

20. DOAG-Konferenz Hinreichende Datenqualität in Geschäftsdaten – viel mehr als Namen- und Adresscleansing

Nürnberg, 21.11.2007

Detlef Apel

1.0



sd&m AG – software design & management – gehört zu den führenden deutschen IT- und Beratungsunternehmen

Geschäftsfelder

- Entwicklung und Integration maßgeschneiderter Informationssysteme für unternehmenskritische Prozesse
- IT-Beratung mit Umsetzungskompetenz

Kunden

- Namhafte Unternehmen und Organisationen, die durch Einsatz individueller Lösungen Wettbewerbsvorteile erlangen

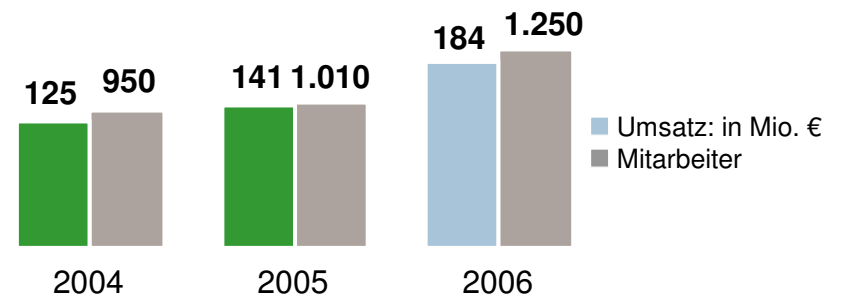
Kernkompetenz

- Gestaltung von IT-Architekturen
- Durchführung komplexer IT-Projekte
- Beurteilung und Einsatz moderner Software-Technik
- Partnerschaftliche Arbeitsweise

25
JAHRE
KOMPETENZ IM
SOFTWARE
ENGINEERING



Unternehmensentwicklung



AGENDA

- **Einleitung**
- Erfolgreiches Datenqualitätsmanagement
- Analyse der Datenqualität durch Data Profiling
- Design der Qualitätsprozesse durch Data Rules
- Korrektur der Daten
- Monitoring mit Data Auditors
- Fazit

Datenqualität wird zum entscheidenden Faktor bei Business-Intelligence-Vorhaben

Schlechte Datenqualität verursacht:

- 1 **Zusatzkosten** (8-25% des Unternehmensumsatzes)
- 2 **Terminverschiebungen** (> 80% aller Projekte mit substantiellen Zeit-/Kostensteigerungen)
- 3 **Imageverlust**
- 4 **Unflexibilität**
- 5 **Falsche Entscheidungen**

Datenqualität ist wichtig!

