Operačný program
Efektívna
verejná správa

Európska únia
Európsky sociálny fond

Praktický návod k meraniu dátovej kvality

Projekt

Zlepšenie využívania údajov vo verejnej správe

KPMG

ÚRAD PODPREDSEDU VLÁDY SR
PRE INVESTÍCIE
A INFORMATIZÁCIU

Tento projekt je podporený z Európskeho sociálneho fondu.

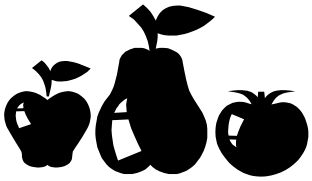
História dokumentu

Verzia	Autori	Dátum	Revízia
1.0	Dávid Igaz Tomáš Horváth Martin Florián	11.10.2019	Finálna verzia dokumentu

Zoznam skratiek

RDBMS	Systém pre riadenie relačnej databáze (Relational Database Management System)
GPS	Global Positioning system – systém poskytujúci údaje o polohe
BP	Biznis pravidlo
KPI	Ukazovateľ výkonnosti alebo kľúčová metrika
SQL	Structured Query Language – skratka pre štandardizovaný štruktúrovaný dopytovací jazyk
CSV	Comma Seperated values – jednoduchý súborový formát určený pre výmenu tabuľkových dát
XML	Extensible Markup Language – rozšíriteľný značkovací jazyk, slúži na výmenu dát medzi aplikáciami
DB	Database – databáza

Obsah



1 Čo potrebujeme (Ingrediencie)

1.1 Dokumenty definujúce vybraný objekt merania

Medzi dokumenty definujúce vybraný objekt merania pre vybraný dataset patria:

1. Karta merania dátovej kvality
2. Aktuálne dátové štandardy
3. Aktuálny dátový model
4. Aktuálne metadáta
5. Aktuálne biznis pravidlá

Všetky tieto dokumenty sú potrebné pri meraní dátovej kvality. Ich absencia výrazne znižuje kvalitu samotného merania.

Bez karty merania dátovej kvality a aktuálnych biznis pravidiel nie je možné merať dátovú kvalitu.

1.1.1 Karta merania dátovej kvality

Vytvorená karta popisuje informácie o vybranom datasete s doplňujúcimi údajmi pre samotné meranie. K týmto informáciám je potrebné získať všetky ďalšie relevantné dokumenty pre potreby merania dátovej kvality (sú uvedené nižšie). Získaním tohto zjednodušeného popisu je zároveň potvrdenie určeného datasetu, ktorého meranie má pre organizáciu zmysel. Informácie v karte merania sa postupne v čase dopĺňajú. Nižšie na (Obrázok 1) je zobrazený pohľad na príklad karty merania dátovej kvality.

Karta merania dátovej kvality		Popis
Identifikačné číslo karty merania		Zatiaľ neurčená konvencia
Názov merania dátovej kvality		Text
Definícia účelu merania		Text
Názov datasetu		Text
Zdroj datasetu		Informačný systém, Databáza, Server a pod.
Poskytovateľ datasetu		Meno, Priezvisko, Kontakt email, Kontakt telefón, OVM, Adresa OVM
Stručná definícia datasetu		Počet tabuliek, Názvy tabuliek, Pre každú tabuľku uviesť aj plný počet atribútov a plný počet záznamov
Výber vzorky záznamov v datasete		Nie (meranie v plnom rozsahu záznamov) / Áno (meranie pre tabuľky XYZ a k nim uvedený redukovaný počet záznamov)
Originálny formát datasetu		Formát súboru / súborov (ak ich je viac rôznorodých, je potrebné špecifikovať jednotlivo)
Dátum a čas exportu datasetu		Formát dátumu: YYYY-MM-DD hh:mm
Dátum a čas aktuálnosti datasetu		Formát dátumu: YYYY-MM-DD hh:mm
Dátum prvého dňa merania		Formát dátumu: YYYY-MM-DD
Dátum posledného dňa merania		Formát dátumu: YYYY-MM-DD
OVM zodpovedný za dataset		Názov, Adresa OVM
Zodpovedný za meranie		Meno, Priezvisko, Kontakt email, Kontakt telefón
Osoby vykonávajúce meranie		Meno, Priezvisko, Kontakt email, Kontakt telefón
Aktuálne dátové štandardy pre dataset		Odkaz na dokument
Aktuálny dátový model pre dataset		Odkaz na dokument
Aktuálne metadáta pre dataset		Odkaz na dokument
Aktuálne biznis pravidlá pre dataset		Odkaz na dokument

Obrázok 1: Karta merania dátovej kvality

1.1.2 Aktuálne dátové štandardy

Získanie aktuálnych dátových štandardov pomáha pochopiť jazyk, v ktorom je vybraný dataset písaný. Sú to pravidlá a smernice, ktoré určujú ako pomenovať dáta, ako ich definovať, ako vytvoriť platné hodnoty a ako špecifikovať biznis pravidlá. Príklady štandardov sú nasledovné:

— konvencia pomenovania tabuliek a polí,

— dátová definícia a konvencia pre písanie biznis pravidiel (Dátový slovník),

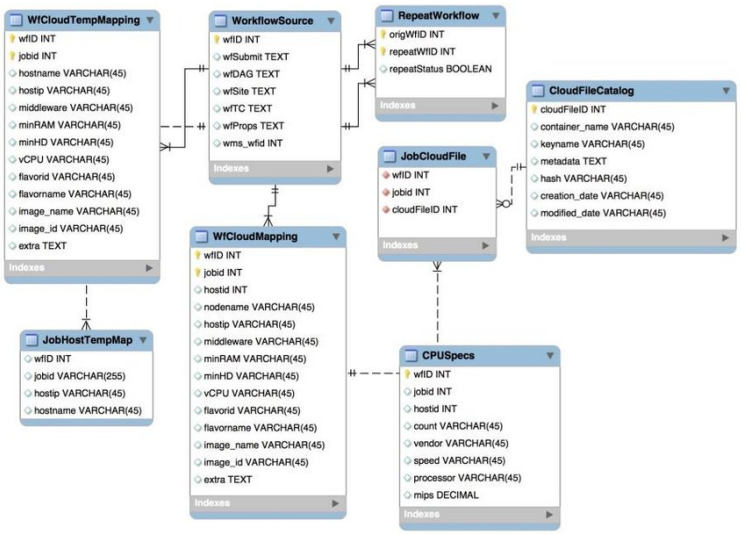
— zostavovanie, dokumentovanie a aktualizovanie zoznamov platných hodnôt,

— výber metódy zápisu a modelovania pre modelovanie dát,

— a pod.

1.1.3 Aktuálny dátový model

Získanie aktuálneho dátového modelu je dôležité z pohľadu pochopenia vybraného datasetu. Je to spôsob vizuálneho znázornenia štruktúry dát (ako sú dáta organizované) v rezorte. Je to zároveň tiež špecifikácia, akou majú byť dáta reprezentované v databáze. V riadení dátovej kvality je dôležité rozumieť databáze obsahujúcej dáta a programy, ktoré ich zachytávajú, ukladajú, manipulujú, transformujú, mažú a zdieľajú. Termíny entita, entitná trieda a atribúty sú hlavnými konceptami. Entitou sa chápe osoba, miesto, udalosť, vec s predmetom záujmu k biznisu. Entitnou triedou sa chápe typ entity resp. súbor tých vecí, ktorých inštancie sú jednoznačne identifikovateľné. Atribútom sa rozumie definícia charakteristík, kvality a vlastností entitných tried (vid nižšie Obrázok 2).



Obrázok 2: Dátový model

1.1.4 Aktuálne metadáta

Získanie aktuálnych metadát, ktoré označujú, popisujú alebo charakterizujú dáta k vybranému datasetu, sú dôležitým prvkov pre pochopenie objektu merania. Zároveň uľahčujú získavanie, interpretáciu alebo použitie informácií.

Nižšie je uvedený príklad metadát (viď Tabuľka 2), ktoré sú zároveň naparované na príklad relevantnej databázovej tabuľky (viď Tabuľka 1).

Databázová tabuľka - príklad:

ZamestnanecID	Meno	Priezvisko	Pohlavie	Dátum začatia	Dátum ukončenia
44	Dávid	Kulich	M	28.02.2018	28.05.2019
45	Tomáš	Vajda	M	15.05.2010	20.09.2019
46	Frederika	Pekná	Ž	01.01.2017	NULL

Tabuľka 1: Databázová tabuľka

Relevantné metadata - príklad:

Stĺpec	Dátový typ	Popis
ZamestanecID	Int	Primárny kľúč tabuľky
Meno	Nvarchar(50)	Meno zamestnanca
Priezvisko	Nvarchar(50)	Priezvisko zamestnanca
Pohlavie	Char(1)	M=muž, Ž=žena, Null=neznáme
Dátum začatia	Date	Dátum nástupu zamestnanca do organizácie
Dátum ukončenia	Date	Dátum ukončenia pracovného pomeru zamestnanca v organizácií

Tabuľka 2: Metadáta

1.1.5 Aktuálne biznis pravidlá

Úlohou je získať alebo vytvoriť aktuálny zoznam biznis pravidiel na úrovni atribútov (stĺpcov) tabuliek príslušného datasetu/datasetov. Pravidlá sa môžu zameriavať na systémové a organizačné biznis pravidlá.

Príklad: Adresné body (GPS súradnice zemepisnej šírky) sa musia nachádzať v rámci intervalu medzi najzápadnejším/najvýchodnejším bodom Slovenskej republiky.

Treba zdôrazniť, že biznis pravidlá sa môžu meniť, aktualizovať a ich organizácia a usporiadanie môže byť veľmi hierarchické a komplexné. Často sa biznis pravidlá vetvia

(jedno pravidlo vychádza z iného (napr. z hlavného pravidla) a tento vzťah musí byť dokumentovaný a riadený). Vo verejnej správe sú biznis pravidlá silne ovplyvňované aj meniacou sa legislatívou.

Ako zdokumentovať biznis pravidlo:

V prvej fáze je podstatné a dôležité zistiť, či sú biznis pravidlá štandardizované. V prípade, že sú, tak je potrebné ich spísať v odporúčanej štruktúre. Nižšie je šablóna na dokumentáciu biznis pravidiel. Platí všeobecné pravidlo:

1 Biznis pravidlo je pre 1 cieľový atribút

Dataset	ID BP	Biznis pravidlo	Zdroj BP	Poznámka	Cieľ BP	ID KPI

Tabuľka 3: Šablóna dokumentácie biznis pravidiel

Šablóna pozostáva zo štruktúrovaných informácií. Je povinné vyplniť všetky polia, pričom pri poli poznámka je vyplnenie voliteľné.

— **Dataset**: názov datasetu z karty merania;

— **ID BP**: identifikačný kód biznis pravidla;

— **Biznis pravidlo**: znenie biznis pravidla;

— **Zdroj BP**: názov zdrojovej tabuľky a zároveň aj názov atribútu v tabuľke pre biznis pravidlo, konvencia: tabuľka (atribút);

— **Poznámka**: poznámka k biznis pravidlu;

— **Cieľ BP**: názov cieľovej tabuľky a zároveň názov cieľového atribútu v tabuľke pre biznis pravidlo, konvencia: tabuľka (atribút);

— **ID KPI**: označenie identifikačného kódu KPI z číselníka kódov (viď Tabuľka 4). Tento atribút hovorí o priradení biznis pravidla k správnemu ID KPI podľa metodiky. O výbere ID KPI je písane nižšie v časti „Ako štandardizovať biznis pravidlo“.

Biznis pravidlá je potrebné dokumentovať a rozširovať podľa potreby.

Ako štandardizovať biznis pravidlo:

V prípade, že biznis pravidlo nie je štandardizované, tak spôsob ako ho štandardizovať je odpoveďou na KPI otázku (viď Tabuľka 4 stĺpec KPI otázka). Medzi KPI otázku napr. patrí:

- povinný / nepovinný atribút;
- formát dát v atribúte (dátum EU, dátum US, konvencia hodnoty (alfanumerické kombinácie a skladanie hodnoty) a pod);
- rozsah dát v atribúte (od do pri číslach) / (počet znakov pri texte);
- očakávaná hodnota (číselník) v atribúte alebo kombináciu atribútov;
- jedinečnosť hodnoty v atribúte (môže byť konkrétna hodnota len raz alebo viackrát?)
- atď.

