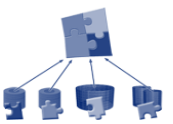
Agenda

0. Datenqualität: die Qualität von Daten?
1. Verantwortung: den Hut aufhaben?
2. Einfluss: wo aber anfangen?
3. Architektur: ist das SAS?
4. Ausblick: wohin geht die Reise?

0 Datenqualität: die Qualität von Daten?

- Zweckeignung: mehr Daten, divergente Erwartungen
- Mangelnde Datenqualität ist mangelnde Qualität der Information: fragwürdige Daten sind ohne Wert, der Benutzer misstraut ihnen *(oder gar „Ihnen“)*
- …ist kein ästhetisches Problem
  – sondern ein existentielles…aber für wen? …die IT?!
- DI macht‘s „schlimmer“: der Mann mit der einen Uhr
- Eine Code-Bibliothek ist kein Regelwerk;
  ein Programmierer ist eine menschliche Ressource
- …und, leider: *„Datenqualität ist kein Projekt!“*

1 Verantwortung: den Hut aufhaben?



- Daten sind das einzige Betriebsmittel, das sich nicht verbraucht: mehr nutzen = Mehrnutzen durch DQM
- DQM (IQM) = DQ-Management = ein Prozess (+Tool)
- *Best Practice im Büro: Googlen, Exceln & Attachen: “Die Wahrheit ist irgendwo dort draußen!”*
- Rückbesinnung: *“Making data personal.“* (Peter Aitken)
- → Wo ist mein *Data Steward*? Gerangel
- Inselreiche und Lobbyarbeit: den Vorstand im Rücken?
- „*DQ, und nu‘?*“ → Leitbild, Vision, Unternehmensziele

Datenqualitätsmanagement : an sich ein IT-technischer Kernprozess

→ Teil jeden Datenflusses im Unternehmen:

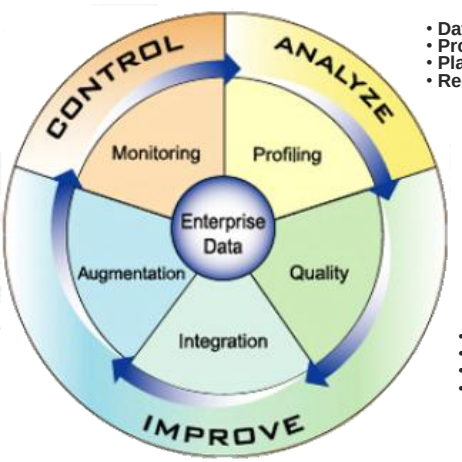
- In allen Datenmanagement-Prozessen
- Laufende Synchronisierung von Daten unterschiedlicher Systeme
- Kernfragestellung bei Fusionen und Migrationen
- Integration externer Daten
- Dateneingabe im Front-Office (Schnittstelle Mensch – System)

2 Einfluss: wo anfangen?

- Wege im Warehousing (wertfrei)
  - DWH mit veredelten Daten (ETL korrigiert)
  - Brutto-wie-Netto (ETL ignoriert)
  - DWH mit unvollständigen Daten (ETL als Firewall)
  - *Alles* ist ein Mart (ETL?!)
  - „Been there, done that!“ (~~ETL~~)
  - DQ-an-der-Quelle (ETL & beyond)
- Die IT liefert Werkzeuge, Know-how, Umgebungen
- … nicht zwingend ausprogrammierte Lösungen
- → z.B. **Profiling**

Die 5 Schritte zur besseren Datenqualität

- Kennzahlen
- Statusprüfung
- Maßnahmencontrolling
- Transparenz



- Dateninformationen
- Problemidentifikation
- Planungsgrundlage
- Reduktion Projektrisiko

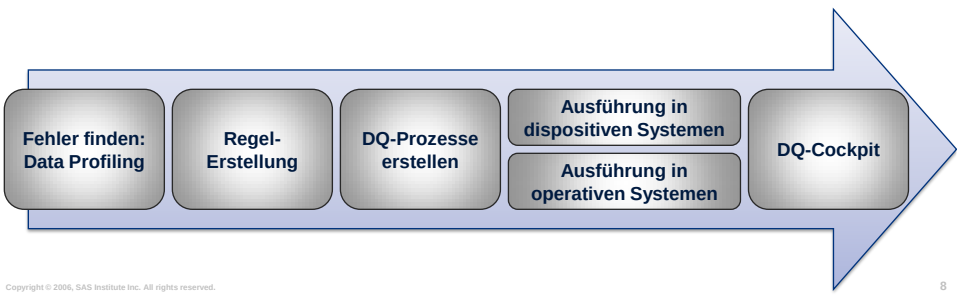
- Datenanreicherung
- Datenergänzung

- Daten Validierung
- Datenabgleich
- Dublettenerkennung
- Haushaltsidentifikation