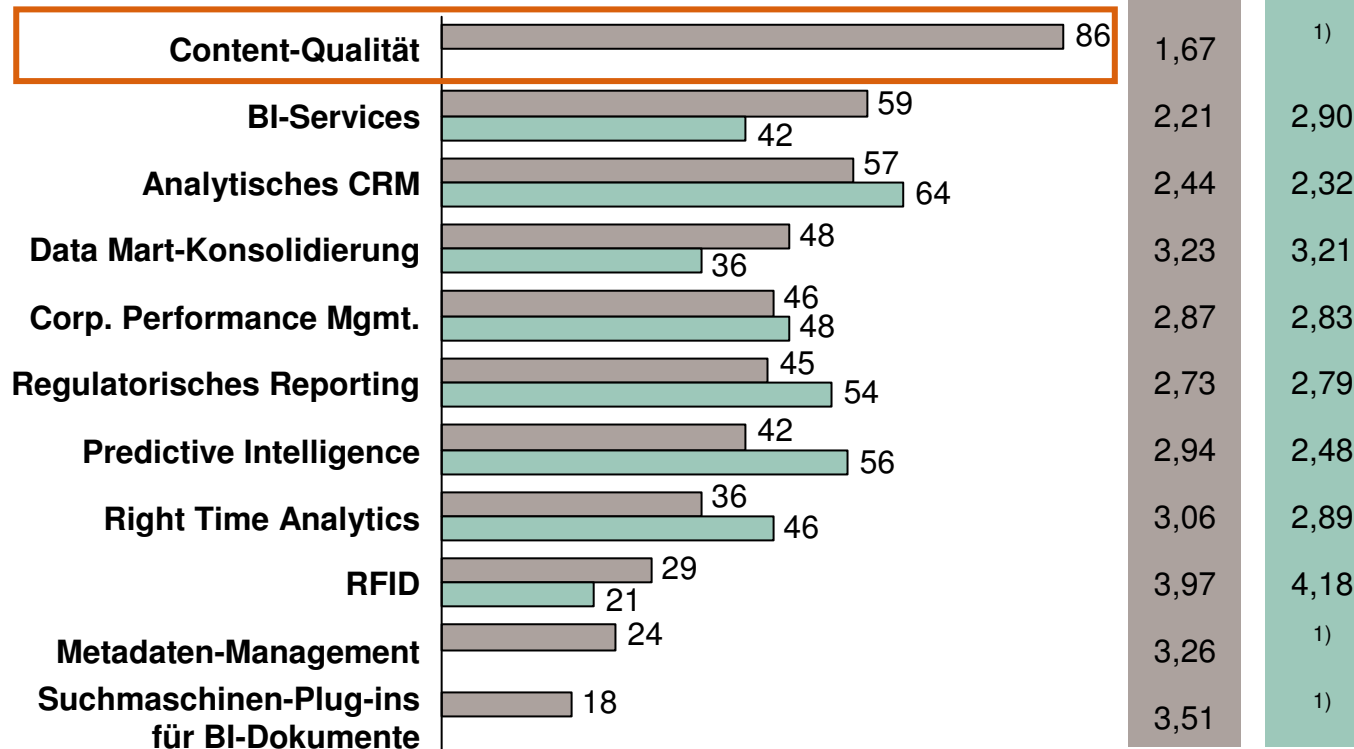
Steigernde Faktoren:

- 6 **Zunehmende Integration von Anwendungen und Prozessen**
- 7 **Vertrauen in BI wird hinfällig, wenn die DQ nicht stimmt**
- 8 **Rasch ändernde Märkte, Gesetze, Compliance etc.**

Das aktuelle BI-Top-Thema ist Datenqualität

Business Intelligence: Bedeutung einzelner Themen [%]

"Wie wichtig werden die folgenden Aspekte im Zusammenhang mit Business Intelligence in den nächsten 12 Monaten sein?"



Die Zeit ist reif für ein erfolgreiches Datenqualitätsmanagement!