Platí všeobecné pravidlo:

1 Biznis pravidlo odpovedá vždy len na 1 KPI

Potom ako je biznis pravidlo štandardizované, tak je možné priradiť ID KPI a následne vložiť do šablóny dokumentácie biznis pravidla.

Nasleduje (Tabuľka 4), ktorá je nástrojom na vytvorenie biznis pravidiel a ich priradenia ku KPI. Obsahuje zoznam počítaných KPI s ich popisom, otázkou, podmienkou a konkrétnym identifikátormi. Tieto KPI sú naparované na ukazovatele a parametre dátovej kvality, pričom pri každom naparovaní biznis pravidla na KPI existuje typ merania určujúci, v ktorej časti sa meranie daného KPI vykoná, či sa jedná o prvé odborné posúdenie, finálne odborné posúdenie, reporting profilingu dát alebo pokročilejšiu analýzu. Tabuľka obsahuje aj možnosť či daný výpočet KPI hodnoty je možné vykonať v Talende.

Pochopenie tabuľky *Zoznam KPI a ich použitie pri meraní* je kľúčové pre meranie dátovej kvality.

Parameter	Ukazovateľ	Typ merania	Možnosť počítať v Talende	KPI ID	KPI	KPI popis	KPI otázka (platí pre atribút alebo skupinu atribútov, odpoveďou je biznis pravidlo)	KPI podmienka merania
Presnosť	Syntaktická presnosť atribútu	Pokročilejšia analýza	Áno	1.2a	Percentuálny podiel záznamov v atribúte tabuľky, ktorý obsahuje hodnoty v súlade s definovaným pravidlom povoľujúcom určené hodnoty.	Za atribút, ktorý má definované biznis pravidlo definujúce očakávané hodnoty napríklad číselník (zoznam hodnôt), kombinácia hodnôt z číselníka (zoznam hodnôt), hodnoty podľa pravidiel gramatiky, veľké malé písmená, intervalové hodnoty napr. od do alebo komplexnejšie intervaly a pod. = definícia povolených hodnôt v biznis pravidle.	Aké sú presne očakávané hodnoty atribútu? Napr. číselník (zoznam hodnôt), kombinácia hodnôt z číselníka (zoznam hodnôt), hodnoty podľa pravidiel gramatiky, veľké malé písmená, intervalové hodnoty napr. od do alebo komplexnejšie intervaly a pod. **POZOR:** pre atribút nesmie existovať zdroj pravdy.	Ak existujú definované očakávané hodnoty, tak sa môže merať KPI 1.2a
	Sémantická presnosť atribútu	Pokročilejšia analýza	Áno	1.3a	Percentuálny podiel záznamov v atribúte tabuľky, ktorý obsahuje hodnotu v súlade s definovaným biznis pravidlom odkazujúcim na zdroj pravdy.	Za atribút, ktorý má definované biznis pravidlo s určeným zdrojom pravdy. Takýto atribút preberá pravidlá syntaktickej presnosti z biznis pravidiel v zdroji pravdy a teda merať syntaktickú presnosť atribútu a syntaktickú presnosť hodnoty v určenom atribúte nemá zmysel. Syntaktická presnosť hodnoty a syntaktická presnosť atribútu má byť potom meraná v zdroji pravdy.	Aký je zdroj pravdy pre atribút a kde sa nachádza zdroju pravdy? Zdrojom pravdy sa myslí iný dátový zdroj, ktorý sa najviac približuje pravdivým hodnotám - realite. **POZOR:** Ak je odpoveď na existenciu zdroja pravdy kladná, meranie syntaktickej presnosti atribútu a syntaktickej presnosti hodnoty nemá prebiehať v určenom atribúte ale v zdroji pravdy.	Ak existuje definovaný zdroj pravdy, tak sa môže merať KPI 1.3a
Konzistent-nosť	Sledovanie konzistentnosti	Prvé odborné posúdenie	Nie	2.1a	Percentuálny podiel atribútov, ktoré majú definované biznis pravidlá.	Za skupinu atribútov v celom meranom datasete, ktoré majú mať definované biznis pravidlo. **POZOR:** atribút bez biznis pravidla nemôže mať meranú dátovú kvalitu.	N/A	N/A

Parameter	Ukazovateľ	Typ merania	Možnosť počítať v Talende	KPI ID	KPI	KPI popis	KPI otázka (platí pre atribút alebo skupinu atribútov, odpoveďou je biznis pravidlo)	KPI podmienka merania
	Dodržiavanie biznis pravidla	Finálne odborné posúdenie	Nie	2.2a	Percentuálny podiel atribútov, ktorých hodnoty plne dodržujú definované biznis pravidlá.	Za skupinu atribútov, ktoré majú mať definované biznis pravidlo. **POZOR:** atribút bez biznis pravidla nemôže mať meranú dátovú kvalitu.	N/A	N/A
		Pokročilejšia analýza	Áno	2.2b	Percentuálny podiel atribútov, ktorých hodnoty plne dodržujú vzťahy medzi atribútmi.	Za skupinu atribútov, ktoré majú mať definovanú procesnú stránku biznis pravidlá. Procesné a chronologické väzby medzi atribútmi tzn. naprogramovaná biznis logika informačných systémov, ktoré napĺňajú a menia dáta jednotlivých atribútov.	Aká je procesná stránka atribútu resp. aké sú vzťahy medzi atribútom a relevantnými atribútmi?	Ak existujú definované väzby medzi atribútmi, tak sa môže merať KPI 2.2b
Správnosť	Dodržiavanie formátu atribútu	Reporting Profilingu dát	Áno	3.1a	Percentuálny podiel záznamov v atribúte tabuľky, ktorý obsahuje hodnotu v požadovanom formáte.	Za atribút, ktorý má definované biznis pravidlo definujúce formát hodnoty.	Aký je formát hodnoty v atribúte?	Ak existuje definovaný formát, tak sa môže merať KPI 3.1a
Kompletnosť	Vyplnenosť povinného údaja	Reporting Profilingu dát	Áno	4.2a	Percentuálny podiel záznamov v atribúte tabuľky, ktorý obsahuje akúkoľvek hodnotu okrem 'null' a prázdnej hodnoty.	Za atribút, ktorý má definované biznis pravidlo definujúce povinne vyplnenú hodnotu. **POZN.:** Výpočet je silne závislý na definovaní očakávaných hodnôt vstupu. Pri nejasnom zadefinovaní významu 'null' a prázdnej hodnoty dôjde k skresleným výsledkom merania dátovej kvality.	Je atribút povinný? **POZOR:** metadáta nemusia obsahovať informácie o všetkých povinných atribútov z pohľadu biznisu. Povinnosť je niekedy ošetrená v samotných vstupoch do databázy, napríklad cez webové formuláre.	Ak existuje definovaná povinnosť vyplnenia, tak sa môže merať KPI 4.2a alebo 4.3a
	Vyplnenosť nepovinného údaja	Reporting Profilingu dát	Áno	4.3a	Percentuálny podiel záznamov v atribúte tabuľky, ktorý obsahuje akúkoľvek hodnotu vrátane prázdnej hodnoty a okrem 'null' hodnoty.	Za atribút, ktorý má definované biznis pravidlo definujúce nepovinne vyplnenú hodnotu. **POZN.**: Výpočet je silne závislý na definovaní očakávaných hodnôt vstupu. Pri nejasnom zadefinovaní významu 'null' a prázdnej hodnoty		

Parameter	Ukazovateľ	Typ merania	Možnosť počítať v Talende	KPI ID	KPI	KPI popis	KPI otázka (platí pre atribút alebo skupinu atribútov, odpoveďou je biznis pravidlo)	KPI podmienka merania
					POZOR: Toto KPI je ako doplnok k meraniu dátovej kvality. Nesmie sa na neho viazať prahová hodnota. Úroveň dátovej kvality je len informatívna hodnota keďže hodnoty atribútu z biznis pohľadu nie sú povinné.	dôjde k skresleným výsledkom merania dátovej kvality.		
Unikátnosť	Unikátnosť hodnoty	Reporting Profilingu dát	Áno	5.2a	Percentuálny podiel záznamov v atribúte tabuľky, ktorý obsahuje rovnakú hodnotu v rámci atribútu len jeden krát. **Pozn.:** Percentuálny podiel distinktných záznamov v atribúte tabuľky sa vyráta ako súčet unikátnych a multiplicitných záznamov.	Za atribút, ktorý má definované biznis pravidlo definujúce jedinečnú hodnotu vo všetkých záznamoch v atribúte.	Majú byť všetky hodnoty atribútu unikátne v rámci atribútu? **POZOR:** metadáta nemusia obsahovať informácie o všetkých pravidlách unikátnosti z pohľadu biznisu. Unikátnosť je niekedy ošetrená v samotných vstupoch do databázy, napríklad cez programové skripty.	Ak existuje definovaná unikátnosť hodnoty, tak sa môže merať KPI 5.2a a 5.2b
				5.2b	Percentuálny podiel záznamov v atribúte tabuľky, ktorý obsahuje rovnakú hodnotu v rámci atribútu viac ako jeden krát. (Jedná sa o multiplicitu hodnoty = duplicita, triplicita a pod.) **Pozn.**: Percentuálny podiel distinktných záznamov v atribúte tabuľky sa vyráta ako súčet unikátnych a multiplicitných záznamov.			
	Unikátnosť záznamov	Pokročilejšia analýza	Áno	5.3a	Percentuálny podiel záznamov za skupinu atribútov, ktorý obsahuje rovnakú kombináciu hodnôt v		Majú byť všetky kombinácie hodnôt skupiny atribútov unikátne v rámci skupiny atribútov?	Ak existuje informácia o unikátnosti

Parameter	Ukazovateľ	Typ merania	Možnosť počítať v Talende	KPI ID	KPI	KPI popis	KPI otázka (platí pre atribút alebo skupinu atribútov, odpoveďou je biznis pravidlo)	KPI podmienka merania
					rámci skupiny atribútov len jeden krát. **Pozn.:** Percentuálny podiel distinktných záznamov za skupinu atribútov sa vyráta ako súčet unikátnych a multiplicitných záznamov.		**POZOR:** metadáta nemusia obsahovať informácie o všetkých pravidlách unikátnosti z pohľadu biznisu. Unikátnosť je niekedy ošetrená v samotných vstupoch do databázy, napríklad cez programové skripty.	záznamu za skupinu atribútov, tak sa môže merať KPI 5.3a a 5.3b
				5.3b	Percentuálny podiel záznamov za skupinu atribútov, ktorý obsahuje rovnakú kombináciu hodnôt v rámci skupiny atribútov, viac ako jeden krát. **(**Jedná sa o multiplicitu záznamu = duplicita, triplicita a pod.) **Pozn.:** Percentuálny podiel distinktných záznamov za skupinu atribútov sa vyráta ako súčet unikátnych a multiplicitných záznamov**.**	Za skupinu atribútov, ktoré majú definované biznis pravidlo definujúce jedinečnú kombináciu hodnôt v rámci skupiny atribútov.		
Referenčná integrita	Kompletnosť referenčného identifikátora	Reporting Profilingu dát	Áno	8.1a	Percentuálny podiel záznamov v atribúte tabuľky, ktorý je zároveň referenčným identifikátorom a obsahuje akúkoľvek hodnotu okrem 'null' a prázdnej hodnoty.	Za atribút, ktorý má definované biznis pravidlo definujúce povinne vyplnenú hodnotu a zároveň je označený za referenčný identifikátor.	Je atribút referenčný identifikátor? Je atribút povinný? (Druhá otázka je identická s ukazovateľom 4.2) **POZOR**: metadáta nemusia obsahovať informácie o všetkých povinných atribútov z pohľadu biznisu. Povinnosť je niekedy ošetrená v samotných vstupoch do databázy, napríklad cez webové formuláre.	Ak existuje potvrdenie, že atribút je referenčný identifikátor a zároveň je povinný, tak sa môže merať KPI 8.1a

Tabuľka 4: Zoznam KPI a ich použitie pri meraní

1.2 Nástroje na meranie dátovej kvality

Je možné zmerať kvalitu dát v rôznych nástrojoch. V našom prípade sa používal Talend a MySQL.

