*Prepoznajmo "**loše**" podatke u ranoj fazi integracije*

Darko Benšić, dbensic@croz.net

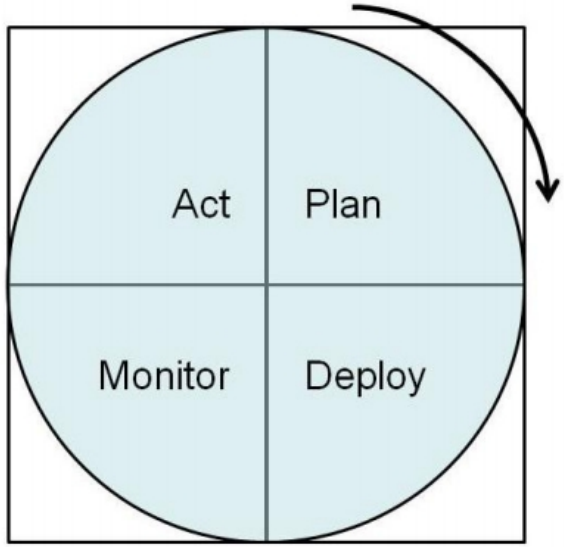
HrOUG 2009, *Rovinj, 13. do 17. listopada 2009.*

Oracle Data Profiling and Quality for Data Integrator

Agenda

- Kvaliteta podataka
- Profiliranje podataka
- Odnos DI – DP - DQ
- Oracle Data Profiling
- DEMO

Kvaliteta podataka - ???



- Kvaliteta podataka -> Proces
- Ciljevi: -> mjerenje vrijednosti
- Data Quality Management (DQM)

- ***Plan ->*** Identifikacija ključnih parametara, procjena trenutnog stanja
- ***Deploy ->*** Prcedure – mjerenje kvalitete
- ***Monitor ->*** Kontinuirano mjerenje podataka (detekcija loših podataka)
- ***Act: ->*** Ispravak anomalija u podacima i poslovnim pravilima

Kvaliteta podataka - aktivnost

Data Quality Management - ***aktivnosti:***

1. Definiranje zahtjeva za kvalitetom podataka
2. Profiliranje i anliziranje kvalitete podataka
3. Definiranje metrike -> mjerenje kvalitete podataka
4. Definiranje poslovnih pravila
5. Testiranje i provjera kvalitete podataka
6. Postavljenje servisa za upravljenje kvalitetom podataka
7. Kontinuirano mjerenje i nadzor kvalitete
8. Upravljenje iznimkama -> poboljšavanje kvalitete
9. Čišćenje i ukljanjenje loših podataka i/ili poslovnih pravila
10. Dizajn i implementacije procedura za mjerenje i nadzor

Profiliranje podataka - ???

> “ ***The good news is we got the data loaded. The bad news is we got the data loaded.”*** *– Jack Olsen*

“Data profiling is the process of examining the data available in an existing data source (e.g. a database or a file) and collecting statistics and information about that data” – Wikipedia

Profiliranje podataka - ???