■ 2007 ■ 2006

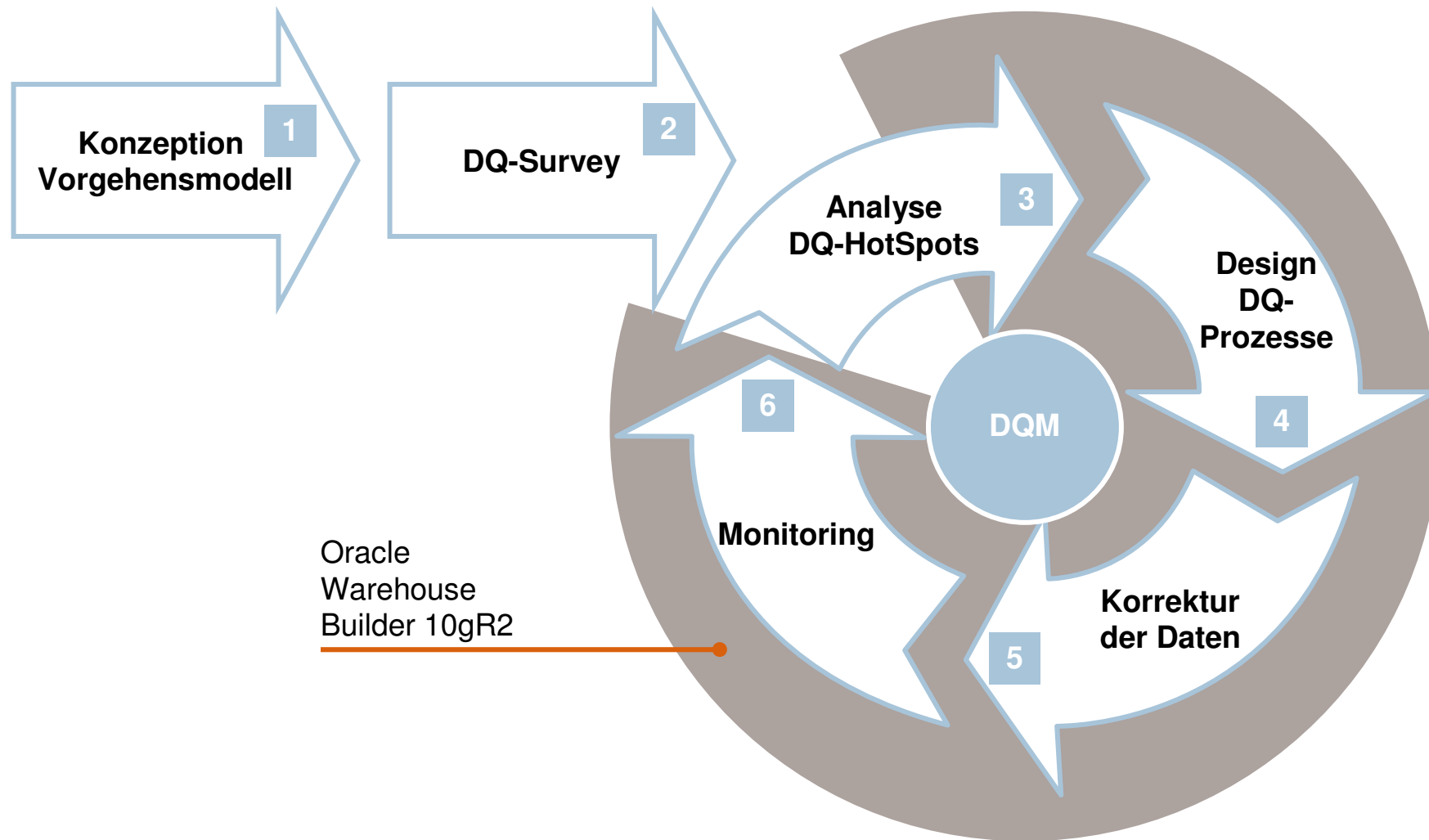
1) Wurde 2006 nicht abgefragt

Quelle: Capgemini IT-Trends 2007; Basis: Befragte, die Business Intelligence für eines der 3 wichtigsten Themen halten (n = 35)

AGENDA

- Einleitung
- **Erfolgreiches Datenqualitätsmanagement**
- Analyse der Datenqualität durch Data Profiling
- Design der Qualitätsprozesse durch Data Rules
- Korrektur der Daten
- Monitoring mit Data Auditors
- Fazit

