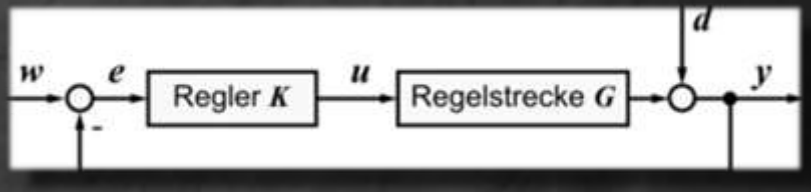
Business Intelligence: Schwerpunkt Datenqualität

Datenqualität in Unternehmen messen



Abstract

Business Intelligence: Schwerpunkt Datenqualität

Daten gewinnen für viele Unternehmen zunehmend an Wert – während früher Daten oft „nur“ eine zentrale Rolle bei der Steuerung der operativen Unternehmensprozesse gespielt haben, werden Daten zunehmend zum Kern des Geschäftsmodells. Damit wird auch die Qualität dieser Daten immer wichtiger. Aber wie können Unternehmen feststellen, wie es um diese Datenqualität steht?

Wenn Sie mich auf diesem Ausflug über die verschlungenen Pfade der Datenqualitätsmessungen begleiten, wird Ihnen hinterher die Wahl des rechten Wegs sicherlich leichter fallen. Die vorgestellten Wegpunkte wurden von mir im industriellen Umfeld bereits alle abgeschritten

Referent:
Bernd Hofner
Diplom-Ingenieur technische Informatik

Herr Hofner jagt seit über 20 Jahren freiberuflich in der Software-Entwicklung. Seine Projektkunden haben ihn gerne als „Schweizer Taschenmesser“ dabei, da er mit seiner ganzheitlichen Denkweise und breiten Erfahrungsbasis flexibel die unterschiedlichsten Rollen im Entwicklungsprozess wahrnehmen kann. Am liebsten sind ihm Projekte, bei denen er alle Entwicklungsstationen, von der Anforderungsanalyse bis zur Inbetriebnahme und Einführung des Systems beim Kunden, selbst aktiv mitgestalten kann . Seine Tätigkeitsschwerpunkte liegen bei Systemanalyse und –design.
Bei der Umsetzung greift er derzeit bevorzugt zu Java, Groovy und HTML5. Er nimmt aber auch gerne Lötkolben und Schraubenzieher in die Hand, um einen Arduino, ein VME-Bus System zur Anlagensteuerung oder ein Server-System für eine Web-Applikation zum Laufen zu bringen.

http://hofner-informatik.de

Agenda

- Datenqualität – Warum denn das?
- Qualität – Was ist das überhaupt?
- Vorgehensmodell „Six Sigma“
- Hersteller & Werkzeuge
- Infrastruktur
- Analyse, Modelle
- Metriken
- Veröffentlichung

Datenqualität – Warum denn das?

26.07.2006 http://www.sueddeutsche.de/

Telekom liefert DSL-Kundin 552 Router

…Die ersten 56 DSL-Router kamen in zwei Raten per **Post**, weitere **496** Exemplare wollte eine Spedition am nächsten morgen per **LKW** abladen…
Grund: u. a. Bezeichnung des Routers im Mengenfeld

11.2.2009 http://www.berliner-zeitung.de/

„Ich muss Ihnen sagen, Sie sind tot“

Wie eine 42-jährige Frau aus Friedrichshain durch die Telekom von ihrem eigenen Ableben erfuhr…
Grund: Totenschein am falschen Kunden abgelegt

01.09.2010 http://www.computerwoche.de/

Der Online-Händler **Amazon.com** hat versehentlich **57.000 Bücher** aus seinen Suchlisten und Verzeichnissen genommen, weil sie **fälschlicherweise als Pornografie** eingestuft worden waren.

01.09.2010 http://www.computerwoche.de/

Laut dem Lieferanten von Wirtschaftsdaten D&B kostet die schlechte Datenqualität die amerikanischen Unternehmen jährlich rund **600 Milliarden Dollar**.

02.06.2015 http://www.information-management.com/

Merrill Lynch Fined for Data Compliance Failure

Das Unternehmen muss ca. 9 Mio$ Strafe zahlen, weil es seinen Wertpapierhandel auf tagelang veralteten Daten abgestützt hatte – zum Schaden seiner Kunden.

Qualität – Was ist das überhaupt?