Profiliranje podataka

-> Proces

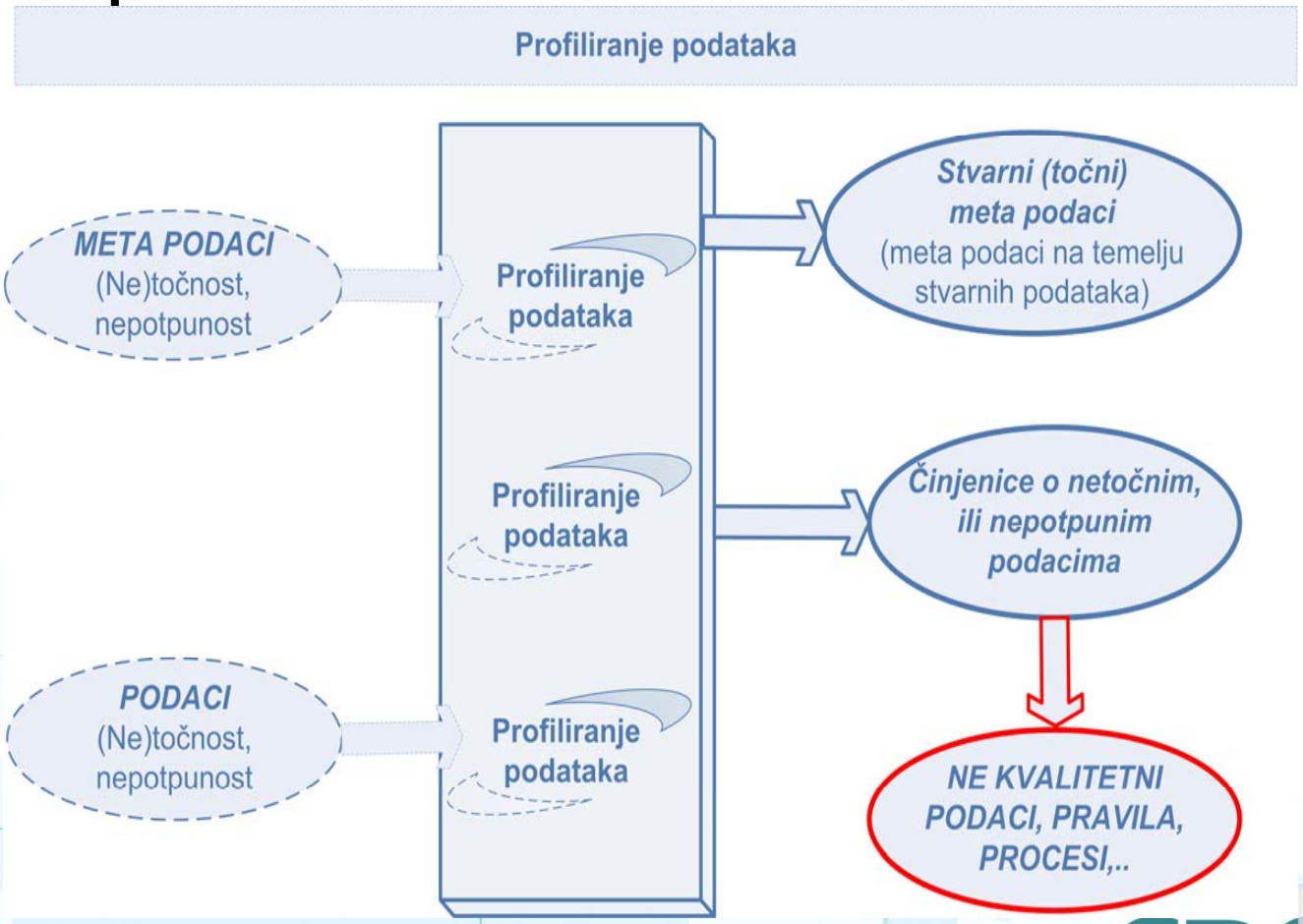
-> Metodologija

-> ***Alat (analiza, nadzor)***

"**Profiliranje podataka** = upoznavanje na praktičan način sa sadržajem i kvalitetom podataka, uspoređujući meta podatke sa stvarnim sadržajem u bazi"

Zadatak	Ručni proces	Automatizirani proces
Upit nad kolonama	1 h do 5 dana	1 min do 4 sata
Odgovor na pitanje	0,5 h do 5 dana po tablici	0,5 h do 2 dana po tablici

Profiliranje podataka - ???



Profiliranje podataka - primjena

Profiliranjem podataka dobivamo ***stvarne*** odgovore na pitanja kao što su:

1. Što se nalazi u određenoj tablici?
2. Koriste li se sve kolone i u kojem postotku?
3. Jesu li kolone adekvatno popunjene?
4. Koje kolone možemo koristiti u DWH?
5. Što točno možemo konvertirati u novi sustav?
6. Koje sve kolone moraju biti transformirane?

Broj telefona:

- (01) 1234-456
- +385 (01) 987-2323
- +385 1 9872 – 323 ...
- 112-233

Profiliranje podataka - proces

Korak 1: Identifikacija procesa, tehnologije i sudionika procesa

Korak 2: Identifikacija podataka koje želimo identificirati

Korak 3: Analiza kolona

Korak 4: Analiza strukture identificiranih tablica

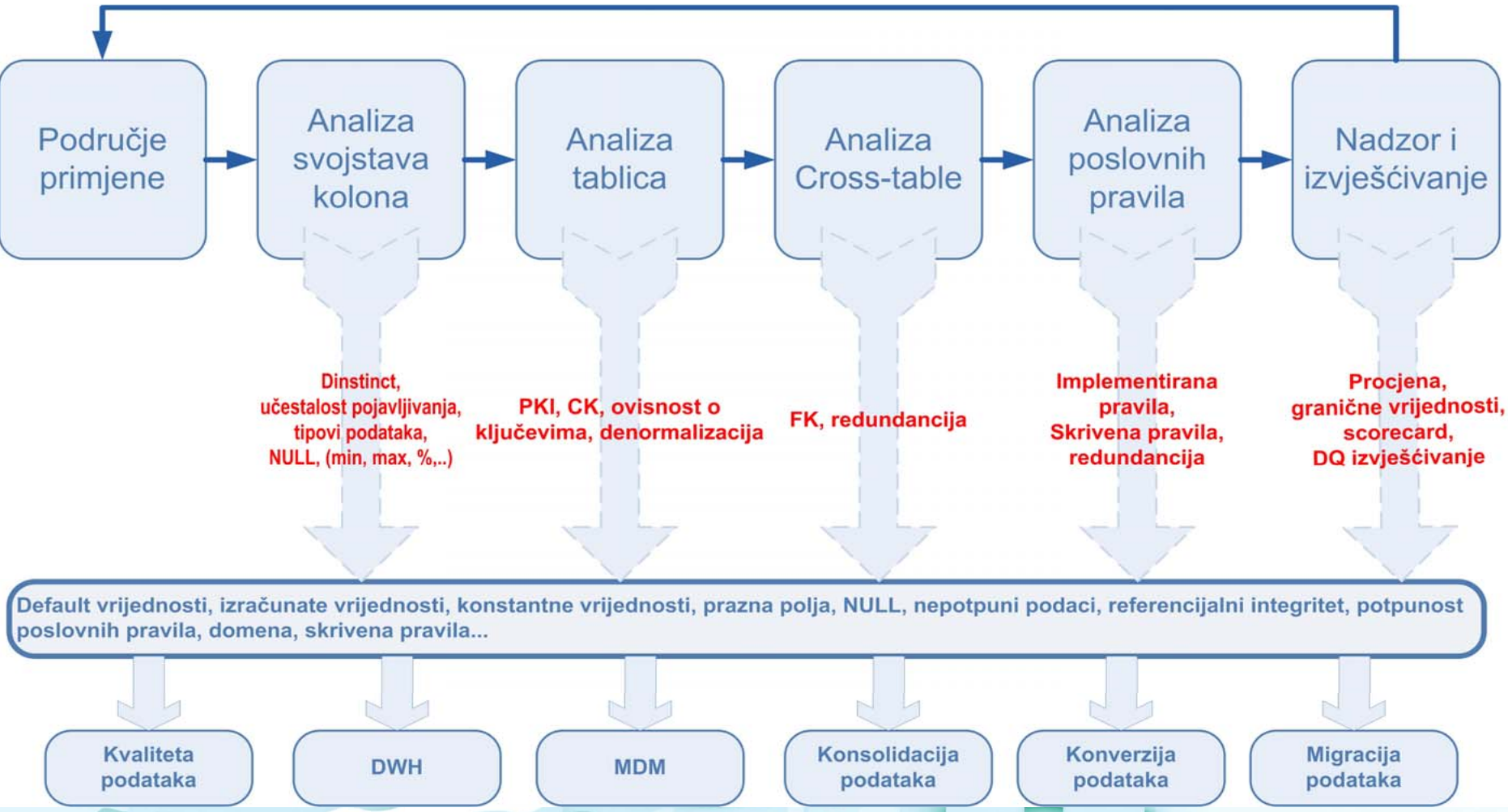
Korak 5: Cross table analiza

Korak 6: Definiranje poslovnih pravila

Korak 7: Automatizacija (nadzor)

Profiliranje podataka - proces

Profiliranje podataka - Analiza i procjena (statistika) o kvaliteti postojećih podataka, istraživanje potencijalnih ili implicitnih relacija između promatranih vrijednosti.