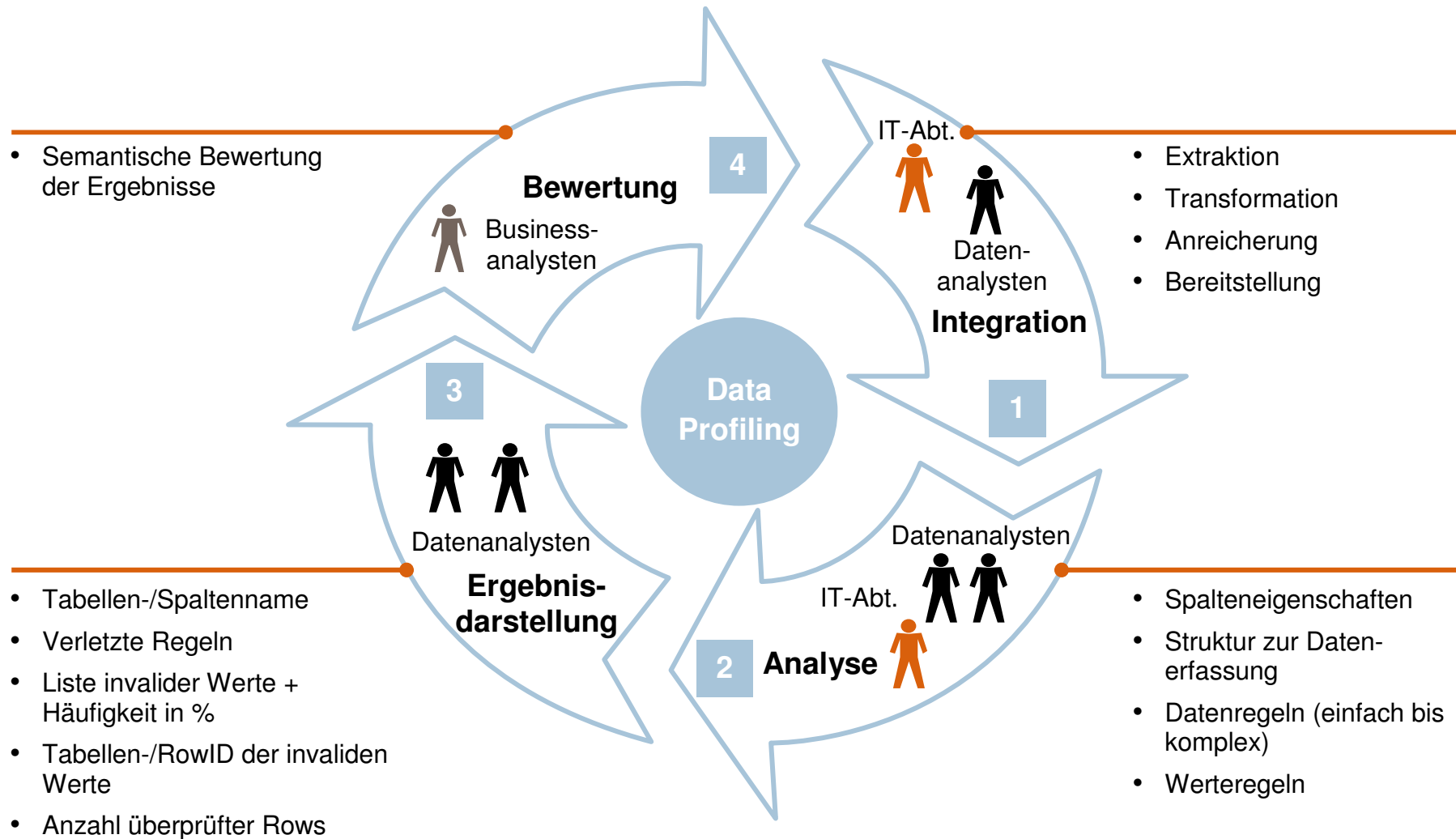
Schlüssel zum erfolgreichen Datenqualitätsmanagement ist ein auf die Aufgabe angepasstes Vorgehensmodell



AGENDA

- Einleitung
- Erfolgreiches Datenqualitätsmanagement
- **Analyse der Datenqualität durch Data Profiling**
- Design der Qualitätsprozesse durch Data Rules
- Korrektur der Daten
- Monitoring mit Data Auditors
- Fazit

Data Profiling ist ein iterativer, zeit- und kostensparender Prozess, aber keine „Wunderwaffe“ oder „Methodenersatz“



Verschiedene Analysemethoden bestimmen die Qualität der Quell- und zugehörigen Metadaten

Unique Key

- Eindeutige Schlüsselattribute?
- Schlüsselattribute NOT NULL?

Functional Dependency

- Beziehungen zwischen Spalten?

Pattern

- Einheitliches Format?

Aggregation

- Defects (z.B. Nullwerte)/Anzahl Datensätze?

Domain

- Attributswerte?
- Verteilung?

Referential Dependency

- Beziehungen zwischen Tabellen?

Data Type

- Bevorzugte Datentypen der Attribute?

User Rule

- Selbst definierte Analysefunktionen

**Jede Methode hat ihren Anwendungsbereich,
im Verbund liegt der Erfolg.**

Die Unique Key-Analyse ist wichtige Grundlage vieler anderen Prüfungen und wird zuerst durchgeführt

Profile Results Canvas

Here are the unique key analysis results for ET_CUSTOMERS, which has 13 columns and 80 rows.

Unique Key	Documented ?	Discovered ?	Local Attribute(s)	# Unique	% Uniq...	Six-Sigma
UK_254	No	Yes	ACCOUNTNR	77	96.2%	3.28
UK_255	No	Yes	CUSTOMER_ID_LEGAC...	80	100%	7.00
UK_256	No	Yes	NAME	78	97.5%	3.46
UK_257	No	Yes	ROAD	72	90%	2.78

Derive Data Rule Remove Data Rule

Data Drill Panel

Here are drill results on ET_CUSTOMERS column ACCOUNTNR related to Unique Key.

