácia tento proces zefektívňuje.

Ak by sa mali zhrnúť výhody správneho popisu a dokumentácie do jedného slova, bola by to **prehľadnosť**. Dobrá dokumentácia prispieva k veľkej miere prehľadnosti vo všetkých oblastiach. Organizácia tak ušetrí množstvo hodín strávených vývojármi, ktoré by inak premrhali opakovaným hľadaním informácií a zdrojov a organizovaním stretnutí.

Ak je cieľom mať dobre popísanú analytickú aplikáciu v podobe modelu strojového učenia, treba sa zaoberať týmito oblasťami:

1. Vízia a plány organizácie,
2. Zdroje a obmedzenia,
3. Použité zdroje údajov, dostupné datasety a vykonané spracovanie (tomu sa venuje aj kapitola 2),
4. Projekty, na ktorých sa v súčasnosti pracuje,
5. Popis samotného modelu strojového učenia,
6. Aktuálny zdrojový kód, ktorý je k dispozícii.

V nasledujúcej časti sa povenujeme každej oblasti osobitne.

1. Vízia a plány organizácie

Toto by mala byť v rámci organizácie vždy priorita číslo jeden. Jasné smerovanie a cieľ analytickej aplikácie ušetria tímu veľa zbytočne vynaloženej energie. Dokumentácia by mala jasne načrtnúť každú vyvíjanú alebo pripravovanú aplikáciu, ako aj jej prípady použitia vrátane používateľov a to, ako zapadne do celkového ekosystému analytických aplikácií a dátovej vrstvy.

Inžinierske výzvy sú veľmi odlišné pre rozličné prístupy k riešeniu daného problému pomocou údajov. Napríklad je veľkým rozdielom, či v danej problematike stačí analyzovať aktuálny stav, alebo aj prognózovať budúci. Jasná vízia budovania analytických aplikácií pomôže celej organizácii smerovať k správnym cieľom. Pomôže tiež vedieť, aké zručnosti si zamestnanci majú osvojovať.

2. Zdroje a obmedzenia

S jasnou víziou treba načrtnúť aj obmedzenia, s ktorými tím v súčasnosti pracuje. Tieto obmedzenia môžu byť orientované na fyzické zdroje (nedostatok pracovnej sily, cloud computing, financie), problémy s údajmi (nedostatok historických údajov na prognózovanie, nízka kvalita údajov) alebo na legislatívne predpisy a metodiky. Niektore obmedzenia môžu vyplývať aj z vnútorného nastavenia organizácie ako splnenie určitých základných kritérií, používanie určitých nástrojov/riešení, integrácia do nastaveného „frameworku“. Ich objasnenie bude kľúčové a mala by sa na ne vzťahovať celá dokumentácia.

3. Informácie o datasetoch (tomu sa venuja aj kapitola 2)

Dokumentácia musí zahŕňať informácie o použitých dátových zdrojoch, ako vyzerá dátové potrubie („pipeline“) a aký druh spracovania sa vykonáva so vstupnými premennými („features“) získanými zo surových údajov.

Tento projekt je podporený z Európskeho sociálneho fondu.

Každá vstupná premenná, ktorá sa používa na strojové učenie či dátovú vedu, by mala mať vlastný popis s informáciami o:

- jej povahe (kategorická / „Boolean“ / číselná atď.),
- zdôvodnenie jej použitia,
- a očakávaný rozsah / rozdelenie.

Toto je tiež dobré miesto **na zdokumentovanie akýchkoľvek predpokladov** a spôsobu, akým ste k nim dospeli.

Ďalším veľmi podceňovaným krokom je stanoviť úlohy, ktoré má algoritmus strojového učenia vykonať, ako aj popísať dôvod, prečo sú potrebné. Tiež pre každú úlohu treba definovať základný výkon pre porovnanie („baseline performance“), ktorú by malo riešenie úlohy prekonať.

4. Projekty, na ktorých sa v súčasnosti pracuje

Dátový analytik, dátový vedec alebo iný špecialista, ktorý pracuje na jednej analytickej aplikácii, by mal mať možnosť pozrieť sa na iné aplikácie, ktoré sa tiež pripravujú. Je to nevyhnutný predpoklad budovania spolupráce medzi tímami. Môže to tiež pomôcť týmto odborníkom vytvárať riešenia s ohľadom na celkový obraz, ako aj zdieľať si efektívnu expertízu a skúsenosti z podobných riešení.

5. Popis samotného modelu strojového učenia

Táto časť predstavuje kľúčovú časť popisu, ktorej sa podrobne venujeme aj v kapitole 5.2. Konkrétne treba zachytiť:

- Zrozumiteľný popis modelu – na čo slúži a ako zapadá do vízie a plánov a ako súvisí s ostatnými analytickými aplikáciami. Tu je potrebné uviesť aj licenciu, s akou sa môže model ďalej používať, ideálne aj s odkazom na ňu. Ďalej je tu rámcový popis toho, o aký model ide a ako bol trénovaný.
- Obmeny modelu, ak existujú, na špecifické účely, ako napríklad rýchlejšie natrénovanie modelu na nových údajoch,
- Zamýšľané prípady použitia,
- Prípady použitia, ktoré model nepokrýva,
- Popis architektúry modelu vrátane jeho grafickej schémy,
- Obdobie, v ktorom bol model trénovaný na vstupných údajoch,
- Čo je vstupom modelu – aký dataset bol použitý na trénovanie a s odkazom na jeho popis (spravidla sa odkazuje na informácie o datasete na vstupe), aké vstupné premenné („features“) boli pre model vytvorené (tieto informácie môžu byť prepojené aj s dátovým katalógom a popisom dostupných datasetov podľa kapitoly 2). Ako aktuálne boli údaje použité na trénovanie.
- Čo je výstupom modelu - aké výstupné premenné („outputs“) boli modelom vytvorené (tieto informácie môžu byť prepojené aj s dátovým katalógom a popisom dostupných datasetov podľa kapitoly 2).
- Aké sú výsledky modelovania (napríklad cez zdokumentované metriky a grafy) a aká je výkonnosť modelu s porovnaním so spomínanou základnou výkonnosťou („baseline performance“). Prípadne podrobný benchmark s inými modelmi.

Tento projekt je podporený z Európskeho sociálneho fondu.

- Zhrnutie etických otázok a obmedzení pri využívaní výstupov modelu v praxi,
- Autori a vlastníci modelu,
- Kontaktné údaje, kam možno nahlásiť zistené chyby v modeli.

Dôsledné spísanie vyššie uvedených aspektov vedie aj k transparentnosti pri využívaní modelov strojového učenia, ako aj k možnosti ich validovať ďalšou stranou, ktorá má možnosť reprodukovať výsledky procesu dátovej vedy alebo modelovania.

6. Aktuálny zdokumentovaný kód k dispozícii

Keď väčšina vývojárov myslí na dokumentáciu, majú v hlave dobre zdokumentovaný zdrojový kód. Existujú tony dobrých príručiek na písanie dobrého, dobre zdokumentovaného kódu. Základným princípom je, že všetky názvy premenných a metód by mali byť popisné. Dokumentácia každej funkcie by mala obsahovať popis premenných, čo daná funkcia robí a jej návratovú hodnotu. Popis sa píše aj pre každú triedu, aby bol ihneď zrejmý jej účel.

Treba tiež dodržiavať jasné návrhové vzory, ako je napríklad „Princíp jednej zodpovednosti“ („The Single Responsibility Principle“)[8]. Tie vedú ku zdrojovému kódu, ktorý sa dá ľahko čítať, pochopiť a upravovať. To členom tímu umožní pracovať na staršom kóde s minimálnym úsilím vynaloženým na to, aby sa snažili pochopiť myšlienkový prúd pôvodného autora kódu.

[8] Zdroj: https://blog.devgenius.io/use-single-responsibility-to-improves-your-machine-learning-pipelines-e724e153680f, Dátum referencie: 28.07.2023

Tento projekt je podporený z Európskeho sociálneho fondu.

4 Výber nástrojov na popisovanie

Bez dátového katalógu môže byť manažment údajov zložitý a časovo náročný proces. A bez takéhoto katalógu a nástrojov na katalogizáciu údajov môžu dátoví analytici, dátoví vedci a iní špecialisti na údaje stráviť hodiny prácou na vyhľadávaní údajov a hľadaním správneho dátového zdroja. Čo je ešte horšie, bez katalogizácie údajov môžu dokonca o existencii konkrétneho datasetu vôbec nevedieť. Tým pádom ich dátová analýza bude buď nepresná, alebo ju ani vôbec nevykonajú v presvedčení, že vytvorenie nedostupného datasetu je príliš veľkou bariérou.

Dátové katalógy sú určené pre veľké organizácie s komplexnou dátovou infraštruktúrou. Sú základným kameňom moderného dátového „stacku“, poskytujú centralizované úložisko na manažment a zdieľanie údajov medzi rôznymi organizáciami, sekciami, oddeleniami a tímami a jednotný prehľad o všetkých dátových aktívach - vrátane citlivých údajov. Dátové katalógy tak pomáhajú aj zabezpečiť dodržiavanie politík na bezpečnosť a ochranu údajov.

Základom úspechu a udržateľnosti dátového katalógu je automatizácia a využívanie algoritmov strojového učenia. Funkcionality pre automatizáciu a strojové učenie v nejakej miere poskytuje väčšina trhom uznaných dátových katalógov, ktoré sú súčasťou moderného dátového „stacku“, popísaného v dokumente „*4.2.1 Koncept pre zavádzanie analytického spracovania údajov do praxe*“. **Problematické je však vybrať jeden najvhodnejší dátový katalóg, ktorý na jednej strane podporuje implementované technológie** (či už technológie Centrálnej integračnej platformy, platformy otvorených údajov alebo Konsolidovanej analytickej vrstvy), **a na druhej strane nevytvára „vendor lock-in“**. Dátový katalóg automatizovane cez vytvorené konektory alebo aplikačné rozhrania zhromažďuje metadáta z databáz, dátových skladov, dátových jazier, systémov BI a iných zdrojov, a používa ich na vytvorenie prehľadného súpisu dátových aktív. Zároveň poskytuje jednotný referenčný bod pre správu biznis metadát, ktoré dokáže spracovať rýchlejšie a efektívnejšie ako staršie typy systémov na správu metadát. **Ako odporúčaná cesta vpred sa javí vyskladať navzájom previazaný ekosystém viacero nástrojov, ktoré budú v maximálnej možnej miere založené na nástrojoch s otvoreným zdrojovým kódom.**

Mnohé organizácie dopĺňajú svoje dátové katalógy aj ďalšími metadátovými nástrojmi - najmä nástrojmi na správu taxonómií (podnikové slovníky) a dátovými slovníkmi, ktoré poskytujú ďalšie informácie, ktoré pomáhajú používateľom pochopiť údaje a ich kontext. Správa taxonómií a ontológií sa vykonáva v nástroji VocBench[9], z ktorého sa publikuje na portál znalosti.gov.sk (spoločne tak predstavujú obdobu podnikového slovníka). Dátový slovník, ktorý okrem iného mapuje dátové prvky integrované cez platformu CIP na pojmy z taxonómií a ontológií, sa momentálne nachádza v platforme Talend. Obom týmto nástrojom sa venujeme v dokumente „*1.1.2 Štandardizácia pre modelovanie údajov*“. V tomto dokumente 1.1.2 sa venujeme aj osobitnej téme dátových katalógov datasetov otvorených údajov, ktoré sú v štandarde Data Catalog Vocabulary (DCAT)[10], a momentálne predstavujú oddelený ekosystém ako dátové katalógy pre dátovú integráciu a analytické spracovanie údajov. Objekty evidencie tak, ako sú momentálne integrované cez platformu CIP a ako ich konzumuje platforma MOU, sú popísané

[9] Zdroj: https://znalosti.gov.sk/vocbench3/#/Home, Dátum referencie: 18.07.2023

[10] Zdroj: https://www.w3.org/TR/vocab-dcat/, Dátum referencie: 18.07.2023

Tento projekt je podporený z Európskeho sociálneho fondu.

v MetaIS – v nástroji Confluence[11]. Pre dátovú integráciu bol v dokumente „*1.1.2 Štandardizácia pre modelovanie údajov*“ vybraný katalóg s otvoreným zdrojovým kódom Apache Atlas, ktorý v nejakej miere dokáže interagovať s nástrojom Talend používaným v rámci platformy CIP. Tento katalóg podporuje aj popisovanie analytických údajov, avšak pre túto oblasť existujú aj lepšie riešenia.

Väčšina dátových katalógov poskytuje nasledujúce základné funkcionality pre dátové tímy, ktoré bližšie popisuje aj Obrázok 4:

- **Vyhľadávanie údajov** (Obrázok 5): Jednou z primárnych funkcií dátového katalógu je vyhľadávanie údajov. Moderný katalóg poskytuje súpis dátových aktív (objektov evidencie a datasetov, ako aj dátových prvkov, ideálne aj dostupných analytických aplikácií) a pomáha používateľom vyhľadávať tieto aktíva na základe ich metadát - údajov o údajoch. Tieto metadáta katalógu môžu obsahovať kľúčové slová, zdroj údajov, typ údajov a ďalšie atribúty, ktoré používateľom umožňujú nájsť správne dátové aktíva pre konkrétny projekt.
- **Prístup k údajom**: Ďalšou dôležitou funkcionalitou katalógu je prístup k údajom. Dátový katalóg poskytuje centralizované miesto na prístup k údajom z viacerých zdrojov údajov v rámci dátovej vrstvy, či už ide o dátové sklady, dátové jazerá alebo iné zdroje. Tým sa zabezpečí, že údaje budú ľahko dostupné tým, ktorí ich potrebujú, a zároveň sa zachová vhodné riadenie bezpečnosti údajov a ochrany súkromia.
- **Manažment údajov:** Dátové katalógy poskytujú centrálne miesto na správu politík a postupov manažmentu údajov a zabezpečujú, aby sa údaje spravovali v súlade s príslušnými predpismi a osvedčenými postupmi.

[11] Zdroj: https://wiki.vicepremier.gov.sk/display/IN/Objekt+evidencie, Dátum referencie: 19.07.2023

Tento projekt je podporený z Európskeho sociálneho fondu.

Obrázok 4:Základné funkcionality softvérov implementujúcich dátový katalóg

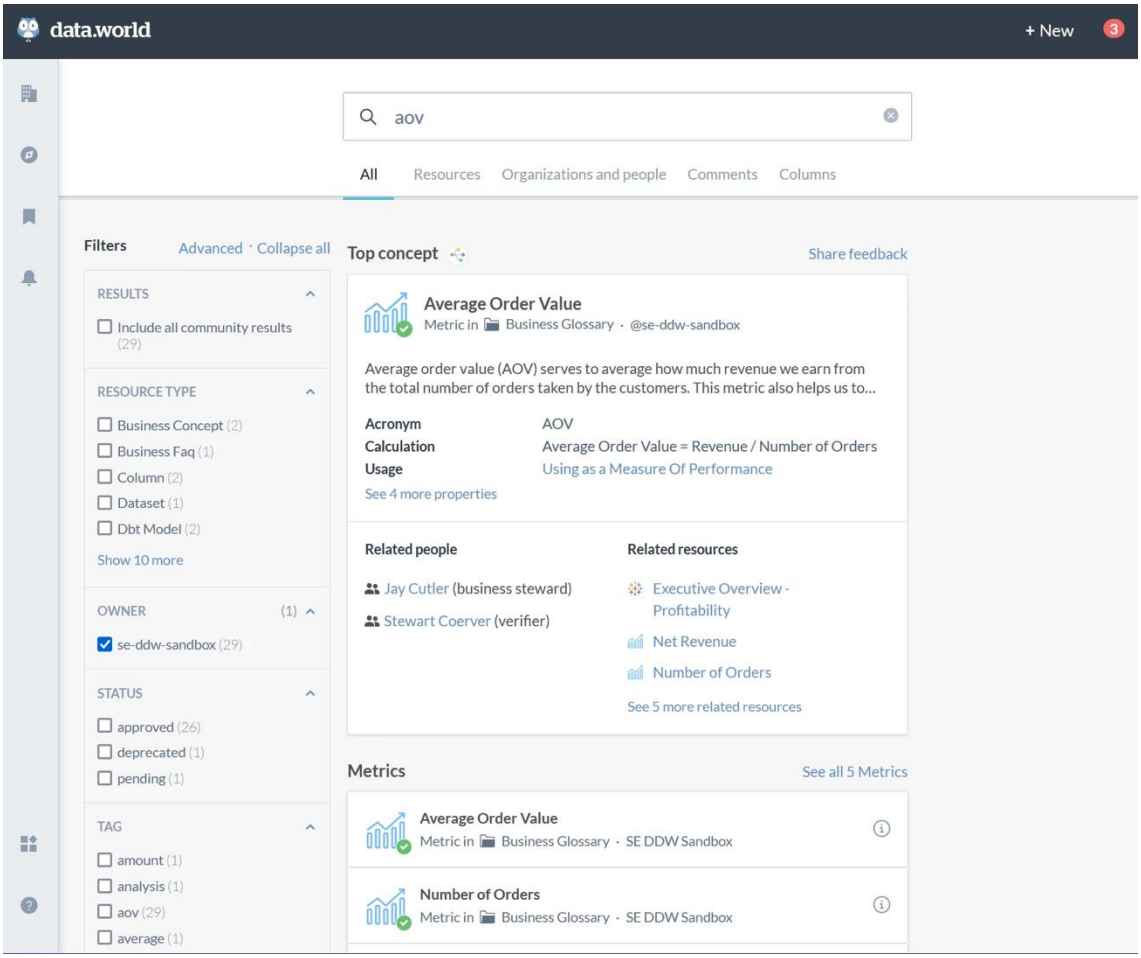
Pokročilé funkcionality dátového katalógu, ktoré už neposkytuje ľubovoľné riešenie, sú nasledovné:

- **Samoobslužná analytika**: Samoobslužná analytika je určená pre konzumentov údajov, ktorí potrebujú samostatne vyhľadávať a pristupovať k dátovým zdrojom. Moderné dátové katalógy by mali poskytovať používateľsky prívetivé vyhľadávanie podobne ako vyhľadávač spoločnosti Google, pričom by mali poskytovať výsledky vyhľadávania na základe presnej zhody metadát, ale aj na základe podobných alebo súvisiacich biznis metadát, ktoré môžu používateľom údajov pomôcť objaviť údaje relevantné pre ich projekty, ktoré predtým nebrali do úvahy alebo o ktorých predtým nevedeli. Samoobslužná analytika navyše umožňuje nájdené údaje jednoducho a priamo použiť v analytických nástrojoch bez potreby písať vlastný zdrojový kód, čo môže ušetriť čas a zdroje dátových inžinierov, dátových vedcov a správcov údajov, ktorí by inak vynaložili veľké množstvo úsilia na samotné plnenie požiadaviek na údaje a ich analýzu.
- **Podpora neštruktúrovaných a streamovaných údajov:** Dnes sa všade stretávame so streamovanými údajmi a neštruktúrovanými údajmi, ktoré nie sú vo formáte klasických tabuliek (napr. JSON, štruktúrami Parquet[12]) a ich objem viditeľne rastie čoraz rýchlejšie. Aj keď dnes tieto technológie v organizácii nepoužívate, hľadajte dátový katalóg, ktorý podporuje vnorené dátové štruktúry a umožňuje integráciu streamovacích technológií v budúcnosti.

