Workshop
Datakwaliteit en Data Profiling voor Testers

Armando Dörsek, Ferran Rohaan, Peter Endema

11 mei 2016

Programma

- **Introductie en opbouw van deze workshop**
- Theoretisch gedeelte
  - Soorten data
  - Kwaliteit
  - Oorzaak en gevolg
  - Data Profiling
- Hands on gedeelte
  - Te bestuderen data
  - Software
  - Opdrachten
- Afronding

verified

CloseSure
ICT - ALL WAYS CONNECTED

Introductie en opbouw van deze workshop

- ❑ Over ons
- ❑ Doelgroep
- ❑ Doelstellingen

verified

CloseSure
ICT - ALL WAYS CONNECTED

Over ons

- Armando Dörsek
  - Verified.nl
  - @adorsek

- Peter Endema
  - Closesure
  - @endemapeter

- Ferran Rohaan
  - Closesure
  - @frohaan

verified

CloseSure
ICT - ALL WAYS CONNECTED

Doelgroep

- Data Steward
- Data Scientist
- Business Analyst
- Information Analyst
- ETL Developer
- Report- and Dashboard Developer
- **Test Analyst**
- **Test Lead, Test Coordinator, Test Manager**
- **QA Manager**

verified

CloseSure
ICT - ALL WAYS CONNECTED

Doelstelling

- Na het volgen van deze workshop beschik je over meer kennis en vaardigheden waarmee je de datakwaliteit kunt bepalen in je project of beheersituatie.

- Met deze kennis en vaardigheden ben je in staat om eerder in het traject analysewerkzaamheden uit te voeren en waar nodig kwaliteitsmaatregelen voor te stellen

- Na deze workshop:
    - Ken je de verschillende soorten data
    - Ken je de belangrijkste oorzaken en gevolgen van gebrekkige datakwaliteit
    - Ken je verschillende dimensies van datakwaliteit
    - Ken je verschillende DQ- en Data Profiling tools
    - Kun je zelfstandig Data Profiling uitvoeren op een dataset

verified CloseSure ICT - ALL WAYS CONNECTED

Programma

- Introductie en opbouw van deze workshop
- **Theoretisch gedeelte**
  - Soorten data
  - Kwaliteit
  - Oorzaak en gevolg
  - Data Profiling
  - ~~Verbeterprogramma's~~
- Hands on gedeelte
  - Te bestuderen data
  - Software
  - Opdrachten
- Afronding

verified

CloseSure
ICT - ALL WAYS CONNECTED

Wat is Data?

- **Data :** 1) Symbols, numbers or other representation of facts; 2) The raw material from which information is produced when it is put in a context that gives it meaning. See also *Information*. (Larry English)

- **Information :** 1) Data in context, i.e., the meaning given to data or the interpretation of data based on its context; 2) the finished product as a result of processing, presentation and interpretation of data. (Larry English)