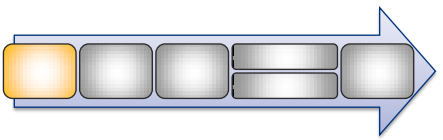
- Datenzusammenführung
- Datenkorrekturen

SAS als durchgängige Plattform für Datenqualität

- **Data Profiling**: Aufspüren von DQ-Problemen in Daten
- Erstellen von **Regeln** zur Behebung und Reporting von DQ-Fehlern
- Erstellen von **Prozessen**, in die die Regeln integriert sind (ETL, ERP…)
- Ausführen von Prozessen in **Batch und (Near-)Real-Time**
- Das **DQ-Cockpit** steuert die Datenqualität im Unternehmen



Data Profiling

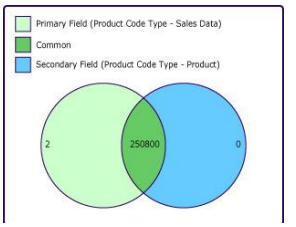


- Analyse der Dateninhalten
- Aufspüren von Fehlermustern, Ausreißern …
- Analyse von Wertebereichen
- Statistische Kennzahlen
- Beliebig erweiterbar um die *analytischen* Möglichkeiten von SAS
  - SAS Enterprise Guide, SAS Enterprise Miner , JMP etc.

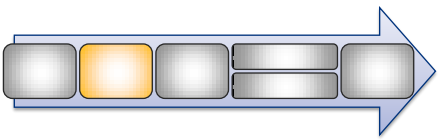
Data Profiling

- **Metadata Validation**
  Prüfung der Datensätze nach verschiedenen Kriterien (z.B. Unique count, Primary Key Candidate, etc.)
- **Pattern Analysis**
  Ermittelt Muster (Pattern) in Feldern, dabei werden Nummern als 9 und Buchstaben als A wiedergegeben.
- **Statistical Analysis**
  Ermittelt Werte wie Min., Max. Mittelwert, Standardabweichung, etc.
- **Frequency Counts**
  Listet die Anzahl von Einträgen auf.
- **Rule Validation**
  Überprüfung der Einhaltung von spezifischen Geschäftsregeln.
- **Relationship Discovery**
  Zeigt an, ob primary/foreign key Relationen konsistent sind und wie hoch die redundante Datenhaltung ist.

METRIC NAME	METRIC VALUE
Data Type	double
Primary Key Candidate	no
Unique Count	1140
Uniqueness	70.11
Pattern Count	(not applicable)
Minimum Value	-223000
Maximum Value	9999999
Minimum Length	(not applicable)
Maximum Length	(not applicable)
Null Count	2
Blank Count	(not applicable)
Actual Type	double
Count	1628
Data Length	53 bit
Mean	114348.170972
Median	4888499.5
Mode	0
Non-Null Count	1626
Nullable	YES
Ordinal Position	7
Decimal Places	0
Standard Deviation	429438.361236
Standard Error	10649.778281