Pre prácu s Talendom je potrebné zabezpečiť prístupy na Talend, prípadne inštaláciou Talend Data Fabric (https://www.talend.com/products/data-fabric/) alebo Talend Open Studio for Data Quality verzia 7.1.1 (https://www.talend.com/products/data-quality/data-quality-release-notes/). Ak by neexistovala možnosť zabezpečiť profesionálnu verziu, meranie je možné vykonať pomocou Open verzie. Avšak Open verzia nepodporuje meranie dátovej kvality v plnom rozsahu a disponuje len niektorými funkcionalitami.

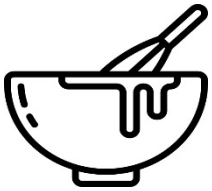
Pre pokročilejších používateľov môže byť zhodná cesta formou programovacieho jazyka SQL (štruktúrovaný dopytovací jazyk) v preferovanom RDBMS, najviac populárne sú MySQL, MS SQL Server, Oracle.

1.3 Požadovaný formát dát objektu merania

Získanie objektu merania je možné docieliť štyrmi spôsobmi:

1. Priamym prístupom na živú databázu – SQL DB connection
2. Exportovaním dát z príslušnej databázy do formátu SQLBAK (backup), nahratie do lokálne SQL databázy a následne pripojenie na túto lokálnu databázu – SQL DB connection
3. Exportovaním dát z príslušnej databázy do formátu CSV, nahratie CSV súboru do lokálnej SQL databázy.
4. Exportovaním dát z aplikácie pripojenej na databázu do formátu XML, parsovanie XML súboru v Talende.

Medzi najideálnejšie scenáre získavania samotných dát je spôsob číslo 1 a 2. Pre spôsob číslo 3 môže nastať znehodnotenie štruktúry dát. Je to spôsobené prevažne textami, ktoré môžu obsahovať bodkočiarku a tým narušiť dátové štruktúry v CSV formáte. Posledný spôsob s použitím formátu XML je najkomplikovanejší, keďže parsovanie komplexnejších súborov bez riadnej dokumentácie štruktúry je veľmi komplikované. Toto platí hlavne v prípade veľkých súborov (1GB+).



2 Ako postupovať (návod)

Nasledujúca kapitola predstaví krok za krokom postup „Kuchárku“, ako merať dátovú kvalitu. Meranie dátovej kvality sa môže vykonať v ľubovoľnom nástroji, dokonca aj v Exceli. Tento návod sa vzťahuje pre nástroje Talend a SQL Server Management Studio. Profiling dát bude prebiehať v Talende a pre pokročilejšiu analýzu sa bude používať SQL Server Management Studio, kde sa skriptujú príkazy v SQL.

Prvá časť merania dátovej kvality pozostáva z profilingu dát, druhá časť z reporting profilingu dát a posledná časť je pokročilejšia analýza. Analýzy vychádzajú z parametrov dátovej kvality a ich ukazovateľov, ku ktorým patria jednotlivé KPI. Meranie dátovej kvality sa kvantifikuje výpočtami KPI hodnôt.

2.1 Postup pre meranie dátovej kvality

Postupnosť krokov merania dát:

1. V prípade Talendu sa pripája k dátovému zdroju (databáza alebo CSV), kvôli prístupu k tabuľkám a stĺpcom, na ktorých sa definujú a vykonávajú analýzy.

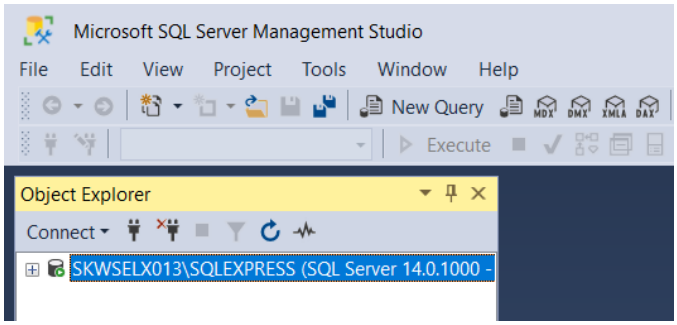
2. Meranie dátovej kvality je vykonané prostredníctvom výpočtov KPI hodnôt. Tieto hodnoty sa získajú pomocou analýzy obsahu databázy, stĺpcovej analýzy, tabuľkovej analýzy a pod. Patrí sem profiling dát a pokročilejšie analýzy.

2.1.1 Pripojenie k dátovému zdroju – v prípade, že ste sa rozhodli používať Talend

Na meranie dátovej kvality sa použil Talend MDM Platform (viď Obrázok 3), v ktorej sa vykoná profiling dát.

Obrázok 3: Talend MDM Platform

Na prácu s dátami a pokročilejšiu analýzu sa použil SQL Server Management Studio (viď Obrázok 4).



Obrázok 4: SQL Server Management Studio

Do SQL Studia sa naimportujú dáta, kvôli jednoduchšej manipulácií, čo bude cenné hlavne pri profilingu dát.

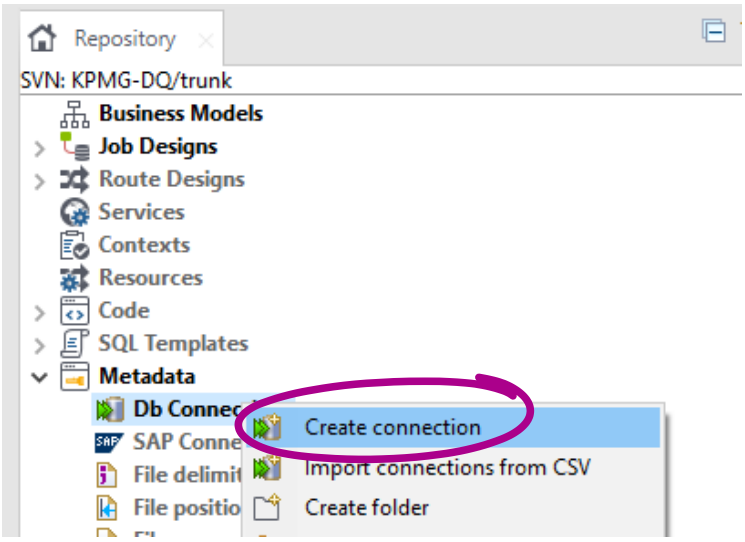
Keď sú dáta naimportované do SQL Studia, môžu sa pripojiť do Talendu. Postup je nasledovný:

Talend postup

1. V Talend MDM Platforme sa zvolí Integration modul. Slúži na získanie dát (viď Obrázok 5).

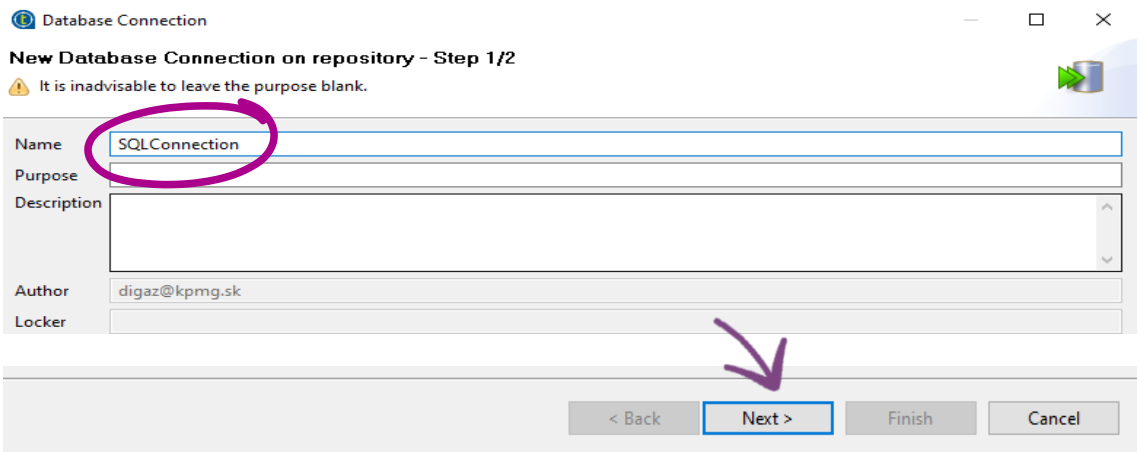
Obrázok 5: Integration modul

2. Cez Metadata v časti Repository sa vytvorí nové pripojenie k databáze (viď Obrázok 6). Metadata ponúka širokú možnosť pripojení, napríklad na databázu, ale aj rôzne ďalšie zdroje alebo importovanie rôznych súborov, ako napríklad CSV, XML a pod.



Obrázok 6: Vytvorenie nového pripojenia k databáze

3. Zobrazí sa okno, v ktorom je treba vyplniť názov pripojenia na databázu. V časti Name sa doplní názov (viď Obrázok 7), ako sa má volať pripojenie na databázu. Vyplnenie mena je povinná podmienka, bez ktorej nie je možné pokračovať ďalej. Pre lepšie chápanie je možné vyplniť aj časť Purpose v zmysle "prečo chcete vytvoriť túto analýzu" a časť Description v zmysle „čo daná analýza bude robiť". Po vyplnení klik na Next.



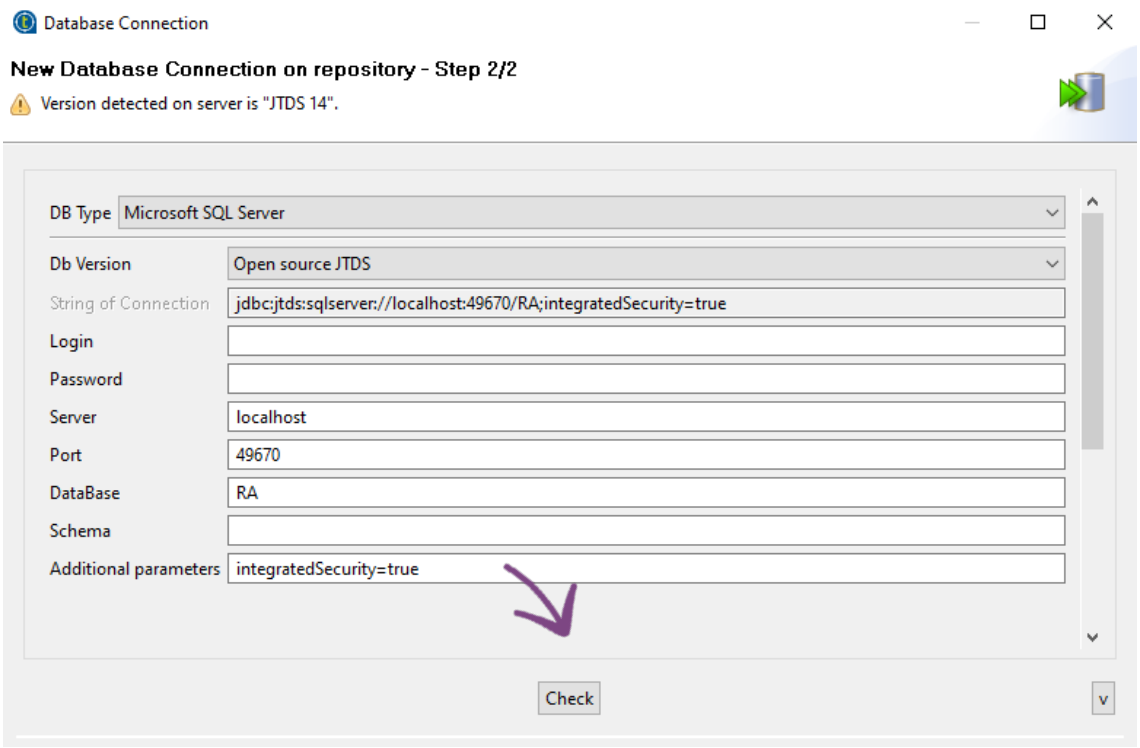
Obrázok 7: Názov pripojenia

4. V ďalšom kroku treba vyplniť údaje na Pripojenie sa na databázu (viď Obrázok 8). Týka sa to parametrov DB Type, DB Version, Server, Port, DataBase a Additional parameters.