- *Bron:* *http://iaidq.org/main/glossary.shtml#D*

Soorten Data

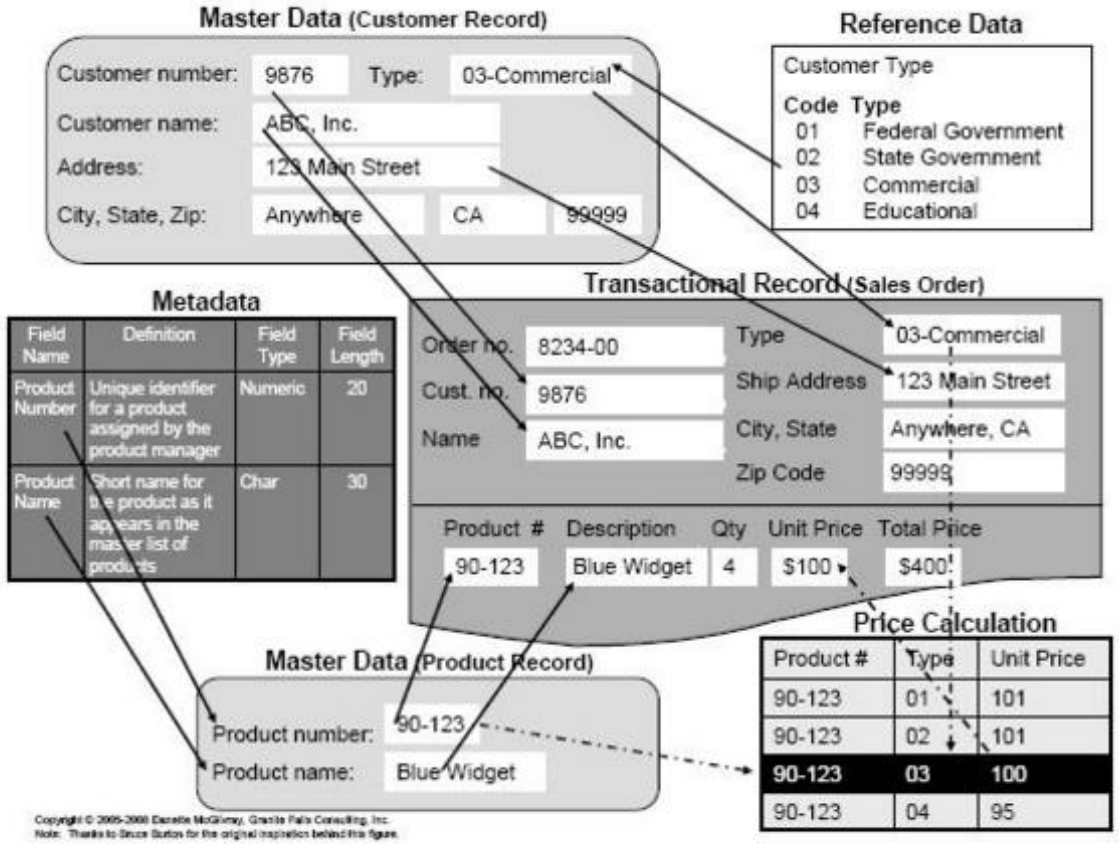
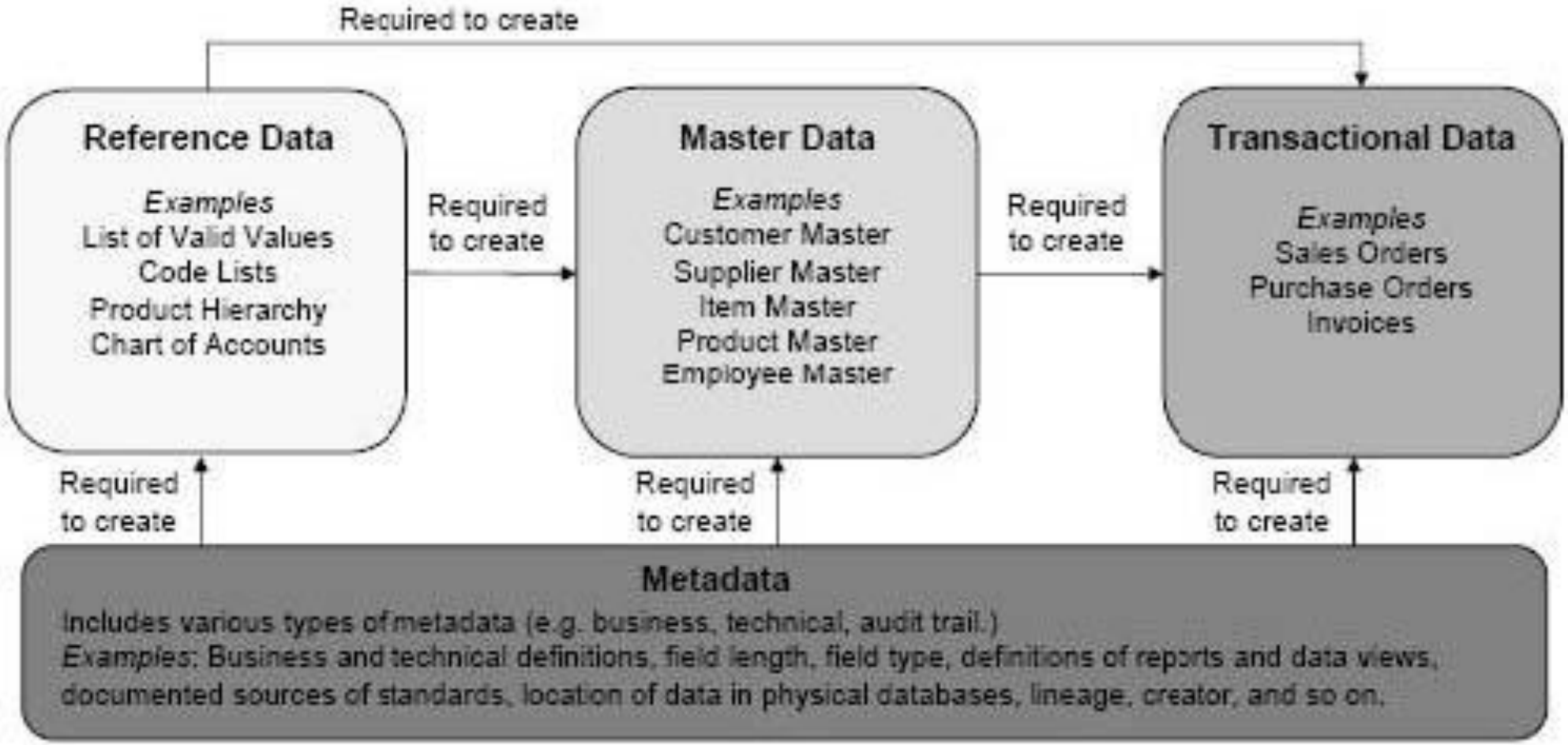


