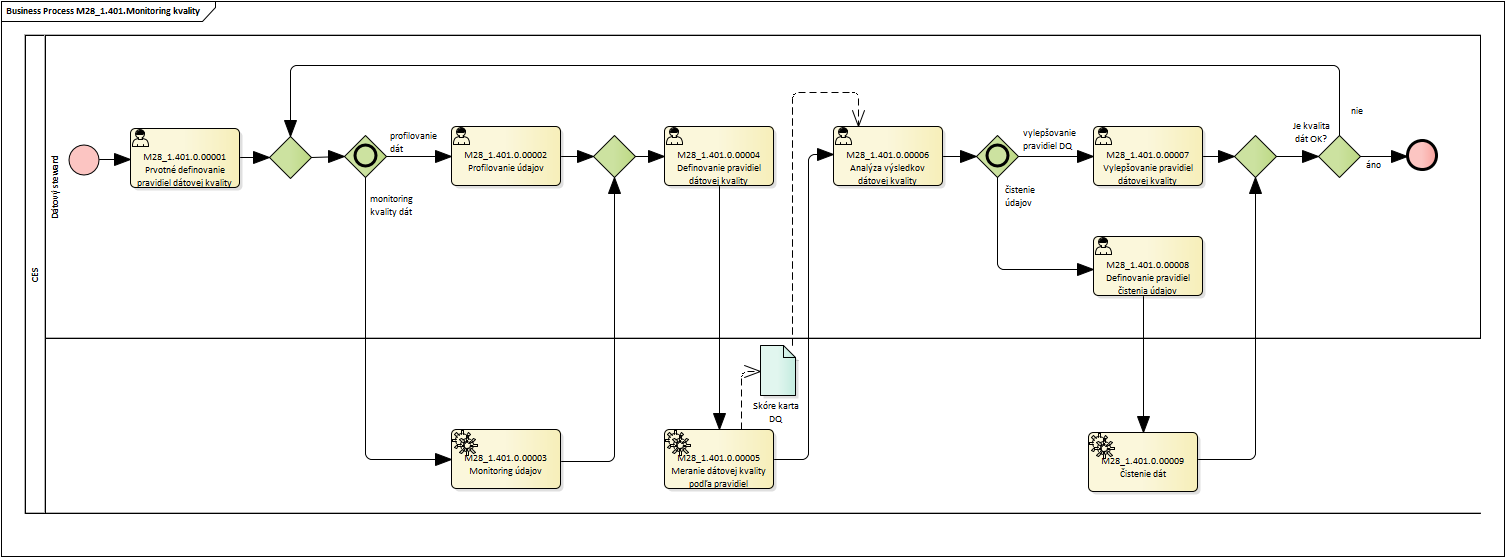
# M28\_1.401.Monitoring kvality



Obrázok 1 – Procesný model

## Popis procesu

Proces dátovej kvality je kontinuálny iteračný proces v kompetencii Správcu CES (dátový kurátor). Začína sa prvotným nadefinovaním pravidiel dátovej kvality a profilovaním údajov, ktoré poskytnú úvodnú predstavu o kvalite sledovaných údajov. Na základe týchto poznatkov je možné zadefinovať presnejšie, adresnejšie pravidlá pre kontrolu a tieto kontroly pravidelne alebo na vyžiadanie vykonávať. Výsledkom každej kontroly je skóre karta, v rámci ktorej sú zobrazované jednotlivé metriky dátovej kvality, rozdelené do jednej zo siedmych dimenzií dátovej kvality. Výsledky kontrol sa historicky evidujú pre analýzu trendov, na základe ktorých je možné opätovne adresnejšie upresniť pravidlá dátovej kvality a rovnako aj navrhnúť pravidlá pre čistenie údajov, na základe ktorých systém zrealizuje parsovanie, štandardizáciu, obohacovanie a opravy údajov.