Profiliranje podataka – Analiza svojstava kolona

Analizom svojstava kolona (eng. Column property analysis)

-> Proces

-> Atomizirane vrijednosti

-> (**ne**) Ispravno

Meta Data:

•Kolona: ID_OSOBA

•Range: 100-99999

Stvarni zapis:

•ID_OSOBA = ”10”

Postupak (proces) prikazuje nam:

- Kolona (Tip podatka, preciznost, (NOT) NULL,)
- Domena (Date: 13.02.1971, ZIP kod)
- Svojstva (Character set, jedinstvenost, restrikcije, CC)
- Naziv kolone (jedinstvenost, ...)
- Preopterećenje polja (legacy sustavi)

Profiliranje podataka – Analiza struktura

Analiza struktura

-> Proces

-> Ovisnost između podatka i meta-podatka

-> (**ne**) Ispravno – izvješća/statistike

Analiza tablice - uključuje promatranje tablice kao zasebnog i neovisnog objekta što uključuje promatranje primarnih ključeva, derived kolona i sl.

Cross table analiza – promatra ovisnost podataka između više tablica što uključuje foregain key, normalizacijske forme (1NF, 2NF i 3NF), denormalizaciju, sinonime primarnih/stranih ključeva, sinonimi redudantnih podataka, te domenski sinonimi.

Profiliranje podataka – Poslovna analiza

Poslovna pravila:

-> Data Rule:
*"KATEGORIJA_DOZVOLE = 'B'
tada TRENUTNI_DATUM – DATUM_RODJENJA > 18 godina"*

-> Process Rule:
"Samo osobi koja je navršila 18 godina institucija smije izdati vozačku dozvolu 'B' kategorije".

Analiza poslovnih vrijednosti (eng. Value Rule Analysis) ???

- **Kardinalnost** -> dupli zapisi
- **Distribucija -** > jedinstveni zapisi
- **Agregacije -** > SUM(), count(*), AVG(), ...

Profiliranje podataka – Nadzor

Vrijeme !!!

-> Promjena strukture (meta podaci)
-> Promjena podataka (CRUD)
-> Promjena poslovnih pravila
-> Integracija podataka
-> Migracija podataka

Krivi podatak!!! – širi se kao virus

-> **Transakcijski sustav**

-> DWH

-> Data Mart

Odnos DI – DQ - DP

Sinergija

-> Integracija podataka (DI)

-> Kvaliteta podataka (DQ)

- Ovisi o DI

-> Profiliranje podataka (DP)

- Planiranje DI i DQ projekata
- Kontionuirano poboljšavanje kavlitete podataka

Rješenje:

-> Kordinacija DI – DQ – DP

Odnos DI – DQ - DP

DI, DQ i DP su :

-> Iterativni procesi (dnevno)

-> ciklusi (dizajn i revizije)

Profiliranje podataka

-> Razvoj:

- Otkrivanje novih (nepoznatih) podatka
- Profiliranje podataka kao uvod u DQ i DI

-> -Implementacija:

- Nadzor podataka (dokaz poboljšanja kvalitete)
- Nadzor promjena u strukturi

Odnos DI – DQ - DP

DI, DQ i DP su :

-> Iterativni procesi (dnevno)

-> ciklusi (dizajn i revizije)

Integracija podataka

-> Razvoj:

- Novi dizajn
- Ažuriranje postojećih integracijskih riješenja

-> -Implementacija:

- Pouzdano dnevno izvršavnje procesa

Odnos DI – DQ - DP

DI, DQ i DP su :

-> Iterativni procesi (dnevno)

-> ciklusi (dizajn i revizije)

Kvaliteta podataka

-> Razvoj:

- Novi dizajn
- Ažuriranje postojećih DQ rješenja

-> -Implementacija:

- Pouzdano dnevno izvršavnje DQ procesa

Oracle Data Profiling and Quality

Oracle Data Profiling and Quality for Data Integrator

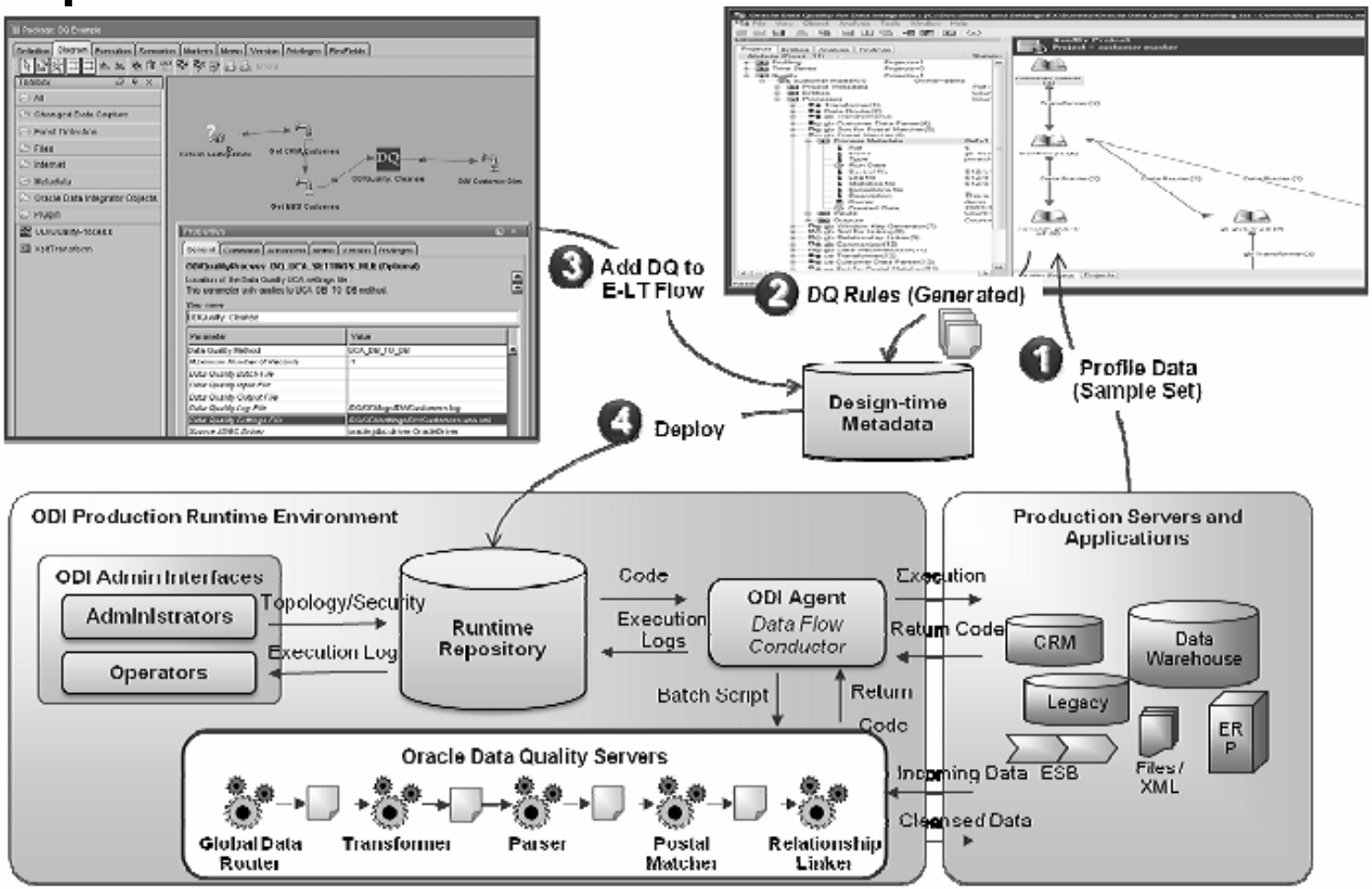
-> Dva produkta

- Oracle Data Profiling
- Oracle Data Quality

-> Tri neovisne logičke cjeline:

- Proces profiliranja (Oracle Data Profiling)
- Proces monitoriranja (Time Series)
- Proces čišćenja (Oracle Data Quality)

Oracle Data Profiling and Quality - arhitektura



Oracle Data Profiling and Quality

-> ***Repozitorij***

- Jedna instalacija -> jedan repozitorij
- jedan repozitorij -> Više Metabase-a
- Konfiguracija
- Sigurnost (korisničke grupe, korisnici)

-> ***Metabase*** (Čuva podatke i meta podatke)

- struktura i postojeće relacije
- Standardi, poslovna pravila
- Statistike, drill-down detalji, domene
- Promjena i trend promjene podatka
- Rezultati procesuiranja kvalitete podataka
- Dokumenti, standardi, izvješća i dr.