Figure 1 – An example of data categories.

- Master Data
- Transactional Data
- Reference Data
- Meta Data

Bron: *Danette McGilvray* , 7-2008, IDQ Newsletter Vol 4 Issue 3.

Soorten Data – Hoe beïnvloeden ze elkaar?



Bron: *Danette McGilvray ,* 7-2008, IDQ Newsletter Vol 4 Issue 3.

Wat is Datakwaliteit?

verified

CloseSure
ICT - ALL WAYS CONNECTED

Definities van datakwaliteit

- **Fit for Use: Voldoet aan de eisen voor gebruik waarvoor ze bedoeld zijn**

- Uitgedrukt in **kwaliteitsattributen**, bijvoorbeeld volgens:
  - TMap Next
  - NEN-ISO 9126 / 25010
  - (6) Kwaliteitsattributen voor Data Kwaliteit > IDQ Model (Valori)
- Of uitgedrukt in **dimensies**, bijvoorbeeld volgens:
  - Danette McGilvray (12 Dimensies)
  - Strong/Lee/Wang (4 categorieën, 15 dimensies)
  - Olsen (6), Loshin (8), Piney Bowes (6), CIHI (5)
  - E.v.a.

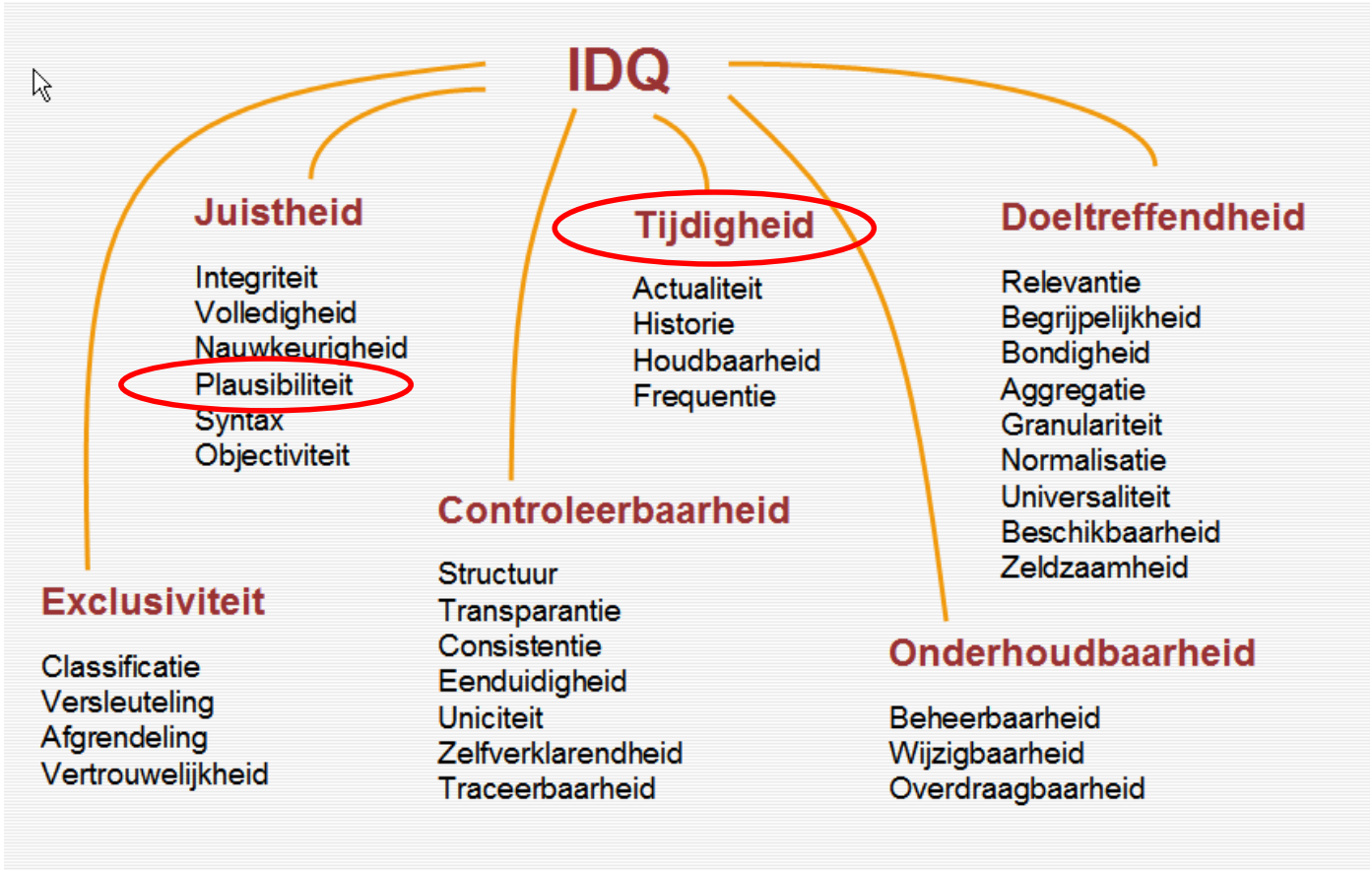
verified CloseSure ICT - ALL WAYS CONNECTED

Wat missen de traditionele kwaliteitsattributen?

verified

CloseSure
ICT - ALL WAYS CONNECTED

IDQ Model voor Datakwaliteit (Valori)



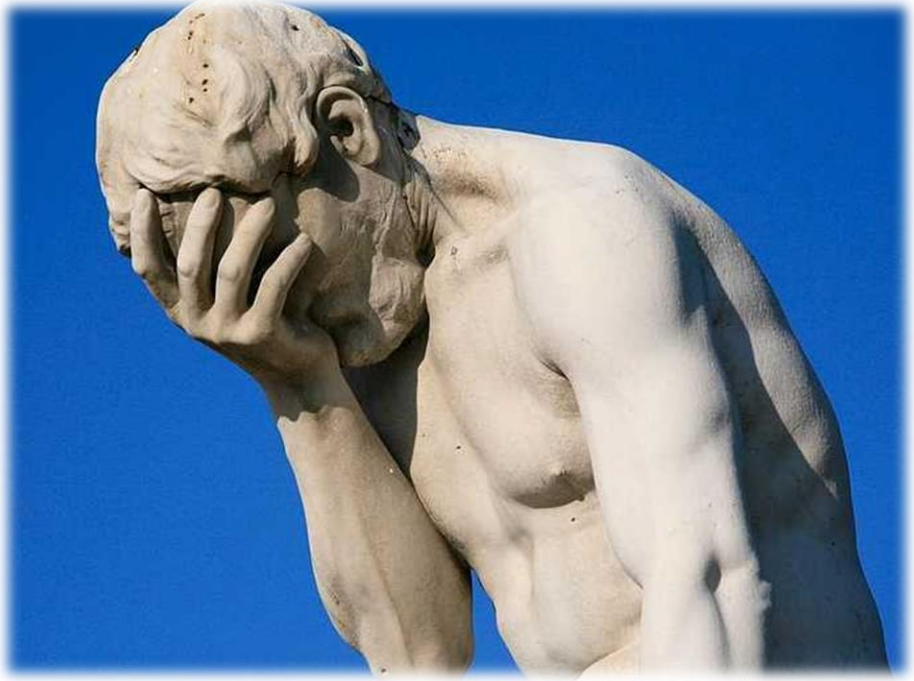
Bron: Valori. Het IDQ model maakt deel uit van het IPS kwaliteitsmodel.