[12] Zdroj: https://parquet.apache.org/, Dátum referencie: 31.07.2023

Tento projekt je podporený z Európskeho sociálneho fondu.

- **Strojové učenie a umelá inteligencia**: Technológie umelej inteligencie a strojového učenia môžu zvýšiť produktivitu dátových tímov prostredníctvom integrácie dátových katalógov, pričom osobitný potenciál sa týka automatizácie vyhľadávania a správy údajov, automatického obohacovania metadát z hľadiska osôb, politík, kontextovo orientovaných polí a všetkých typov vzťahov medzi objektami evidencie, datasetmi, ale i samotnými dátovými prvkami.



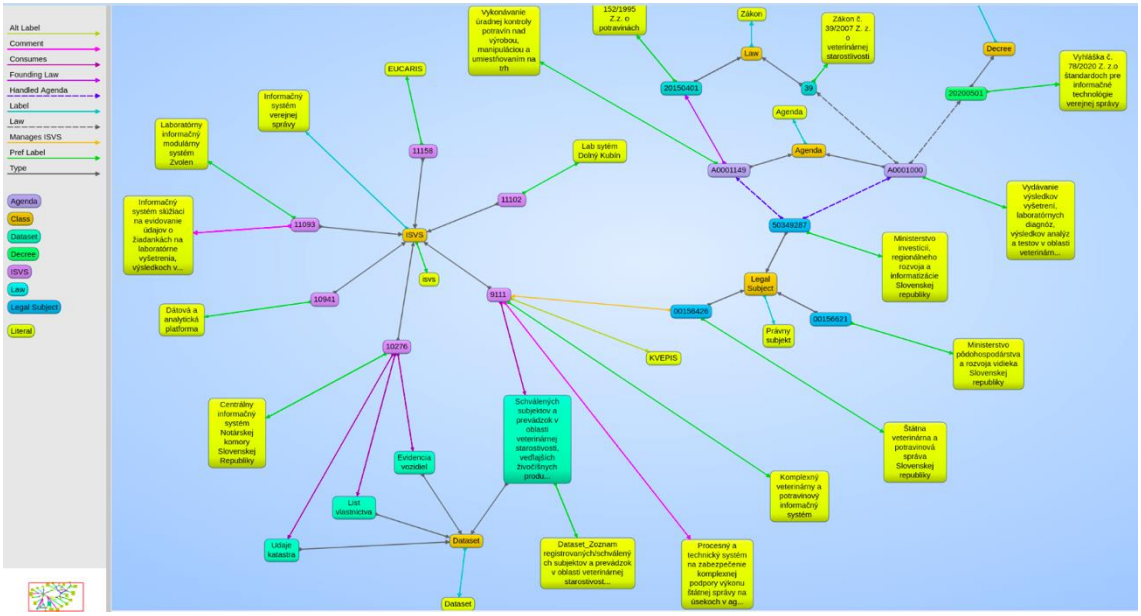
Obrázok 5: Príklad výsledkov vyhľadávania dátových aktív zobrazených prostredníctvom katalógu údajov data.world.

Prečo by mal byť dátový katalóg založený na znalostnom grafe?

Ako už bolo spomenuté, základným účelom a teda aj funkcionalitou dátového katalógu údajov je umožniť zamestnancom získať lepší prehľad o údajoch ako celku a vedieť efektívne vyhľadávať dátové aktíva. Na splnenie tohto účelu musí dátový katalóg vytvárať a spravovať množiny dátových aktív a vzťahov medzi nimi vo verejnej správe a poskytovať jednotný pohľad na dátový ekosystém poskytovateľom údajov (napr. správcom IS VS) a konzumentom údajov (napr. dátovým vedcom a dátovým analytikom). Tieto množiny dátových aktív zahŕňajú tabuľky a stĺpce databázy, pojmy zo slovníka, analytické aplikácie ako zostavy z informačných panelov BI alebo modely strojového učenia. Kľúčovým poznatkom je, že správa vzťahov by mala byť základným nástrojom dátového katalógu. A práve na správu vzťahov najlepšie fungujú znalostné grafy.

Tento projekt je podporený z Európskeho sociálneho fondu.

Čo je to znalostný graf?



Obrázok 6: Príklad znalostného grafu podľa CMÚ

Znalostné grafy sú spôsobom organizácie a reprezentácie informácií v strojovo čitateľnom formáte. Model grafu znalostí predstavuje súbor reálnych pojmov (zobrazených ako uzly) a vzťahov (zobrazených ako hrany) vo forme grafu, ktorý sa používa na prepojenie a integráciu údajov pochádzajúcich z rôznych zdrojov. Preklenujú medzeru medzi údajmi a ich významom, spájajú podnikovú terminológiu a kontext s údajmi a umožňujú prístup k údajom prostredníctvom všeobecne zrozumiteľného jazyka, čím výrazne zlepšujú vyhľadávanie, objavovanie nových dátových aktív, prehľadnosť a presnosť.

Výhody budovania katalógu údajov na základe znalostného grafu

Postavením dátového katalógu na znalostnom grafe možno ten istý model grafu rozšíriť na všetky nové dátové aktíva, ktoré sa časom získajú alebo vytvoria. A tieto nové dátové aktíva je možné ľahko prepojiť so zvyškom ako aj s ich kontextom.

Naproti tomu dátový katalóg založený na tradičnej relačnej technológii je nepružný a rýchlo zastará. To znamená, že podpora nových typov zdrojov údajov a dátových aktív môže trvať mesiace, čo ide na úkor efektivity využívania údajov, ktoré si ich konzumenti nevedia dohľadať. Tým pádom organizácie, ktoré budujú dátový katalóg na relačných architektúrach, takmer z definície nemôžu byť dátovo orientované.

Veľkou výhodou v tejto téme je, že základ takéhoto znalostného grafu, ktorého niektoré dátové entity sú zachytené aj v CMÚ, a teda aj v podnikovom slovníku, už existuje a je popísaný v dokumente „*1.1.7 Analýza toku údajov*“.

Finálny výber dátového katalógu alebo dátových katalógov treba prispôsobiť výberu kľúčových technológií KAV a podporiť pilotom, ktorý overí jeho možnosti rozširovania a užitočnosť aj v širšom kontexte dátového programu. Tiež kľúčovým parametrom bude udržateľnosť aktuálnosti informácií o dátových aktívach, aby si zvolené riešenie nevyžadovalo príliš veľa ľudskej práce.

V dokumente „*1.1.2 Štandardizácia pre modelovanie údajov*“ sa hodnotili rôzne katalógy patriace do tretej generácie na základe kritérií zosumarizovanej v tabuľke (Tabuľka 2), v ktorej bol jednoznačným víťazom Apache Atlas. Tento dátový katalóg má aj tú výhodu, že podporuje znalostné grafy, keďže ukladá dáta do databázy JanusGraphs – ide o distribuovanú, open source, masívne škálovateľnú databázu grafov s dátovým modelom grafov s vlastnosťami („property graph“). V tomto dokumente pridávame ďalšieho konkurenta **OpenMetadata**[13]**,** s ktorým sa experimentuje v rámci CIP.

Tabuľka 2: Kritériá pre porovnanie dátových katalógov tretej generácie

Skupina kritérií	Kritérium	ATLAS	OpenMetadata	UBER DATABOOK
Cena	Open source	✔	✔	✔
Vlastnosť	Slovník a podnikové metadáta	✔	✔	
	Dokumentácia	✔	(✔) - neúplná	-
	Lineage	✔	✔	-
	Upozornenie na aktualizácie	✔	✔	-
	Vlastníctvo	✔	✔	
	Čerstvosť (Freshness)	-	✔	
	Ukážka údajov	-	✔	-
	Stĺpcové štatistiky	-	✔	-
Pipeline integrácie	Integrácia API	✔	✔	-
	Integrácia správ	✔	✔	-
	Integrácia s Rangerom	✔	-	
	Integrácia s Hive	✔	✔	✔
	Kompatibilita s Talendom	✔	(✔) – potreba implementácie	✔
	Push based	✔	-	

[13] Zdroj: https://open-metadata.org, Dátum referencie: 25.07.2023

Tento projekt je podporený z Európskeho sociálneho fondu.

Skupina kritérií	Kritérium	ATLAS	OpenMetadata	UBER DATABOOK
	Pull based	-	✔	
Flexibilita	Definujte nové typy metadát	✔	✔	
	Importovať/exportovať metadáta	✔	-	

OpenMetadata bol inšpirovaný poznatkami získanými pri budovaní metadátovej infraštruktúry spoločnosti Uber, ktorú možno považovať za prvú iteráciu OpenMetadata. Systém metadát spoločnosti Uber obsahuje interné nástroje, ako je Databook[14] (tiež ho sumarizuje aj Tabuľka 2). Suresh Srinivas, zakladateľ spoločnosti OpenMetadata, vo svojom blogu s oznámením uviedol dôvody, prečo interný systém spoločnosti Uber nebol sám o sebe otvorený, a tým pádom musel byť dátový katalóg OpenMetadata vybudovaný od základov. Dôvody v zásade vychádzajú z myšlienky zabezpečiť, aby priority spoločnosti Uber a open-source komunity neboli v procese vývoja takéhoto nástroja v rozpore. OpenMetadata je jedným z najnovších prírastkov do prostredia dátových katalógov tretej generácie s otvoreným zdrojovým kódom, ktoré zahŕňa ďalšie nástroje ako Amundsen, DataHub, Apache Atlas, ktorým sa venujeme v dokumente „*1.1.2 Štandardizácia pre modelovanie údajov*“. Závery z tohto dokumentu sú platné, avšak v dôslednom pilotnom riešení a testovaní treba finálne vybrať dátový katalóg medzi Apache Atlas (ktorý čiastočne podporuje aj nástroj Talend v rámci CIP) a **OpenMetadata, ktorého hlavnou výhodou je dostupnosť mnohých konektorov**, avšak okrem Talendu[15]. **Nevýhodou OpenMetadata je, že hoci deklaruje grafovú reprezentáciu metadát, nie je implementovaný cez grafovú databázu, ale cez MySQL a elasticsearch** (viď Obrázok 7).

[14] Zdroj: https://atlan.com/uber-databook-metadata-catalog/?ref=/openmetadata-explained/, Dátum referencie: 25.07.2023

[15] Zdroj: https://docs.open-metadata.org/v1.1.0/connectors, Dátum referencie: 25.07.2023

Tento projekt je podporený z Európskeho sociálneho fondu.

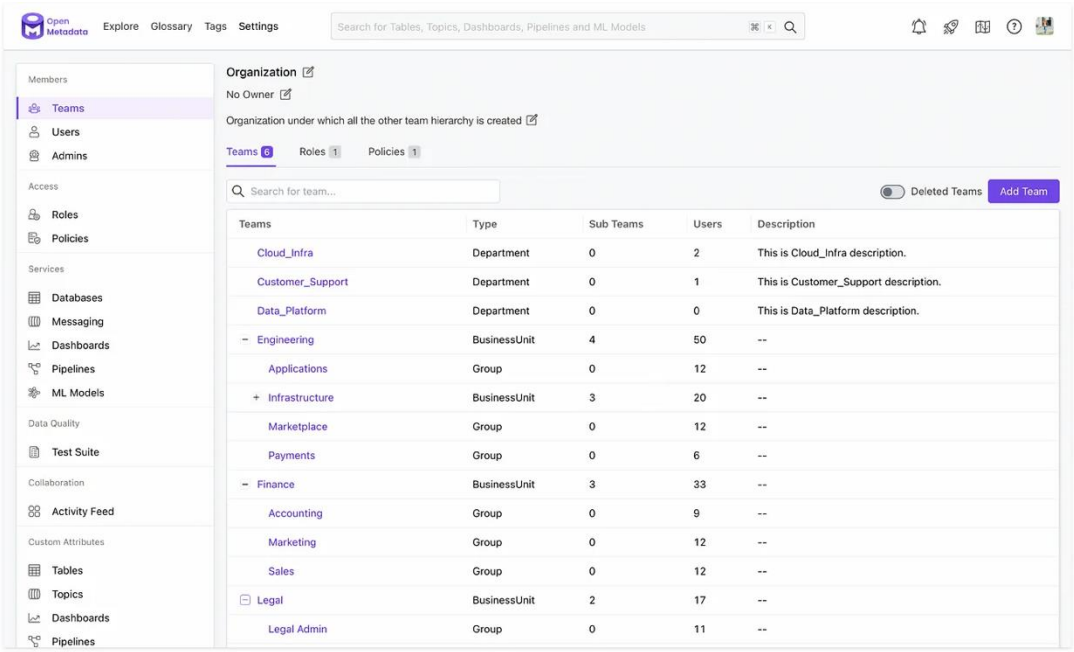
Obrázok 7: Komponenty dátového katalógu OpenMetadata

Na rozdiel od systému Amundsen je OpenMetadata skôr čiernou skrinkou. Jeho vnútorný systém má päť hlavných komponentov (viď Obrázok 7), ktoré fungujú ako jeden celok, takže by sa mal ľahšie nastavovať a používať. Kľúčové komponenty OpenMetadata zahŕňajú nasledujúce:

1. **Používateľské rozhranie (UI)** - centrálne miesto, kde môžu používatelia prehľadávať všetky údaje a spolupracovať na nich.
2. **Ingestion framework** (rámec na prijímanie údajov) - zásuvný rámec na integráciu nástrojov a prijímanie metadát do úložiska metadát. Rámec pre prijímanie údajov už podporuje známe dátové sklady. Kompletný zoznam a dokumentáciu o podporovaných službách možno nájsť v časti o konektoroch[16].

[16] Zdroj: https://docs.open-metadata.org/v1.1.0/connectors, Dátum referencie: 25.07.2023

3. **Metadátové API** - na vytváranie a konzumáciu metadát postavených na schémach pre používateľské rozhrania a na integráciu nástrojov, systémov a služieb[17].
4. **Úložisko metadát** - uchováva graf metadát, ktorý spája dátové aktíva a metadáta generované používateľmi a nástrojmi.
5. **Metadátové schémy** - definujú základné abstrakcie a slovník pre metadáta so schémami pre typy, entity a vzťahy medzi entitami. Toto je základ otvoreného metadátového štandardu – „Open MetaData Standard“, ktoré sú žiaľ založené na JSON, čo podobne ako pri ostatných dátových katalógoch nie je v súlade so štandardom CMÚ. Viac informácií o schémach metadát možno nájsť v časti o koncepte schém na oficiálnej webovej lokalite OpenMetadata[18].



Obrázok 8: Ukážka podnikového slovníka v OpenMetada

Výhody dátového katalógu OpenMetadata

- Poskytuje veľké množstvo pripravených konektorov - približne 30 vytvorených konektorov. Je však možné, že niektoré z nich budú časom zastarané alebo nebudú fungovať podľa očakávania.
- Sťahuje automatizovane metadáta aj pre analytické aplikácie vo forme informačných panelov a reportov z nástrojov ako Looker, Power BI a Tableau ako aj z nástrojov na implementáciu dátovej integrácie cez messaging ako Kafka (tento nástroj sa testoval aj v rámci CIP).

[17] Zdroj: https://docs.open-metadata.org/swagger.html#section/APIs/API-Organization, Dátum referencie: 25.07.2023

[18] Zdroj: https://docs.open-metadata.org/v1.1.0/main-concepts/metadata-standard/schemas, Dátum referencie: 25.07.2023

Tento projekt je podporený z Európskeho sociálneho fondu.