„…die Gesamtheit von Merkmalen einer Einheit bezüglich ihrer Eignung, festgelegte und vorausgesetzte Erfordernisse zu erfüllen.“
ISO alt

„…Grad, in dem ein Satz inhärenter Merkmale Anforderungen erfüllt“
ISO neu

„..ein mehrdimensionales Maß für die Eignung von Daten, den mit ihrer Erfassung, Generierung oder Nutzung verbundenen Zweck zu erfüllen.“
Dr. Volker Würthele - ~~www.iiiq.org~~

„…wenn der Kunde zurückkommt und nicht die Ware“
Hermann Tietz, Hertie-Gründer

„…ist das beste Rezept“
Dr. Oetker

Quellen:
http://de.wikipedia.org/wiki/Qualit%C3%A4t
http://www.aphorismen.de/zitat/112002

Definition Datenqualität

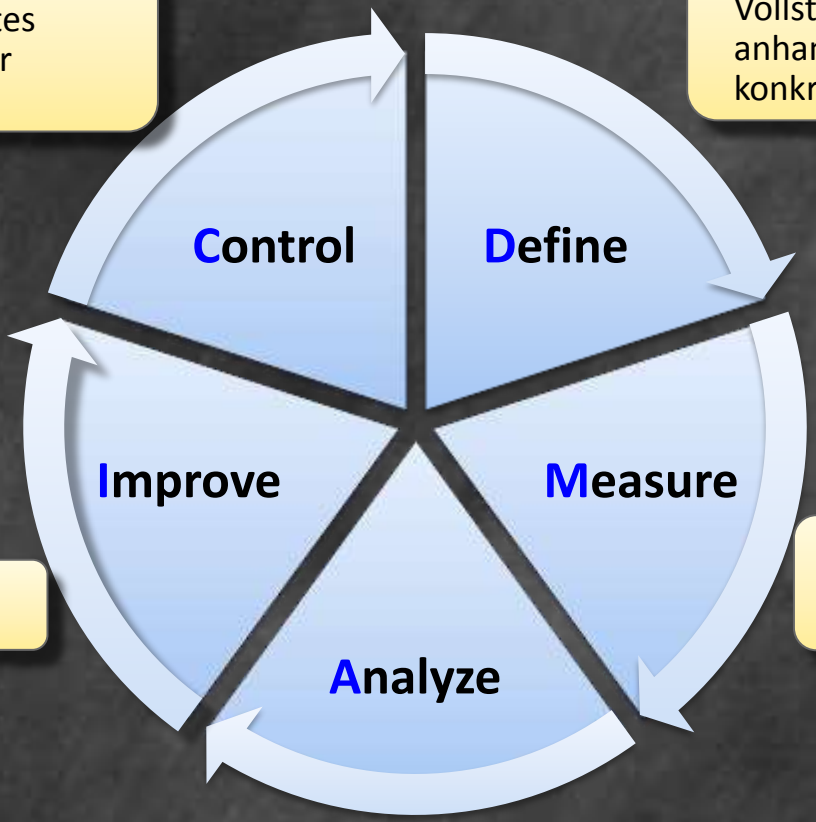
Nicht alles davon ist (leicht) messbar

- **Korrektheit**
  Die Daten müssen mit der Realität übereinstimmen.
- **Konsistenz:**
  Ein Datensatz darf in sich und zu anderen Datensätzen keine Widersprüche aufweisen.
- **Zuverlässigkeit**
  Die Entstehung der Daten muss nachvollziehbar sein.
- **Vollständigkeit**
  Ein Datensatz muss alle notwendigen Attribute enthalten.
- **Genauigkeit**
  Die Daten müssen in der jeweils geforderten Exaktheit vorliegen (Beispiel: Nachkommastellen).
- **Aktualität**
  Alle Datensätze müssen jeweils dem aktuellen Zustand der abgebildeten Realität entsprechen.
- **Eindeutigkeit**
  Jeder Datensatz muss eindeutig interpretierbar sein.
- **Redundanzfreiheit**
  Innerhalb der Datensätze dürfen keine Dubletten vorkommen.
- **Relevanz**
  Der Informationsgehalt von Datensätzen muss den jeweiligen Informationsbedarf erfüllen.
- **Einheitlichkeit**
  Die Informationen eines Datensatzes müssen einheitlich strukturiert sein.
- **Verständlichkeit**
  Die Datensätze müssen in ihrer Begrifflichkeit und Struktur mit den Vorstellungen der Fachbereiche übereinstimmen.