verified CloseSure ICT - ALL WAYS CONNECTED

Top 5 Data Quality Issues

- Incomplete Data
- Incorrect / Wrong Data
- Aging Data
- Duplicate Data
- Data Reconciliation between Sources

Bron: Oracle



verified

CloseSure
ICT - ALL WAYS CONNECTED

Oorzaken Datakwaliteitsissues

- Fouten in de **Bron**
  - Fouten bij de **invoer** in de bron
    - Bv. onvoldoende validatie op schermen
    - Ander gebruik van een attribuut door verschillende afdelingen
  - Fouten in de **verwerking** binnen de bron
    - Ingevoerde gegevens worden niet (volledig) opgeslagen
    - Gegevens raken verloren of verminkt bij verwerking
- **Conversies, migraties e.d**.
  - B.v. upgrades van bronsystemen, import/export van datasets, “harde updates” door functioneel- of technisch beheer zonder toepassing van de juiste validaties, aangepaste naamgeving van attributen door de tijd heen, verwerking naar het datawarehouse
- **Verouderen** van data

verified CloseSure ICT - ALL WAYS CONNECTED

Kosten van slechte datakwaliteit (1)

Poor quality customer data costs U.S. companies $611 billion annually in postage, printing and staff expense

(TDWI)

At least 25% of critical data within Fortune 1000 companies will be inaccurate

(Gartner)

The cost of poor data quality can reach as high as 15% to 25% of operating profit

(TDWI)

50% to 80% of computerized criminal records in the U.S. were found to be inaccurate, incomplete, or ambiguous.

Strong, Lee and Wang (1997)

verified

CloseSure
ICT - ALL WAYS CONNECTED

Kosten van slechte datakwaliteit (2)

(…) Now I'm going to try the bad data quality tag line "Get your **$3 trillion** back America", which is the estimated cost of bad data on the US government according to Hollis Tibbets. For those who are more interested in their own bottom line it would go more along the lines of "Get your $611 Billion back America". Which is TDWI's estimates of how much solely bad *customer* data costs U.S. businesses each year. These numbers are high enough that according to Gartner, "Fortune 1000 enterprises will lose more money in operational inefficiency due to data quality issues than they will spend on data warehouse and customer relationship management (CRM) initiatives."

The SiriusDecisions **1-10-100 Rule** by W. Edwards Deming is a good starting place to finding the source of the bad data costs. The rule is simple, it costs about $1 to verify a record as it is entered, about $10 dollars to fix it later, and $100 if nothing is done, as the ramifications of the mistakes are felt over and over again.

IT departments are also affected with as much as **50% of their IT budget** going toward "information scrap and rework" related to dealing with poor data quality (Larry English). These statistics are what help form a popular opinion of data experts, namely Thomas Redman, Jack Olson and Larry English, that **approximately 15-45% of operating expense** of almost all organizations are wasted due to data quality issues.

verified CloseSure ICT - ALL WAYS CONNECTED

http://www.datablueprint.com/cost-of-bad-data-by-the-numbers/
(Chelsea Wilson, 10-3-2014)

Opbrengsten van goede Datakwaliteit

- "… a data quality strategy and targeted data quality improvement efforts that solve conflicts at the source **can lead to a 25% increase converting** inquiries to marketing-qualified leads" (In: destinationcrm.com)

- Voorbeeld uit de praktijk bij een Nederlandse ziektekostenverzekeraar:
  Voor elk geverifieerde BSN draagt de overheid EUR 1000,- bij, voor elk niet-geverifieerde BSN in het backoffice systeem mist de verzekeraar dus EUR 1000 per verzekerde. Ergo: bij slechts 100 (alsnog) geverifieerde BSN's ontvangt de organisatie een extra EUR 100.000 (!). De moeite waard…