   Kroky:

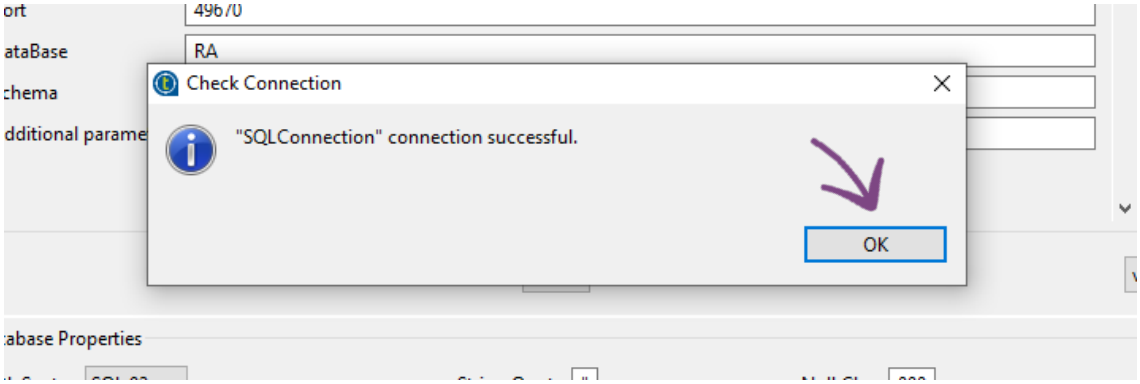
   — Vyberie sa „Microsoft SQL Server“ ako *DB Type*
   — Vyberie sa „Open source JTDS“ ako *Db Version*
   — Špecifikuje sa meno servera (v tomto prípade „localhost“)
   — Špecifikuje sa *port*
   — Špecifikuje sa názov *DataBase* (tu má byť názov vašej databázy v SQL Studiu)
   — Nezabudnúť na dôležitý parameter *Additional parameters* vyplniť ako „integratedSecurity=true“.

Po vyplnení parametrov sa otestuje pripojenie kliknutím na Check.



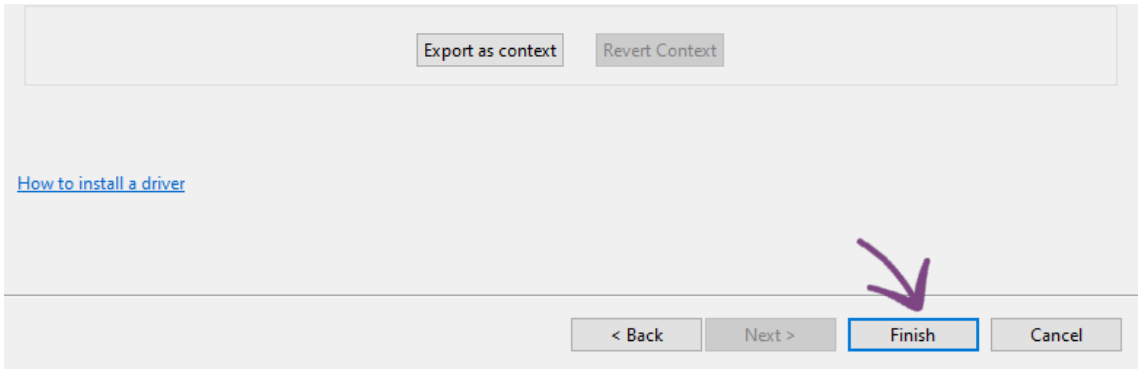
Obrázok 8: Vyplnenie údajov pripojenia sa na databázu

5. Ak bolo všetko správne vyplnené, tak sa zobrazí hlásenie, že „SQLConnection pripojenie bolo úspešne založené“ (viď Obrázok 9). Klik na OK . V prípade, že nastane problém, je potrebné stiahnuť dodatočne „jtds-1.3.1-dist.zip“ a obrátiť sa na IT podporu.



Obrázok 9: Kontrola založeného pripojenia

6. Keď je pripojenie na databázu založené, klik na Finish a časť pripojenie ku dátovému zdroju je hotová (viď Obrázok 10).



Obrázok 10: Hotové pripojenie

Pripomíname, že je lepšie sa pripojiť priamo na živú databázu a z nej merať dátovú kvalitu. Otázkou je možnosť získania prístupu.

2.1.2 Meranie dátovej kvality

Gratulujeme, podarilo sa Vám úspešne pripojiť k dátovému zdroju. Teraz nás čaká zábavnejšia časť, v ktorej sa bude merať kvalita dát.

2.1.2.1 Prvé odborné posúdenie: ukazovateľ Sledovanie konzistentnosti

Generický postup

Meranie prvého KPI: Ukazovateľ sledovania konzistentnosti je odporúčaný krok, ktorým treba začať. Znamená to vypočítať percentuálny podiel atribútov, ktoré majú definované biznis pravidlá. Meranie je vykonané odborným posúdením a prakticky ide o analyzuje zoznamu biznis pravidiel, ktoré sú popísané v časti (1.1.5). Najprv treba určiť kľúčové atribúty. Zoznam biznis pravidiel sa analyzuje na úrovni celého datasetu pre vybrané atribúty. Výsledkom prvého merania bude vyplnená šablóna nižšie.

Dataset	Parameter	Ukazovateľ	Význam KPI 2.1a	Zmeraná hodnota KPI 2.1a
	Konzistentnosť	Sledovanie konzistentnosti	Percentuálny podiel atribútov, ktoré majú definované biznis pravidlá.	

Tabuľka 5: Šablóna prvého merania KPI: ukazovateľ Sledovanie konzistentnosti

Šablóna pozostáva zo štruktúrovaných informácií:

— Dataset: **je potrebné vyplniť**
— Parameter: je už vyplnený
— Ukazovateľ: je už vyplnený
— Význam KPI 2.1a: je už vyplnený a popisuje význam KPI
— Zmeraná hodnota KPI 2.1a: **je potrebné vyplniť**

V praxi nie je potrebné aby mali všetky atribúty definované biznis pravidlá. Je potrebné rozlišovať tie atribúty, u ktorých nie je potreba merať dátovú kvality. Pokiaľ tieto „pre potreby merania nepotrebné“ atribúty nemajú definované biznis pravidlá, je to akceptovateľný stav.

2.1.2.2 *Profiling dát*

Je základné meranie, pri ktorom je nutná znalosť biznis pravidiel. Profiling nám dá prvý pohľad ako dáta vyzerajú a čo obsahujú. Hodnoty z tohto merania sa využívajú pri reportingu profilingu dát.

Pri profilingu dát sa vykonáva stĺpcová analýza. Patrí sem:

— Jednoduchá štatistika
  - Row Count – celkový počet záznamov
  - Null Count – počet záznamov s hodnotou Null
  - Distinct Count – počet odlišných hodnôt v záznamoch
  - Unique Count – počet jedinečných hodnôt v záznamoch
  - Duplicate Count – počet hodnôt, ktoré sa vyskytujú v záznamoch aspoň 2-krát (tento viacnásobný výskyt sa nazýva multiplicita a patria sem duplicity, triplicity a podobne)
  - Blank Count – počet záznamov s prázdnou hodnotou (nevyplnenou hodnotou)

— Frekvencia hodnôt – zistí sa frekvencia výskytu hodnôt v atribúte

— Vzory hodnôt – zistia sa frekvencie výskytu vzorov hodnôt v atribúte

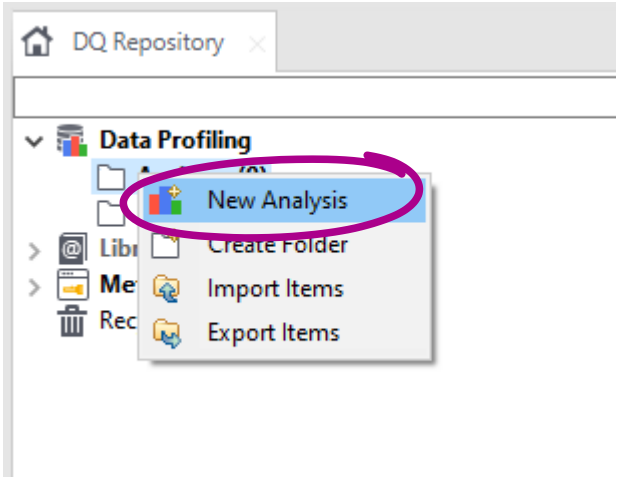
Základná analýza je vytvorená nad jednotlivými stĺpcami tabuľky. Jej kroky sú nasledovné:

1. V Talend MDM Platforme sa zvolí Profiling modul (viď Obrázok 11). Slúži na analýzu dát.

Talend postup

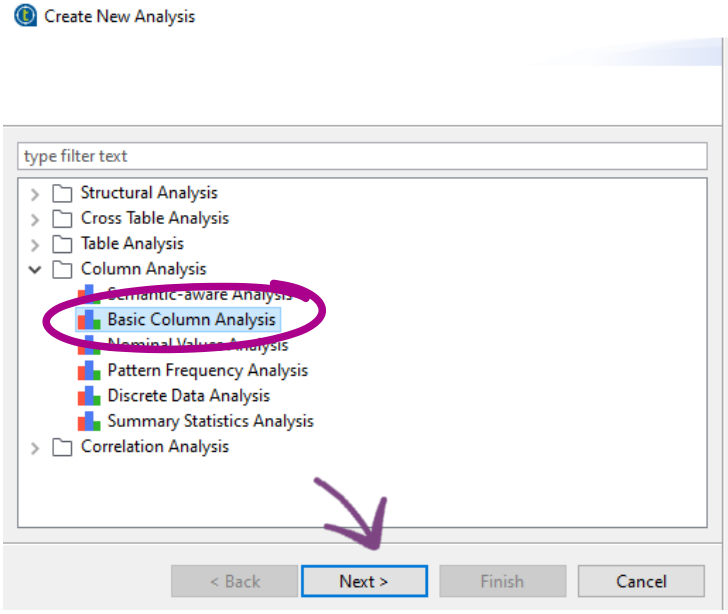
Obrázok 11: Profiling modul

2. Cez Data Profiling v časti Repository sa vytvorí Nová Analýza (viď Obrázok 12). Okrem vytvorenia novej analýzy je možné vytvoriť aj nový priečinok (odporúčané v prípade analýzy rôznych datasetov pre poriadok).