# Roly a aktivity

## Tabuľka rolí a aktivít

|  |  |  |
| --- | --- | --- |
|  | Proces | M28\_1.301.Monitoring kvality |
|  | Org./Odd. | Dátový steward |
| **Rola** |  | **Dátový steward** |
|  | M28\_1.401.0.00001 | Prvotné definovanie pravidiel dátovej kvality |
|  | M28\_1.401.0.00002 | Profilovanie údajov |
|  | M28\_1.401.0.00004 | Definovanie pravidiel dátovej kvality |
|  | M28\_1.401.0.00006 | Analýza výsledkov dátovej kvality |
|  | M28\_1.401.0.00007 | Vylepšovanie pravidiel dátovej kvality |
|  | M28\_1.401.0.00008 | Definovanie pravidiel čistenia údajov |
| **Rola** |  | **Systém SAP** |
|  | M28\_1.401.0.00003 | Monitoring údajov |
|  | M28\_1.401.0.00005 | Meranie dátovej kvality podľa pravidiel |
|  | M28\_1.401.0.00009 | Čistenie údajov |
| **Vstup do procesu** | Zmena zamestnanca |  |
| **Výstup z procesu** | Vylepšené pravidlá dátovej kvality | a zlepšená dátová kvalita na základe čistenia dát. |

Tabuľka 1 - roly a aktivity

## Popis vykonaných aktivít

|  |  |
| --- | --- |
| **Aktivita** | **Popis** |
| **M28\_1.401.0.00001** | **Prvotné definovanie pravidiel dátovej kvality**  Na začiatku je potrebné zadefinovať a nastaviť úvodné pravidla pre kontrolu dátovej kvality. Pravidlá sa najprv zadefinujú voči parametrom a až následne sa tieto parametre mapujú voči konkrétnym atribútom vybraných tabuliek, teda jedno pravidlo môže kontrolovať viacero rôznych zdrojových údajov. |
| **M28\_1.401.0.00002** | **Profilovanie údajov**  Profilovanie poskytuje prvotný prehľad o celkovej kvalite údajov a zahŕňa v sebe:   * Profilovanie jednotlivých atribútov/stĺpcov kmeňových údajov   + prehľad dátových typov stĺpcov   + detekcia obsahu jednotlivých atribútov, napr. dátum, titul, email, adresné údaje (ulica, číslo, mesto, región)   + minimálna, maximálna (abecedne/numericky) a priemerná hodnota pre jednotlivé atribúty   + minimálna, maximálna a priemerná dĺžka reťazcov v hodnotách jednotlivých atribútov   + vyplnenosť jednotlivých atribútov – percentuálne zastúpenie výskytov hodnôt NULL, blank, medzera a 0   + počty jedinečných výskytov (distinct) hodnôt a ich distribúcia * Profilovanie závislostí   + identifikovanie závislostí v údajoch prehľadávaním hodnôt v jednom alebo viacerých stĺpcoch, ktoré sú závislé na hodnote v primárnom stĺpci. * Profilovanie adries   + určovanie kvality adries ich stotožňovaním voči registru adries * Profilovanie redundancií a jedinečnosti   + určenie miery prekrývania hodnôt alebo ich duplicít medzi viacerými stĺpcami (redundacia) a v rámci jedného stĺpca (jedinečnosť) |
| **M28\_1.401.0.00003** | **Monitoring údajov**  Nastavené pravidla pre kontrolu kvality dát sú pravidelne alebo na vyžiadanie spúšťané, čím sa zabezpečuje pravidelný monitoring kvality údajov. |