verified CloseSure ICT - ALL WAYS CONNECTED

Impact van slechte datakwaliteit

- **Ontwikkelfase**
  - Documentatie (IA/BA) bijstellen
  - Herstelwerkzaamheden brondata, ETL en/of Rapportages
  - Ontwikkelwerkzaamheden en (her)testen loopt uit
- **Implementatiefase**
  - Ontevreden eindgebruikers, heropleveringen, herstel van data in bronsystemen, datawarehouses etc., uitloop van implementatie
  - Extra inspanningen van functioneel beheerders en kennishebbers
- **Productiefase**
  - Afgebroken laadprocessen, overschrijding van deadlines, herstelwerkzaamheden
  - Extra belasting van ontwikkel- en beheerteam, analisten en beheerders van bronnen
  - Verlies van klanten, boetes van toezichthouders, reputatieschade

verified CloseSure ICT - ALL WAYS CONNECTED

Dus…

verified

CloseSure
ICT - ALL WAYS CONNECTED

Data Profiling

- Methode voor data analyse die inzicht levert in de **waarden, structuur en kwaliteit** van de gegevens
- Voldoet de data aan de bedoelingen van de organisatie (fit for use)
- Zou integraal onderdeel moeten zijn van het (ETL) analyse- en ontwerpproces (…maar is vaak een ondergeschoven kindje)
    - Veel bouwwerk te vermijden door bv. tijdig signaleren van uitzonderingen
    - Controles, cleansing, verrijken en herstellen te vermijden (mits de bron het aanpakt)

- Wanneer?
    - Tijdens ontwerp- en ontwikkelproces (vooraf) incl. testen
    - Tijdens beheerfase (periodiek, permanent) incl. testen

verified CloseSure ICT - ALL WAYS CONNECTED

Data Profiling

- Het **datatype** en het **domein**, de mogelijk waarden vaststellen (cf. documentatie)
- Zonder oordeel vaststellen van de **daadwerkelijke** voorkomens in de bronsystemen:

  51% “Man”

  40% “Vrouw”

  5% “Jongen”

  2% “Meisje”

  1% “Onbekend”

  0.7% NULL

  0.1% “X”

  0.1% “-”

  0.1% 99999

- Acties bepalen

verified CloseSure ICT - ALL WAYS CONNECTED

Data Profiling: Waarmee?

- Nu vaak in de praktijk: ad hoc **SQL queries**

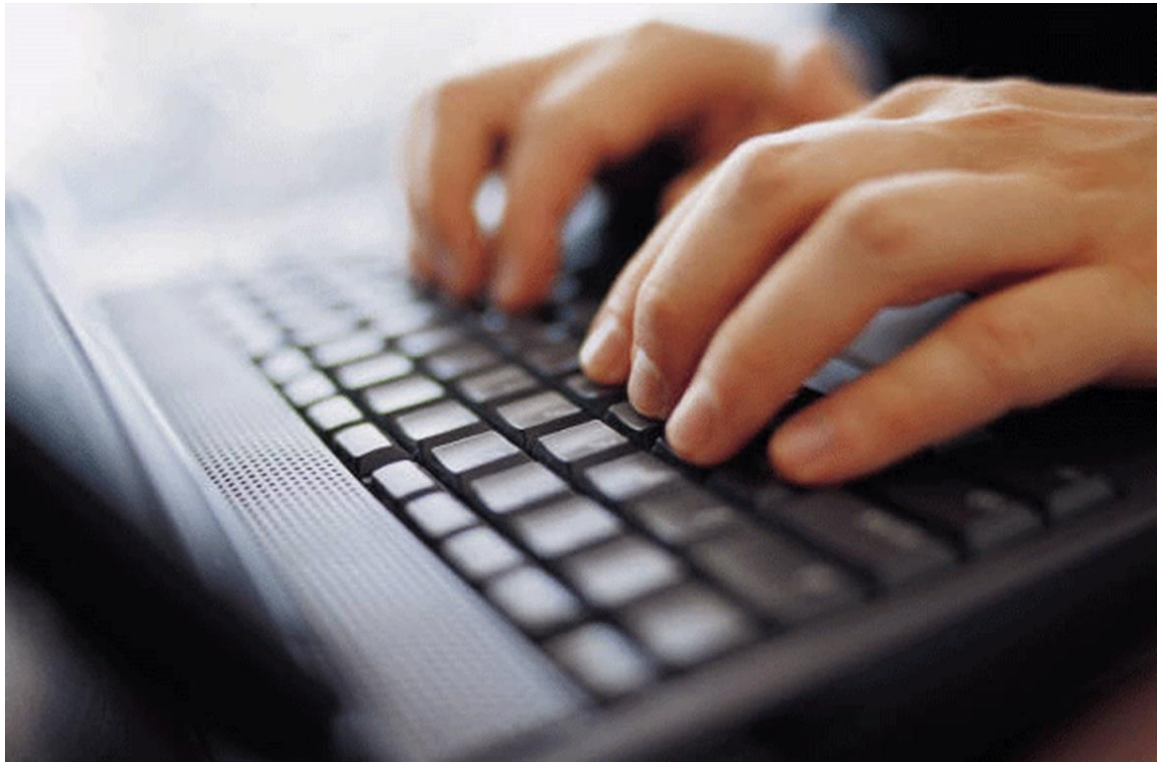
- Ook beschikbaar:
  - Features in BI/ETL suites, zoals Informatica DQ en SAP BO Information Steward, Talend Data Quality en losse (open source) tools zoals **DataCleaner**
  - Uitbreiding van suites voor subsetting en anonimisering, zoals DatProf Analyzer (In Beta).