- Podporuje mnoho dátových aktív, nielen datasety (tu je priestor aj zjednotiť katalogizáciu analytických údajov s katalogizáciou objektov evidencie[19]), ale napríklad aj databázy, informačné panely, modely strojového učenia (tu[20] je ukážka, ako to môže vyzerať) a „pipelines“.
- Ponúka profilovanie údajov ako súčasť preberania metadát pomocou samostatného konfiguračného súboru. Profily údajov umožňujú kontrolovať nulové hodnoty v nenulových stĺpcoch, duplicity v jedinečnom stĺpci atď. Prostredníctvom poskytovaných popisných štatistík možno lepšie pochopiť rozloženie údajov v stĺpcoch. Avšak nie všetky konektory podporujú existujúce požitie dátových profilov.
- Podporuje označovanie PII údajov, aj automatizovanie v rámci profilovania.
- Podporuje dátovú „lineage“, avšak nie u všetkých kontektorov. Dátovú „lineage“ možno pridať ručne prostredníctvom používateľského rozhrania alebo rozhrania API alebo ju nástroj môže odvodiť zo zdroja, ak je podporovaný. Lineage môže mať ľubovoľná entita, ako sú tabuľky (datasety), „pipelines“, informačné panely atď. Tabuľky môžu mať Lineage na úrovni stĺpcov.
- Veľmi aktívny kanál Slack, kde môžu súčasní používatelia OpenMetadata klásť otázky a hľadať pomoc u vývojárov, ktorí pracujú alebo predtým pracovali na vývoji samotného nástroja. Tiež online možno nájsť veľa návodov, ako napríklad tu[21] alebo tu[22].
- K dispozícii sú aj stretnutia a ukážky najnovších verzií od vývojárov, ktorí vysvetľujú nové funkcie a spôsob ich používania.
- Najdôležitejšie je, že ide o otvorený zdrojový kód a je vydaný pod licenciou Apache, verzia 2.0. Takže ho môže používať každý a zároveň ho možno ďalej rozširovať a vylepšovať.

Nevýhody dátového katalógu OpenMetadata

- Najväčším problémom by mohla byť jeho (trochu nedostatočná) dokumentácia, takže trochu pripomína čiernu skrinku. Dokumentácia nie je aktuálna alebo nie sú správne vysvetlené ďalšie kroky rozvoja nástroja. Avšak existuje spomínaná komunita aj na kanáli Slack, kde sa dá hľadať odpovede ako aj zúčastňovať sa stretnutí komunity.
- Hoci nástroj deklaruje grafovú reprezentáciu metadát, nie je implementovaný cez grafovú databázu.
- Potreba vytvoriť v nástroji vlastný podnikový slovník založený na hierarchii, klasifikácii, synonymách a súvisiacich pojmoch (viď Obrázok 8), čo duplikuje informácie obsiahnuté v spomínanom využívanom podnikovom slovníku VocBench – treba prísť s riešením na mieru, ako informácie v týchto dvoch nástrojoch automatizovane synchronizovať. Je možné robiť import podnikového slovníka cez formát CSV.

---

[19] Zdroj: https://wiki.vicepremier.gov.sk/display/IN/OBJEKTY+EVIDENCIE, Dátum referencie: 31.07.2023

[20] Zdroj: https://towardsdatascience.com/how-to-bring-custom-ml-models-into-openmetadata-969311a16d91, Dátum referencie: 31.07.2023

[21] Zdroj: https://faisal-22508.medium.com/openmetadata-data-catalog-that-works-bfc24a76d69d, Dátum referencie: 31.07.2023

[22] Zdroj: https://ajithshetty28.medium.com/know-your-data-using-openmetadata-cddef8ecea25, Dátum referencie: 31.07.2023

- Potenciálne problémy s niektorými hotovými funkcionalitami - nemusia fungovať podľa očakávaní alebo aspoň podľa predstáv.
- Aby bolo možné nahrať niektoré požadované údaje, nastavenie si vyžaduje určité skúsenosti s vývojom alebo pomoc dátových inžinierov, pretože inak to môže byť pre bežného používateľa, ktorý chce len používať údaje v rámci dátového katalógu, príliš náročné. Hoci OpenMetadata propaguje jednoduchý spôsob implementácie konektorov pomocou používateľského rozhrania, často ich je potrebné vykonať ručne s určitými vývojárskymi zručnosťami.
- Implementované funkcionality pre jednotlivé konektory nemusia podporovať všetky vlastnosti dátového katalógu ako napríklad profilovanie metadát či „lineage“.

4.1.1 Princípy dizajnu a architektúry dátového katalógu

Nasledujúce princípy dizajnu a architektúry, ktoré nasleduje aj dátový katalóg OpenMetada, by mal dodržiavať akýkoľvek zvolený dátový katalóg:

1. Jednotný model metadát založený na schémach,
2. Otvorené a štandardizované rozhrania API pre integrácie,
3. Rozšíriteľnosť metaúdajov,
4. Automatizované nahrávanie metadát („pull-based metadata ingestion“),
5. Grafové ukladanie metaúdajov,

1. Jednotný model metadát

Organizácie pracujú s rôznymi zdrojmi údajov na rôzne účely. Tieto zdroje údajov majú svoju architektúru prispôsobenú konkrétnym prípadom použitia; niektoré sú orientované na neštruktúrované údaje vo formáte JSON, iné uchovávajú geolokačné údaje atď. Keďže sa tieto zdroje údajov líšia v spôsobe ukladania údajov, je prirodzené, že sa odlišne ukladajú aj základné metadáta.

Ak je snahou umožniť objavovanie nielen analytických, ale akýchkoľvek údajov („data discovery“) v rámci celej verejnej správy a podporiť plošný manažment údajov vrátane dátovej „lineage“, je nevyhnutné mať jednotný model metadát. Vďaka tomu bude možné centralizovane konfigurovať a udržiavať rôzne integrácie. S jednotným modelom metadát bude tiež jednoduché vystaviť metadáta na konzumáciu internými mikroslužbami a externými aplikáciami. Tu je diagram z blogu spoločnosti OpenMetadata, ktorý znázorňuje takéto nastavenie.

2. Otvorené a štandardizované rozhrania API pre integrácie

Jednotný model metadát pomáha dátovému katalógu umožniť lepšiu integráciu s rôznymi zdrojmi údajov. To spolu s otvorenými rozhraniami API založenými na dobre zdokumentovaných a široko akceptovaných štandardoch schém pomáha dátovému katalógu sprístupniť jednotný dátový model pre rôzne nadväzujúce aplikácie, akými sú napríklad analytické aplikácie, nástroje pre dátovú integráciu alebo engine pre riadenie dátovej kvality.

Otvorené rozhrania Open API sa musia opierať o rovnakú silne typovú, dobre štruktúrovanú a komentovanú schému, napríklad podľa špecifikácie JSON Schema.

Tento projekt je podporený z Európskeho sociálneho fondu.

Rovnakú špecifikáciu je dobré použiť aj na definovanie testov kvality údajov, respektíve ju treba prispôsobiť štandardu údajov.

3. Rozšíriteľnosť metaúdajov

Organizácia a jej procesy a priority sa neustále menia. Aby bolo možné vyhovieť meniacim sa požiadavkám, musí byť model metadát dostatočne flexibilný na to, aby zvládol všetky dodatočné dátové pvky, uzly a ďalšie polia.

To znamená, že jednotný model metadát možno koncepčne rozdeliť na dve časti - základný model metadát a rozšírený model metadát.

Základný model metadát pozostáva zo všetkých metadát, ktoré sú spoločné pre viaceré zdroje údajov, a rozšírený model metadát sa postará o všetky úpravy špecifické pre jednotlivé zdroje údajov. Dátový katalóg teda musí byť navrhnutý tak, aby bol rozšíriteľný.

4. Automatizované nahrávanie metadát („Pull-based metadata ingestion“)

Väčšina systémov na nahrávanie metadát je založená na princípe „pull“, čo znamená, že za extrakciu metadát je zodpovedný metadátový engine, a nie zdroj údajov. Niektoré metadátové katalógy, ako napríklad DataHub, podporujú načítavanie metadát na báze „push“ aj „pull“.

OpenMetadata zvolil prístup založený na „pull“, pretože autori OpenMetadata veria, že žiadny systém metadát nemôže byť založený výlučne na „push“. Za touto voľbou stojí myšlienka, že od zdrojov údajov nemožno očakávať, že budú tlačiť údaje do systému agregácie metadát. Úloha extrakcie a transformácie metadát do jednotného metadátového modelu pripadá na dátový katalóg, podobne ako to robí nástroj ETL pri vytváraní dátových jazier a dátových skladov.

5. Grafové úložisko pre metadáta

Dátový katalóg musí využívať prístup centralizovaného ukladania **metadát, kde sú aktívne organizované ako znalostný graf** spájajúci údaje so všetkými tímami, nástrojmi a procesmi.

To umožňuje organizáciám vytvárať, udržiavať a využívať znalostný graf metadát, ktorý následne môžu využívať nadväzujúce aplikácie a umožniť tak mnohé funkcionality s pridanou hodnotou, ako napríklad katalogizáciu údajov, manažment údajov („data governance“), dátovú „lineage“, automatizovanú kvalitu údajov a testovanie, profilovanie údajov, pozorovateľnosť údajov („data observability“) atď.

4.2 Nástroje pre dátovú „lineage“

Ďalším dôležitým aspektom manamžnetu údajov je dátová „lineage“, ktorú by mal moderný dátový katalóg umožniť používateľovi bez problémov používať. Dátová „lineage“ sa vzťahuje na historický záznam o dátovom aktíve, od jeho pôvodu až po súčasný stav, všetky transformácie a procesy, ktorými prešiel (viac v kapitole 2.1).

Na čo sa zamerať pri výbere nástrojov na sledovanie dátovej „lineage“

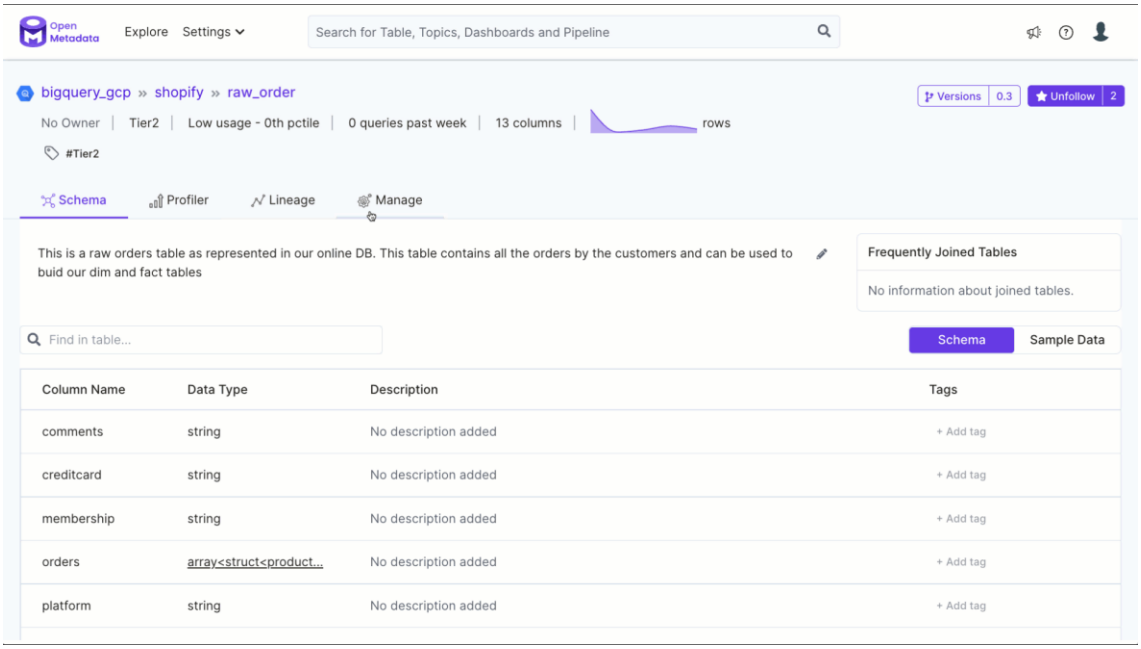
Tento projekt je podporený z Európskeho sociálneho fondu.

Ručné zhromažďovanie metadát a dokumentovanie dátovej „lineage“ si vyžaduje značné investície do zdrojov. Je tiež náchylné na chyby, čo môže spôsobiť veľké problémy, najmä keď sa organizácie čoraz viac spoliehajú na analýzu údajov pri plnení si svojej agendy. V dôsledku toho napomáha snahám o manažment údajov hľadanie nástrojov, ktoré spravujú reprezentácie dátovej „lineage“ a automaticky ich mapujú v rámci celej verejnej správy.

Ak padne rozhodnutie pokročiť v procese hodnotenia technológií pre prípadný nákup, treba sa poobzerať po nástrojoch pre dátovú „lineage“, ktoré dokážu nasledovné:

- natívne pristupovať k širokej škále zdrojov údajov, databáz, systémov a dátových produktov, skúmať metadáta, ktoré obsahujú, a zhromažďovať ich na účely manažmentu údajov, čoraz častejšie pomocou algoritmov umelej inteligencie a strojového učenia;
- agregovať získané metadáta do centralizovaného úložiska;
- odvodzovať typy údajov a priraďovať spoločné použitia referenčných údajov z podnikového slovníka a iných zdrojov k dátovým prvkom z rôznych systémov;
- poskytovať zjednodušené prezentácie agregovaných metadát koncovým používateľom a podporovať spoločné úsilie o validáciu opisov metadát;
- zdokumentovať koncové mapovanie toku údajov v systémoch organizácie.
- vytvárať vizualizované reprezentácie dátovej „lineage“;
- zahŕňať API pre vývojárov, ktoré môžu použiť pri vytváraní systémov, ktoré sa môžu pýtať na záznamy o dátovej „lineage“;
- vytvoriť inverzný index na mapovanie názvov dátových prvkov na ich použitie v rôznych fázach spracovania;
- ponúkať možnosť vyhľadávania na rýchle sledovanie toku údajov od miesta ich vzniku až po ich následné ciele,
- umožniť používateľom sledovať toky údajov smerom dopredu aj dozadu.

Tento projekt je podporený z Európskeho sociálneho fondu.



Obrázok 9: Ukážka dátovej „lineage“ v nástroji Open Metadata

Dodávatelia dátovej „lineage“

Nástroje na dokumentovanie a správu dátovej „lineage“ ponúka niekoľko rôznych typov dodávateľov vrátane nasledujúcich:

- veľkí dodávatelia IT riešení, ktorí predávajú platformy na manažment údajov, ako napríklad IBM, Informatica, Microsoft, Oracle, SAP a SAS, ako aj poskytovatelia cloudových platforiem AWS a Google Cloud;
- dodávatelia softvéru so širokým portfóliom produktov, ktoré zahŕňajú nástroje na správu a riadenie údajov, ako sú Hitachi Vantara, OneTrust, Precisely a Quest Software;
- dodávatelia, ktorí sa zameriavajú na správu a riadenie údajov, ako napríklad ASG Technologies, Ataccama, Boomi, Collibra, Semarchy, Syniti a **Talend**;
- špecialistov na správu metadát a dátovú „lineage“, ako sú Alex Solutions, Manta a Octopai, a
- dodávatelia nástrojov pre dátové katalógy, ako sú Alation, Atlan, Data.world a OvalEdge,
- dátová „lineage“ je aj súčasťou nástrojov s otvoreným zdrojovým kódom ako Apache Atlas a Open Metadata (viď Obrázok 9).

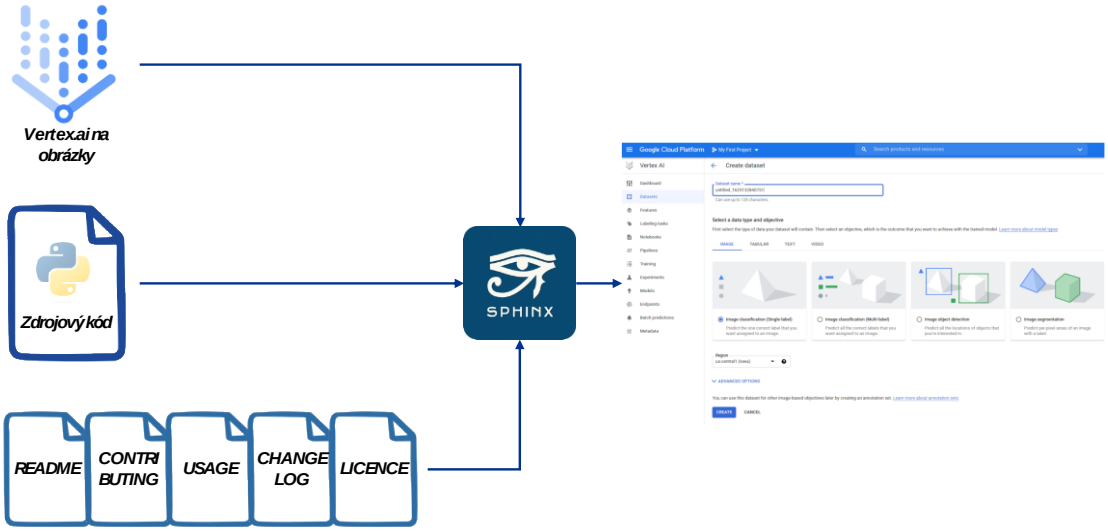
Dodávatelia, ktorí ponúkajú samoobslužný softvér na prípravu údajov pre dátových vedcov a analytické tímy, ako napríklad DataRobot a Trifacta spoločnosti Alteryx, tiež podporujú funkcionality dátovej „lineage“, rovnako ako rôzni dodávatelia nástrojov BI a analytických nástrojov na použitie v rámci aplikácií, ktoré na nich bežia.

Tento projekt je podporený z Európskeho sociálneho fondu.

4.2.1 Nástroje na popisovanie analytických aplikácií

4.2.1.1 Popisovanie analytických aplikácií s vlastným zdrojovým kódom alebo skriptom

Základnými črtami nástrojov pre efektívne popisovanie sú automatizácia a dostupnosť výstupu čitateľného pre čo najširší okruh používateľov, čo vedie k potrebe minimalizovať licencované nástroje, minimálne na oblasť konzumovania obsahu. Keďže analytické aplikácie a algoritmy strojového učenia sa najčastejšie programujú alebo skriptujú v jazyky Python (viac v dokumente „*4.2.1 Koncept pre zavádzanie analytického spracovania údajov do praxe*“), budeme sa v tejto časti venovať nástrojom dostupným pre tento ekosystém. Proces popisovania zahŕňa skombinovanie zdrojového kódu, dokumentácie (podľa kapitoly 3) a obrázkov a ich vypublikovanie v podobe interaktívneho obsahu ako webovej stránky, ako naznačuje Obrázok 10.



