

Prepoznamo "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

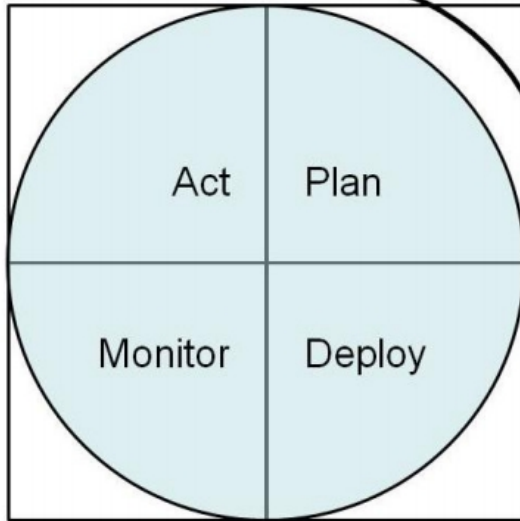
CROZ

Agenda

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

CROZ

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 analiziranje kvalitete podataka
3. Definiranje metrike -> mjerenje kvalitete podataka
4. Definiranje poslovnih pravila
5. Testiranje i provjera kvalitete podataka
6. Postavljanje servisa za upravljanje kvalitetom podataka
7. Kontinuirano mjerenje i nadzor kvalitete
8. Upravljanje iznimkama -> poboljšavanje kvalitete
9. Čišćenje i ukljanjenje loših podataka i/ili poslovnih pravila
10. Dizajn i implementacije procedura za mjerenje i nadzor

CROZ

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



CROZ

Profiliranje podataka - ???

Profiliranje podataka

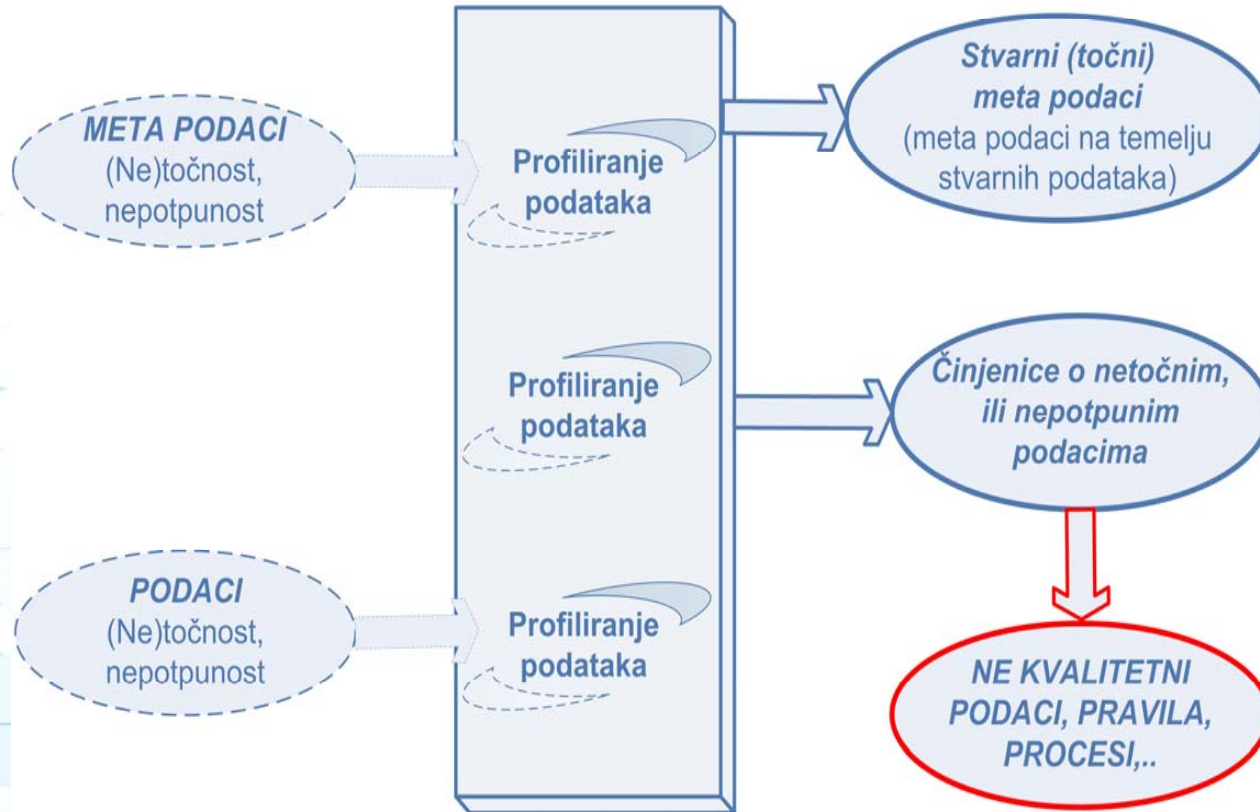
- > Proces
- > Metodologija
- > **Alat (analiza, nadzor)**