| **M28\_1.401.0.00004** | **Definovanie pravidiel dátovej kvality**  Pravidlá sa najprv zadefinujú voči parametrom a až následne sa tieto parametre mapujú voči konkrétnym atribútom vybraných tabuliek, teda jedno pravidlo môže kontrolovať viacero rôznych zdrojových údajov.  Každé pravidlo je potrebné priradiť do niektorej zo 7 dimenzií dátovej kvality:   * **kompletnosť** – pravidlo zachytáva záznamy s chýbajúcimi alebo nekompletnými údajmi. Aj údaje s prázdnymi nepovinnými poľami môžu byť kompletné, ak spĺňajú zadefinované pravidlá (napr. musí byť vyplnené jedno z dvoch nepovinných polí) * **presnosť** – pravidlo určuje mieru, do akej údaje reflektujú reálny svet alebo popisovanú udalosť (napr. dátum narodenia musí byť v minulosti) * konzistentnosť – pravidlá určujú, do akej miery korešpondujú rovnaké údaje naprieč viacerými systémami/oblasťami a či sú tieto údaje synchronizované (napr. status je uzavretý a nemá aktívne položky) * **jedinečnosť** – pravidlá na detekciu duplicitných záznamov (napr. rovnaký názov, IČO) * **integrita** – pravidlá detegujú dodržanie potrebných prepojení v údajoch a zaručujú, že údaje je možné sledovať a prepojiť s inými údajmi * **časová dostupnosť** – pravidlo definuje, či sú požadované informácie dostupné v požadovanom alebo potrebnom čase * **konformnosť** – pravidlo sleduje, či údaje dodržiavajú definované typy, veľkosti, formáty alebo masky |
| **M28\_1.401.0.00005** | **Meranie dátovej kvality podľa pravidiel**  Vytvorené pravidlá je taktiež možné použiť v skóre kartách, každá pre určitú kľúčovú oblasť údajov (kmeňový záznam materiálu, obchodný partner, ... ). Do jednotlivých oblastí skóre karty (dimenzií dátovej kvality) je možné pridávať jednotlivé pravidlá a ováhovať ich v prípade viacerých pravidiel s rovnakou dimenziou. |
| **M28\_1.401.0.00006** | **Analýza výsledkov dátovej kvality**  Po vygenerovaní skóre karty je dochádza k jej analyzovaniu a vyhodnocovaniu voči zadefinovaným pravidlám. |
| **M28\_1.401.0.00007** | **Vylepšovanie pravidiel dátovej kvality**  Na základe analýzy výsledkov dátovej kvality je možné následne upraviť existujúceho pravidlá, alebo vytvoriť nové, adresnejšie pravidlá za účelom zlepšenia dátovej kvality. |
| **M28\_1.401.0.00008** | **Definovanie pravidiel čistenia údajov**  Systém bude poskytovať aj nástroje pre čistenie údajov. V nich je možné definovať pravidlá pre čistenie, ktoré možné pravidelne aplikovať na dáta. Medzi pravidlá pre čistenie dát patria:   * **parsovanie** – výber rôznych častí reťazcov z textu podľa definovaných pravidiel * **štandardizácia** – zjednotenie názov a skratiek v texte podľa používateľom nadefinovaných slovníkov s pravidlami napr.   Ulica = ulica, ul, ul., Ul, Ul., street, str., str, utca, ...   * **obohacovanie** – doplnenie textov podľa pravidiel * **opravy údajov** – detekcia a oprava gramatických chýb a preklepov * **identifikácia duplicít.** |
| **M28\_1.401.0.00009** | **Čistenie údajov**  Samotné čistenie údajov je realizované systémom na základe používateľom zadefinovaných pravidiel pre čistenie. |