Obrázok 10: Prehľad procesu popisovania analytickej aplikácie v nástroji

Príkladom takého nástroja je výkonný a ľahko použiteľný nástroj s otvoreným zdrojovým kódom Spinx[23], ktorý slúži na automatické generovanie dokumentácie. Tento nástroj komunita používajúca Python veľmi intenzívne využíva. Dokáže generovať vynikajúcu štruktúrovanú dokumentáciu. Existuje niekoľko alternatív, ako napríklad MkDocs[24], Doxygen[25], pdoc[26] a ďalšie, ale Sphinx zostáva silným konkurentom. Hlavné funkcionality takého nástroja sú nasledovné:

- podpora viacerých výstupných formátov: HTML, PDF, obyčajný text, EPUB, TeX atď.
- automatické generovanie dokumentácie,

[23] Zdroj: https://www.sphinx-doc.org/en/master/, Dátum referencie: 25.07.2023

[24] Zdroj: https://www.mkdocs.org, Dátum referencie: 25.07.2023

[25] Zdroj: https://www.doxygen.nl, Dátum referencie: 25.07.2023

[26] Zdroj: https://pdoc.dev, Dátum referencie: 25.07.2023

Tento projekt je podporený z Európskeho sociálneho fondu.

- automatické generovanie odkazov,
- prispôsobenie dokumentácie s využitím loga, obrázkov a „markdown“ obsahu,
- podpora viacerých jazykov,
- rôzne rozšírenia k dispozícii.

Príklad využitia tohto nástroja na generovanie dokumentácie pre modely strojového učenia sa nachádzajú na tomto odkaze[27]. Sphinx je k dispozícii aj s množstvom ďalších rozšírení, ktoré možno použiť na to, aby bola dokumentácia ešte atraktívnejšia.

Ďalším prístupom ku generovaniu čitateľných kariet s popisom modelu bez ad hoc manuálnej práce sú karty DAG[28]. Sú inšpirované interaktívnymi kartami modelov od spoločnosti Google[29]. DAG[30] predstavuje smerový acyklický graf, ktorým je popísaná „pipeline“ strojového učenia, napríklad vo frameworkoch ako Tensorflow[31] alebo Pytorch[32]. Cieľom je, aby sa karty automaticky vytvárali zo zdrojového kódu a aby boli interaktívne. Spomínané karty modelov od spoločnosti Google väčšinou uspokojujú prípad použitia medzi podnikom a koncovými používateľmi, keď je jeden model celosvetovo dostupný vo veľkom meradle prostredníctvom verejného rozhrania. My však potrebujeme karty primárne určené pre interné publikum vo verejnej správe, kde je ťažké sledovať nové analytické aplikácie, do modelov sa vkladajú značné doménové znalosti a do školenia a testovania je zapojených niekoľko služieb a frameworkov.

Dostupný zdrojový kód[33] možno použiť na vygenerovanie karty DAG, ako ukazuje Obrázok 11. Nevýhodou tohto prístupu je, že musia byť splnené nasledujúce predpoklady:

- Používanie Metaflow[34] je rámec s otvoreným zdrojovým kódom, ktorý umožňuje rýchlo a jednoducho vytvárať a spravovať reálne projekty dátovej vedy a strojového učenia.
- Konto na Weights & Biases[35] s platným kľúčom API – ide o platformu, ktorá podporuje takzvané „MLOps“ – ide o sledovanie trénovania modelov a experimentov, manažment a verzionovanie údajov, nasadzovanie modelov do prevádzky, sledovanie ich výkonnosti a aktualizovanie modelov, spolupráca tímov.

Skript na vytváranie kariet je celkom jednoduchý. Ak je k dispozícii:

---

[27] Zdroj: https://towardsdatascience.com/document-your-machine-learning-project-in-a-smart-way-35c68aa5fc0e, Dátum referencie: 28.07.2023

[28] Zdroj: https://towardsdatascience.com/dag-card-is-the-new-model-card-70754847a111, Dátum referencie: 28.07.2023

[29] Zdroj: https://modelcards.withgoogle.com/face-detection, Dátum referencie: 28.07.2023

[30] Zdroj: https://towardsdatascience.com/machine-learning-execution-is-a-directed-acyclic-graph-211e5e1e6c57, Dátum referencie: 31.07.2023

[31] Zdroj: https://www.tensorflow.org, Dátum referencie: 31.07.2023

[32] Zdroj: https://pytorch.org, Dátum referencie: 31.07.2023

[33] Zdroj: https://github.com/jacopotagliabue/dag-card-is-the-new-model-card, Dátum referencie: 28.07.2023

[34] Zdroj: https://metaflow.org, Dátum referencie: 28.07.2023

[35] Zdroj: https://wandb.ai/site, Dátum referencie: 28.07.2023

Tento projekt je podporený z Európskeho sociálneho fondu.

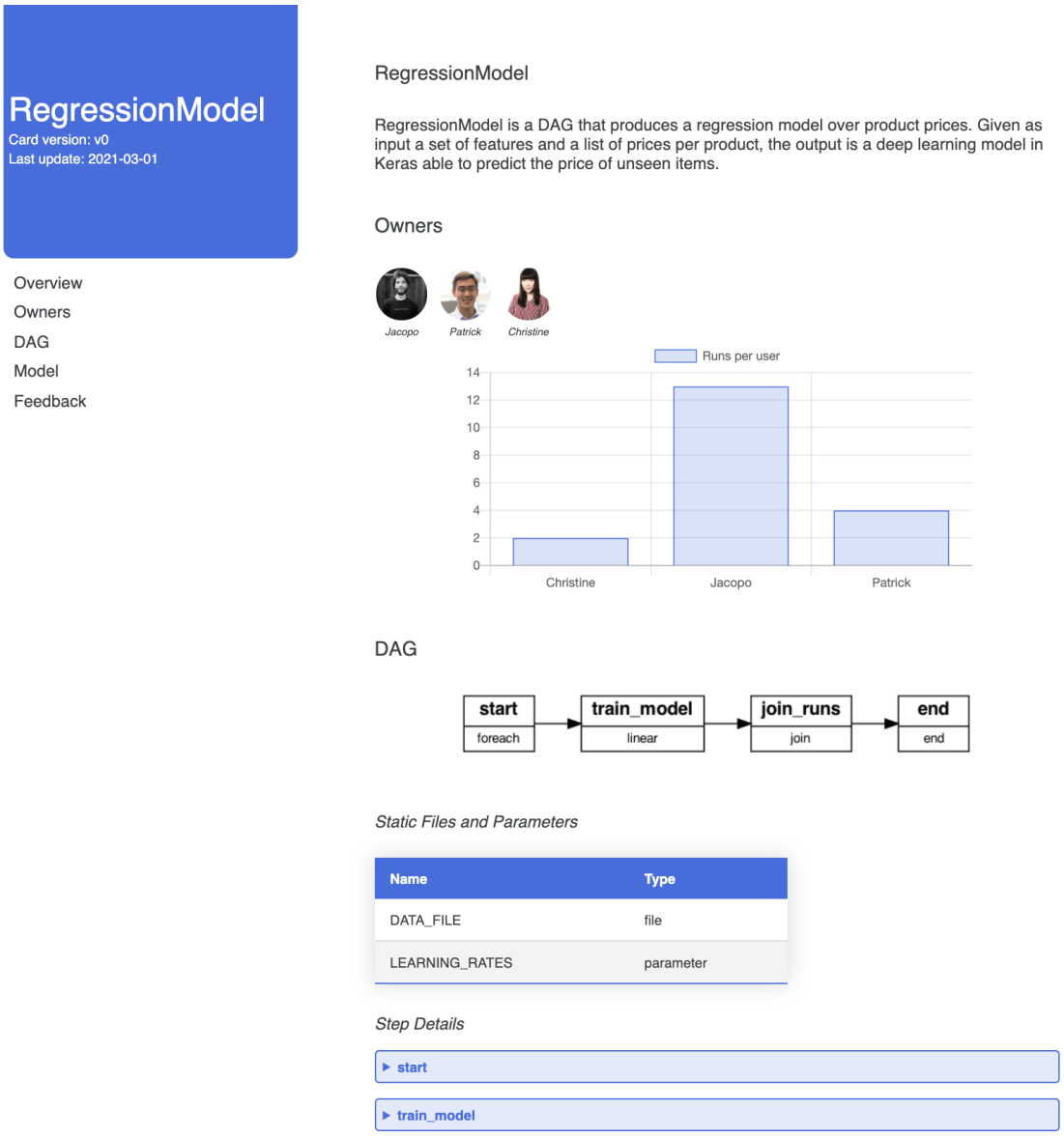
- HTML (Jinja) šablóna[36] - šablóny Jinja možno skladať ako matriošky, čo predstavuje principiálny spôsob rozšírenia používateľského rozhrania na oveľa viac prípadov použitia,
- trieda Metaflow (a konfiguračný súbor na inštanciovanie klienta[37]);
- kľúč API pre platformu Weights & Biases;

skript zavolá príslušné služby, získa informácie o DAG a modeli, pripraví množstvo JSON súborov a nakoniec zostaví statickú stránku HTML na použitie. Inými slovami, skript na tvorbu kariet prevezme šablónu Jinja a vyplní sloty volaním API z Metaflow, Weights & Biases a prípadne ďalších služieb, pričom výstupom je jednoduchá statická stránka HTML, ako ukazuje Obrázok 11.

[36] Zdroj: https://jinja.palletsprojects.com/en/2.11.x/api/#basics, Dátum referencie: 28.07.2023

[37] Zdroj: https://docs.metaflow.org/metaflow/client, Dátum referencie: 28.07.2023

Tento projekt je podporený z Európskeho sociálneho fondu.



Obrázok 11: Pohľad na kartu DAG s vlastníkmi, úlohami, architektúrou modelu, vstupnými datasetmi a parametrami.

Obsah stránky karty DAG je rozdelený do piatich hlavných častí (Obrázok 11):

1. **Prehľad**: vysokoúrovňový opis karty DAG,
2. **Vývojári**: časť obsahujúca vývojárov/vlastníkov karty DAG a graf zobrazujúci rozdelenie behov („runs“) medzi nimi, teda ako intenzívne daný model vyvíjali a používali.
3. **Schéma DAG**: vizuálny opis DAG a sekcie s podrobnosťami o jednotlivých krokoch (vytvorené prostredníctvom vlastnej reprezentácie Metaflow).
4. **Model**: sekcie uvádzajúce metriky pre posledných K behov modelu, cestu do cloudového úložiska (pre ML inžinierov, ktorí chcú obnoviť presný artefakt), náčrt architektúry a graf strát/epoch.

Tento projekt je podporený z Európskeho sociálneho fondu.

5. **Testy**: okrem kontrol kódu (napr. veľkosť vektora) a všeobecných kvantitatívnych testov (napr. presnosť) sú testy správania sa modelu navrhnuté tak, aby boli kontrolou správnosti kvalitatívneho správania modelu a sondou výkonnosti v prípadoch, ktoré sa považujú za dôležité pre nasadenie (napr. výkonnosť modelu na malej, ale kľúčovej množine vstupov). Prirodzeným rozšírením pre zložitejšie modely by bolo rozdelenie presnosti podľa rozdelenia vstupného priestoru (povedzme pravdepodobnosť nesplácania dlhov podľa pohlavia a veku): zatiaľ čo ľudská analýza je neprekonateľná, testy správania sa modelu pôsobia ako silná ochrana pred zjavnými (ale stále nie zriedkavými) chybami.

Tento koncept popisu modelov cez karty DAG predstavuje prvý vstup do vytvorenia zložitejšieho „toolchainu“ na základe potrieb analytických tímov a reálnej skúsenosti z prostredia verejnej správy.

Ďalším dôležitým a veľmi využívaným nástrojom je **Pickle**[38] v prostredí jazyka Python. Pomocou tohto nástroja sa dá uložiť výsledný dataset a použiť ho na trénovanie viacerých modelov alebo možno uložiť samotný model strojového učenia a otestovať ho na viacerých datasetoch bez toho, aby bolo potrebné model znovu trénovať. V prvom prípade stačí odovzdať dáta s navrhnutými vstupnými premennými funkciou do nástroja Pickle a uložiť ich v binárnom formáte. Potom sa tento dataset vo formáte Pickle dajú načítať späť, kedykoľvek ste pripravení vykonať modelovanie. V druhom prípade sa vytrénovaný model uloží a načíta späť vždy, keď budú k dispozícii nové údaje, na ktorých sa dá otestovať, pričom už netreba opäť realizovať niekoľkohodinové alebo aj dlhšie trénovanie modelu.

[38] Zdroj: https://towardsdatascience.com/why-turn-into-a-pickle-b45163007dac, Dátum referencie: 28.07.2023

Tento projekt je podporený z Európskeho sociálneho fondu.

5 Príklad správneho popisu

5.1 Príklad správneho popisu analytického datasetu

Tento popis vychádza z konceptu popisu Dátovej karty pre datasety učené na strojové učenie, ako bolo popísané v úvode kapitoly 2. Takýto popis (dokumentáciu) možno vytvoriť v MetaIS (Confluence), ktorý vytvára vhodný priestor na spoluprácu.

1. Názov datasetu (skratka)	*Napíšte krátke zhrnutie opisujúce dataset (maximálne 200 slov). Uveďte informácie o obsahu a téme údajov, zdrojoch a motivácii vytvorenia datasetu, prínosoch a problémoch alebo prípadoch použitia, na ktoré je vhodný.*
LINK DATASET	AUTOR(I) DÁTOVEJ KARTY
Uveďte odkaz na samotný dataset a na jeho záznam v dátovom katalógu:	*Vyberte **jednu rolu** pre každého autora:* *(Poznámka k používaniu: Vyberte najvhodnejšiu možnosť, ktorá opisuje úlohu autora pri vytváraní dátovej karty).*
<Odkaz na dataset> <Odkaz na záznam v dátovom katalógu>	**Meno a Priezvisko, kontakt:** (vlastník / prispievateľ / manažér) **Meno a Priezvisko, kontakt:** (vlastník / prispievateľ / manažér) **...**

2. Autorstvo - Poskytovatelia		
ORGANIZÁCIA(-E) POSKYTOVATEĽA	SEKTOR A ODVETVIE	KONTAKTNÉ ÚDAJE
Uveďte názvy inštitúcií alebo organizácií zodpovedných za zverejnenie datasetu:	*Vyberte **všetky príslušné** sektory a odvetvia, do ktorých organizácie poskytovateľov patria:*	*Uveďte kontaktné údaje poskytovateľa:*

Tento projekt je podporený z Európskeho sociálneho fondu.

Názov organizácie alebo inštitúcie	Verejný sektor - rezort Súkromný sektor – odvetvie (napríklad telekomunikácie, digitálny marketing, IT atď.) Akademický sektor – zameranie inštitúcie (prírodovedné vedy, technické vedy atď.) Neziskový sektor – Zameranie MNO (napríklad digitálne zručnosti, sociálna pomoc apod) Jednotlivec alebo iné (uveďte)	**Meno a Priezvisko:** <Uveďte meno a priezvisko poskytovateľa> **Zaradenie:** <Uveďte zaradenie a oddelenie> **Kontakt:** <Uveďte kontaktné údaje> **Webová stránka:** <Uveďte webovú stránku datasetu, ak je k dispozícii.>

Autorstvo - Vlastníci datasetu

ODDELENIE(A)	KONTAKTNÉ ÚDAJE	AUTOR(I)
Uveďte názov oddelenia alebo tímu, ktorý vlastní dataset:	*Poskytnite spôsoby na kontaktovanie vlastníkov datasetu:*	*Uveďte údaje o všetkých autoroch spojených s datasetom:* *(Poznámka k používaniu: Uveďte afiliáciu a rok, ak sa líši od inštitúcií alebo organizácií poskytovateľa)*
Názov oddelenia alebo tímu	**Vlastník(-ci) datasetu:** <Uveďte mená vlastníkov datasetu> **Afiliácia:** <Uveďte afiliáciu vlastníkov datasetu> **Kontakt:** <Uveďte e-mail vlastníka datasetu> **Skupinový e-mail:** <Uveďte odkaz na mailing-list pre tím vlastníka datasetu>. **Webová stránka:** <Uveďte odkaz na webovú stránku oddelenia alebo tímu vlastníka datasetu>	Meno, titul, afiliácia, RRRR Meno, titul, afiliácia, RRRR Meno, titul, afiliácia, RRRR Meno, titul, afiliácia, RRRR

Tento projekt je podporený z Európskeho sociálneho fondu.

3. Prehľad o datasete a jeho charakteristikách

Popis obsahu datasetu	*Uveďte krátky opis obsahu datasetu v bodoch.* *<Uveďte odkazy, ak sú k dispozícii>* ***Ďalšie poznámky:*** *<Pridajte sem>*

SUBJEKT(-Y) ÚDAJOV	CHARAKTERISTIKA DATASETU
*Vyberte **všetky príslušné** subjekty obsiahnuté v datasete:*	*Uveďte stručnú charakteristiku datasetu:* *(Použite dodatočné poznámky, aby ste uviedli príslušné informácie, a odkazy na tabuľky s podrobnejším členením)*
• Citlivé údaje o ľuďoch (PII) • Údaje o ľuďoch (nejde o PII) • Údaje o prírodných javoch • Údaje o miestach a objektoch • Synteticky generované údaje • Údaje o systémoch alebo službách a ich správaní • Neznáme • Iné (uveďte prosím)	<Uveďte nadpis pre uvedenú tabuľku alebo vizualizáciu.> Veľkosť súboru údajov: 123456 MB Počet inštancií: 123456 Počet premenných: 123456 Označené triedy: 123456 Počet štítkov („labels“): 123456789 Priemerný počet štítkov na inštanciu: 12,34 Algoritmické štítky: 123456789 Ručne vytvorené štítky: 123456789 Iné charakteristiky: 123456 **Ďalšie poznámky:** <Pridajte sem>

Tento projekt je podporený z Európskeho sociálneho fondu.