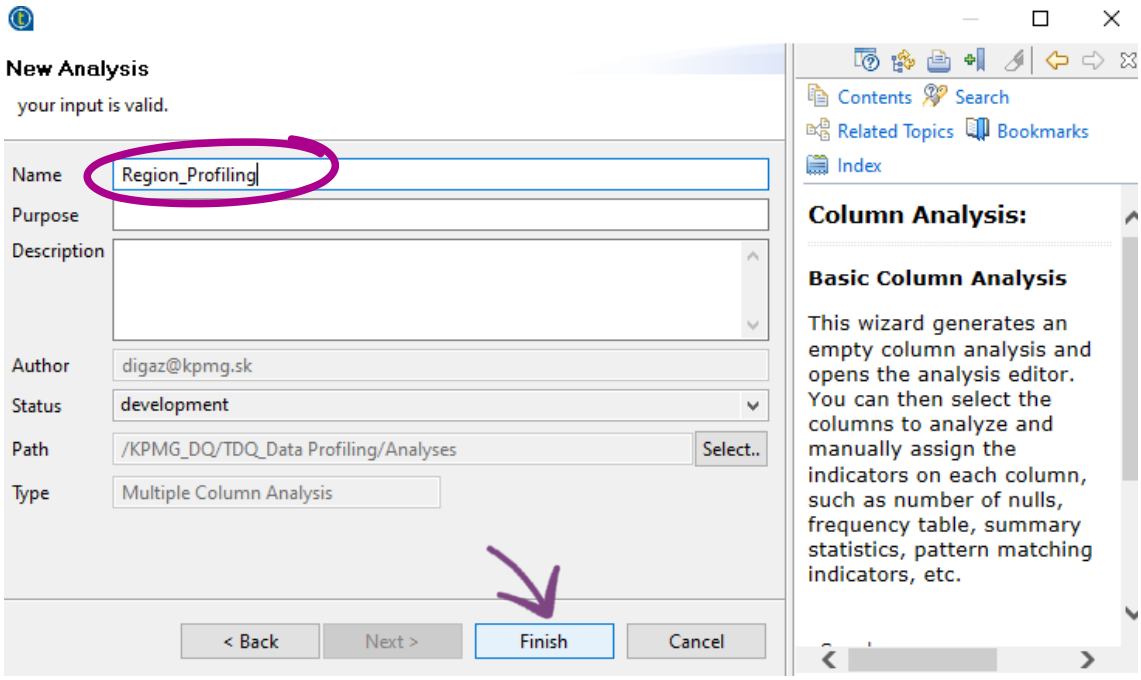
Obrázok 12: Vytvorenie novej analýzy

3. Pre profilovanie dát nad stĺpcami sa vyberie Basic Column Analysis (viď Obrázok 13) a prejde sa ďalej kliknutím na Next. Okrem analýzy nad stĺpcami sa môže analyzovať aj nad tabuľkami alebo naprieč tabuliek, prípadne korelačná analýza alebo analýza štruktúry.



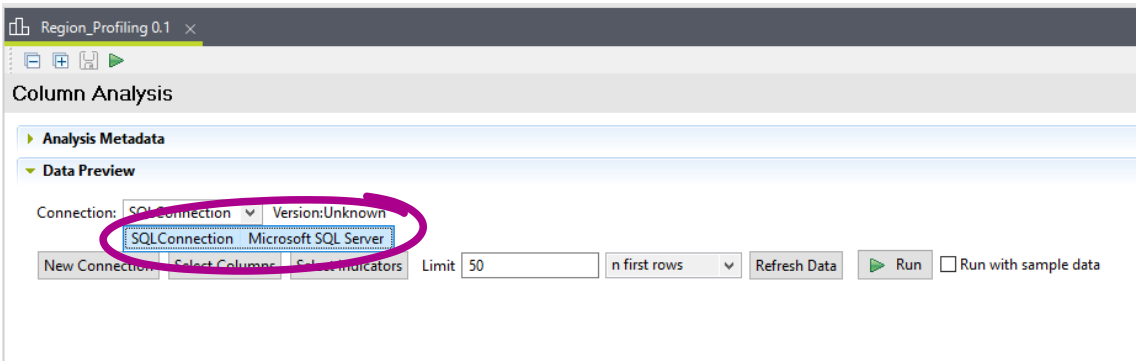
Obrázok 13: Výber basic column analysis

4. Zobrazí sa okno New Analysis (viď Obrázok 14), v ktorom je treba vyplniť názov analýzy, v tomto prípade „Region_Profiling“. Zaužívaná konvencia písania je najprv názov dát a potom aktivita, ktorá sa bude vykonávať oddelené znakom _. Po vyplnení klik na Finish.



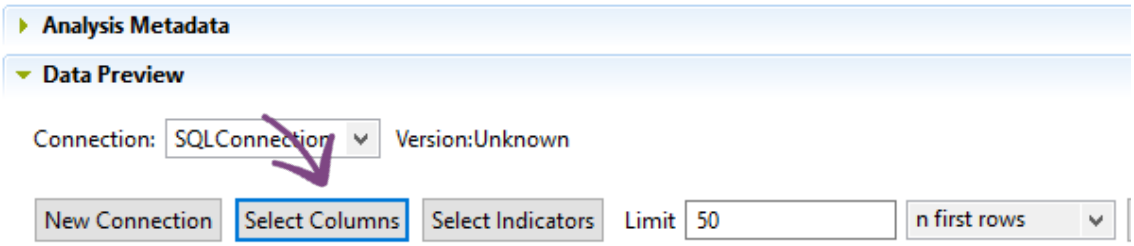
Obrázok 14: Vyplnenie názvu analýzy

5. Teraz treba vybrať dátový zdroj, nad ktorým sa má analýza vykonať (viď Obrázok 15). Kliknútím na Connection sa vyberie potrebný zdroj.



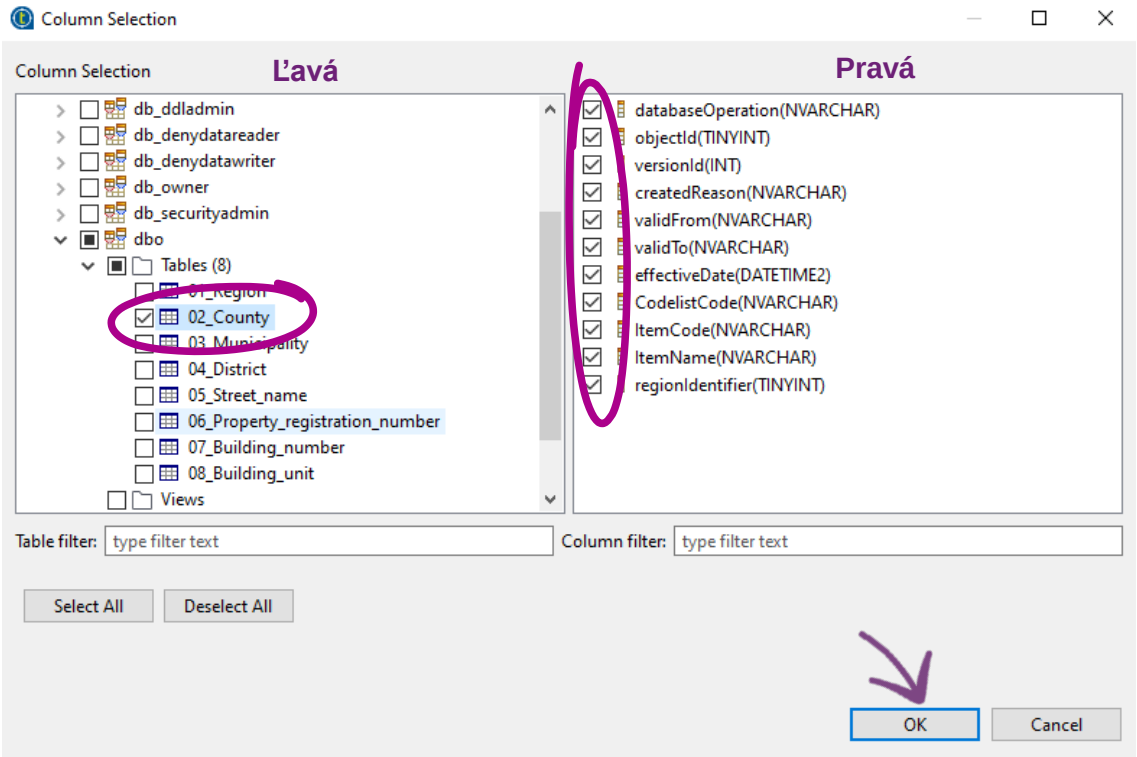
Obrázok 15: Výber dátového zdroja

6. Nasleduje krok výberu stĺpcov kliknutím na Select Columns (viď Obrázok 16).



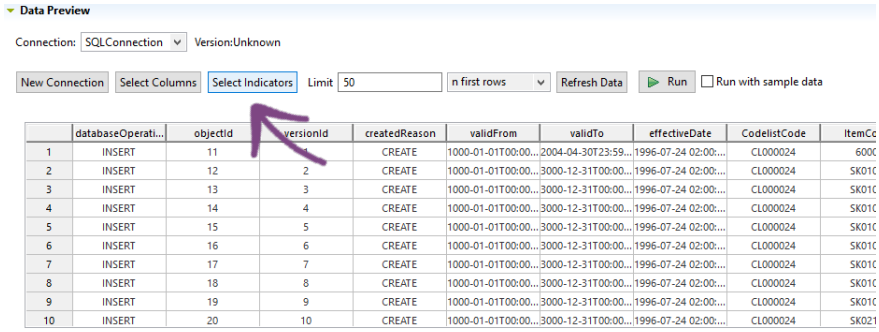
Obrázok 16: Výber stĺpcov

7. V ľavej časti je potrebné sa preklikať k tabuľke (v tomto prípade 02_County), ktorá obsahuje stĺpce, ktoré sa majú analyzovať. V pravej časti sa vyberajú konkrétne stĺpce zakliknutím. Vyberajú sa len tie stĺpce, ktoré sa budú analyzovať (viď Obrázok 17). Následne klik na OK.



Obrázok 17: Konkrétny výber stĺpcov

8. Po výbere stĺpcov nasleduje krok výberu indikátorov resp. ukazovateľov kliknutím na Select Indicators (viď Obrázok 18).