verified

CloseSure
ICT - ALL WAYS CONNECTED

Programma

- Introductie en opbouw van deze workshop
- Theoretisch gedeelte
  - Soorten data
  - Kwaliteit
  - Oorzaak en gevolg
  - Data Profiling
- **Hands on gedeelte**
  - Te bestuderen data
  - Software
  - Opdrachten
- Afronding



verified

CloseSure
ICT - ALL WAYS CONNECTED

Hands on gedeelte

- Te bestuderen data
  - Voorbeeld database gebaseerd op Microsoft's demo database Adventureworks
  - Verbinding via RDP, zie USB-stick met voorgeconfigureerde RDP-client.
- Gebruikte software
  - SQL tool (op remote desktop)
  - Data profiling tool: Datacleaner (op remote desktop)
- Opdrachten workshop
  - Losse opdrachten, ook op USB beschikbaar
  - Uitvoeren en telkens aansluitend plenair te bespreken
  - Stel gerust vragen!

verified CloseSure ICT - ALL WAYS CONNECTED

Cases

Wifi:
Voorjaarsevenement2016

Wachtwoord:
congres2016

verified

CloseSure
ICT - ALL WAYS CONNECTED

Programma

- Introductie en opbouw van deze workshop
- Theoretisch gedeelte
  - Soorten data
  - Kwaliteit
  - Oorzaak en gevolg
  - Data Profiling
- Hands on gedeelte
  - Te bestuderen data
  - Software
  - Opdrachten
- **Afronding**

verified

CloseSure
ICT - ALL WAYS CONNECTED

Opvolging

- Nu:
  - Als nuttig ervaren?
  - Pluspunten en suggesties welkom
- Straks:
  - Ervaringen in je project?
  - Ideeën voor cases?
  - Andere of nieuwe tools?

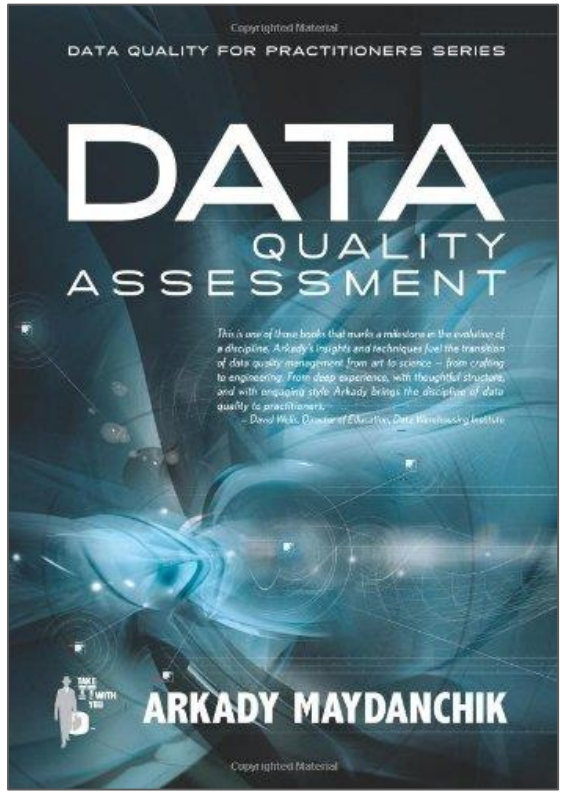
Deel het met ons!