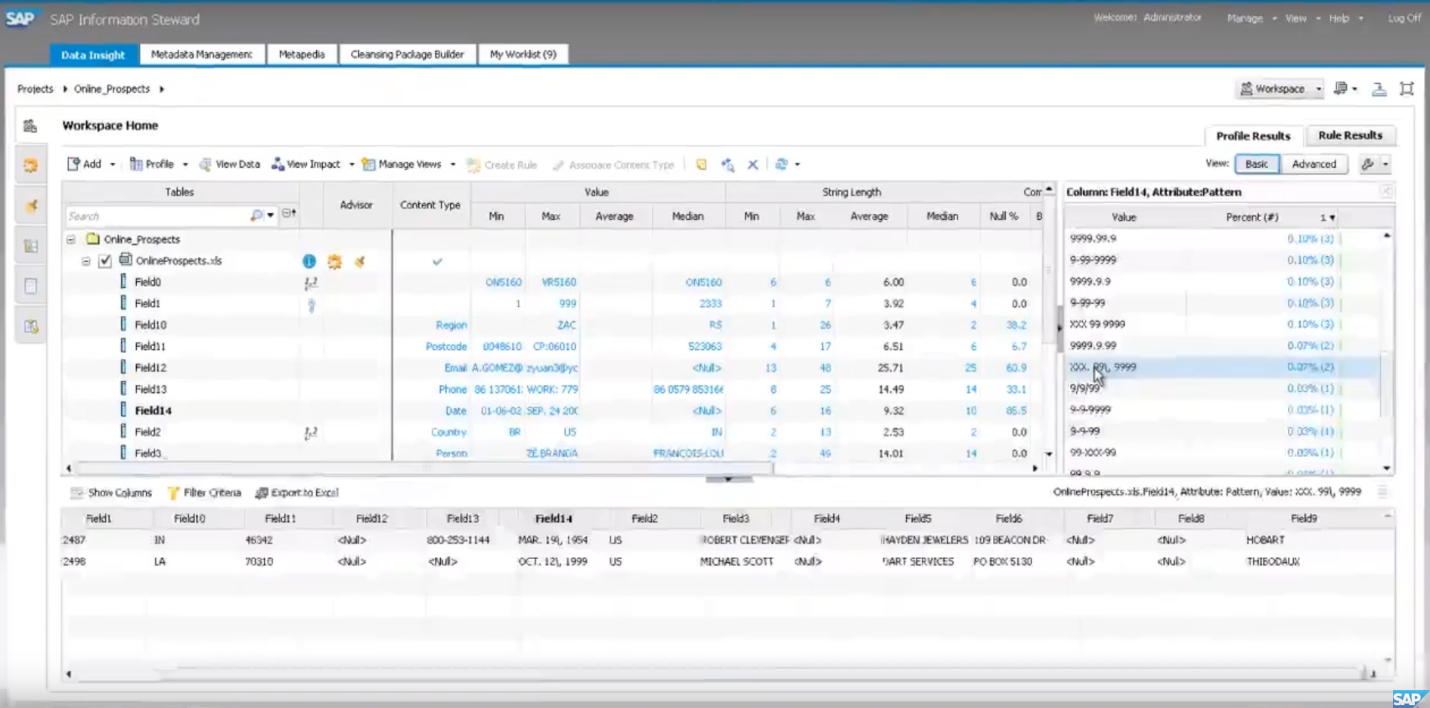
Tabuľka 2 Zoznam aktivít

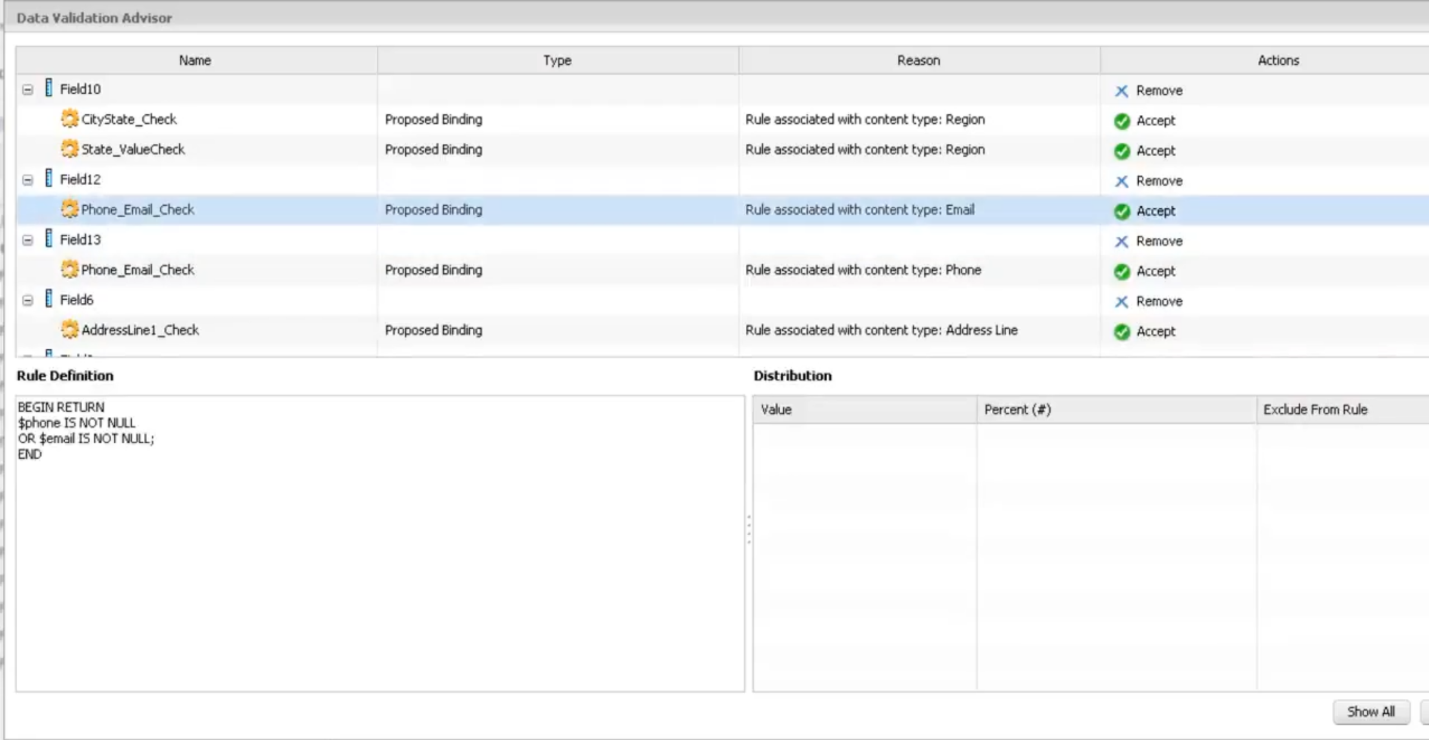
# Business workflow

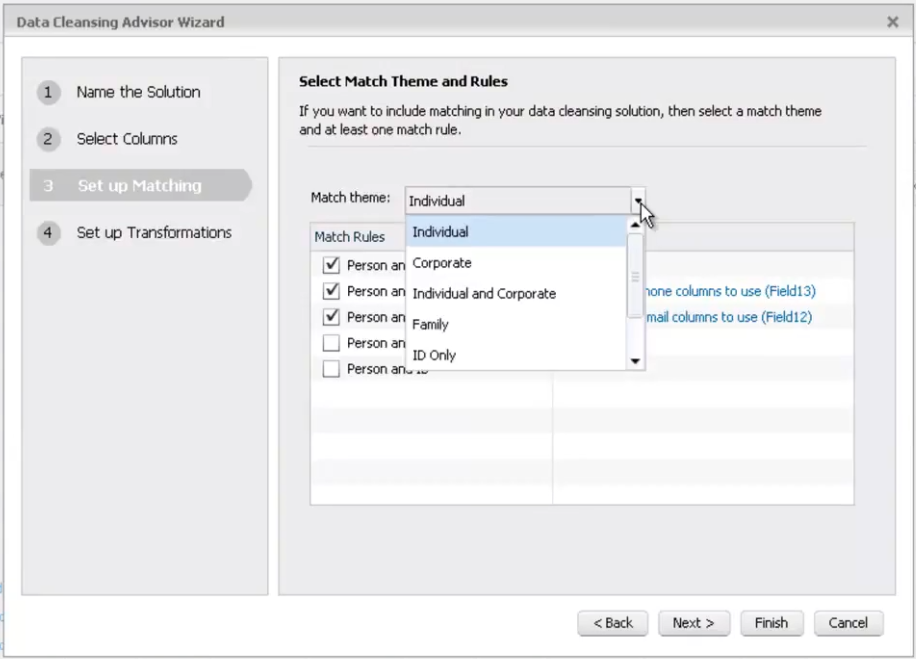
Keďže proces merania dátovej kvality je iteračný, kontinuálny a v kompetencii len jednej roly (Dátový steward), nie je potrebné podporiť ho workflowom. Nižšie sú znázornené obrazovky, pokrývajúce dostupnú funkcionalitu pre monitoring dátovej kvality.