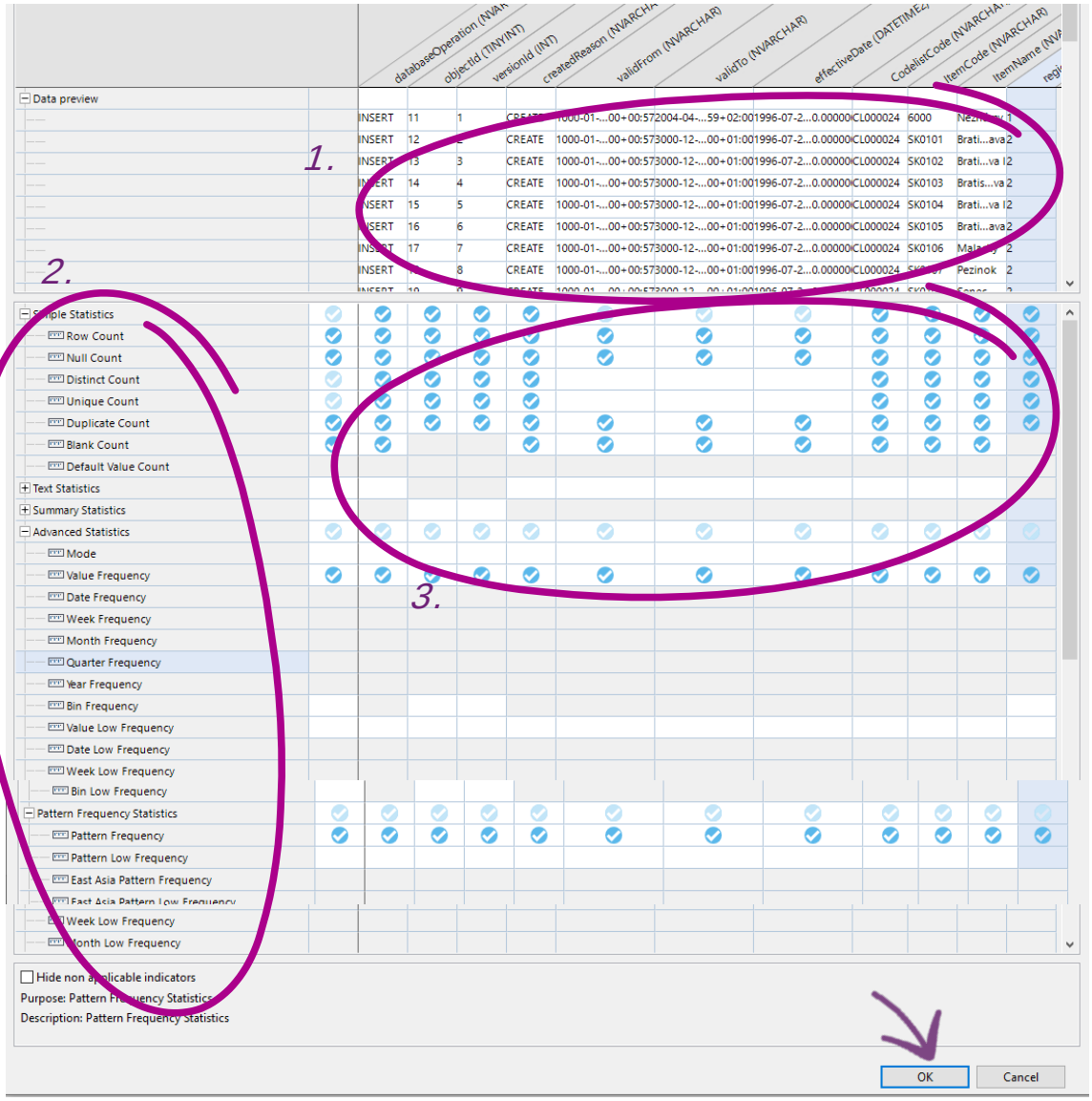


	databaseOperati...	objectId	versionId	createdReason	validFrom	validTo	effectiveDate	CodelistCode	ItemCo
1	INSERT	11	1	CREATE	1000-01-01T00:00:...	2004-04-30T23:59:...	1996-07-24 02:00:...	CL000024	600(
2	INSERT	12	2	CREATE	1000-01-01T00:00:...	3000-12-31T00:00:...	1996-07-24 02:00:...	CL000024	SK01(
3	INSERT	13	3	CREATE	1000-01-01T00:00:...	3000-12-31T00:00:...	1996-07-24 02:00:...	CL000024	SK01(
4	INSERT	14	4	CREATE	1000-01-01T00:00:...	3000-12-31T00:00:...	1996-07-24 02:00:...	CL000024	SK01(
5	INSERT	15	5	CREATE	1000-01-01T00:00:...	3000-12-31T00:00:...	1996-07-24 02:00:...	CL000024	SK01(
6	INSERT	16	6	CREATE	1000-01-01T00:00:...	3000-12-31T00:00:...	1996-07-24 02:00:...	CL000024	SK01(
7	INSERT	17	7	CREATE	1000-01-01T00:00:...	3000-12-31T00:00:...	1996-07-24 02:00:...	CL000024	SK01(
8	INSERT	18	8	CREATE	1000-01-01T00:00:...	3000-12-31T00:00:...	1996-07-24 02:00:...	CL000024	SK01(
9	INSERT	19	9	CREATE	1000-01-01T00:00:...	3000-12-31T00:00:...	1996-07-24 02:00:...	CL000024	SK01(
10	INSERT	20	10	CREATE	1000-01-01T00:00:...	3000-12-31T00:00:...	1996-07-24 02:00:...	CL000024	SK021

Obrázok 18: Výber ukazovateľov

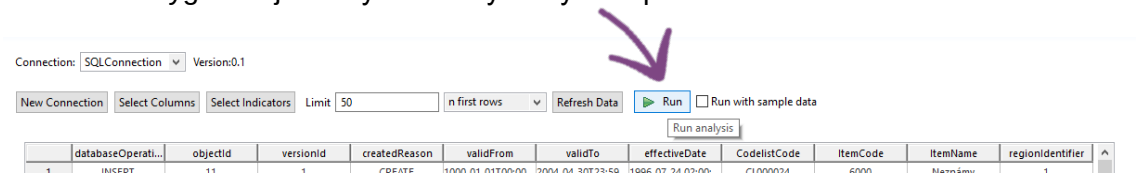
9. Prišiel veľmi dôležitý krok, ktorým je výber indikátorov resp. ukazovateľov (viď Obrázok 19). V tejto časti sa rieši základná štatistika, frekvencia hodnoty, vzor hodnôt a podobne.

Pohľad 1 v hornej časti (Obrázok 19) zobrazuje dáta. **Pohľad 2** ponúka zoznam možností profilovania dát. V **pohľade 3** sa zaklikávajú žiadané indikátory, resp. ukazovatele.



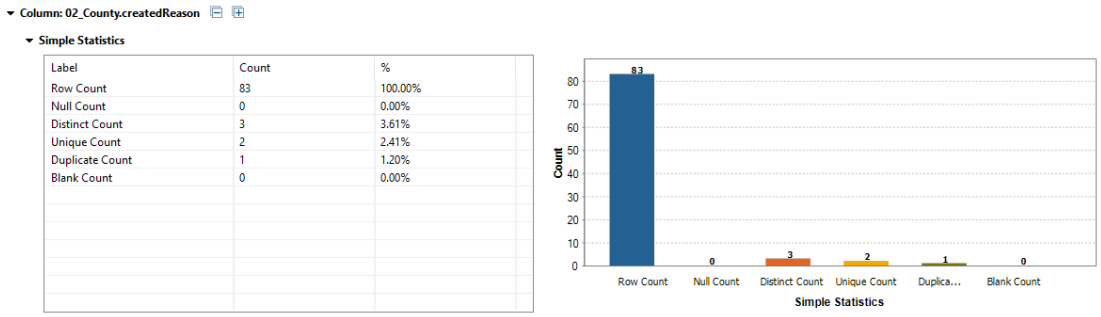
Obrázok 19: Výber konkrétnych ukazovateľov

10. Zásadný krok, kde sa spustí analýza kliknutím na Run (viď Obrázok 20). V tejto fáze sa vygeneruje analýza nad vybranými stĺpcami.



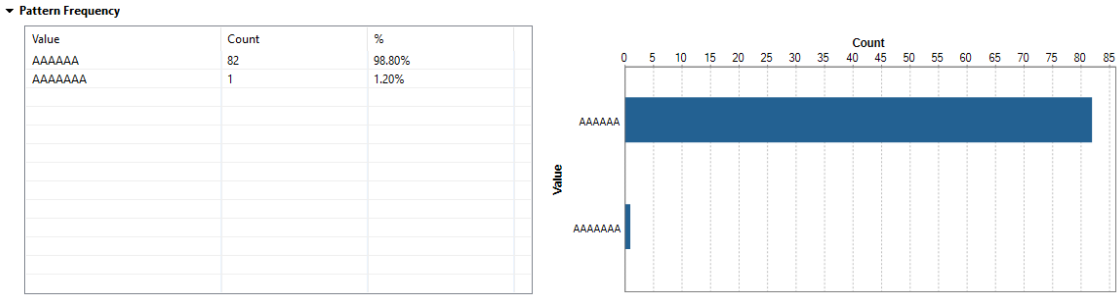
Obrázok 20: Spustenie analýzy

Vygenerovaná analýza, po ukončení *RUN* fázy, predstavuje súbor grafov príslušných stĺpcov, resp. atribútov. Obrázok 21 zobrazuje základná štatistiku, ktorá obsahuje hodnoty počtu riadkov, null hodnôt, odlišných, jedinečných, duplicitných a prázdnych hodnôt.



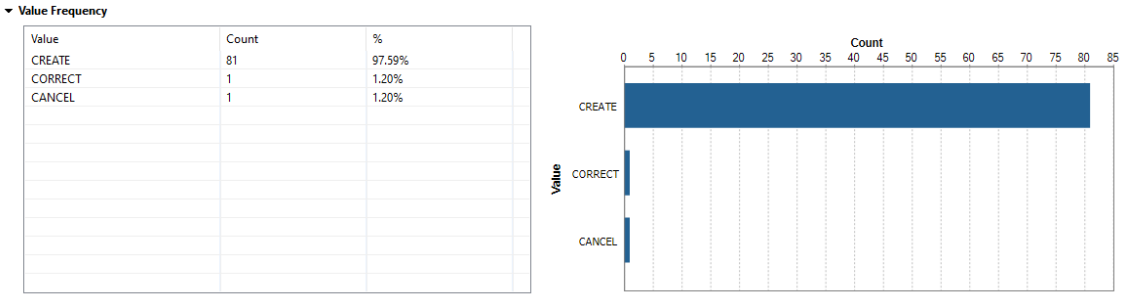
Obrázok 21: Príklad základnej štatistiky

Obrázok 22 zobrazuje frekvencia vzoru, ktorá má 2 vzory pričom prvý má 6 znakov s frekvenciou 82 a druhý vzor má 7 znakov a vyskytuje sa len raz.



Obrázok 22: Príklad frekvencie vzoru

Obrázok 23 zobrazuje frekvenciu hodnoty, ktorá má 3 hodnoty s vypočítanou frekvenciou výskytu.



Obrázok 23: Príklad frekvencie hodnoty

Generický postup

Hodnoty z profilingu dát sa zapíšu do odporúčanej šablóny profilingu dát:

Dataset	Tabuľka	Atribút	Row Count	Null Count	Distinct Count	Unique Count	Duplilcate Count	Blank Count	Pattern Count

Tabuľka 6: Šablona profilingu dát

Šablóna pozostáva zo štruktúrovaných informácií na zhromažďovanie počtu hodnôt:

- Dataset – názov datasetu z karty merania,
- Tabuľka – názov tabuľka
- Atribút – názov atribútu,
- Row Count – celkový počet záznamov,
- Null Count - počet záznamov s hodnotou Null,
- Distinct Count - počet odlišných hodnôt v záznamoch,
- Unique Count - počet jedinečných hodnôt v záznamoch,
- Duplicate Count - počet hodnôt, ktoré sa vyskytujú v záznamoch aspoň 2-krát (tento viacnásobný výskyt sa nazýva multiplicita a atria sem duplicity, triplicity a podobne),
- Blank Count - počet záznamov s prázdnou hodnotou (nevyplnenou hodnotou),
- Pattern Count – frekvencia vybraného výskytu vzoru,

2.1.2.3 *Reporting profilingu dát*

V tejto časti sa počítajú KPI hodnoty pre ukazovatele dátovej kvality. Pričom hodnoty z časti Profiling dát (2.1.2.2) sa využívajú na ich výpočet. Medzi počítané KPI patrí:

1. **Ukazovateľ dodržanie formátu atribútu: KPI 3.1a** Percentuálny podiel záznamov v atribúte tabuľky, ktorý obsahuje hodnotu v požadovanom formáte. Pri výpočte sa nepočítajú hodnoty Null.
   - ✓ Vypočíta sa ako (frekvencia vzoru/počet riadkov)*100, aby sa určil správny vzor atribútu je dôležité poznať biznis pravidlo. V tomto prípade sa vyberal formát, ktorý zodpovedal najpravdepodobnejšiemu výskytu vzoru, pretože biznis pravidlo nebolo známe.
2. **Ukazovateľ vyplnenosť povinného údaja: KPI 4.2a** Percentuálny podiel záznamov v atribúte tabuľky, ktorý obsahuje akúkoľvek hodnotu okrem 'null' a prázdnej hodnoty.
   - ✓ Vypočíta sa ako (Počet riadkov – Počet Null hodnôt – Počet prázdnych hodnôt)/(Počet riadkov)*100. Aby sa určili správne atribúty pre výpočet vyplnenosti povinných údajov, je dôležité poznať biznis pravidlo. V tomto prípade sa vybrali všetky atribúty.
3. **Ukazovateľ vyplnenosť nepovinného údaja: KPI 4.3a** Percentuálny podiel záznamov v atribúte tabuľky, ktorý obsahuje akúkoľvek hodnotu vrátane prázdnej hodnoty a okrem 'null' hodnoty.
   - ✓ Vypočíta sa ako (Počet riadkov – Počet Null hodnôt)/(Počet riadkov)*100. Aby sa určili správne atribúty pre výpočet vyplnenosti nepovinných údajov, je dôležité poznať biznis pravidlo. V tomto prípade sa vybrali všetky atribúty.
4. **Ukazovateľ unikátnosť hodnoty: KPI 5.2a** Percentuálny podiel záznamov v atribúte tabuľky, ktorý neobsahuje rovnakú hodnotu v rámci atribútu viac ako jeden krát.
   - ✓ Vypočíta sa ako (Počet unikátnych hodnôt)/(Počet riadkov)*100. Aby sa určili správne atribúty pre výpočet unikátnosti hodnôt údajov, je dôležité poznať biznis pravidlo. V tomto prípade sa vybrali všetky atribúty.
5. **Ukazovateľ unikátnosť hodnoty: KPI 5.2b** Percentuálny podiel záznamov v atribúte tabuľky, ktorý obsahuje rovnakú hodnotu v rámci atribútu viac ako jeden krát.
   - ✓ Vypočíta sa ako (Počet multiplicitných hodnôt)/(Počet riadkov)*100. Aby sa určili správne atribúty pre výpočet multiplicity hodnoty údajov, je dôležité poznať biznis pravidlo. V tomto prípade sa vybrali všetky atribúty.
6. **Ukazovateľ Kompletnosť referenčného identifikátora: KPI 8.1a** Percentuálny podiel záznamov v atribúte tabuľky, ktorý je zároveň referenčným identifikátorom a obsahuje akúkoľvek hodnotu aj 'null' okrem prázdnej hodnoty.
   - ✓ Vypočíta sa ako (Počet unikátnych hodnôt)/(Počet riadkov)*100 pre vybraný identifikátor. Aby sa určili správny identifikátor pre výpočet kompletnosti referenčného identifikátora, je dôležité poznať biznis

pravidlo. V tomto prípade sa vybrali identifikátori, ktoré zodpovedali a boli považované za identifikátory tabuliek.