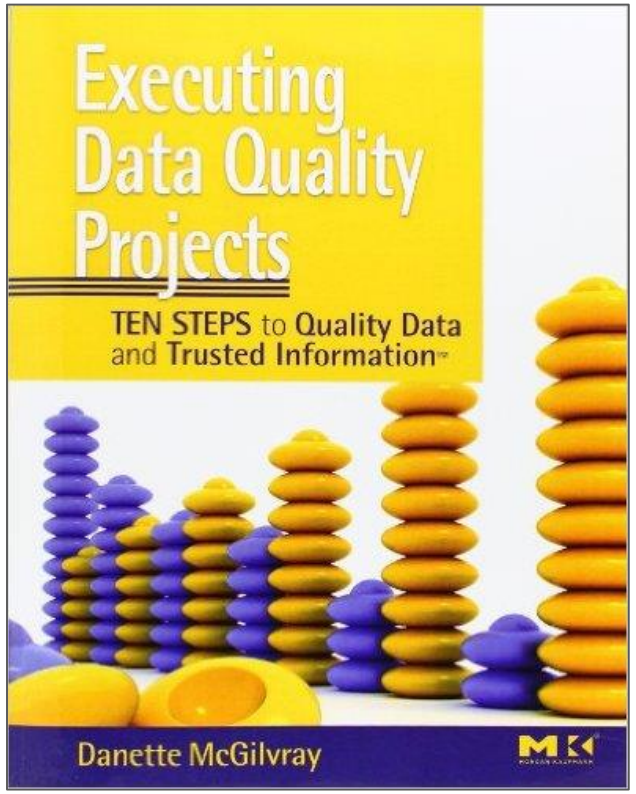
verified

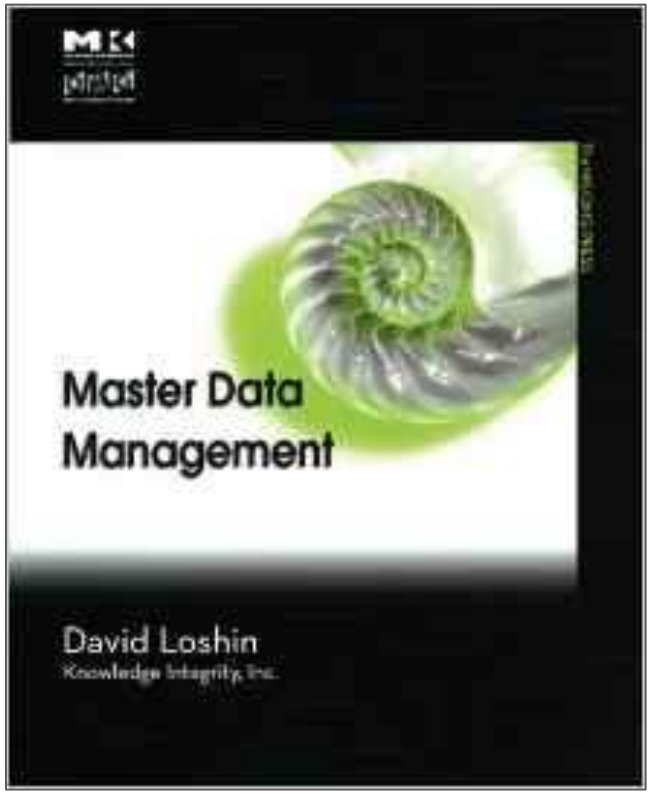
CloseSure
ICT - ALL WAYS CONNECTED

Leessuggesties

- Instituten zoals TDWI, DMBOK en specialisten zoals







verified

CloseSure
ICT - ALL WAYS CONNECTED

Dank jullie wel!

- Armando Dörsek
  - Verified.nl
  - @adorsek
  - armando.dorsek@verified.nl
  - www.verified.nl

- Peter Endema
  - Closesure
  - @endemapeter
  - peter.endema@closesure.nl
  - www.closesure.nl

- Ferran Rohaan
  - Closesure
  - @frohaan
  - ferran.rohaan@closesure.nl
  - www.closesure.nl

verified

CloseSure
ICT - ALL WAYS CONNECTED