POPISNÁ ŠTATISTIKA

Uveďte základné popisné štatistiky pre každú premennú. Pomocou dodatočných poznámok zaznamenajte všetky ďalšie dôležité informácie alebo úvahy.
Poznámka k používaniu: Niektoré štatistiky budú relevantné pre číselné údaje, ale nie pre textové reťazce.

<Uveďte nadpis pre uvedenú tabuľku alebo vizualizáciu.>

Štatistika	Názov premennej	Názov premennej	Názov premennej	Názov premennej
počet				
priemer				
štandardná odchýlka				
minimum				
25%				
50%				
75%				
maximum				
mód				

Ďalšie poznámky: <Pridajte sem>

CITLIVOSŤ ÚDAJOV

TYP(Y) CITLIVOSTI	PREMENNÉ S CITLIVÝMI ÚDAJMI	BEZPEČNOSŤ A OCHRANA SÚKROMIA
*Vyberte **všetky príslušné** typy údajov prítomné datasete:*	*Uveďte premenné v datasete, ktoré obsahujú (C)/PII, a uveďte, či ich zber bol úmyselný alebo neúmyselný.*	*Zhrňte opatrenia alebo kroky na spracovanie citlivých údajov v tomto datasete.*

Tento projekt je podporený z Európskeho sociálneho fondu.

	Pomocou dodatočných poznámok zaznamenajte všetky ďalšie dôležité informácie alebo úvahy.	*Pomocou dodatočných poznámok zaznamenajte všetky ďalšie dôležité informácie alebo úvahy.*
• Obsah používateľa • Metadáta používateľa • Údaje o činnosti používateľa • Identifikovateľné údaje • Citlivé údaje (C) • PII, • Podnikové údaje • Údaje o zamestnancoch • Pseudonymizo-vané údaje • Anonymné údaje • Údaje o zdraví • Údaje o deťoch • Žiadne • Iné (uveďte prosím)	**Zámerne zozbierané citlivé údaje** (C/PII boli zozbierané v rámci procesu tvorby datasetu.) **Názov premennej** / **Popis** <Názov premennej> / <Typ C/PII> <Názov premennej > / <Typ C/PII> ... / ... **Neúmyselne zozbierané citlivé údaje** (C/PII neboli explicitne zozbierané v rámci procesu tvorby datasetu, ale možno ich odvodiť pomocou metód.) **Názov premennej** / **Popis** <Názov poľa> / <Typ C/PII> <Názov poľa> / <Typ C/PII> ... / ...	<Zhrňte tu. V prípade potreby uveďte odkazy a metriky.> **<Metóda>:** [popis] **<Metóda>:** [popis] **<Metóda>:** [popis] **Ďalšie poznámky:** <Pridajte sem>

Tento projekt je podporený z Európskeho sociálneho fondu.

TYP(Y) RIZIKA	ODKAZY NA ĎALŠIU DOKUMENTÁCIU	RIZIKO(-Á) A MITIGÁCIA(-E)
*Vyberte **všetky príslušné** typy rizík, ktoré sú spojené s datasetom:*	*Uveďte odkaz(-y) na dokumentáciu týkajúcu sa citlivých údajov v datasete:*	*Zhrňte kroky prijaté na identifikáciu a zmiernenie rizík vyplývajúcich z PII alebo citlivých informácií.*
• Priame riziko • Nepriame riziko • Reziduálne riziko • Žiadne známe riziká • Iné (uveďte)	**<Názov odkazu alebo typ dokumentu>:** [Odkaz] **<Názov odkazu alebo typ dokumentu>:** [Odkaz] **<Názov odkazu alebo typ dokumentu>:** [Odkaz]	<Zhrňte tu. V prípade potreby uveďte odkazy a metriky.> **<Typ rizika>:** [Popis + mitigácia] **<Typ rizika>:** [Popis + mitigácia] **Ďalšie poznámky:** <Pridajte sem>

VERZIA A ÚDRŽBA DATASETU		
STAV ÚDRŽBY	PODROBNOSTI O VERZII	PLÁN ÚDRŽBY
Vyberte jednu z nich:	*Uveďte podrobnosti o tejto verzii datasetu:*	*Zhrňte plán údržby datasetu:*

Tento projekt je podporený z Európskeho sociálneho fondu.

Pravidelne aktualizované (Nové verzie datasetu boli alebo budú naďalej sprístupňované.) **Aktívne udržiavané** (Nebudú žiadne nové verzie, ale tento dataset sa bude aktívne udržiavať, vrátane aktualizácií údajov.) **Obmedzená údržba** (Údaje sa nebudú aktualizovať, ale budú sa riešiť všetky technické problémy.) **Vyradené** (Tento dataset je zastaraný alebo sa už neudržiava.)	**Aktuálna verzia:** 1.0 **Posledná aktualizácia:** MM/RRRR **Dátum vydania:** MM/RRRR	<V prípade potreby uveďte odkazy a metriky.> **Verzionovanie:** <Uveďte informácie o kritériách pre verzionovanie datasetu.> **Aktualizácie:** <Uveďte informácie o kritériách obnovy alebo aktualizácie datasetu.> **Chyby:** <Uveďte informácie o tom, ako sa chyby riešia alebo spracúvajú.> **Spätná väzba:** <Uveďte informácie o poskytovaní spätnej väzby.> **Ďalšie poznámky:** <Pridajte sem>
	ĎALŠIE PLÁNOVANÉ AKTUALIZÁCIE	OČAKÁVANÁ ZMENA (ZMENY)
Tento riadok vyplňte pri bežiacej aktualizácií alebo údržbe:	*Uveďte podrobnosti o ďalšej plánovanej aktualizácii:*	*Zhrňte aktualizácie, ktoré sa očakávajú pri ďalšej aktualizácii.*
	Dotknutá verzia: 1.0 **Ďalšia aktualizácia údajov:** MM/RRRR **Ďalšia verzia:** 1.1 **Ďalšia aktualizácia verzie:** MM/RRRR	**Aktualizácie údajov:** <Podľa potreby uveďte odkazy, grafy a vizualizácie.> **Aktualizácie súboru údajov:** < Podľa potreby uveďte odkazy, grafy a vizualizácie.>

Tento projekt je podporený z Európskeho sociálneho fondu.

4. Príklad dátových bodov

<table>
<tr><th>PRIMÁRNA MODALITA ÚDAJOV</th><th>VÝBER VZORIEK BODOV</th><th>PREMENNÉ V DATASETE</th></tr>
<tr><td>Vyberte jednu z nich:</td><td>Poskytnite linky na dátové body alebo ukážky:</td><td>Uveďte premenné v dátových bodoch a ich opisy.

(Poznámka k používaniu: Popíšte každú premennú v dátovom bode. Prípadne použite na ukážku príklad.)</td></tr>
<tr><td>• Obrazové údaje
• Textové údaje
• Tabuľkové údaje
• Zvukové údaje
• Údaje o videu
• Časové rady
• Údaje v grafe
• Geopriestorové údaje
• Multimodálna doprava (uveďte)
• Neznáme
• Iné (uveďte prosím)</td><td>[Demo Link]

[Odkaz na typické dátové body]

[Odkaz na odľahlý dátový bod]

[Odkaz na iný dátový bod]

[Odkaz na iný dátový bod]</td><td><Uveďte nadpis pre uvedenú tabuľku alebo vizualizáciu, ak sa používa.>
<table><tr><th>Názov premennej</th><th>Typ premennej</th><th>Popis</th></tr><tr><td><Názov premennej></td><td><Hodnota premennej ></td><td><Popis></td></tr><tr><td><Názov premennej ></td><td><Hodnota premennej ></td><td><Popis></td></tr><tr><td>...</td><td>...</td><td>...</td></tr></table>Ďalšie poznámky: <Pridajte sem></td></tr>
<tr><th>TYPICKÝ DÁTOVÝ BOD</th><th></th><th>ODĽAHLÝ DÁTOVÝ BOD</th></tr>
<tr><td>Uveďte príklad typického dátového bodu a opíšte, čo ho robí typickým.

Pomocou dodatočných poznámok zaznamenajte všetky ďalšie dôležité informácie alebo úvahy.</td><td></td><td>Uveďte príklad odľahlého dátového bodu a opíšte, čo ho robí atypickým.

Pomocou dodatočných poznámok zaznamenajte všetky ďalšie dôležité informácie alebo úvahy.</td></tr>
</table>

Tento projekt je podporený z Európskeho sociálneho fondu.

<Zhrňte tu. Uveďte všetky kritériá typickosti dátového bodu> { "výška": 1,76, "váha": 78,9, }, **Ďalšie poznámky:** <Pridajte sem>	<Zhrňte tu. Uveďte všetky kritériá odľahlosti dátového bodu> { "výška": 10,53, "váha": 88,9, }, **Ďalšie poznámky:** <Pridajte sem>

Tento projekt je podporený z Európskeho sociálneho fondu.

5. Motivácia vytvorenia datasetu a zámer jeho využívania - Motivácia

ÚČEL(Y)	OBLASŤ(-I) POUŽITIA	MOTIVAČNÝ(É) FAKTOR(Y)
Vyberte jednu z nich:	*Uveďte zoznam kľúčových oblastí použitia, pre ktoré bol dataset navrhnutý:* *(Poznámka k použitiu: Použite kľúčové slová oddelené čiarkami.)*	*Uveďte hlavné motivácie na vytvorenie alebo kurátorovanie tohto datasetu:*
• Monitorovanie • Plánovanie • Lepšie služby • Výskum • Iné (uveďte)	*Napríklad: „Detekcia objektov“, „Klasifikácia“, „Hľadanie vzorov“, „Identifikácia podobností“, „Prognózovanie“.* „kľúčové slovo“, „kľúčové slovo“, kľúčové slovo“	*Napríklad:* *- Zvýšenie demografickej diverzity údajov na trénovanie snímok pre modely detekcie objektov.* *- Včasnejšia dostupnosť odhadu vývoja priemyselnej výroby.* <Zhrňte motiváciu tu. V prípade potreby uveďte odkazy.>

Zamýšľané využitie

POUŽITIE(A) DATASETU	VHODNÉ PRÍPADY POUŽITIA	NEVHODNÝ(É) PRÍPAD(Y) POUŽITIA
Vyberte jednu z nich:	*Zhrňte známe vhodné a plánované prípady použitia tohto datasetu.* *Pomocou dodatočných poznámok zachyťte všetky špecifické vzory, na ktoré by si*	*Zhrňte známe nevhodné a nezamýšľané prípady použitia tohto datasetu.* *Pomocou dodatočných poznámok zachyťte všetky špecifické vzory, na ktoré by si*

Tento projekt je podporený z Európskeho sociálneho fondu.

	mali konzumenti dať pozor, alebo iné dôležité informácie či úvahy.	*mali konzumenti dať pozor, alebo iné dôležité informácie či úvahy.*
• Bezpečné na použitie pri plínovaní • Bezpečné na použitie vo výskume • Podmienené použitie - niektoré nebezpečné aplikácie • Len schválené použitie • Iné (uveďte prosím)	**[Vhodný prípad použitia]** : <Zhrňte tu. V prípade potreby uveďte odkazy.> **[Vhodný prípad použitia]**: <Zhrňte tu. V prípade potreby uveďte odkazy.> **[Vhodný prípad použitia]**: <Zhrňte tu. V prípade potreby uveďte odkazy.> **Ďalšie poznámky:** <Pridajte sem>	**[Nevhodný prípad použitia]** : <Zhrňte tu. V prípade potreby uveďte odkazy.> **[Nevhodný prípad použitia]**: <Zhrňte tu. V prípade potreby uveďte odkazy.> **[Nevhodný prípad použitia]**: <Zhrňte tu. V prípade potreby uveďte odkazy.> **Ďalšie poznámky:** <Pridajte sem>

Tento projekt je podporený z Európskeho sociálneho fondu.

6. Prístup, doba uchovávania a vymazanie datasetu - Prístup

TYP PRÍSTUPU	ODKAZ(Y) NA DOKUMENTÁCIU	PREDPOKLAD(Y)
Vyberte jednu z nich:	*Uveďte odkazy, ktoré opisujú dokumentáciu na prístup k tomuto datasetu:*	*Opíšte všetky požadované školenia alebo predpoklady na prístup k tomuto datasetu.*
• Interné - neobmedzené • Interné - obmedzené • Externý - otvorený prístup • Iné (uveďte prosím)	[URL adresa webovej stránky súboru údajov] [Github URL]	Napríklad, Tento dataset vyžaduje členstvo v [konkrétnych] databázových skupinách: • Absolvujte [Povinné školenie] • Prečítajte si [Zásady používania údajov] • Iniciovanie žiadosti o údaje podaním [odkaz]
	ODKAZ(Y) NA POLITIKU	ZOZNAM(Y) RIADENIA PRÍSTUPU
	Uveďte odkaz na zásady prístupu:	*Zoznam a zhrnutie všetkých zoznamov riadenia prístupu súvisiacich s týmto datasetom. V prípade potreby uveďte odkazy.*
	• Priama adresa URL na stiahnutie • Adresa URL iného úložiska Kód na prevzatie údajov # ...	**[Zoznam riadenia prístupu]:** <Napíšte sem zhrnutie a poznámky.> **[Zoznam riadenia prístupu]:** <Napíšte sem zhrnutie a poznámky.> **Ďalšie poznámky:** <Pridajte sem>

Tento projekt je podporený z Európskeho sociálneho fondu.

Doba uchovávania

TRVANIE	ZHRNUTIE POLITIKY
Zadajte dobu, počas ktorej sa tento dataset môže uchovávať:	*Zhrňte politiku uchovávania tohto datasetu.*
<Uveďte dĺžku trvania v dňoch, mesiacoch alebo rokoch.>	**ID plánu uchovávania:** <napíšte tu> **Zhrnutie:** <na tomto mieste napíšte zhrnutie a poznámky>
PRÍRUČKA PROCESU	VÝNIMKA(-Y)
Zhrňte všetky požiadavky a súvisiace kroky na uchovávanie datasetu. *Pomocou dodatočných poznámok zaznamenajte všetky ďalšie dôležité informácie.*	*Zhrňte všetky výnimky a súvisiace kroky na uchovávanie da. V prípade potreby uveďte odkazy.* *Pomocou dodatočných poznámok zaznamenajte všetky ďalšie dôležité informácie.*
Napríklad Tento súbor údajov je v súlade so [štandardnými usmerneniami] **Ďalšie poznámky:** <Pridajte sem>	**Kód výnimky:** `ANONYMNÉ_ÚDAJE` / `ZAMESTNANECKÉ_ÚDAJE` / `OTVORENÉ_ÚDAJE` / `INTERNÉ_ _ÚDAJE` / `SIMULOVANÉ_TESTOVACIE_ ÚDAJE` **Zhrnutie:** <Napíšte sem zhrnutie a poznámky.> **Ďalšie poznámky:** <Pridajte sem>

Vymazanie datasetu

TRVANIE	ZHRNUTIE UDALOSTI VYMAZANIA

Tento projekt je podporený z Európskeho sociálneho fondu.

Zadajte trvanie, po ktorom sa má tento dataset vymazať:	*Zhrňte postupnosť udalostí a prípustné spracovanie pri vymazávaní údajov.*
<Uveďte dĺžku trvania v dňoch, mesiacoch alebo rokoch.>	**Postupnosť udalostí vymazania a spracovania:** <Súhrn prvej udalosti.> <Súhrn druhej udalosti.> < Súhrn tretej udalosť.> **Ďalšie poznámky:** <Pridajte sem>
PRIJATEĽNÉ SPÔSOBY VYMAZANIA	ZÁVÄZKY PO VYRADENÍ
Uveďte prijateľné spôsoby vymazania:	*Zhrňte postupnosť povinností po udalosti vymazania.*
<Napíšte prijateľné spôsoby vymazania.> <Napíšte prijateľné spôsoby vymazania.> <Napíšte prijateľné spôsoby vymazania.>	**Postupnosť záväzkov po vyradení:** <Zhrňte prvú povinnosť tu.> <Zhrňte druhú povinnosť tu.> <Zhrňte tretiu povinnosť tu.> **Ďalšie poznámky:** <Pridajte sem>
PREVÁDZKOVÉ POŽIADAVKY	VÝNIMKY A OSLOBODENIA
Uveďte všetky prevádzkové požiadavky na integráciu vymazania:	*Zhrňte všetky výnimky a súvisiace kroky k udalosti vymazania.*
<Napíšte sem prvú požiadavku.> <Napíšte druhú požiadavku.> <Napíšte tretiu požiadavku.>	**Chyba výnimky politiky:** [chyba] **Zhrnutie:** <Napíšte sem zhrnutie a poznámky.> **Ďalšie poznámky:** <Pridajte sem>

Tento projekt je podporený z Európskeho sociálneho fondu.