“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



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



CROZ

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

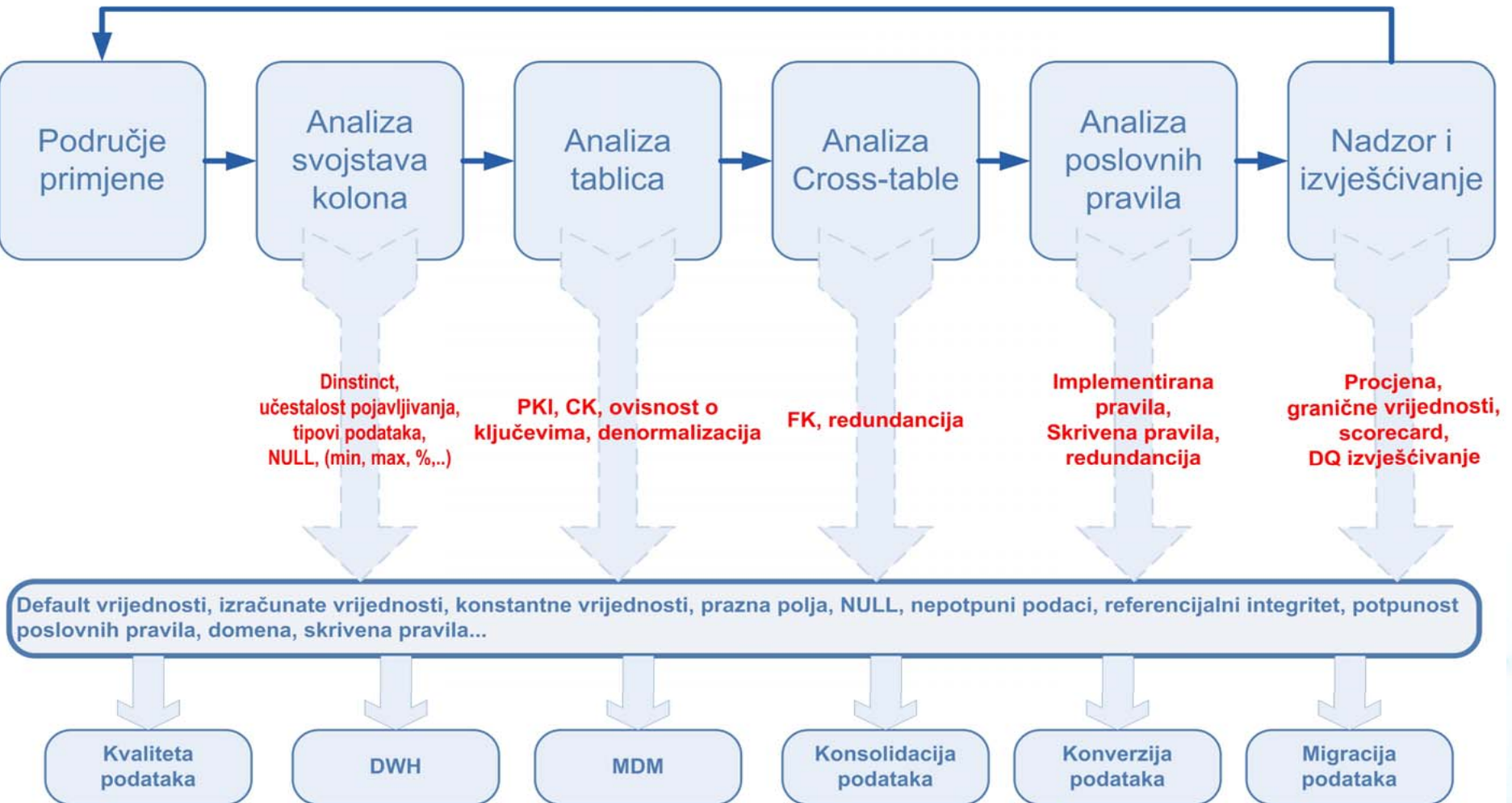
Korak 7: Automatizacija (nadzor)