Distinct values: All

ACCOUNTNR	# Rows	% of 80
72	43	1 1.3%
73	22	1 1.3%
74	78	1 1.3%
75	47	2 2.5%
76	40	2 2.5%
77	3	2 2.5%

Displaying 77 Rows out of 77 more

Rows for the selected distinct value:

ACCOUNTNR	CONTACT	CREDITLINE	CUSTOMER_ID...	NAME	PROFILE	QUARTER
40	Fam. Stöppler	50000	40-1-2	Rhein-Blick	C	1
40	Gaby Hippeli	100000	46-10-5	Cassiopeia	B	1

Displaying 2 Rows out of 2 more

Schlüsselattribut

Nicht eindeutig

Eindeutigkeit nicht definiert

Drei doppelte ACCOUNTNR

ACCOUNTNR doppelt vergeben

AGENDA

- Einleitung
- Erfolgreiches Datenqualitätsmanagement
- Analyse der Datenqualität durch Data Profiling
- **Design der Qualitätsprozesse durch Data Rules**
- Korrektur der Daten
- Monitoring mit Data Auditors
- Fazit

Data Rules legen die zulässigen Werte und Beziehungen fest

Arten von Data Rules

- Domain List: Werteliste
- Domain Pattern List: Patterns (reguläre Ausdrücke)
- Domain Range: Wertebereich
- Common Format: Spezielle Formate (z.B. Telefonnummer)
- No Nulls: Verbot von NULL-Werten
- Functional Dependency: Abhängigkeiten innerhalb eines Datenobjekts
- Unique Key: Eindeutigkeit von Attributen/-gruppen
- Referential Dependency: Referentielle Integrität zwischen Datenobjekten
- Name und Adresse: Validierung von Namen und Adressen
- Custom: Eigene SQL-Expression

Here are the domain analysis results for ET_CUSTOMERS, which has 13 columns and 80 rows.

Columns	Found Domain	% Compliant	Six-Sigma
NAME	.	0%	-6.25
PROFILE	B A C D	100%	7.00

Buttons: Derive Data Rule, Remove Data Rule

Applied Rules:

Name	Rule	Type	Description
<input checked="" type="checkbox"/>	PROFILE_RULE	DERIVED_DATA_RULE...	Domain List
<input type="checkbox"/>			Liste zulässiger Profile

Bindings:

Parameter	Binding
VALUE	% PROFILE

Apply Rule

... aus Profiling

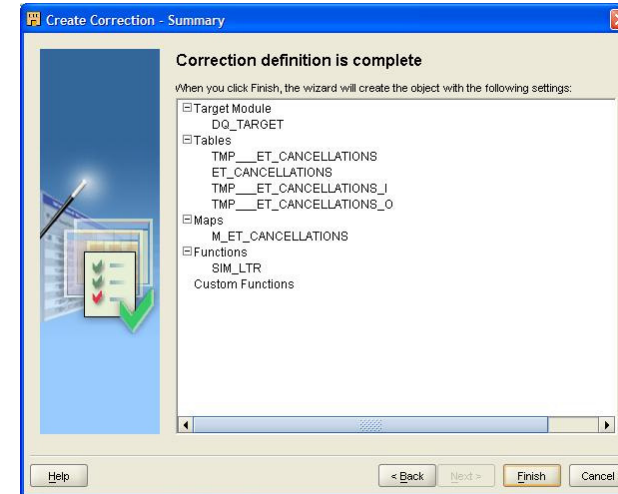
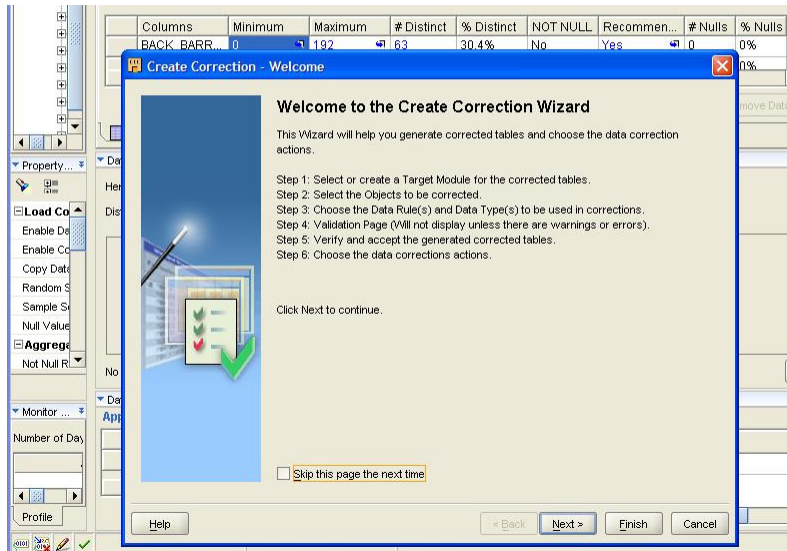
... selbst erstellt