7. Pôvod a uchovávanie datasetu („Provenance“) - zber

POUŽITÉ METÓDY	PODROBNOSTI O METODIKE	POPIS(Y) ZDROJA
*Vyberte **všetky príslušné** metódy použité na zber údajov:*	*Uveďte opis každej použitej metódy zberu.* *(Poznámka k používaniu: Duplikujte a vyplňte nasledujúce údaje pre typ metódy zberu.)*	*Uveďte opis každého zdroja údajov.*
• API • Umelo vytvorené • Crowdsourco-vané - Platené • Crowdsourco-vané - Dobrovoľník • Zo zberu dodávateľa • Scrapované alebo crawled • Prieskum, formuláre alebo ankety • Prevzaté z iných existujúcich datasetov • Neznáme • Bude určené • Iné (uveďte prosím)	**<Metóda zberu>** **Zdroj:** <Popíšte tu. Uveďte odkazy, ak sú k dispozícii.> **Platforma:** [Názov platformy], <Popíšte platformu. V prípade potreby uveďte odkazy.> **Považuje sa tento zdroj za citlivý alebo vysoko rizikový?** [Áno / Nie] **Dátumy zberu:** [MMM RRRR - MMM RRRR] **Primárna modalita zozbieraných údajov:** *Poznámka k používaniu: Vyberte jeden pre tento typ zberu.* • Obrazové údaje • Textové údaje • Tabuľkové údaje • Zvukové údaje • Video • Časové rady • Grafové údaje • Priestorové údaje • Neznáme • Multimodálna doprava (uveďte) • Iné (uveďte prosím)	**[Zdroj]:** <Popíšte tu. V prípade potreby uveďte odkazy, príklady údajov, metriky, vizualizácie.> **[Zdroj]:** <Popíšte tu. V prípade potreby uveďte odkazy, príklady údajov, metriky, vizualizácie.> **[Zdroj]:** <Popíšte tu. V prípade potreby uveďte odkazy, príklady údajov, metriky, vizualizácie.> **Ďalšie poznámky:** <Pridajte sem>

Tento projekt je podporený z Európskeho sociálneho fondu.

	Frekvencia aktualizácie zozbieraných údajov: *Poznámka k používaniu: Vyberte jeden pre tento typ zberu.* • Ročne • Štvrťročne • Mesačne • Dvojtýždenne • Týždenne • Denne • Každú hodinu • Statické údaje • Iné (uveďte prosím) **Ďalšie odkazy pre tento zber:** • [Politika prístupu] • [Politika vymazania] • [Politika uchovávania] **Ďalšie poznámky:** <Pridajte sem>	
SPÔSOB ZBERU	INTEGRÁCIA ÚDAJOV	SPRACOVANIE ÚDAJOV
Vyberte všetky príslušné:	*Uveďte všetky premenné zozbierané z rôznych zdrojov a uveďte, či boli zahrnuté alebo vylúčené z datasetu.* *(Poznámka k používaniu: Duplikujte a vyplňte nasledujúce údaje pre každý zdroj.)*	*Zhrňte, ako sa údaje z rôznych zdrojov alebo metód agregovali, spracovali alebo prepojili.* *(Poznámka k použitiu: Duplikujte a vyplňte nasledujúce údaje pre každý zdroj alebo metódu zberu.)*
Statické údaje (Údaje boli zozbierané raz z jedného alebo viacerých zdrojov.) **Streamované**	**<Zdroj>** **Zahrnuté premenné** (Premenné, ktoré boli zozbierané a sú zahrnuté v datasete.) **Názov** | **Popis**	**<Metóda zberu alebo zdroj>** **Popis:** <Popíšte tu. V prípade potreby uveďte odkazy.> **Použité metódy:** <Popíšte tu. V

Tento projekt je podporený z Európskeho sociálneho fondu.

(Údaje sa získavajú priebežne z jedného alebo viacerých zdrojov.)

Dynamické

(Údaje sa pravidelne aktualizujú z jedného alebo viacerých zdrojov.)

Iné (uveďte prosím)

premennej	
<Názov premennej >	<Popíšte tu. V prípade potreby uveďte odkazy, príklady údajov, metriky, vizualizácie.>
...	...

Ďalšie poznámky: <Pridajte sem>

Vylúčené premenné

(Premenné, ktoré boli zozbierané, ale sú vylúčené z datasetu.)

Názov premennej	Popis
<Názov premennej >	<Popíšte tu. V prípade potreby uveďte odkazy, príklady údajov, metriky, vizualizácie.>
...	...

Ďalšie poznámky: <Pridajte sem>

prípade potreby uveďte odkazy.>

Nástroje alebo knižnice: <Popíšte tu. V prípade potreby uveďte odkazy.>

Ďalšie poznámky: <Pridajte sem>

Kritériá zberu

VÝBER ÚDAJOV	ZAČLENENIE ÚDAJOV	VYLÚČENIE ÚDAJOV

Tento projekt je podporený z Európskeho sociálneho fondu.

Zhrňte kritériá výberu údajov.	*Zhrňte kritériá zahrnutia údajov.*	*Zhrňte kritériá vylúčenia údajov.*
<Metóda zberu alebo zdroj>: <Zahrňte kritériá výberu údajov. Uveďte odkazy, ak sú k dispozícii.> **<Metóda zberu alebo zdroj>:** <Zahrňte kritériá výberu údajov. Uveďte odkazy, ak sú k dispozícii.> **Ďalšie poznámky:** <Pridajte sem>	**<Metóda zberu alebo zdroj>:** <Zahrňte tu kritériá zahrnutia údajov. Uveďte odkazy, ak sú k dispozícii.> **<Metóda zberu alebo zdroj>:** <Zahrňte tu kritériá zahrnutia údajov. Uveďte odkazy, ak sú k dispozícii.> **Ďalšie poznámky:** <Pridajte sem>	**<Metóda zberu alebo zdroj>:** <Zahrňte tu kritériá vylúčenia údajov. Uveďte odkazy, ak sú k dispozícii.> **<Metóda zberu alebo zdroj>:** <Zahrňte tu kritériá vylúčenia údajov. Uveďte odkazy, ak sú k dispozícii.> **Ďalšie poznámky:** <Pridajte sem>

Tento projekt je podporený z Európskeho sociálneho fondu.

8. Citlivé atribúty

ODÔVODNENIE	*Opíšte motiváciu, zdôvodnenie, úvahy alebo prístupy, ktoré viedli k tomu, že tento dataset obsahuje uvedené citlivé atribúty.* *Zhrňte, prečo alebo ako to môže ovplyvniť používanie datasetu.* *< Podľa potreby uveďte odkazy, tabuľky a médiá>*
CITLIVÝ(É) ATRIBÚT(Y)	ZÁMER
*Vyberte **všetky atribúty,** ktoré sú zastúpené (priamo alebo nepriamo) v datasete.*	*Vypíšte premenné v datasete, ktoré obsahujú citlivé atribúty, a uveďte, či ich zber bol zámerný alebo neúmyselný.*
• Vek • Sociálno-ekonomický status • Pobyt • Jazyk • Národnosť • Náboženstvo • Zdravotný stav • Skúsenosti • Výška platu • Iné (uveďte)	**Zámerne zozbierané atribúty** (Citlivé atribúty boli označené alebo zozbierané v rámci procesu tvorby datasetu.) **Ďalšie poznámky:** <Pridajte sem> **Neúmyselne zozbierané atribúty** (Citlivé atribúty neboli explicitne zbierané v rámci procesu tvorby datasetu, ale možno ich odvodiť pomocou ďalších metód.)

Názov premennej	**Popis**
<Názov premennej>	<Popis>
...	...

Názov premennej	**Popis**
<Názov	<Popis>

Tento projekt je podporený z Európskeho sociálneho fondu.

premennej>	
...	...

Ďalšie poznámky: <Pridajte sem>

ZDROJ(E)	PODROBNOSTI O METODIKE
Uveďte zdroje citlivých údajov.	*Opíšte metódy použité na zber citlivých atribútov v datasete.* *(Poznámka k použitiu: Duplikujte a vyplňte nasledujúce údaje pre každý citlivý atribút.)*
[Citlivý atribút]: Zdroje **[Citlivý atribút]:** Zdroje **[Citlivý atribút]:** Zdroje **Ďalšie poznámky:** <Pridajte sem>	**[Citlivý atribút]** **Metóda:** <Popíšte tu metódu zberu. V prípade potreby uveďte odkazy.> **Úloha zberu:** <Popíšte úlohu. V prípade potreby uveďte odkazy.> **Platformy, nástroje alebo knižnice:** [Platforma, nástroj alebo knižnica]: <Napíšte popis.> [Platforma, nástroj alebo knižnica]: <Napíšte popis.> [Platforma, nástroj alebo knižnica]: <Napíšte popis.> **Ďalšie poznámky:** <Pridajte sem>

POPISNÉ ŠTATISTIKY

Uveďte základné opisné štatistiky pre každý citlivý atribút a v nadpise uveďte kľúčové informácie.

Pomocou dodatočných poznámok zaznamenajte všetky ďalšie dôležité informácie alebo úvahy.

(Poznámka k použitiu: Duplikujte a vyplňte nasledujúce údaje pre každý citlivý atribút.)

Citlivý atribút

<Uveďte nadpis pre uvedenú tabuľku alebo vizualizáciu.>

Tento projekt je podporený z Európskeho sociálneho fondu.

[Citlivý atribút]	Štítok alebo trieda	Štítok alebo trieda	Štítok alebo trieda	Štítok alebo trieda
Počet	123456	123456	123456	123456
[Štatistika]	123456	123456	123456	123456
[Štatistika]	123456	123456	123456	123456
[Štatistika]	123456	123456	123456	123456

Ďalšie poznámky: <Pridajte sem>

RIZIKO(-Á) A MITIGÁCIA

Zhrňte systémové alebo reziduálne riziká, očakávania výkonnosti, kompromisy a výhrady z dôvodu citlivých atribútov v tomto datasete.

Pomocou dodatočných poznámok zaznamenajte všetky ďalšie dôležité informácie alebo úvahy.

Poznámka k použitiu: Duplikujte a vyplňte nasledujúce údaje pre každý citlivý atribút.)

[Citlivý atribút]

<Zhrňte tu. V prípade potreby uveďte odkazy a metriky.>

<Typ rizika>: [Popis + mitigácia]

<Typ rizika>: [Popis + mitigácia]

<Typ rizika>: [Popis + mitigácia]

Kompromisy, výhrady a iné úvahy: <Zhrňte tu. V prípade potreby uveďte vizualizácie, metriky alebo odkazy.>

Ďalšie poznámky: <Pridajte sem>

Tento projekt je podporený z Európskeho sociálneho fondu.

9. Rozšírené používanie

Použitie s inými údajmi

ÚROVEŇ BEZPEČNOSTI	ZNÁME BEZPEČNÉ DATASETY ALEBO TYPY ÚDAJOV	DOBRÁ PRAX
Vyberte jednu z nich:	*Uveďte zoznam známych datasetov alebo typov údajov a zodpovedajúcich transformácií, s ktorými je možné tento dataset bezpečne spojiť alebo agregovať.*	*Zhrňte osvedčené postupy používania tohto datasetu s inými datasetmi alebo typmi údajov.*
• Bezpečné používanie s inými údajmi • Podmienečne bezpečné použitie s inými údajmi • Nemal by sa používať s inými údajmi • Neznáme • Iné (uveďte)	**Dataset alebo typ údajov:** <Zhrňte tu. V prípade potreby uveďte vizualizácie, metriky alebo odkazy.> **Dataset alebo typ údajov:** <Zhrňte tu. V prípade potreby uveďte vizualizácie, metriky alebo odkazy.> **Dataset alebo typ údajov:** <Zhrňte tu. V prípade potreby uveďte vizualizácie, metriky alebo odkazy.>	<Zhrňte tu. V prípade potreby uveďte vizualizácie, metriky, názorné príklady alebo odkazy.> **Ďalšie poznámky:** <Pridajte sem>
	ZNÁME NEBEZPEČNÉ DATASETY ALEBO TYPY ÚDAJOV	OBMEDZENIE(-A) A ODPORÚČANIE(-A)
Tento riadok vyplňte, ak ste vybrali možnosť "Podmienečne bezpečné použitie s inými datasetmi" alebo	*Uveďte zoznam známych datasetov alebo typov údajov a zodpovedajúcich transformácií, ktoré **nie je možné prepojiť' alebo agregovať'** s týmto datasetom.*	*Zhrňte obmedzenia datasetu, ktoré predstavujú predvídateľné riziká pri spojení datasetu s inými datasetmi.*

Tento projekt je podporený z Európskeho sociálneho fondu.

"Nemal by sa používať s inými datasetmi":		
	Dataset alebo typ údajov: <Zhrňte tu. V prípade potreby uveďte vizualizácie, metriky alebo odkazy.> **Dataset alebo typ údajov:** <Zhrňte tu. V prípade potreby uveďte vizualizácie, metriky alebo odkazy.> **Dataset alebo typ údajov:** <Zhrňte tu. V prípade potreby uveďte vizualizácie, metriky alebo odkazy.>	<Zhrňte tu. V prípade potreby uveďte odkazy a metriky.> **<Typ obmedzenia>:** [Súbor údajov alebo typ údajov, popis a odporúčanie.] **<Typ obmedzenia>:** [Súbor údajov alebo typ údajov, popis a odporúčanie.] **Ďalšie poznámky:** <Pridajte sem>

Použitie v modeloch strojového učenia alebo umelej inteligencie

POUŽITIA DATASETU	ZÁSADNÁ(É) PREMENNÁ(É)	USMERNENIE(-A) NA POUŽÍVANIE
Vyberte všetky príslušné:	*Opíšte všetky zásadné premenné alebo vzťahy medzi jednotlivými premennými, ktoré sú explicitne uvedené.* *Uveďte odkazy na servery, kde si konzumenti môžu sami preskúmať údaje.*	*Zhrňte usmernenia alebo zásady používania, ktoré by mali konzumenti poznať.* *Zahrňte odkaz na modely, ktoré dataset už využívajú.*
• Trénovanie • Testovanie • Validácia • Vývojové alebo produkčné použitie • Ladenie modelu • Iné (uveďte)	**Ukážka datasetu v praxi:** [Odkaz na server alebo demo.] **<Názov zásadnej premennej>:** <Popíšte tu. V prípade potreby uveďte odkazy, príklady údajov, metriky, vizualizácie.> <Uveďte nadpis pre uvedenú tabuľku alebo vizualizáciu.> **Ďalšie poznámky:** <Pridajte sem>	**Pokyny na používanie:** <Zhrňte tu. V prípade potreby uveďte odkazy.> **Kroky schvaľovania:** <Zhrňte tu. V prípade potreby uveďte odkazy.> **Oponent:** <Uveďte meno oponenta použití, ktoré sa týkajú tohto datasetu.> **Odkaz na model:** <Prelinkovanie s kartou modelu podľa kapitoly 5.2>

Tento projekt je podporený z Európskeho sociálneho fondu.

10. Transformácie

Túto časť vyplňte, ak sa pri vytváraní súboru údajov použili nejaké transformácie.

POUŽITÁ(-É) TRANSFORMÁCIA(-E)	TRANSFORMOVANÁ PREMENNÁ(-É)	POUŽITÁ KNIŽNICA (KNIŽNICE) A METÓDA (METÓDY)
*Vyberte **všetky príslušné** transformácie, ktoré boli použité.*	*Uveďte premenné v datasete, ktoré boli transformované.* *(Poznámka k používaniu: Duplikujte a vyplňte nasledujúce údaje pre každý typ transformácie)*	*Uveďte opis metód použitých na transformáciu alebo spracovanie datasetu.* *(Poznámka k použitiu: Duplikujte a vyplňte nasledujúce údaje pre každý použitý typ transformácie.)*
• Detekcia anomálií • Čistenie nezhodujúcich sa hodnôt • Čistenie chýbajúcich hodnôt • Konverzia dátových typov • Agregácia údajov • Zníženie dimenzionality • Spájanie vstupných zdrojov • Redakcia alebo anonymizácia • Iné (uveďte prosím)	**<Typ transformácie>** **Názov premennej** / **Zdroj a cieľ** <Názov premennej> / <Zdroj: Cieľ> <Názov premennej> / <Zdroj: Cieľ> ... / ... **Ďalšie poznámky:** <Pridajte sem>	**<Typ transformácie>** **Metóda:** <Popíšte tu metódu transformácie. V prípade potreby uveďte odkazy.> **Platformy, nástroje alebo knižnice:** [Platforma, nástroj alebo knižnica]: <píšte tu popis> [Platforma, nástroj alebo knižnica]: <píšte tu popis> [Platforma, nástroj alebo knižnica]: <píšte tu popis> **Výsledky transformácie:** <Uveďte výsledky, výstupy a opatrenia prijaté v dôsledku transformácie. Ak je to možné, uveďte vizualizácie.> **Ďalšie poznámky:** <Pridajte sem>

Tento projekt je podporený z Európskeho sociálneho fondu.

Rozdelenie transformácií podľa typu

Vyplňte príslušné riadky.

ČISTENIE CHÝBAJÚCICH HODNÔT ALEBO INÝ TYP	POUŽITÉ METÓDY	POROVNÁVACIE ZHRNUTIE
V ktorých premenných údajov chýbali hodnoty? Koľko?	*Ako boli vyčistené chýbajúce hodnoty? Aké ďalšie možnosti sa zvažovali?*	*Prečo sa chýbajúce hodnoty čistili touto metódou (a nie inými)? Uveďte porovnávacie grafy zobrazujúce stav pred a po očistení chýbajúcich hodnôt.*
<Zhrňte tu. Uveďte odkazy, ak sú k dispozícii.> **Názov premennej:** Počet alebo opis **Názov premennej:** Počet alebo opis **Názov premennej:** Počet alebo opis	<Zhrňte tu. V prípade potreby uveďte odkazy.> **Platformy, nástroje alebo knižnice:** **[Platforma, nástroj alebo knižnica]:** <Napíšte popis.> **[Platforma, nástroj alebo knižnica]:** <Napíšte popis.> **[Platforma, nástroj alebo knižnica]:** <Napíšte popis.>	<Zhrňte tu. Uveďte odkazy, tabuľky, vizualizácie, ak sú k dispozícii> <Uveďte nadpis pre uvedenú tabuľku alebo vizualizáciu.> **Názov premennej** / **Rozdiel** <Názov premennej> / <Pred: Po> <Názov premennej> / <Pred: Po> ... / ... **Ďalšie poznámky:** <Pridajte sem>

Tento projekt je podporený z Európskeho sociálneho fondu.