CROZ

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 foreign key, normalizacijske forme (1NF, 2NF i 3NF), denormalizaciju, sinonime primarnih/stranih ključeva, sinonimi redundantnih podataka, te domenski sinonimi.

Profiliranje podataka – Poslovna analiza

Poslovna pravila:

-> Data Rule:

"KATEGORIJA DOZVOLE = 'B'
tada $TRENTUTNI_DATUM - DATUM_RODJENJA > 18$
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(), ...

CROZ

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

CROZ

Odnos DI – DQ - DP

Sinergija

-> Integracija podataka (DI)

-> Kvaliteta podataka (DQ)

- Ovisi o DI

-> Profiliranje podataka (DP)

- Planiranje DI i DQ projekata
- Kontinuirano poboljšavanje kvalitete 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:
 - o Otkrivanje novih (nepoznatih) podataka
 - o Profiliranje podataka kao uvod u DQ i DI
- > -Implementacija:
 - o Nadzor podataka (dokaz poboljšanja kvalitete)
 - o Nadzor promjena u strukturi

Odnos DI – DQ - DP

DI, DQ i DP su :

- > Iterativni procesi (dnevno)
- > ciklusi (dizajn i revizije)

Integracija podataka

- > Razvoj:
 - o Novi dizajn
 - o Ažuriranje postojećih integracijskih rješenja
- > -Implementacija:
 - o Pouzdano dnevno izvršavanje procesa

Odnos DI – DQ - DP

DI, DQ i DP su :

- > Iterativni procesi (dnevno)
- > ciklusi (dizajn i revizije)

Kvaliteta podataka

- > Razvoj:
 - o Novi dizajn
 - o Ažuriranje postojećih DQ rješenja
- > -Implementacija:
 - o Pouzdano dnevno izvršavanje DQ procesa