Data Rules werden im Profiling, in der Korrektur, im Cleansing und Auditing verwendet.

AGENDA

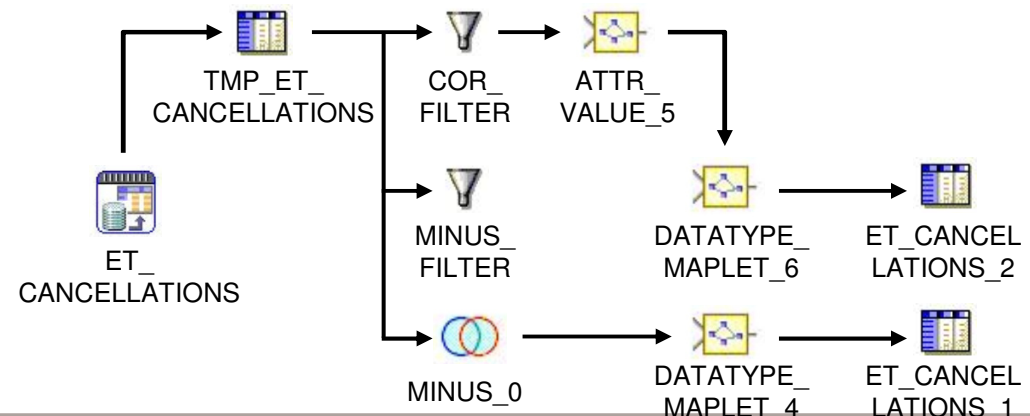
- Einleitung
- Erfolgreiches Datenqualitätsmanagement
- Analyse der Datenqualität durch Data Profiling
- Design der Qualitätsprozesse durch Data Rules
- **Korrektur der Daten**
- Monitoring mit Data Auditors
- Fazit

Auf Basis der Data Profiling Ergebnisse werden die Korrekturen automatisch generiert



Korrekturarten

- Ignore
- Report
- Cleanse
 - Remove
 - Custom– Soundex
 - Set to Min
 - Set to Max Mode
 - Similarity
 - Merge
 - Set to