Pozn.: Počet odlišných (distinct) hodnôt sa rovná súčtu unikátnych a multiplicitných (duplicate) hodnôt.

Pri výpočte percentuálnych podielov ukazovateľov (výpočet KPI je nutná znalosť biznis pravidiel. Vypočítané hodnoty sa zapíšu do odporúčanej šablóny pre reporting profilingu dát nižšie:

Generický postup

Dataset	Tabuľka	Atribút	Ukazovateľ dodržanie formátu atribútu: KPI 3.1a	Ukazovateľ vyplnenosť povinného údaja: KPI 4.2a	Ukazovateľ vyplnenosť nepovinného údaja: KPI 4.3a	Ukazovateľ unikátnosť hodnoty: KPI 5.2a	Ukazovateľ unikátnosť hodnoty: KPI 5.2b (multipliticta)	Kompletnosť referenčného identifikátora: KPI 8.1a
			Percentuálny podiel záznamov v atribúte tabuľky, ktorý obsahuje hodnotu v požadovanom formáte	Percentuálny podiel záznamov v atribúte tabuľky, ktorý obsahuje akúkoľvek hodnotu okrem 'null' a prazdnej hodnoty.	Percentuálny podiel záznamov v atribúte tabuľky, ktorý obsahuje akúkoľvek hodnotu vrátane prazdnej hodnoty a okrem 'null' hodnoty.	Percentuálny podiel záznamov v atribúte tabuľky, ktorý neobsahuje rovnakú hodnotu v rámci atrubútu viac ako jeden krát.	Percentuálny podiel záznamov v atribúte tabuľky, ktorý obsahuje rovnakú hodnotu v rámci atribútu viac ako jeden krát.	Percentuálny podiel záznamov v atribúte tabuľky, ktorý je zároveň referenčným identifikátorom a obsahuje akúkoľvek hodnotu aj 'null' okrem prazdnej hodnoty.

Tabuľka 7: Šablóna reportingu profilingu dát

Šablóna obsahuje KPI hodnoty a je finálnym výstupom časti profilingu dát. KPI hodnoty by sa mali merať pravidelne, pre identifikáciu zlepšenia dátovej kvality.

2.1.2.4 Pokročilejšia analýza

V tejto časti sa tiež počítajú KPI hodnoty pre ukazovatele dátovej kvality. Na rozdiel od reportingu profilingu dát, kde stačia získané hodnoty z profilingu, tu sa hodnoty získavajú pomocou používania pokročilejších techník. Prakticky návod sa nezameriava na detailný popis všetkých pokročilejších techník, ktoré sú vykonané v nástroji ako napríklad Talend, alebo aj formou písania SQL dotazov v SQL Management Studiu. Na pokročilejšie techniky je potrebná určitá úroveň programovania. Cieľom je pochopiť, ktoré KPI hodnoty sa počítajú v pokročilejšej analýze. Medzi počítané KPI patrí:

1. **Ukazovateľ Syntaktická presnosť atribútu: KPI 1.2a** – Percentuálny podiel záznamov v atribúte tabuľky, ktorý obsahuje hodnoty v súlade s definovaným pravidlom povoľujúcom určené hodnoty. Výpočet KPI hodnôt pomocou skriptu alebo Talend riešenia.

2. **Ukazovateľ Sémantická presnosť atribútu: KPI 1.3a** – Percentuálny podiel záznamov v atribúte tabuľky, ktorý obsahuje hodnotu v súlade s definovaným biznis pravidlom odkazujúcim na zdroj pravdy. Výpočet KPI hodnôt pomocou skriptu alebo Talend riešenia.

3. **Ukazovateľ Dodržiavanie biznis pravidla: KPI 2.2b** - Percentuálny podiel atribútov, ktorých hodnoty plne dodržujú vzťahy medzi atribútmi. Výpočet KPI hodnôt pomocou skriptu alebo Talend riešenia.

4. **Ukazovateľ Unikátnosť záznamov: KPI 5.3a** - Percentuálny podiel záznamov za skupinu atribútov, ktorý obsahuje rovnakú kombináciu hodnôt v rámci skupiny atribútov len jeden krát. Výpočet KPI hodnôt pomocou skriptu alebo Talend riešenia.

5. **Ukazovateľ Unikátnosť záznamov: KPI 5.3b** - Percentuálny podiel záznamov za skupinu atribútov, ktorý obsahuje rovnakú kombináciu hodnôt v rámci skupiny atribútov, viac ako jeden krát. (Jedná sa o multiplicitu záznamu = duplicita, triplicita a pod.). Výpočet KPI hodnôt pomocou skriptu alebo Talend riešenia.

Vypočítané hodnoty sa zapisujú do odporúčanej šablóny pre pokročilejšiu analýzu nižšie:

Generický postup

Dataset	Tabuľka	Atribút	Ukazovateľ Syntaktická presnosť atribútu: KPI 1.2a	Ukazovateľ Sémantická presnosť atribútu: KPI 1.3a	Ukazovateľ Dodržiavanie biznis pravidla: KPI 2.2b	Ukazovateľ Unikátnosť záznamov: KPI 5.3a	Ukazovateľ Unikátnosť záznamov: KPI 5.3b
			Percentuálny podiel záznamov v atribúte tabuľky, ktorý obsahuje hodnoty v súlade s definovaným pravidlom povoľujúcom určené hodnoty.	Percentuálny podiel záznamov v atribúte tabuľky, ktorý obsahuje hodnotu v súlade s definovaným biznis pravidlom odkazujúcim na zdroj pravdy.	Percentuálny podiel atribútov, ktorých hodnoty plne dodržujú vzťahy medzi atribútmi.	Percentuálny podiel záznamov za skupinu atribútov, ktorý obsahuje rovnakú kombináciu hodnôt v rámci skupiny atribútov len jeden krát.	Percentuálny podiel záznamov za skupinu atribútov, ktorý obsahuje rovnakú kombináciu hodnôt v rámci skupiny atribútov, viac ako jeden krát. (Jedná sa o multiplicitu záznamu = duplicita, triplicita a pod.)

Tabuľka 8: Šablóna pokročilejšej analýzy

Šablóna obsahuje KPI hodnoty a je finálnym výstupom časti pokročilejšia analýza. KPI hodnoty by sa mali merať pravidelne, pre identifikáciu zlepšenia dátovej kvality.

Príklady pokročilejších techník

V tejto časti sú predstavené reálne príklady, ktorých KPI hodnoty sú vypočítané buď Talend riešením alebo pomocou skriptov v SQL. Slúžia na lepšie pochopenie predstavy pokročilejšej analýzy.

1. Ukazovateľ Syntaktická presnosť atribútu: KPI 1.2a

Biznis pravidlo:

Adresné body (zemepisná šírka) sa musia nachádzať v rámci intervalu medzi najzápadnejším/najvýchodnejším bodom Slovenskej republiky.

Pokročilejšia analýza:

Zistia sa koordináty najzápadnejšieho a najvýchodnejšieho bodu Slovenskej republiky. Ak adresné body, patria do intervalu (medzi hodnoty najzápadnejšieho a najvýchodnejšieho bodu), tak spĺňajú definované pravidlo.

SQL postup

SQL:

```
/*
Zemepisná šírka
     Min = 47.73369167
     Max = 49.61373056
*/

-- Počet vyplnených adresných bodov je 1 179 601
SELECT COUNT(*)
  FROM [RA].[dbo].[7_Agrregated_Vchody_Orientačné čísla (Building_Number)]
  WHERE AxisL IS NOT NULL;

-- Počet AxisB hodnôt menších alebo rovných ako min(zemepisnej šírky) je 75

SELECT AxisB, AxisL
  FROM [RA].[dbo].[7_Agrregated_Vchody_Orientačné čísla (Building_Number)]
  WHERE AxisB <= 47.73369167
  ORDER BY AxisB DESC;

-- Počet AxisB hodnôt väčších alebo rovných ako max(zemepisnej šírky) je 1

SELECT AxisB, AxisL
  FROM [RA].[dbo].[7_Agrregated_Vchody_Orientačné čísla (Building_Number)]
  WHERE AxisB >= 49.61373056
  ORDER BY AxisB DESC;
```

(Počet adresných bodov, ktoré nepatria intervalu)/(celkový počet adresných bodov)*100 = 76/1179601*100 = 0,006%.

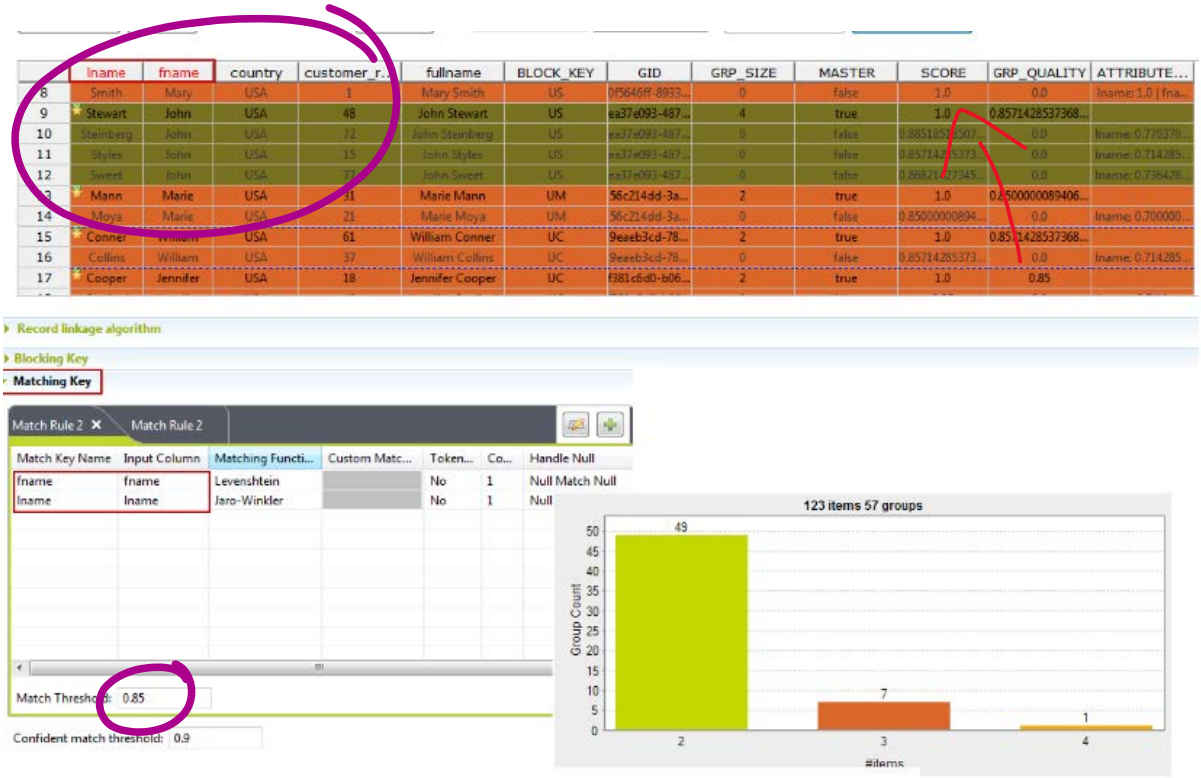
Treba získať adresné body, ktoré spĺňajú biznis pravidlo, teda 99,99% adresných bodov. Výsledná hodnota sa zapíše do šablóny Tabuľka 8.