Obrázok, na ktorom je text, snímka obrazovky, vnútri

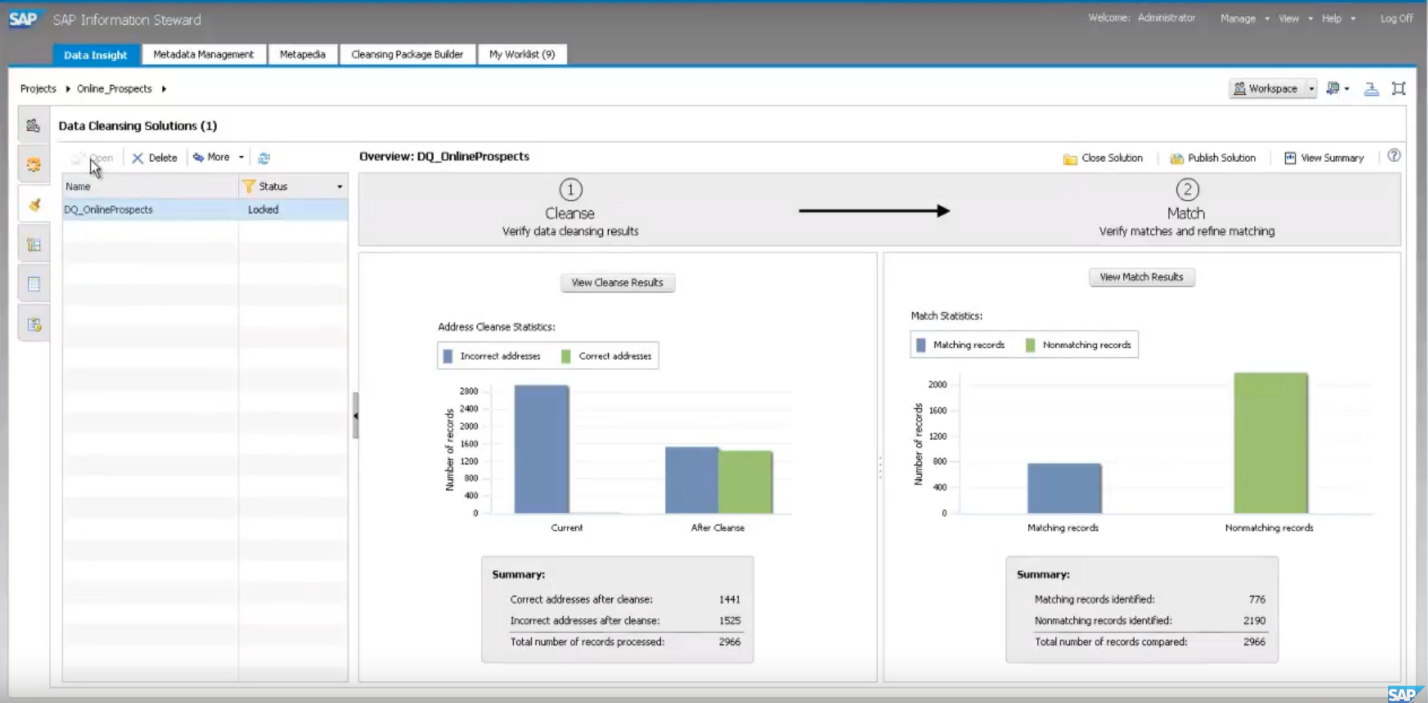
Automaticky generovaný popis*Obrázok 2 – Profilovanie údajov*

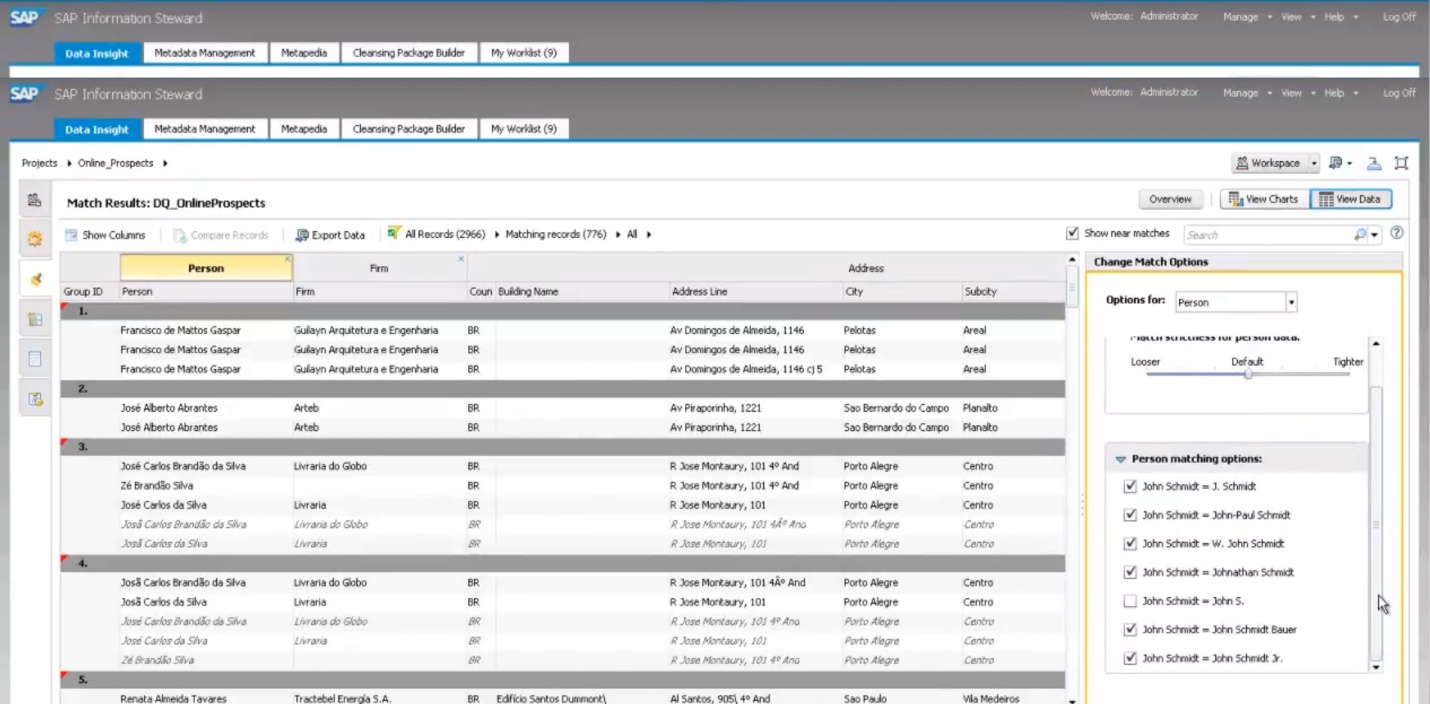
Obrázok 3 – Profilovanie údajov - zobrazenie výskytov formátov dátumov (vpravo) s konkrétnymi záznamami (dolu)