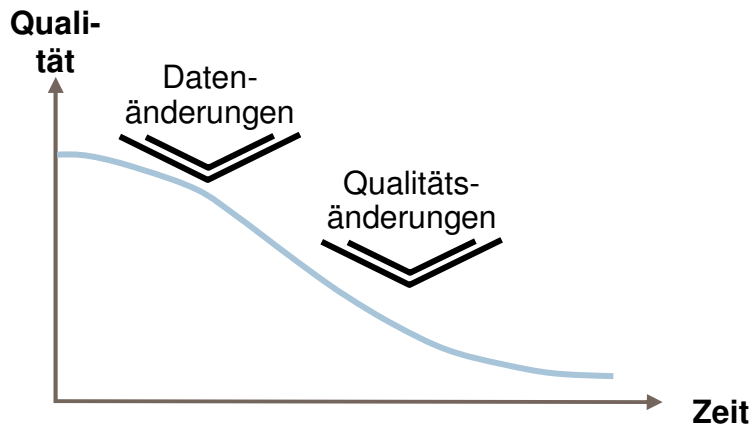


AGENDA

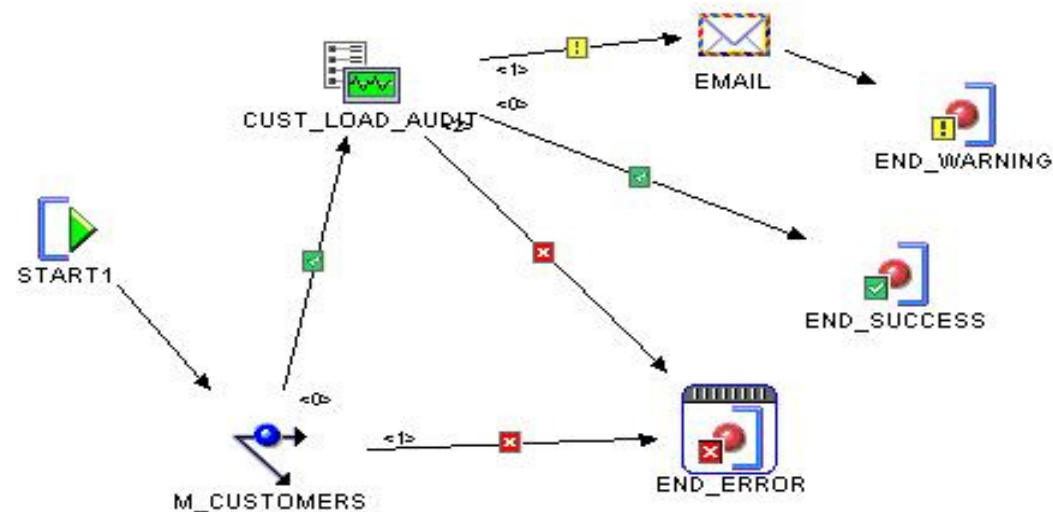
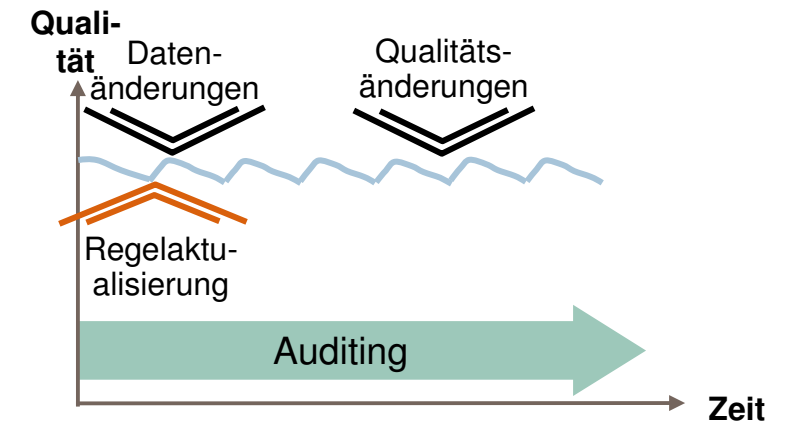
- Einleitung
- Erfolgreiches Datenqualitätsmanagement
- Analyse der Datenqualität durch Data Profiling
- Design der Qualitätsprozesse durch Data Rules
- Korrektur der Daten
- **Monitoring mit Data Auditors**
- Fazit

Nur durch stetiges, automatisches Monitoring kann die Datenqualität erhalten werden

Datenqualitätslifecycle (ohne Auditing)



Datenqualitätslifecycle (mit Auditing)



AGENDA

- Einleitung
- Erfolgreiches Datenqualitätsmanagement
- Analyse der Datenqualität durch Data Profiling
- Design der Qualitätsprozesse durch Data Rules
- Korrektur der Daten
- Monitoring mit Data Auditors
- **Fazit**

Die Zeit ist (über)reif für korrekte Unternehmensdaten mit Hilfe des Oracle Warehouse Builders (1/2)

- Vorgehensmodell zum Datenqualitätsmanagement
 - Unabdingbare Voraussetzung!
 - Angepasst an die Aufgabe
 - Managementunterstützung durch den CxO ist zwingend erforderlich
- Data Profiling
 - Keine Wunderwaffe, sondern ein nützliches Hilfsmittel
 - Fehler findet man auch ohne Data Profiling – mit aber schneller
 - Intuition, Spürsinn und Prozesskenntnisse sind unverzichtbar
 - Ergebnisse können für Korrekturen, Cleansing und Auditing weiterverwendet werden.
- Data Rules
 - Bilden die Basis für das Cleansing und Auditing
 - Generierbar aus dem Data Profiling, aber auch eigene Rules möglich
 - Vielseitig einsetzbar