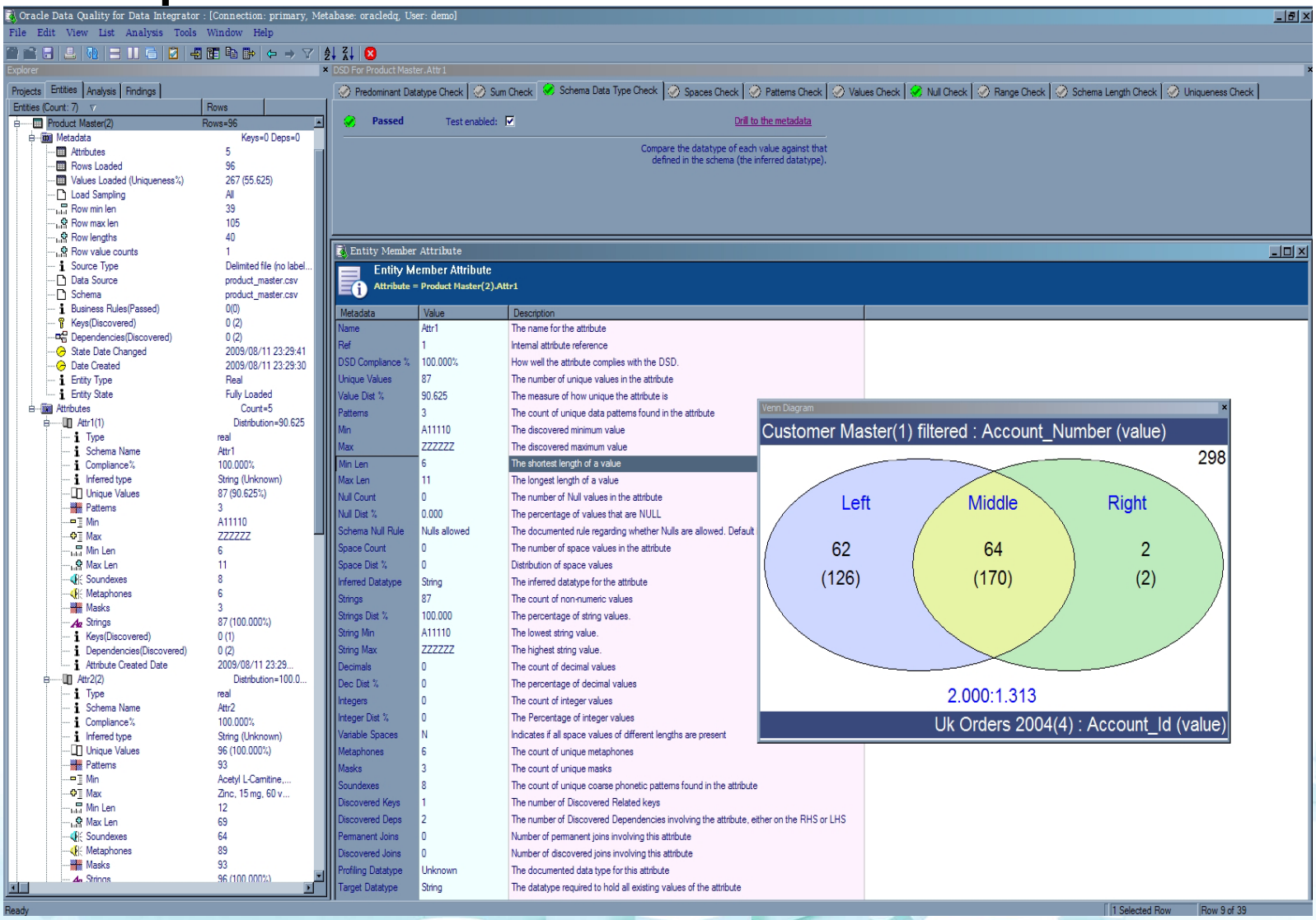
Oracle Data Profiling and Quality

Osnovna funkcionalnost:

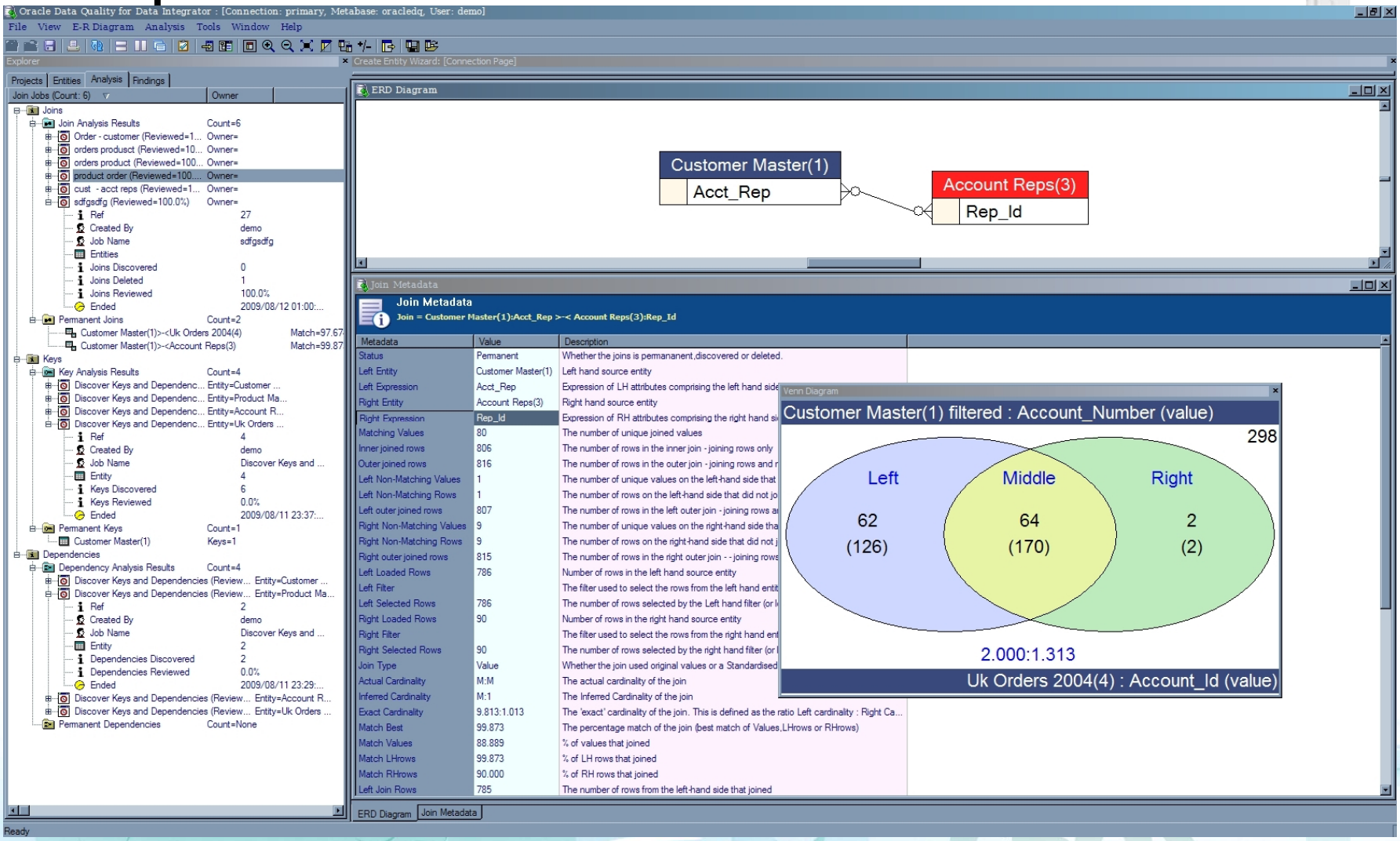
1. *Istraživanje podataka* – obuhvaća gotovo sva potrebna svojstva u procesu profiliranja koja smo objasnili u ovom radu.
2. *Istraživanje ralacija* unutar entiteta - što uključuje istraživanje postojećih i pronađenih alternativnih ključeva, pronalaženje ovisnosti, te istraživanje relacija između promatranih entiteta.
3. *Provjera standarda* - provjerava naznačene standarde koje smo prethodno definirali.
4. *Postavljanje poslovnih pravila* - Oracle Data Profiling alata omogućuje dodavanje poslovnih pravila kao što su:

"***Samo osobi koja je navršila 18 godina institucija smije izdati vozačku dozvolu 'B' kategorije***"

Oracle Data Profiling and Quality



Oracle Data Profiling and Quality



Oracle Data Profiling

H V A L A !