2. Ukazovateľ Sémantická presnosť atribútu: KPI 1.3a

V prípade, že chceme zmerať zhodu datasetu so zdrojom pravdy použijeme Techniku **Približnej zhody podobností** – slúži na vytvorenie skupín podobných záznamov pomocou algoritmu Levenshtein alebo Jaro Winkler. V príklade sú použité 2 kľúče pre zhodu mena a priezviska. Algoritmus hľadá približnú zhodu vyjadrenú v percentách napr. zhoda od 85%. Záznam, ktorý má 100% zhodu s menom a priezviskom získava titul master, ku ktorému sú pridelené záznamy so zhodou min. 85%. Tým sa vytvorí skupina podobných záznamov (viď Obrázok 24).

Talend:

Talend postup



Obrázok 24: Príklad približnej zhody podobnosti

Výsledná hodnota KPI je 85% a zapíše sa do šablóny Tabuľka 8.

3. Ukazovateľ Dodržiavanie biznis pravidla: KPI 2.2b

Biznis pravidlo:

Každé aktívne konanie (stav začaty_proces_konania) má mať prideleného len aktívneho správcu.

Pokročilejšia analýza:

Zistí sa počet aktívnych alebo začatých konaní a z nich sa zistí počet podaní, ktoré nemajú prideleného správcu alebo pridelený správca nie je aktívny. Hodnota, ktorá dodržiava biznis pravidlo je výslednou KPI hodnotou.

SQL postup

SQL:

```
-- Počet všetkých začatých konaní a procesov je 5168
SELECT [RegisterUpadcov_Core].[core].[konanie].*,
        [RegisterUpadcov_Core].[core].[spravca].[stav] as spravca_stav
     FROM [RegisterUpadcov_Core].[core].[konanie]
     LEFT JOIN
     [RegisterUpadcov_Core].[core].[spravca]
     ON
[RegisterUpadcov_Core].[core].[konanie].[spravca]=[RegisterUpadcov_Core].[core].[
spravca].[id]
          -- Podmienka 1: Konanie alebo proces je zacaty*/
          WHERE ([RegisterUpadcov_Core].[core].[konanie].[stav_konania_ru] =
'ZACATY_PROCES_KONANIA' OR
[RegisterUpadcov_Core].[core].[konanie].[stav_konania_ru] = 'ZACATE_KONANIE')


-- Počet aktívnych konaní a procesov s neaktívnym správcom je 185
-- Vyber celu tabuľku konaní a prilep k nej stav správcu
SELECT [RegisterUpadcov_Core].[core].[konanie].*,
        [RegisterUpadcov_Core].[core].[spravca].[stav] as spravca_stav
     FROM [RegisterUpadcov_Core].[core].[konanie]
     LEFT JOIN
     [RegisterUpadcov_Core].[core].[spravca]
     ON
[RegisterUpadcov_Core].[core].[konanie].[spravca]=[RegisterUpadcov_Core].[core].[
spravca].[id]
          -- Podmienka 1: Konanie alebo proces je začatý
          WHERE ([RegisterUpadcov_Core].[core].[konanie].[stav_konania_ru] =
'ZACATY_PROCES_KONANIA'
                    OR
[RegisterUpadcov_Core].[core].[konanie].[stav_konania_ru] = 'ZACATE_KONANIE')
          -- Podmienka 2: stav prideleného správcu nie je aktívny
          AND [RegisterUpadcov_Core].[core].[spravca].[stav] != 'AKTIVNY'
```

(Počet aktívnych konaní)/(Počet začatých konaní)*100 = 185/5168*100= 3,57%

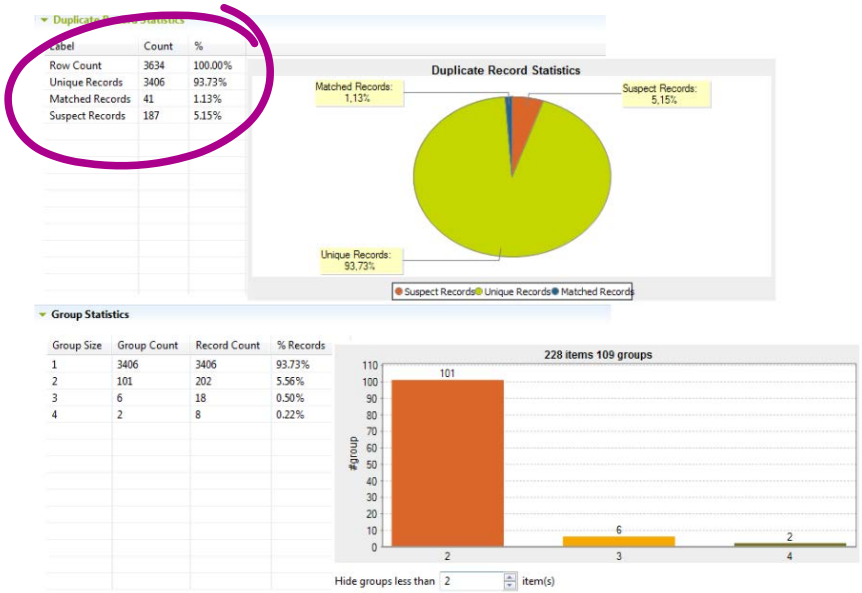
Pomer konaní, ktorá spĺňajú biznis pravidlo je 96,43%. Táto hodnota sa zapíše do šablóny Tabuľka 8.

4. Ukazovateľ Unikátnosť záznamov: KPI 5.3a

V prípade, že chceme posúdiť dataset z pohľadu unikátnosti záznamov, použijeme techniku **Presnej zhody**. Slúži napríklad na porovnanie celkového obrazu o unikátnych záznamov v datasete, viď obrázok nižšie.

Talend:

Talend postup



Obrázok 25: Príklad presnej zhody

Výslednou KPI hodnotou je percentuálny pomer *Unique records* 93,73%. Táto hodnota sa zapíše do šablóny Tabuľka 8.

5. Ukazovateľ Unikátnosť záznamov: KPI 5.3b

V tomto prípade, je postup rovnaký ako pri KPI 5.3a akurát, že podľa popisu KPI sa tu jedná o multiplicitu záznamu tzn. Riešia sa tu duplilcity, triplicity, kvadriplicity a pod.

Pri meraní dátovej kvality nezáleží na nástroji, ktorý použijete. Použite to, s čím sa vám lepšie pracuje.

2.1.2.5 *Finálne odborné posúdenie: Ukazovateľ dodržiavania biznis pravidla*

Meranie posledného KPI: Ukazovateľ dodržiavanie biznis pravidla, je krokom, ktorým treba skončiť meranie dátovej kvality. Znamená to vypočítať percentuálny podiel atribútov, ktorých hodnoty plne dodržujú definované biznis pravidlá. KPI 2.2a je merané odborným posúdením a prakticky analyzuje výsledky všetkých doterajších meraní. Toto KPI 2.2a sumarizuje výsledky merania. Výsledkom posledného merania bude vyplnená šablóna nižšie:

Generický postup

Dataset	Parameter	Ukazovateľ	KPI 2.2a	Zmeraná hodnota KPI 2.2a
	Konzistentnosť	Dodržiavanie biznis pravidla	Percentuálny podiel atribútov, ktorých hodnoty plne dodržujú definované biznis pravidlá.	

Tabuľka 9: Šablóna posledného merania KPI: ukazovateľ dodržiavanie biznis pravidla

Šablóna pozostáva zo štruktúrovaných informácií:

— Dataset: **je potrebné vyplniť**
— Parameter: je už vyplnený
— Ukazovateľ: je už vyplnený
— Význam KPI 2.2a: je už vyplnený a popisuje význam KPI
— Zmeraná hodnota KPI 2.1a: **je potrebné vyplniť**

3 Zhrnutie výstupov

Sumarizácia výsledkov merania dátovej kvality pozostáva z niekoľkých hlavných výstupov. Medzi tieto výstupy patria:

- Plne vyplnená karta merania
- Ďalšie relevantné dokumenty, využité pri meraní dátovej kvality (Aktuálne dátové štandardy, Aktuálny dátový model, Aktuálne metadáta, Aktuálne biznis pravidlá), kde práve Aktuálne biznis pravidlá patria medzi najdôležitejšie a povinne dodávané
- Výstup z prvého odborného posúdenie: ukazovateľ Sledovanie konzistentnosti
- Výstup z profilingu dát
- Výstup z reportingu profilingu dát
- Výstup z pokročilej analýzy
- Výstup z finálneho odborného posúdenie: Ukazovateľ dodržiavania biznis pravidla

Každý zo spomenutých výstupov má svoj účel a význam pre zhodnotenie výsledkov merania. Štruktúra a forma týchto výstupov je rovnako dôležitá ako samotný obsah. Je to hlavne z dôvodu potreby porovnávania a komplexného vyhodnocovania výsledkov merania a sledovania trendu zlepšenia dátovej kvality. Nižšie je popísaná sumarizácia štruktúry, formy a účelu jednotlivých výstupov:

Plne vyplnená karta merania

Proces vypĺňania karty merania začína od prvého kroku merania dátovej kvality a končí posledným krokom. Je prítomná počas celého priebehu merania a podľa potreby sa priebežne aktualizuje. Obsahuje základné informácie o samotnom meraní a objekte merania. Forma karty je popísaná v kapitole 1.1.1 Karta merania dátovej kvality.

Ďalšie relevantné dokumenty, využité pri meraní dátovej kvality

Aktuálne dátové štandardy, Aktuálny dátový model a Aktuálne metadáta slúžia pre potreby tvorenia alebo aktualizáciu zoznamu aktuálnych biznis pravidiel. Zároveň pomáhajú k lepšiemu pochopenie objektu merania. Tieto dokumenty nemajú presne definovanú formu a štruktúru pre potreby merania. Toto neplatí pre Aktuálne biznis pravidlá, bez ktorých nie je možné merať dátovú kvalitu podľa definovanej metodiky. Forma biznis pravidiel je popísaná v kapitole 1.1.5 Aktuálne biznis pravidlá.

Výstup z prvého odborného posúdenie: ukazovateľ Sledovanie konzistentnosti

Tento výstup zobrazuje výsledok prvého kroku celkového merania dátovej kvality zvoleného objektu merania. Štruktúra informácií je bližšie popísaná v kapitole 2.1.2.1 Prvé odborné posúdenie: ukazovateľ Sledovanie konzistentnosti.

Výstup z profilingu dát

Výsledky z merania profilingu dát sú potrebné pre výpočet KPI v nasledujúcom reportingu profilingu dát. Kapitola 2.1.2.2 Profiling dát popisuje formu výstupu.

Výstup z reportingu profilingu dát

Tento výstup transformuje výsledky z profilingu dát do hodnôt KPI z merania dátovej kvality. Štruktúra informácií vo výstupe je bližšie popísaná v kapitole 2.1.2.3 Reporting profilingu dát.

Výstup z pokročilej analýzy

Ide o výpočty výsledkov merania dátovej kvality pre pokročilejšie analýzy. Rovnako ako vo výstupe z reportingu profilingu dát, aj tu je zobrazený výsledok merania vo forme hodnôt KPI. Štruktúra informácií vo výstupe je bližšie popísaná v kapitole 2.1.2.4 Pokročilejšia analýza.

Výstup z finálneho odborného posúdenie: Ukazovateľ dodržiavania biznis pravidla

Posledný výstup v poradí zobrazuje výsledok posledného kroku celkového merania dátovej kvality zvoleného objektu merania. Štruktúra informácií je bližšie popísaná v kapitole 2.1.2.5 Finálne odborné posúdenie: Ukazovateľ dodržiavania biznis pravidla.

Štruktúru a formu hlavných výstupov je možné prebrať aj z prílohy č. 1 Šablóny (Excel), kde sa nachádzajú potrebné šablóny.

4 Zoznamy

4.1 Zoznam tabuliek

4.2 Zoznam obrázkov

5 Prílohy

Príloha č. 1: Šablóny (Excel)