REZIDUÁLNE A INÉ RIZIKO(-Á)	OPATRENIE(-A) DOHĽADU	ĎALŠIE ÚVAHY
Aké riziká vznikli v dôsledku tejto transformácie? Ktoré riziká boli mitigované?	*Aké opatrenia dohľadu vrátane dodatočného testovania, prešetrovania a schvaľovania boli prijaté v dôsledku tejto transformácie?*	*Aké ďalšie úvahy boli vykonané?*
<Zhrňte tu. V prípade potreby uveďte odkazy a metriky.> **<Typ rizika>:** [Popis + mitigácia] **<Typ rizika>:** [Popis + mitigácia] **<Typ rizika>:** [Popis + mitigácia]	<Zhrňte tu. Uveďte odkazy, ak sú k dispozícii.>	<Zhrňte tu. Uveďte odkazy, ak sú k dispozícii.>

Tento projekt je podporený z Európskeho sociálneho fondu.

11. Anotácie a označovanie

Túto časť vyplňte, ak sa pri vytváraní datasetu vykonali nejaké úlohy anotácie človekom alebo algoritmom.

TYP ANOTÁCIE	ANOTAČNÁ CHARAKTERISTIKA(-Y)	POPIS(Y) ANOTÁCIE
*Vyberte **všetky príslušné** typy anotácií alebo metódy použité na anotáciu datasetu:*	*Opíšte príslušné charakteristiky anotácií, ako je uvedené. V prípade metrík kvality zvážte zahrnutie presnosti, konzenzuálnej presnosti, IRR, XRR s príslušnou granularitou (napr. v rámci celého datasetu, podľa anotátora, podľa anotácie atď.).* *(Poznámka k používaniu: Duplikujte a vyplňte nasledujúce údaje pre každý typ anotácie.)*	*Uveďte opisy anotácií použitých na datasete. Ak je to možné, uveďte odkazy a označte použité platformy, nástroje alebo knižnice.* *(Poznámka k používaniu: Duplikujte a vyplňte nasledujúce údaje pre každý typ anotácie.)*
• Cieľ anotácie v údajoch • Strojovo generované anotácie • Anotácie ľuďmi (expert) • Anotácie ľuďmi (laická verejnosť) • Anotácie ľuďmi (zamestnanci) • Anotácie ľuďmi (dodávatelia) • Anotácie ľuďmi (crowdsourcing) • Anotácie ľuďmi (outsourcované/riadené) tímy • Neoznačené • Iné (uveďte prosím)	**(Typ anotácie)** <Uveďte nadpis pre uvedenú tabuľku alebo vizualizáciu.> • Počet jedinečných anotácií • Celkový počet anotácií • Priemerný počet anotácií na príklad • Počet anotátorov na príklad • [Metrika kvality na granularitu] • ... **Ďalšie poznámky:** <Pridajte sem>	**(Typ anotácie)** **Popis:** <Popis vytvorených anotácií (štítky, hodnotenia). Uveďte, ako bola vytvorená alebo kto je jej autorom.> **Odkaz:** <relevantný odkaz URL.> **Platformy, nástroje alebo knižnice:** [Platforma, nástroj alebo knižnica]: <Napíšte popis.> [Platforma, nástroj alebo knižnica]: <Napíšte popis.> **Ďalšie poznámky:** <Pridajte sem>

Tento projekt je podporený z Európskeho sociálneho fondu.

	POPISNÁ(-É) ŠTATISTIKY ANOTÁCIÍ	ÚLOHA(-Y) ANOTÁCIE
	Uveďte rozdelenie anotácií pre každú anotáciu alebo triedu anotácií pomocou nižšie uvedeného formátu. *(Poznámka k používaniu: Duplikujte a vyplňte nasledujúce údaje pre každý typ anotácie.)*	*Zhrňte každý typ úlohy spojený s anotáciami v datasete.* *(Poznámka k používaniu: Duplikujte a vyplňte nasledujúce údaje pre každý typ úlohy.)*
	(Typ anotácie) <Uveďte nadpis pre uvedenú tabuľku alebo vizualizáciu.> Anotácie (alebo trieda 12345 (20%) Anotácie (alebo trieda 12345 (20%) Anotácie (alebo trieda 12345 (20%) **Ďalšie poznámky:** <Pridajte sem>	**(Typ úlohy)** **Opis úlohy:** <Zhrňte tu. Uveďte odkazy, ak sú k dispozícii.> **Pokyny k úlohe:** <Zhrňte tu. Uveďte odkazy, ak sú k dispozícii.> **Použité metódy:** <Tu zhrňte. Uveďte odkazy, ak sú k dispozícii.> **Politika posudzovania medzi hodnotiteľmi:** <Zhrňte tu. Uveďte odkazy, ak sú k dispozícii.> **Časté otázky:** <Zhrňte tu. Uveďte odkazy, ak sú k dispozícii.> **Ďalšie poznámky:** <Pridajte sem>

Tento projekt je podporený z Európskeho sociálneho fondu.

Ľudskí anotátori

Túto časť vyplňte, ak boli použité anotácie ľuďmi.

POPIS(Y) ANOTÁTORA	ÚLOHA(-Y) ANOTÁTORA
Uveďte stručný opis každej skupiny anotátorov, ktorí vykonávajú úlohu anotácie. *Pomocou dodatočných poznámok zaznamenajte všetky ďalšie dôležité informácie alebo úvahy.* *(Poznámka k používaniu: Duplikujte a vyplňte nasledujúce údaje pre každý typ anotácie.)*	*Uveďte stručný opis každej úlohy anotácie.* *(Poznámka k používaniu: Duplikujte a vyplňte nasledujúce údaje pre každý typ úlohy.)*
(Typ anotácie) **Typ úlohy.** <Zhrňte tu. Uveďte odkazy, ak sú k dispozícii> **Počet jedinečných anotátorov:** <Zahrňte sem. Ak sú k dispozícii, uveďte odkazy.> **Odbornosť anotátorov:** <Zhrňte tu. Uveďte odkazy, ak sú k dispozícii.> **Opis anotátorov:** <Zahrňte sem. Uveďte odkazy, ak sú k dispozícii.> **Jazykové rozdelenie anotátorov:** <Zhrňte tu. Uveďte odkazy, ak sú k dispozícii.> **Geografické rozmiestnenie anotátorov:** <Zhrňte tu. Uveďte odkazy, ak sú k dispozícii.> **Zhrnutie pokynov pre anotáciu:** <Zhrňte tu. Ak sú k dispozícii, uveďte odkazy.> **Zhrnutie častých otázok:** <Zhrňte tu. Uveďte odkazy, ak sú k dispozícii.> **Anotačné platformy:** <Zhrňte tu. Uveďte odkazy, ak sú k dispozícii.> **Ďalšie poznámky:** <Pridajte sem>	**(Typ úlohy)** **Opis úlohy:** <Zhrňte tu. Uveďte odkazy, ak sú k dispozícii.> **Pokyny k úlohe:** <Zhrňte tu. Uveďte odkazy, ak sú k dispozícii.> **Použité metódy:** <Tu zhrňte. Uveďte odkazy, ak sú k dispozícii.> **Politika posudzovania medzi hodnotiteľmi:** <Zhrňte tu. Uveďte odkazy, ak sú k dispozícii.> **Časté otázky:** <Zhrňte tu. Uveďte odkazy, ak sú k dispozícii.> **Ďalšie poznámky:** <Pridajte sem>

Tento projekt je podporený z Európskeho sociálneho fondu.

12. Typy validácie

Túto časť vyplňte, ak boli údaje v datasete overené počas vytvárania datasetu alebo po ňom.

METÓDA(Y)	ROZPAD	POPIS(Y)
Vyberte všetky príslušné:	*Uveďte opis premenných a dátových bodov, ktoré boli validované.* *(Poznámka k používaniu: Duplikujte a vyplňte nasledujúce údaje pre každý typ validátora.)*	*Uveďte opis metód použitých na overenie datasetu.* *(Poznámka k používaniu: Duplikujte a vyplňte nasledujúce údaje pre každý typ validátora.)*
• Overovanie typu údajov • Overenie rozsahu a obmedzení • Overovanie kódu/krížových odkazov • Štruktúrovaná validácia • Overenie konzistentnosti • Nevalidované • Iné (uveďte)	**(Typ overenia)** **Počet overených dátových bodov:** 12345 Uveďte nadpis pre uvedenú tabuľku alebo vizualizáciu. **Overené premenné:** Premenná – počet overených inštancií Premenná – počet overených inštancií Premenná – počet overených inštancií	**(Typ overenia)** **Metóda:** <Popíšte tu metódu validácie. V prípade potreby uveďte odkazy.> **Platformy, nástroje alebo knižnice:** [Platforma, nástroj alebo knižnica]: <Napíšte popis.> [Platforma, nástroj alebo knižnica]: <Napíšte popis.> **Výsledky validácie:** <Uveďte výsledky, výstupy a opatrenia prijaté v dôsledku validácie. Ak je to možné, uveďte vizualizácie.> **Ďalšie poznámky:** <Pridajte sem>

Opis ľudských validátorov

Túto časť vyplňte, ak bol dataset validovaný pomocou ľudských validátorov

Tento projekt je podporený z Európskeho sociálneho fondu.

CHARAKTERISTIKA(Y)	POPIS(Y)
Uveďte charakteristiky skupiny(-í) validátorov. Použite ďalšie poznámky na zachytenie akýchkoľvek ďalších relevantných informácií alebo úvah.	*Uveďte stručný opis fondu(-ov) validátorov. Použite ďalšie poznámky na zachytenie akýchkoľvek ďalších relevantných informácií alebo úvah.* *(Poznámka k používaniu: Duplikujte a vyplňte nasledujúce údaje pre každý typ validátora.)*
(Typ overenia) • Počet unikátnych validátorov: 12345 • Počet príkladov na validátora: 123456 • Priemerné náklady/úloha/validátor: EUR • Poskytnutá odborná príprava: Áno/Nie • Požadované odborné znalosti: Áno/Nie	**(Typ overenia)** **Popis validátora:** <Zhrňte tu. Uveďte odkazy, ak sú k dispozícii.> **Poskytovaná odborná príprava:** <Zhrňte tu. Uveďte odkazy, ak sú k dispozícii.> **Kritériá výberu validátora:** <Zhrňte tu. Uveďte odkazy, ak sú k dispozícii.> **Poskytovaná odborná príprava:** <Zhrňte tu. Uveďte odkazy, ak sú k dispozícii.> **Ďalšie poznámky:** <Pridajte sem>

Tento projekt je podporený z Európskeho sociálneho fondu.

13. Metódy vzorkovania

Ak dataset využíva nejaké metódy vzorkovania, vyplňte nasledujúcu časť.

POUŽITÉ METÓDY	CHARAKTERISTIKA(Y)	KRITÉRIÁ ODBERU VZORIEK
*Vyberte **všetky príslušné** metódy použité pri vytváraní tohto datasetu:*	*Uveďte charakteristiky každej použitej metódy vzorkovania.* *(Poznámka k použitiu: Duplikujte a vyplňte nasledujúce údaje pre každú použitú metódu vzorkovania.)*	*Opíšte kritériá používané na výber vzoriek údajov z predchádzajúcich zdrojov.* *Pomocou dodatočných poznámok zaznamenajte všetky ďalšie dôležité informácie alebo úvahy.*
• Zhlukové vzorkovanie • Náhodný výber vzoriek • Viacstupňové vzorkovanie • Retrospektívne vzorkovanie • Stratifikované vzorkovanie • Systematické vzorkovanie • Vážené vzorkovanie • Neznáme • Nevzorkované • Iné (uveďte prosím)	**(Typ odberu vzorky)** <Uveďte nadpis pre uvedenú tabuľku alebo vizualizáciu.> • Pôvodný zdroj: [Napíšte sem] • Celkový počet odobratých vzoriek: 123m • Veľkosť vzorky: 123 • Uplatnená prahová hodnota: 123k jednotiek • Vzorkovacia frekvencia: 123 • Priemerná hodnota vzorky: 123 • Štandardná odchýlka: 123 • Rozdelenie vzoriek: 123 • Odchýlka pri výbere vzorky: 123 • ... **Ďalšie poznámky:** <Pridajte sem>	• **Metóda odberu vzoriek:** <Zhrňte tu. V prípade potreby uveďte odkazy.> • **Metóda odberu vzoriek:** <Zhrňte tu. V prípade potreby uveďte odkazy.> • **Metóda odberu vzoriek:** <Zhrňte tu. V prípade potreby uveďte odkazy.>

Tento projekt je podporený z Európskeho sociálneho fondu.

Tento projekt je podporený z Európskeho sociálneho fondu.

5.2 Príklad správneho popisu modelu strojového učenia

Verzia karty modelu: 0.0_YYYY

Licencia: Apache 2.0

Názov modelu: Prognóza 12.B

Model: [Odkaz na model]
Dokumentácia: [Odkaz na podrobnú dokumentáciu]

Autori vzorových kariet: [meno a kontakt]

Napíšte zhrnutie opisujúce váš model v niekoľkých vetách. Uveďte informácie o type modelu a úlohách, motiváciu modelu a problémy alebo prípady použitia, pre ktoré je vhodný. Aký je prínos používania tohto modelu?

Prehľad modelu

ARCHITEKTÚRA MODELU	VSTUP(Y)	VÝSTUP(Y)
Tu opíšte architektúru modelu. Uveďte aj schému.	*Uveďte opis (s potrebnými špecifikáciami) vstupných premenných poskytnutých modelu pre tvorbu výstupov.*	*Uveďte opis (s potrebnými špecifikáciami) výstupných premenných z modelu pre dané vstupy.*

Tento projekt je podporený z Európskeho sociálneho fondu.

Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti.	Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti alebo príklady. **Uveďte prelinkovanie s datasetom (kapitola 5.1) a dátovým katalógom.**	Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti alebo príklady. **Uveďte prelinkovanie s datasetom (kapitola 5.1) a dátovým katalógom.**
Používanie modelu		
PRÍPADY POUŽITIA	VÝHODY	ZNÁME NÁSTRAHY
Kde sa tento model používal alebo používa? Na aké prípady použitia? Uveďte odkazy, kde sa možno dozvedieť viac.	*Prečo by si používatelia mohli vybrať tento model v porovnaní s inými? Svoju odpoveď doložíte metrikami alebo výsledkami výkonnosti*	*Existujú nejaké známe nástrahy pri používaní tohto modelu, ktorým sa dá predísť?*
Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti alebo príklady.	Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti alebo príklady.	Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti alebo príklady.
Tvorcovia modelov		

Tento projekt je podporený z Európskeho sociálneho fondu.

SPÔSOB KONTAKTOVANIA AUTOROV	AUTOR(I) MODELU	CITÁCIE (AK SA JEDNÁ O VEDECKÚ ČINNOSŤ) A ODKAZY
Ako možno kontaktovať vlastníkov a autorov modelu v prípade otázok týkajúcich sa modelu?	*Napíšte mená všetkých autorov spojených s modelom. Uveďte afiliáciu a rok, pričom použite formát Meno, Priezvisko, Titul, Afiliácia, RRRR:*	*Ak je k dispozícii, uveďte odkaz na váš model, napríklad na platforme Github; inak uveďte, že nie je k dispozícii.*
Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie údaje o autoroch.	Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie údaje o autoroch.	Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti alebo formát bibtex/citácie.

Typ systému

POPIS SYSTÉMU	ZÁVISLOSTI OD ZDROJA ÚDAJOV	NADVÄZUJÚCE ZÁVISLOSTI
Je to samostatný model alebo je určený na použitie ako súčasť systému s inými modelmi? V prípade potreby uveďte odkazy.	*Ak si model vyžaduje špecifické vstupy, odkiaľ by mali pochádzať? Existujú nejaké špecifické kroky predbežného spracovania, ktoré by sa mali použiť? V prípade potreby uveďte odkazy.*	*Ak sa výstupy modelu môžu dostať do iného systému, kam by mali ísť? Existujú nejaké špecifické kroky následného spracovania, ktoré by sa mali použiť? V prípade potreby uveďte odkazy.*

Tento projekt je podporený z Európskeho sociálneho fondu.

Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti alebo príklady.	Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti alebo príklady. **Prepojte s popisom datasetu podľa kapitoly 5.1, v časti o zdroji a potrebnom spracovaní, ak je to relevantné.**	Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti alebo príklady.

Rámce implementácie

HARDVÉR A SOFTVÉR PRE TRÉNOVANIE	HARDVÉR A SOFTVÉR NA NASADENIE
Opíšte hardvér a softvér použitý na trénovanie modelu.	*Opíšte hardvér a softvér použitý na nasadenie modelu.*
• Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti alebo príklady.	• Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti alebo príklady.
POŽIADAVKY NA VÝPOČET PRE FINE TUNING*	POŽIADAVKY NA INFERENCIU*

Tento projekt je podporený z Európskeho sociálneho fondu.