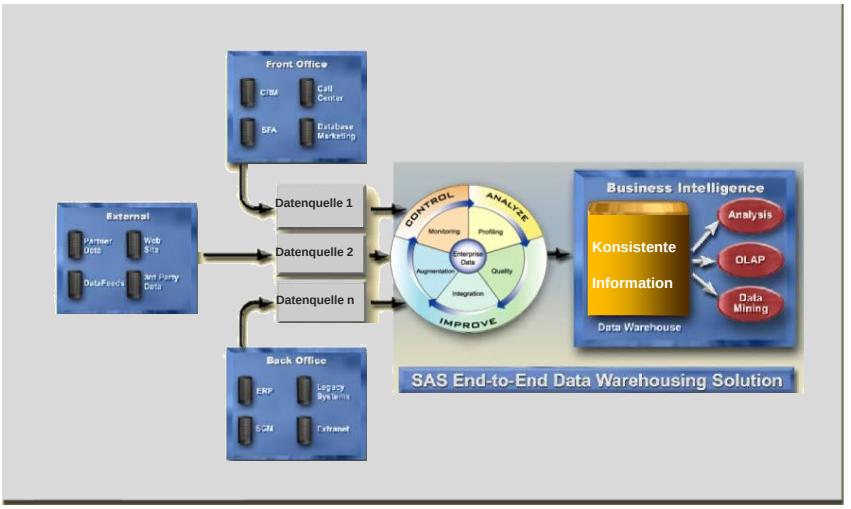


Regelerstellung

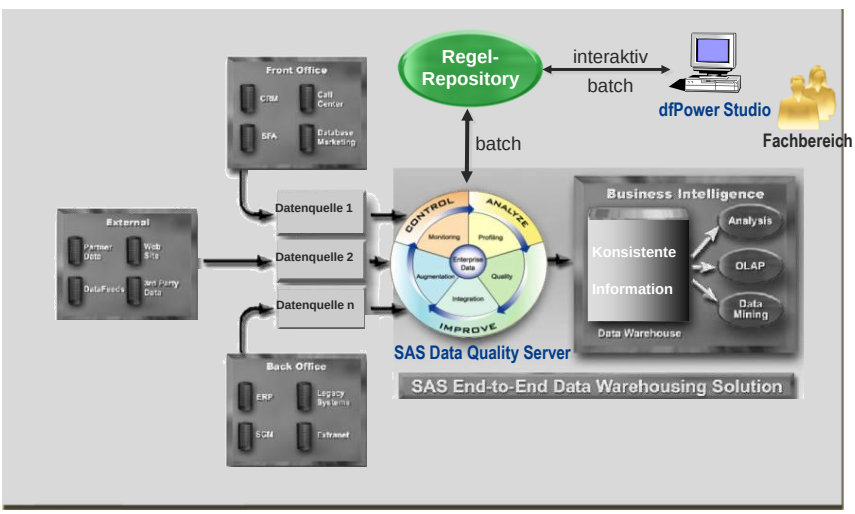


- Regeln…
  - korrigieren Fehler
  - melden Fehler (Fehlerberichte)
  - führen Daten zusammen (erzeugen Match-Codes)
  - standardisieren Schreibweisen
  - reichern an mit Informationen
- Erstellen von Datenbereinigungsregeln
- Ablegen von Business Regeln in einem Regel-Repository
- **Zielgruppe:** Die Regeln wird meist durch **Fachbereiche** entsprechend fachlicher Vorgaben erstellt und gepflegt.

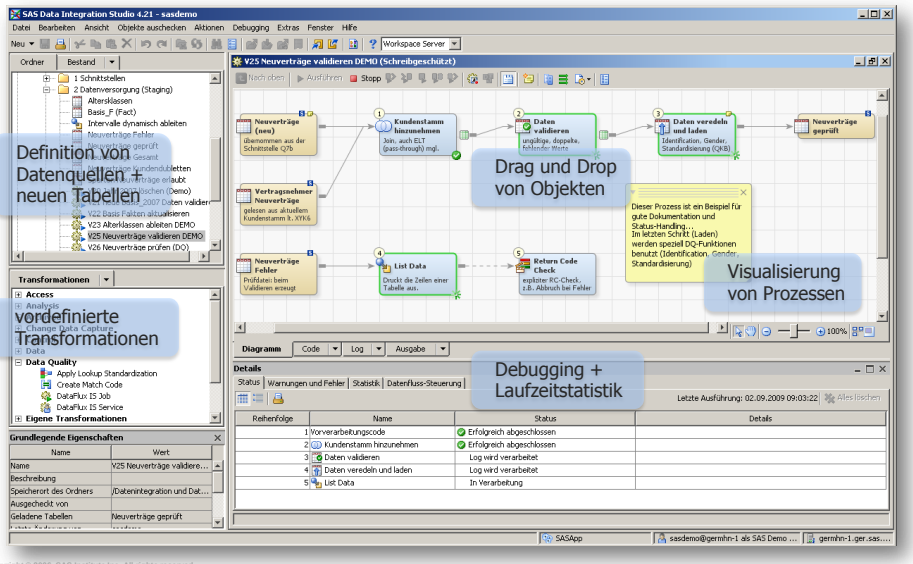
Datenqualität im DI/ETL-Prozess
Die SAS DQ Solution: Qualität, traditionell dispositiv