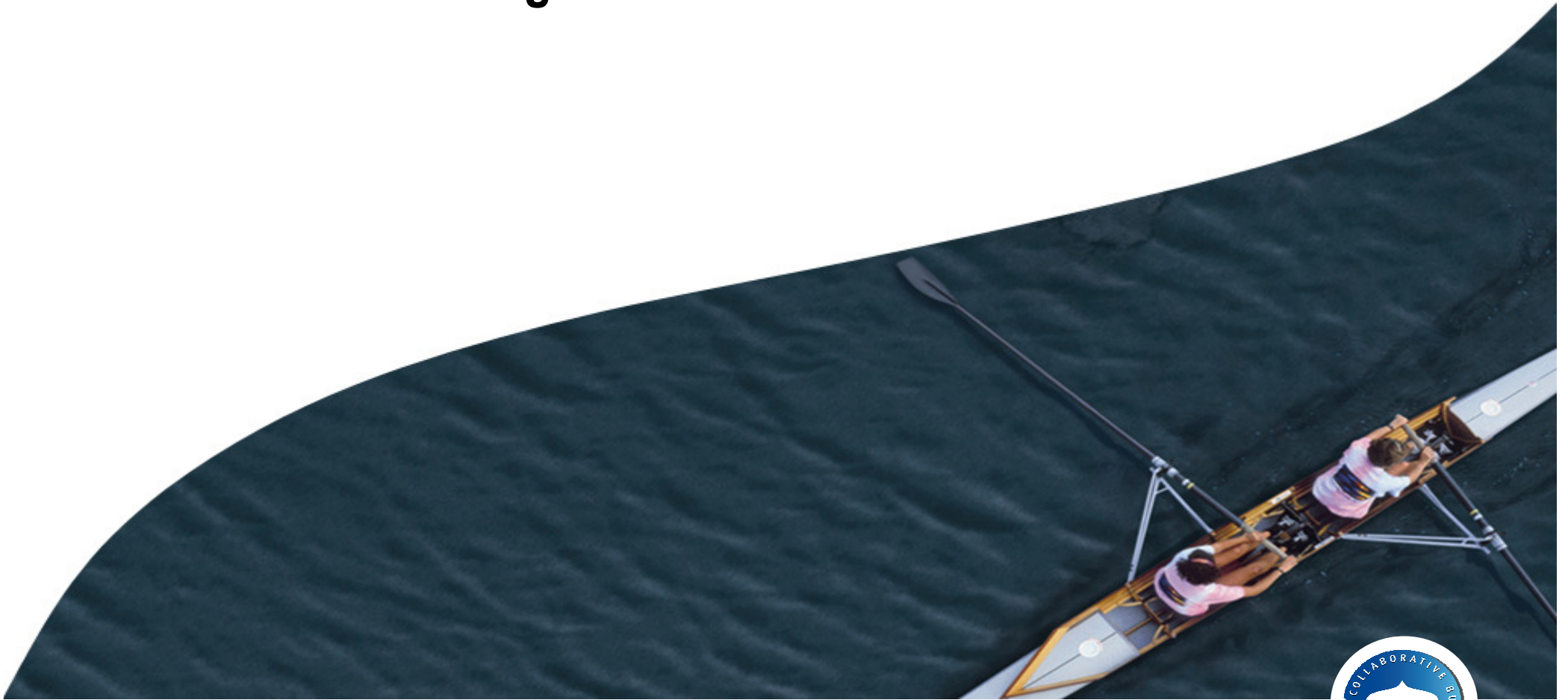
Die Zeit ist (über)reif für korrekte Unternehmensdaten mit Hilfe des Oracle Warehouse Builders (2/2)

- Data Auditoren
 - Ermöglichen ständige Überwachung der Datenqualität
 - Verwenden ebenfalls Data Rules
 - Ziel: Frühzeitiges Erkennen der Verschlechterung der Datenqualität, um rechtzeitig Gegenmaßnahmen einzuleiten
- Sonstiges
 - Es gibt auch andere Anbieter.
 - Der Warehouse Builder hat gegenüber Konkurrenzprodukten Stärken und Schwächen.
 - Im Vergleich „relativ“ günstig

Fragen und Antworten



Gemeinsam. Energien freisetzen.



GEMEINSAM. ENERGIEN FREISETZEN.