Opíšte nasledujúce požiadavky na výpočty. V prípade potreby uveďte, že nie sú k dispozícii alebo nie sú relevantné. Nevymazávajte žiadne možnosti.	*Opíšte nasledujúce požiadavky na výpočty. V prípade potreby uveďte, že nie sú k dispozícii alebo nie sú relevantné. Nevymazávajte žiadne možnosti.*
Počet čipov: [uveďte podrobnosti] Čas trénovania (hodiny alebo dni): [uveďte podrobnosti] Celkový výpočet (operácie s „floating points“): [uveďte podrobnosti] Meraný výkon (TFLOPS/s): [uveďte podrobnosti] Spotreba energie (MWh): [uveďte podrobnosti]	Počet čipov: [uveďte podrobnosti] Čas trénovania (hodiny alebo dni): [uveďte podrobnosti] Celkový výpočet (operácie s „floating points“): [uveďte podrobnosti] Meraný výkon (TFLOPS/s): [uveďte podrobnosti] Spotreba energie (MWh): [uveďte podrobnosti]

**Modelované podľa Patterson, David, et al. "Carbon emissions and large neural network training." arXiv preprint arXiv:2104.10350 (2021).*

Tento projekt je podporený z Európskeho sociálneho fondu.

Charakteristika modelu

INICIALIZÁCIA MODELU		STAV MODELU		ŠTATISTIKY MODELU	
Opíšte, ako bol model inicializovaný. Uveďte informácie o tom, či sa model trénoval z náhodnej inicializácie, alebo sa dolaďoval z predtrénovaného modelu?		*Je model statický, alebo sa pretrénováva na základe dynamických údajov? Ak sa tento model opakovane trénuje, uveďte frekvenciu aktualizácií a dátum vydania najnovšej verzie.*		*Aká je veľkosť modelu? Zahrňte atribúty ako počet váh a vrstiev, ak sa jedná o model založený na neurónových sieťach.*	
Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti alebo príklady.		Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti alebo príklady.		Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti alebo príklady.	
Epochy na trénovanie	[uveďte podrobnosti]	**Názov datasetu**	[uveďte podrobnosti] [odkaz na jeho spôsob aktualizácie a verzie v dátovej karte podľa kapitoly 5.1]	**Veľkosť modelu**	[uveďte podrobnosti]

Tento projekt je podporený z Európskeho sociálneho fondu.

Základná rýchlosť učenia („learning rate“)	[uveďte podrobnosti]	**Verzia**	[uveďte podrobnosti]	**Váhy**	[uveďte podrobnosti]
Metóda	[uveďte podrobnosti]	**Dátum vydania**	DD.MM.RRRR: HH	**Vrstvy**	[uveďte podrobnosti]
Strata	[uveďte podrobnosti]	**Frekvencia aktualizácie**	[uveďte podrobnosti] Najbližšia aktualizácia modelu: DD.MM.RRRR: HH	**Latencia**	[uveďte podrobnosti]

„PRUNING“ - OREZANIE	KVANTIZÁCIA	„DIFFERENTIAL PRIVACY“
Je váš model orezaný? Ak áno, aká je úroveň riedkosti nasadeného modelu?	*Je váš model kvantizovaný? Ak áno, aká je bitová reprezentácia nasadeného modelu?*	*Ak existujú, opíšte techniky zavedené na ochranu súkromia.*
Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti alebo príklady.	Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti alebo príklady.	Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti alebo príklady.

Tento projekt je podporený z Európskeho sociálneho fondu.

Metódy	[uveďte podrobnosti]	**Metódy**	[uveďte podrobnosti]
Štruktúrovanie	[uveďte podrobnosti]	**Predkvantizovaná reprezentácia**	[uveďte podrobnosti]
Úroveň „sparsity“	[uveďte podrobnosti]	**Reprezentácia „End Bit“**	[uveďte podrobnosti]
Počet parametrov pri „sparsity“	[uveďte podrobnosti]	**Hardvér**	[uveďte podrobnosti]
Presnosť pri konečnej „sparsity“ po trénovaní	[uveďte podrobnosti]		
„Perplexity pri konečnej	[uveďte podrobnosti]		

Tento projekt je podporený z Európskeho sociálneho fondu.

„sparsity“ po trénovaní			

Výsledky hodnotenia modelu

Výsledky celkového hodnotenia modelu

Zdokumentujte svoje celkové hodnotenie výkonnosti modelu.

PROCES HODNOTENIA	VÝSLEDKY HODNOTENIA
Opíšte všetky zásadné faktory v procese hodnotenia celkovej výkonnosti modelu.	*Zhrňte a prepojte výsledky hodnotenia tejto analýzy.*

Tento projekt je podporený z Európskeho sociálneho fondu.

Metriky: Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti alebo príklady. **Hodnotiaci dataset:** Napíšte sem. Ak je to možné/potrebné, uveďte odkaz na ďalšie podrobnosti alebo príklady. **Pripojte odkaz na hodnotiaci dataset, ideálne zdokumentovaný v dátovej karte.** **Proces:** Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti alebo schémy.	Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti alebo príklady.

Výsledky hodnotenia podskupín

Zdokumentujte svoje čiastkové hodnotenie pre konkrétnu podskupinu vstupov (napr. spravodlivosť). Duplikujte túto časť (podskupina, proces hodnotenia a údaje, výsledky hodnotenia) pre každú hodnotenú podskupinu.

HODNOTENÁ PODSKUPINA	PROCES HODNOTENIA A ÚDAJE	VÝSLEDKY HODNOTENIA
Ktorá podskupina bola hodnotená?	*Opíšte všetky zásadné faktory v procese čiastkového hodnotenia výkonnosti modelu. Uveďte všetky predpoklady prijaté pri rozdeľovaní údajov v datasete.*	*Existujú nejaké známe nástrahy tohto modelu, ktorým sa dá predísť?*

Tento projekt je podporený z Európskeho sociálneho fondu.

Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti alebo príklady.	Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti alebo príklady.	Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti alebo príklady.
Výsledky hodnotenia spravodlivosti		
KRITÉRIÁ SPRAVODLIVOSTI	METRIKY SPRAVODLIVOSTI A ZÁKLADNÁ ÚROVEŇ	VÝSLEDKY SPRAVODLIVOSTI
Ako ste definovali spravodlivosť modelu? Opíšte cieľové kritériá spravodlivosti, ktoré ste chceli splniť alebo optimalizovať pred spustením.	*Opíšte metriky a základnú úroveň spravodlivosti, na základe ktorých prezentujete výsledky spravodlivosti, a spôsob ich výpočtu.*	*Opíšte výsledky analýzy spravodlivosti. Uveďte všetky konkrétne upozornenia alebo body, ktoré by ste chceli pre používateľov modelu zdôrazniť.*
Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti alebo príklady.	Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti alebo príklady.	Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti alebo príklady.

Tento projekt je podporený z Európskeho sociálneho fondu.

Používanie modelu a obmedzenia

CITLIVÉ PRÍPADY POUŽITIA	OBMEDZENIA	ETICKÉ ASPEKTY A RIZIKÁ
Existujú prípady použitia, pri ktorých by sa nasadenie tohto modelu považovalo za citlivé?	*Aké faktory môžu obmedziť výkonnosť modelu? Aké podmienky musia byť splnené, aby bolo možné model používať?*	*Aké etické faktory zohľadnili autori modelu? Boli identifikované nejaké riziká? Aké opatrenia na zmiernenie alebo nápravu boli prijaté? Ak je to možné, uveďte odkaz na ďalšie dokumenty.*
Použitie: Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti alebo príklady. **Nevyhnutné školenie:** Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti alebo príklady.	**Vstupné podmienky:** Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti alebo príklady. **Nástrahy pri používaní výstupov:** Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti alebo príklady.	**Výskum a vývoj:** Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti alebo príklady. **Nasadenie:** Napíšte sem. Ak je to potrebné, uveďte odkaz na ďalšie podrobnosti alebo príklady.

Tento projekt je podporený z Európskeho sociálneho fondu.

Pojmy a definície uvedené v tejto karte modelu

Tento priestor využite na uvedenie rozšírení a definícií akronymov, pojmov alebo odborných termínov používaných v celej karte modelu. Ak je to možné, použite štandardné definície (napr. MLCC Glosár), ktoré by mali byť zavedené aj v podnikovom slovníku. Ak je dostupný, uveďte zdroj definície. Ak používate výklad, úpravu alebo modifikáciu štandardnej definície na účely karty modelu alebo samotného modelu, uveďte aj svoj výklad. Všeobecne platné definície musia byť súčasťou podnikového slovníka.

[TERMÍN]	[TERMÍN]	[TERMÍN]
Definícia: <píšte tu> Prepojenie s podnikovým slovníkom: <písať a odkaz> Zdroj: <písať a odkaz> Výklad: <píšte tu>	Definícia: <píšte tu> Prepojenie s podnikovým slovníkom: <písať a odkaz> Zdroj: <písať a odkaz> Výklad: <píšte tu>	Definícia: <píšte tu> Prepojenie s podnikovým slovníkom: <písať a odkaz> Zdroj: <písať a odkaz> Výklad: <píšte tu>

Tento projekt je podporený z Európskeho sociálneho fondu.

Úvahy o modeli

Tento priestor použite na uvedenie akýchkoľvek ďalších informácií o modeli, ktoré neboli zachytené na karte modelu.

[Názov]	Tu napíšte poznámky
[Názov]	Tu napíšte poznámky
[Názov]	Tu napíšte poznámky

Tento projekt je podporený z Európskeho sociálneho fondu.

6 Zhrnutie najlepšej praxe v oblasti dokumentovania dátových aktív pre budúci rozvoj

Dátové katalógy môžu byť výkonnými platformami na podporu efektívneho manažmentu údajov. Bez správnej metodiky katalogizácie údajov a ďalších dátových aktív však môže ich potenciál a investície do nich vyjsť na zmar. Vzhľadom na tento fakt ponúka táto kapitola prehľad 10 najlepších postupov pre katalogizáciu údajov.

1. Zdokumentujte skutočne všetky dátové aktíva

Údaje sú všade a nachádzajú sa aj v textových súboroch, tabuľkách, analytických aplikáciách a v ďalších analytických produktoch. Hoci údaje môžu byť roztrúsené, nedá sa ani začať riešiť problém s údajmi, kým sa všetko nezinventarizuje. Každý člen tímu by mal byť vyškolený, aby premýšľal o všetkých možných miestach, kde sa môžu nachádzať údaje. Následne treba zabezpečiť, aby bol každý kúsok týchto diskrétnych údajov katalogizovaný udržateľným spôsobom.

2. Spravujte dátové toky

Nástroje na určovanie dátovej „lineage“ a pôvodu údajov sú dobré, ale väčšina z nich mapuje tok údajov v rámci známej domény alebo množiny domén. Dobrý dátový katalóg, ktorý je podporený zisťovaním tokov údajov (viac o tokoch údajov sa nachádza aj v dokumente „*1.1.7 Analýza tokov údajov*“), často identifikuje toky medzi rôznorodými datasetmi. Takéto toky pomôžu odhaliť pohyb údajov v rámci verejnej správy, ktorý nemusí byť dobre známy. Tieto toky sa potom dajú skontrolovať z hľadiska platnosti. Preto je správa dátových tokov dobrým postupom na vytvorenie efektívneho dátového katalógu.

3. Využívajte stále viac dynamické údaje v reálnom čase

Ako bolo spomenuté v kapitole 3.1.1 používanie dynamických údajov v reálnom čase jedným zo základných kameňov úspechu analytického reportovania. Interakcia s údajmi v reálnom čase prostredníctvom dynamických vizualizácií zabezpečí, že sa dá reagovať na akýkoľvek potenciálny problém v dátových tokoch tak, ako sa deje, a nie až spätne, kedy už mohol spôsobiť neželané dôsledky. Navyše práca s údajmi v reálnom čase, ktoré sú v súlade s nastavenými cieľmi, zlepší rozhodovanie v danom okamihu.

4. Prioritizujte citlivé údaje

Jedným z hlavných účelov efektívneho dátového katalógu je pomôcť identifikovať umiestnenie citlivých údajov. V scenároch, v ktorých sa rovnaké citlivé údaje nachádzajú na viacerých miestach, môže pomôcť identifikovať nadbytočné údaje. Správa citlivých a redundantných údajov tak umožňuje minimalizovať priestor na narušenie ochrany súkromia a vytvoriť spoľahlivú ochranu údajov pred akýmkoľvek vonkajším útokom.

5. Zvážte neštruktúrované údaje

Neštruktúrované údaje (dokumenty, webové stránky, e-maily, obsah sociálnych sietí, mobilné údaje, obrázky, audio a videá) sú údaje, ktoré nezodpovedajú dátovému modelu a nemajú ľahko identifikovateľnú štruktúru. Nie sú vhodné pre bežné relačné databázy.

Tento projekt je podporený z Európskeho sociálneho fondu.

Dátový katalóg môže pomôcť zviditeľniť implicitné štruktúry údajov. To možno dosiahnuť prepracovaním celkovej štruktúry údajov na základe požiadavky tímu alebo organizácie. Preto zohľadnenie "neštruktúrovaných" údajov môže byť pre každý dátový katalóg veľmi dôležité.

6. Priraďte ľahko objaviteľné („discoverable“) názvy a popisy

Dobrý názov a výstižný opis umožnia, aby boli údaje ľahšie objaviteľné (súvisí s „data discovery“ – ako jednoducho nájsť to, čo hľadám, v záplave údajov) pre príslušných členov tímu. Opis môže uvádzať alternatívne názvy pre ten istý objekt objekt evidencie alebo dátový prvok a pomáhať pri vytváraní komplexnej ontológie údajov.

7. Zaobchádzajte s tabuľkami dátového jazera odlišne

V relačných databázach môžu byť údaje rozložené vo viacerých tabuľkách. Dátové jazerá však majú tendenciu zhlukovať množstvo údajov do jednotlivých súborov. V oblasti BI sa v jednom dátovom súbore môžu ukladať miery a dimenzie spoločne, a nie oddelene. To platí aj pre systémy, ktoré reprezentujú údaje ako tabuľky v databáze. To môže spôsobiť, že údaje budú horšie vyhľadateľné, ale dátové katalógy musia tento problém priamočiaro vyriešiť.

8. Poskytovanie transparentných hodnotení

„Crowd-sourcované“ hodnotenia, pochvaly a negatívne hodnotenia v dátovom katalógu môžu používateľom pomôcť rýchlejšie získať relevantné a spoľahlivé informácie. To si však vyžaduje prísne štandardy a metodické postupy. Údaje by nemali dostať päťhviezdičkové hodnotenie, ak nespĺňajú veľmi prísne kritériá kvality v súlade so štandardom. Rovnako by dobré údaje nemali byť hodnotené zle. Používatelia musia mať v hodnotenia dôveru, inak im nebudú veriť. Preto by dátová kancelária mala zabezpečiť, aby boli normy jednotné a presné.

9. Vytvorte dátové jazero, nie „bažinu“

Katalogizácia všetkého, čo sa nachádza v dátovom jazere (ak sa tento prístup k ukladaniu dát zvolí pre Konsolidovanú analytickú vrstvu, viac v dokumente *„4.2.1 Koncept pre zavádzanie analytického spracovania údajov do praxe“*), umožní usporiadať ho a urobiť ho použiteľným. Keď je jazero skatalogizované, dajú sa v ňom vytvoriť zóny a urobiť z neho miesto, kde môžu používatelia získať údaje, a nie len miesto, kde ich môžu odložiť.

10. Zaveďte pravidlá na validáciu údajov

Prísne pravidlá validácie údajov môžu pomôcť overiť, či údaje zodpovedajú definíciám v dátovom katalógu. Takýto proces zabezpečuje kvalitu údajov a slúži ako kontrola kvalitatívneho hodnotenia hviezdičkami (ktoré môže ale aj nemusí byť prepojené s hodnotením pre prelinkované údaje[39]). Využívanie pravidiel validácie v dátovom katalógu vzbudzuje dôveru medzi používateľmi údajov.

11. Využívajte stále viac techniky strojového učenia

[39] Zdroj: https://5stardata.info/en/, Dátum referencie: 24.07.2023

Tento projekt je podporený z Európskeho sociálneho fondu.

Manuálne katalogizovanie je dnes vzhľadom na zvýšený objem údajov nemožné. Katalogizácia sa jednoducho nikdy nedokončí alebo nebude držať krok s príchodom nových údajov alebo rôznymi aktualizáciami v dátovej vrstve. Strojové učenie („Machine Learning (ML)“) je však sľubným nástrojom s ohľadom na udržateľnosť katalogizovania veľkého objemu rôznorodých údajov. Modely ML dokážu identifikovať typy údajov a vzťahy. To pomáha budovať katalóg naprieč viacerými datasetmi a systémami. Takisto šíri dátové značky naprieč väčším počtom objektov evidencie a datasetov rýchlejšie ako manuálny katalóg.

Tento projekt je podporený z Európskeho sociálneho fondu.

Error! Reference

Contact us

Name Surname

Sector name

T +44 20 0000 0000

E name.surname@kpmg.com

Name Surname

Sector name

T +44 20 0000 0000

E name.surname@kpmg.com

Name Surname

Sector name

T +44 20 0000 0000

E name.surname@kpmg.com

Some or all of the services described herein may not be permissible for KPM audit clients and their affiliates or related entities.

www.kpmg.com

Note

This Standard Independence disclaimer text in **blue** should be included on all Tax and Advisory service materials, which discuss KPMG's tax and advisory services, for both internal and external distribution.

The disclaimer must be included in a visible location and in same font size and color as content within the document. It shall not be included within the footer or as part of the KPMG copyright and disclaimer statements on the last page of materials.

The information contained herein is of a general nature and is not intended to address the circumstances of any particular individual or entity. Although we endeavour to provide accurate and timely information, there can be no guarantee that such information is accurate as of the date it is received or that it will continue to be accurate in the future. No one should

Tento projekt je podporený z Európskeho sociálneho fondu.

act on such information without appropriate professional advice after a thorough examination of the particular situation.

The KPMG name and logo are trademarks used under license by the independent member firms of the KPMG global organization.