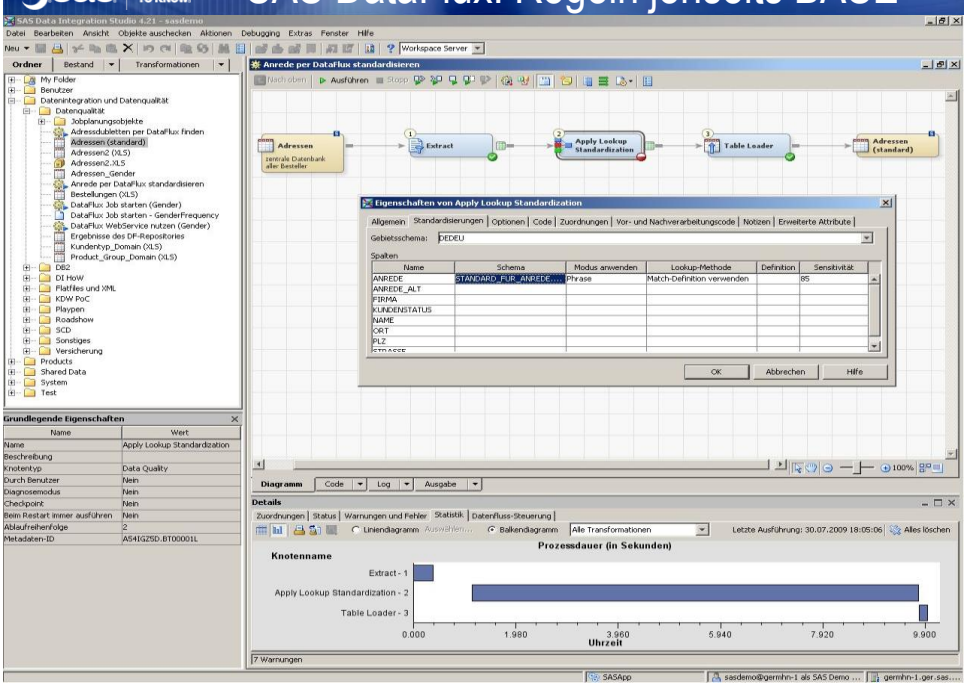
Datenqualität im DI/ETL-Prozess
Lösungsarchitektur: Regelbasiert seitens Fachbereich



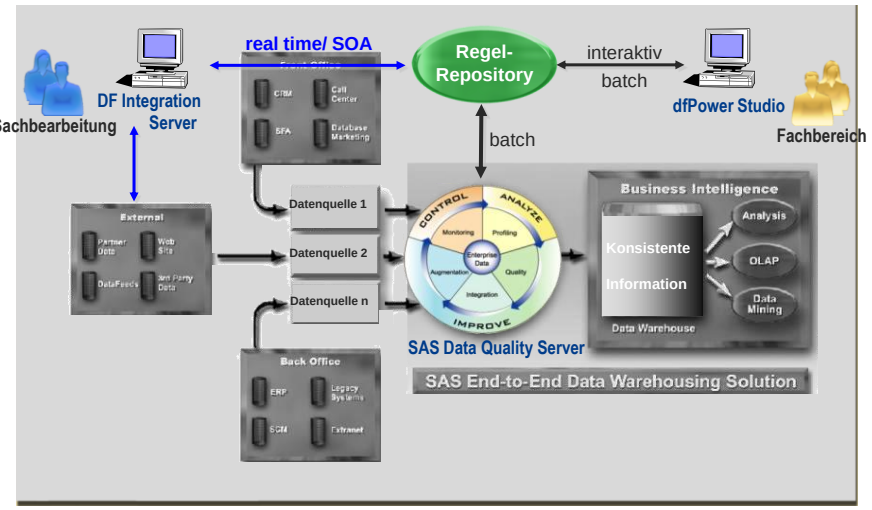
❸ Architektur: ist das SAS?