Oracle Data Profiling and Quality

Oracle Data Profiling and Quality for Data Integrator

-> Dva produkta

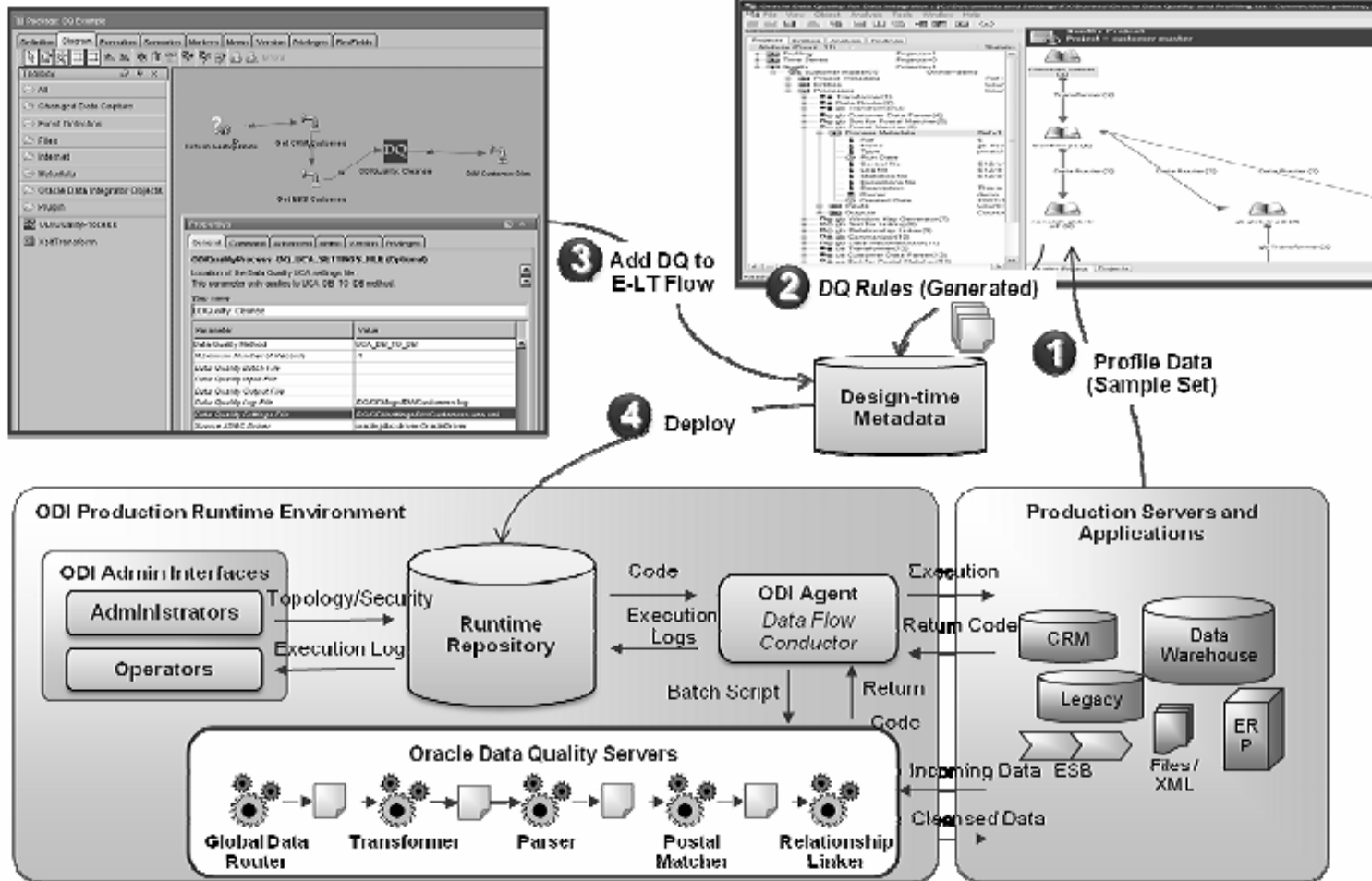
- Oracle Data Profiling
- Oracle Data Quality

-> Tri neovisne logičke cjeline:

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

CROZ

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.

CROZ

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 Quality for Data Integrator : [Connection: primary, Metabase: oracleled, User: demo]