Obrázok 4 – Validačné pravidlá



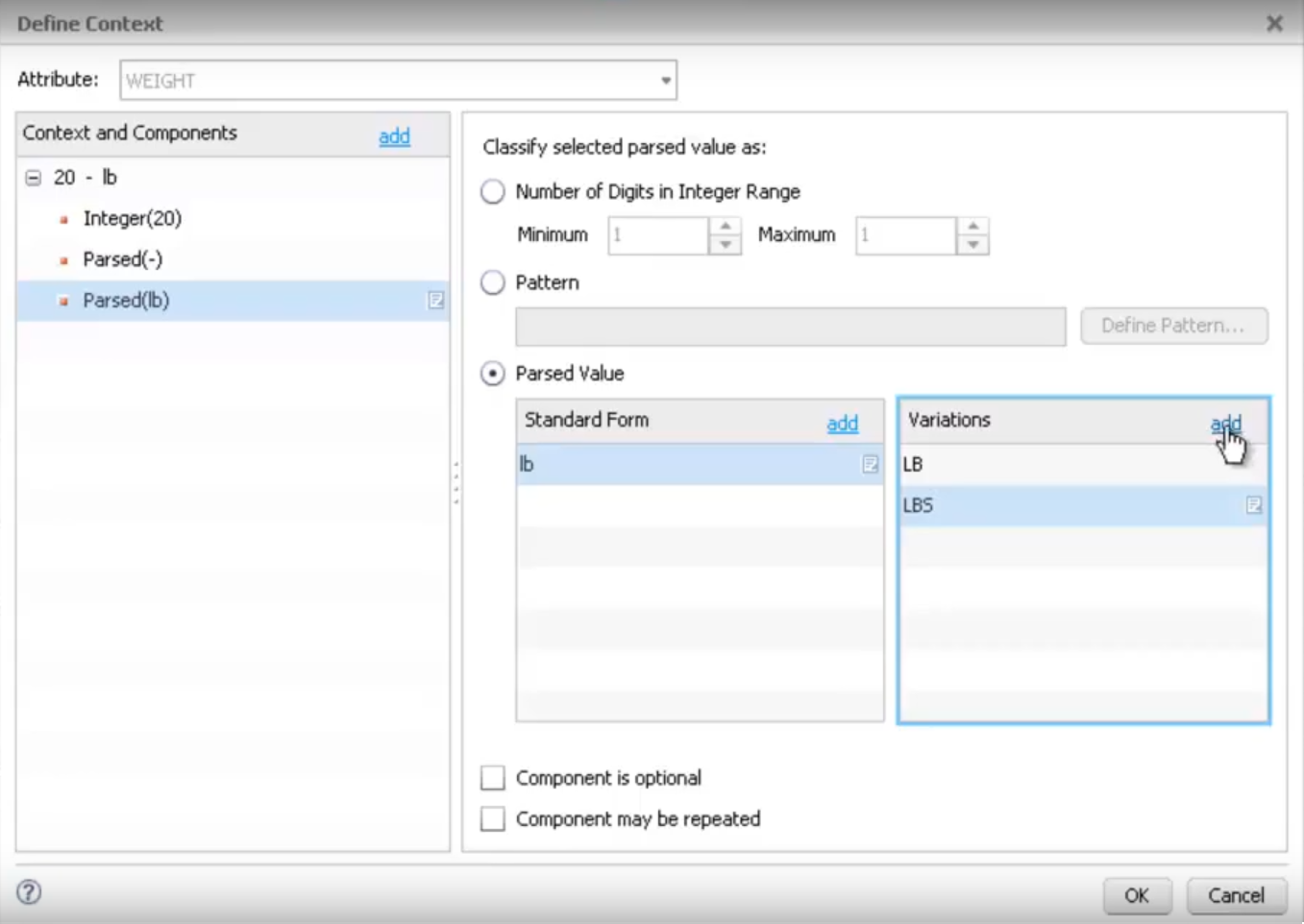
Obrázok 5 – Čistenie údajov – tvorba pravidla s návrhom scenárov pre stotožňovanie/deduplikáciu