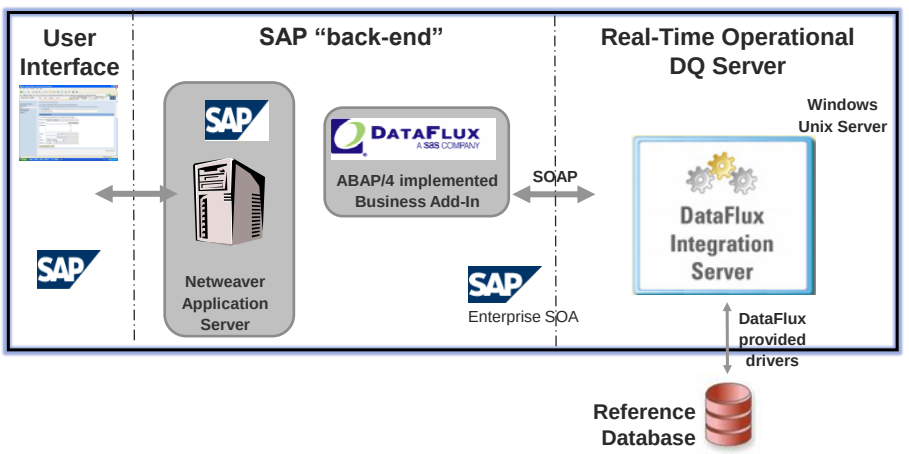
SAS DataFlux: Regeln jenseits BASE



Datenqualitätsregeln als zentrale Basis
Nutzung im Dispositiven und Operativem gleichermaßen

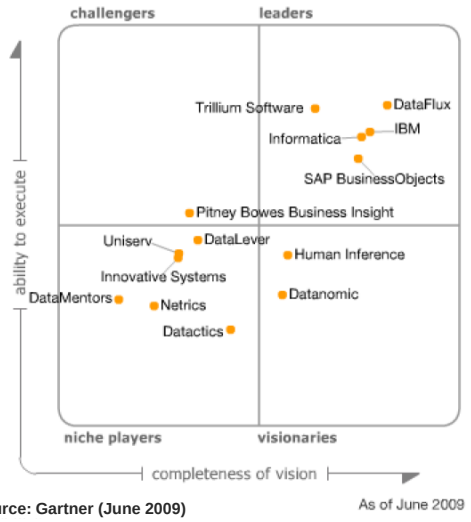


Beispiel Real-Time DQ (vs. “klassischem” SAS)



Gartner: Data Quality Magic Quadrant 2009

Figure 1. Magic Quadrant for Data Quality Tools



Gartner, Inc. "Magic Quadrant for Data Quality Tools" by Ted Friedman and Andreas Bitterer. 9 June 2009.

Source: Gartner (June 2009)

Vorteile der SAS Lösung

- Plattform statt Tool, Strategie anstelle Flickwerk
- Regeln transparent und erweiterbar, auch „grüne Wiese“
- Regeln auf Basis von Rollen, „kleben“ also nicht an einzelnen Variablen oder sind tief im ETL-Code verborgen
- Regeln werden vom Fachbereich „implementiert“ und daher verantwortet (=DQ-Sponsor und Owner!)
- integriert in SAS (auf beliebiger Ebene)
- skalierbar hinsichtlich Daten und Anwendungsgebieten
- erweiterbar (eine QKB für alle Architekturen)

4 Ausblick: wohin geht die Reise?

- Anspruch dieses Netzwerkes
- SAS, Markt, Trends
  - Metadaten generieren Base SAS…
  - ELT: Teradata und andere
  - Storage, Data Federation, Grid, Sa(a)s
  - DI als Faktor der Wertschöpfung

sas forum Deutschland 2009 Mannheim

KONFERENZ FÜR BUSINESS ANALYTICS UND BUSINESS INTELLIGENCE

STEUERN - OPTIMIEREN - ERNEUERN

16.-17. September 2009, Mannheim

HOME ABSTRACT AGENDA MITTWOCH AGENDA DONNERSTAG

Abstract

Donnerstag, 17. September 2009 – 11:30 Uhr
Raum Schönberg

IT und Fachbereich
Gemeinsam zur optimalen Datenqualität mit SAS und DataFlux
Georg Franzke - SAS Deutschland,
Eric Ecker - DataFlux Corp.

Datenqualität ist für das Scheitern der meisten Data Warehouse- & Migrationsprojekte verantwortlich. Sichern Sie sich ab und erweitern Sie Ihre Projekte um umfangreiche Datenqualitätsfunktionen. Mit dem Marktführer DataFlux zeigen wir Ihnen in diesem Vortrag, wie einfach und schnell der Fachbereich DQ-Regeln erstellt und die IT diese in eine SAS Umgebung einbindet. Eine künftige unternehmensweite Nutzung der erstellten Datenqualitätsregeln in SAP oder einer SOA-Umgebung sichert Ihnen die Sichtbarkeit im gesamten Unternehmen und verhilft Ihrem Projekt zu zusätzlichem Erfolg."

zurück zur Agenda