File Edit View List Analysis Tools Window Help

Explorer x DSD For Product Master.Atr1

Projects Entities Analysis Findings Rows

Entities Count: 7 Product Master(2) Rows=96

Metadata Keys=0 Deps=0

- Attributes 5
- Rows Loaded 96
- Values Loaded (Uniqueness%) 267 (55.625)
- Load Sampling All
- Row min len 39
- Row max len 105
- Row lengths 40
- Row value counts 1
- Source Type Delimited file (no label...)
- Data Source product_master.csv
- Schema product_master.csv
- Business Rules(Passed) 0(0)
- Keys(Discovered) 0 (2)
- Dependencies(Discovered) 0 (2)
- State Date Changed 2009/08/11 23:29:41
- Date Created 2009/08/11 23:29:30
- Entity Type Real
- Entity State Fully Loaded
- Attributes Count=5
- Attr1(1) Distribution=90.625
- Type real
- Schema Name Atr1
- Compliance% 100.000%
- Inferred type String (Unknown)
- Unique Values 87 (90.625%)
- Patterns 3
- Min A11110
- Max ZZZZZZ
- Min Len 6
- Max Len 11
- Soundexes 8
- Metaphones 6
- Masks 3
- Strings 87 (100.000%)
- Keys(Discovered) 0 (1)
- Dependencies(Discovered) 0 (2)
- Attribute Created Date 2009/08/11 23:29:...
- Attr2(2) Distribution=100.0...
- Type real
- Schema Name Atr2
- Compliance% 100.000%
- Inferred type String (Unknown)
- Unique Values 96 (100.000%)
- Patterns 93
- Min Acetyl L-Carnitine,...
- Max Zinc, 15 mg, 60 v...
- Min Len 12
- Max Len 69
- Soundexes 64
- Metaphones 89
- Masks 93
- Strings 96 (100.000%)

Predominant Datatype Check Sum Check Schema Data Type Check Spaces Check Patterns Check Values Check Null Check Range Check Schema Length Check Uniqueness Check

Passed Test enabled: [Drill to the metadata](#)

Compare the datatype of each value against that defined in the schema (the inferred datatype).

Entity Member Attribute

Entity Member Attribute
Attribute = Product Master(2).Atr1

Metadata	Value	Description
Name	Atr1	The name for the attribute
Ref	1	Internal attribute reference
DSD Compliance %	100.000%	How well the attribute complies with the DSD.
Unique Values	87	The number of unique values in the attribute
Value Dist %	90.625	The measure of how unique the attribute is
Patterns	3	The count of unique data patterns found in the attribute
Min	A11110	The discovered minimum value
Max	ZZZZZZ	The discovered maximum value
Min Len	6	The shortest length of a value
Max Len	11	The longest length of a value
Null Count	0	The number of Null values in the attribute
Null Dist %	0.000	The percentage of values that are NULL
Schema Null Rule	Nulls allowed	The documented rule regarding whether Nulls are allowed. Default
Space Count	0	The number of space values in the attribute
Space Dist %	0	Distribution of space values
Inferred Datatype	String	The inferred datatype for the attribute
Strings	87	The count of non-numeric values
Strings Dist %	100.000	The percentage of string values.
String Min	A11110	The lowest string value.
String Max	ZZZZZZ	The highest string value.
Decimals	0	The count of decimal values
Dec Dist %	0	The percentage of decimal values
Integers	0	The count of integer values
Integer Dist %	0	The Percentage of integer values
Variable Spaces	N	Indicates if all space values of different lengths are present
Metaphones	6	The count of unique metaphones
Masks	3	The count of unique masks
Soundexes	8	The count of unique coarse phonetic patterns found in the attribute
Discovered Keys	1	The number of Discovered Related keys
Discovered Deps	2	The number of Discovered Dependencies involving the attribute, either on the RHS or LHS
Permanent Joins	0	Number of permanent joins involving this attribute
Discovered Joins	0	Number of discovered joins involving this attribute
Profiling Datatype	Unknown	The documented data type for this attribute
Target Datatype	String	The datatype required to hold all existing values of the attribute

Venn Diagram

Customer Master(1) filtered : Account_Number (value) 298

Left: 62 (126) Middle: 64 (170) Right: 2 (2)

2.000:1.313

Uk Orders 2004(4) : Account_Id (value)

Ready | 1 Selected Row | Row 9 of 39

Oracle Data Profiling and Quality

Oracle Data Quality for Data Integrator : [Connection: primary, Metabase: oracledq, User: demo]