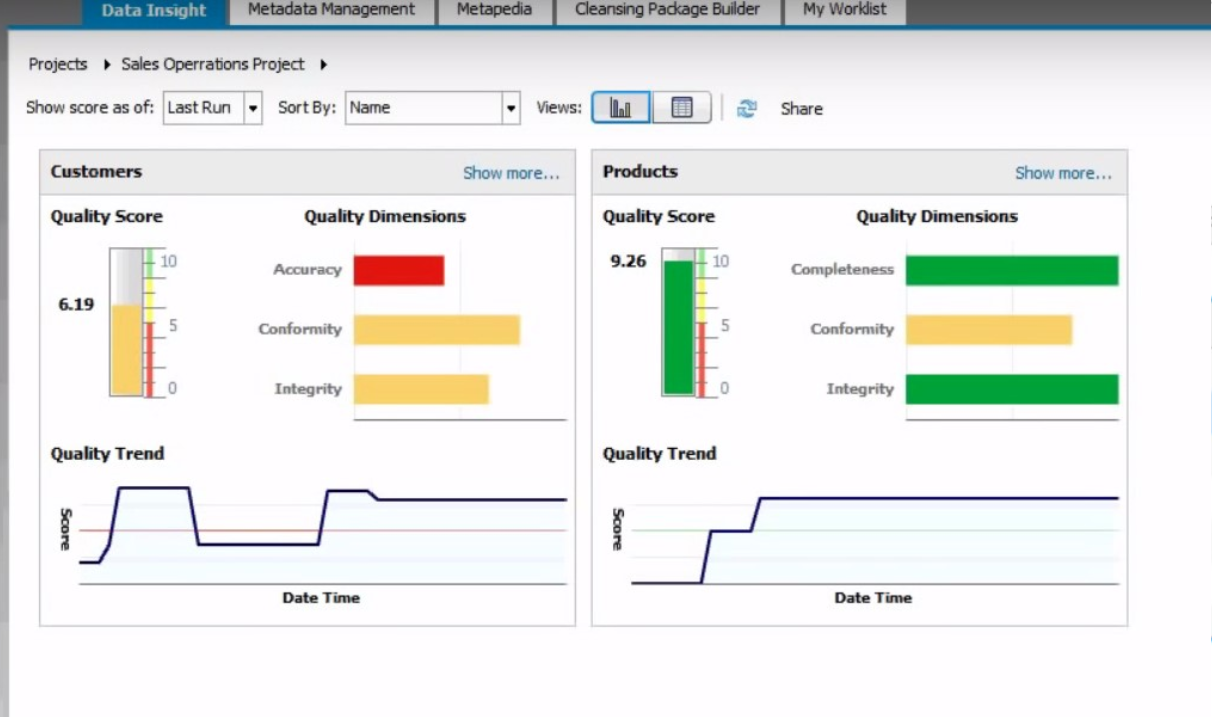
Obrázok 6 – Čistenie (1) a deduplikácia (2) údajov

Obrázok 7 – Stotožňovanie osôb

Obrázok, na ktorom je stôl

Automaticky generovaný popisObrázok 8 – Výber zlatého záznamu

Obrázok 9 – Parsovanie a štandardizácia materiálu

Obrázok 10 – Skóre karta dátovej kvality