File View E-R Diagram Analysis Tools Window Help

Explorer Create Entity Wizard: [Connection Page]

Projects Entities Analysis Findings

Join Jobs (Count: 6)

Join Analysis Results Count=6

- Order - customer (Reviewed=1... Owner=)
- orders product (Reviewed=10... Owner=)
- orders product (Reviewed=100... Owner=)
- product order (Reviewed=100... Owner=)
- cust - acct reps (Reviewed=1... Owner=)
- sdfgsdfg (Reviewed=100.0%) Owner=

Permanent Joins Count=2

- Customer Master(1)-<Uk Orders 2004(4) Match=97.67
- Customer Master(1)-<Account Reps(3) Match=99.87

Keys Count=4

- Discover Keys and Dependenc... Entity=Customer ...
- Discover Keys and Dependenc... Entity=Product Ma...
- Discover Keys and Dependenc... Entity=Account R...
- Discover Keys and Dependenc... Entity=Uk Orders ...

Permanent Keys Count=1

- Customer Master(1) Keys=1

Dependencies Count=4

- Discover Keys and Dependencies (Review... Entity=Customer ...)
- Discover Keys and Dependencies (Review... Entity=Product Ma...

ERD Diagram

```

    graph LR
      CM[Customer Master(1)] --- AR[Account Reps(3)]
      CM --> AR
      subgraph CM
        CM --> Acct_Rep[Acct_Rep]
      end
      subgraph AR
        AR --> Rep_Id[Rep_Id]
      end
  
```

Join Metadata

Join = Customer Master(1):Acct_Rep >-< Account Reps(3):Rep_Id

Metadata	Value	Description
Status	Permanent	Whether the joins is permanent, discovered or deleted.
Left Entity	Customer Master(1)	Left hand source entity
Left Expression	Acct_Rep	Expression of LH attributes comprising the left hand side
Right Entity	Account Reps(3)	Right hand source entity
Right Expression	Rep_Id	Expression of RH attributes comprising the right hand side
Matching Values	80	The number of unique joined values
Inner joined rows	806	The number of rows in the inner join - joining rows only
Outer joined rows	816	The number of rows in the outer join - joining rows and r
Left Non-Matching Values	1	The number of unique values on the left-hand side that
Left Non-Matching Rows	1	The number of rows on the left-hand side that did not jo
Left outer joined rows	807	The number of rows in the left outer join - joining rows a
Right Non-Matching Values	9	The number of unique values on the right-hand side tha
Right Non-Matching Rows	9	The number of rows in the right hand side that did not j
Right outer joined rows	815	The number of rows in the right outer join - joining rows
Left Loaded Rows	786	Number of rows in the left hand source entity
Left Filter		The filter used to select the rows from the left hand ent
Left Selected Rows	786	The number of rows selected by the Left hand filter (or l
Right Loaded Rows	90	Number of rows in the right hand source entity
Right Filter		The filter used to select the rows from the right hand ent
Right Selected Rows	90	The number of rows selected by the right hand filter (or l
Join Type	Value	Whether the join used original values or a Standardised
Actual Cardinality	M:M	The actual cardinality of the join
Inferred Cardinality	M:1	The Inferred Cardinality of the join
Exact Cardinality	9.813:1.013	The 'exact' cardinality of the join. This is defined as the ratio Left Ca...
Match Best	99.873	The percentage match of the join (best match of Values, LHrows or RHrows)
Match Values	88.889	% of values that joined
Match LHrows	99.873	% of LH rows that joined
Match RHrows	90.000	% of RH rows that joined
Left Join Rows	785	The number of rows from the left-hand side that joined

Venn Diagram

Customer Master(1) filtered : Account_Number (value)

298

Left: 62 (126)

Middle: 64 (170)

Right: 2 (2)

2,000:1,313

Uk Orders 2004(4) : Account_Id (value)

ERD Diagram Join Metadata

Ready



Oracle Data Profiling

H V A L A !

CROZ