Quelle: Computerwoche 12.10.2011 Martin Bayer – Gute Daten Schlechte Daten

Qualitätsmanagement nach Six Sigma - DMAIC

1. Definition von Konsistenz- und Vollständigkeitsmessungen anhand von Prozessvorgaben und konkreten Problemstellungen.



2. Durchführung der Messung und Aufbereitung der Ergebnisse.

3. Detailanalyse der Messergebnisse auf Probleme und Ableitung von Maßnahmen. Ggf. Nachjustierung von Messregeln.

4. Umsetzung der geplanten Maßnahmen.

5. Prüfung des Maßnahmenerfolgs mit nächster Messung. Ggf. erneutes Analysieren und Einleiten einer Korrekturmaßnahme.

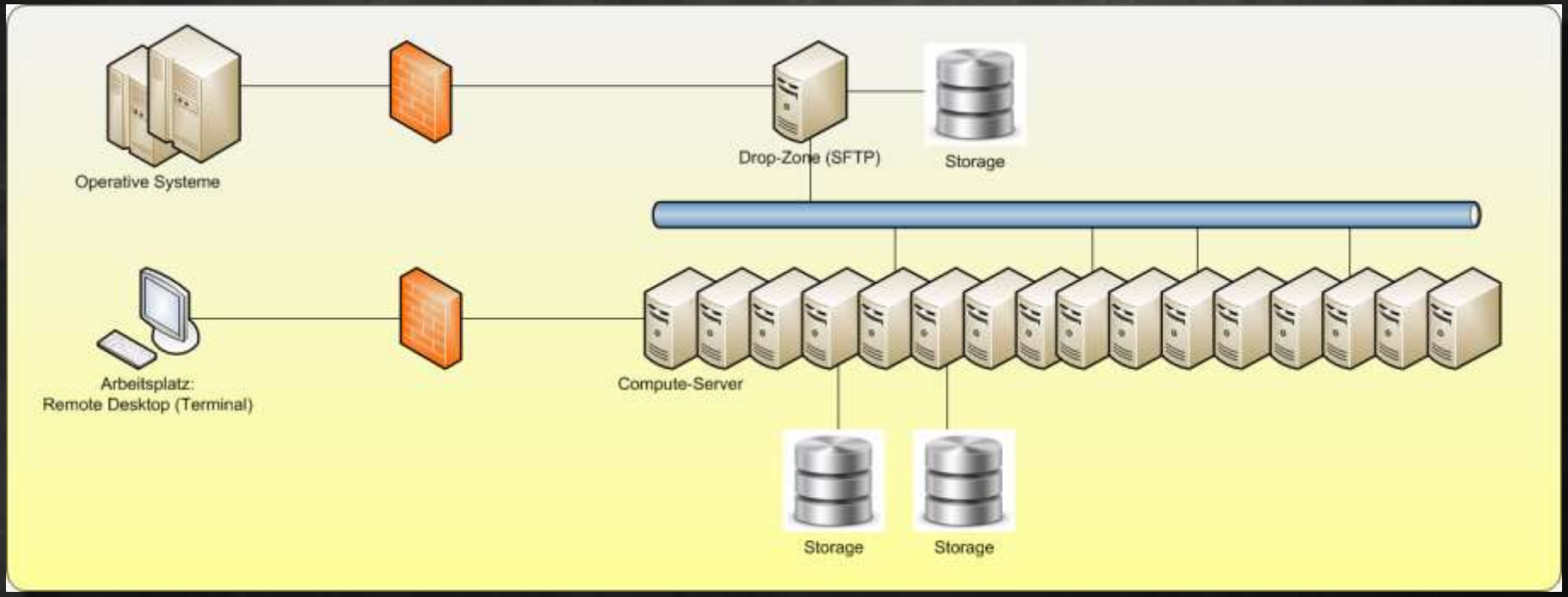
Womit?

Tool- & Technologie-Zoo

- Zeilenorientierte Datenbanken (Oracle, MySQL)
- Map/Reduce Ökosysteme (Apache Hadoop, Pig, Hive, …)
- In-Memory Datenbanken (SAP HANA)
- Spaltenorientierte Datenbanken (Actian Vectorwise, Sybase iQ)
- Integrierte Storage-/Datenbank Lösungen (Oracle Exadata)
- Proprietäre Software-Stacks (Miosoft, Trillium, Pentaho)
- …

Wo? - DQ-Infrastruktur

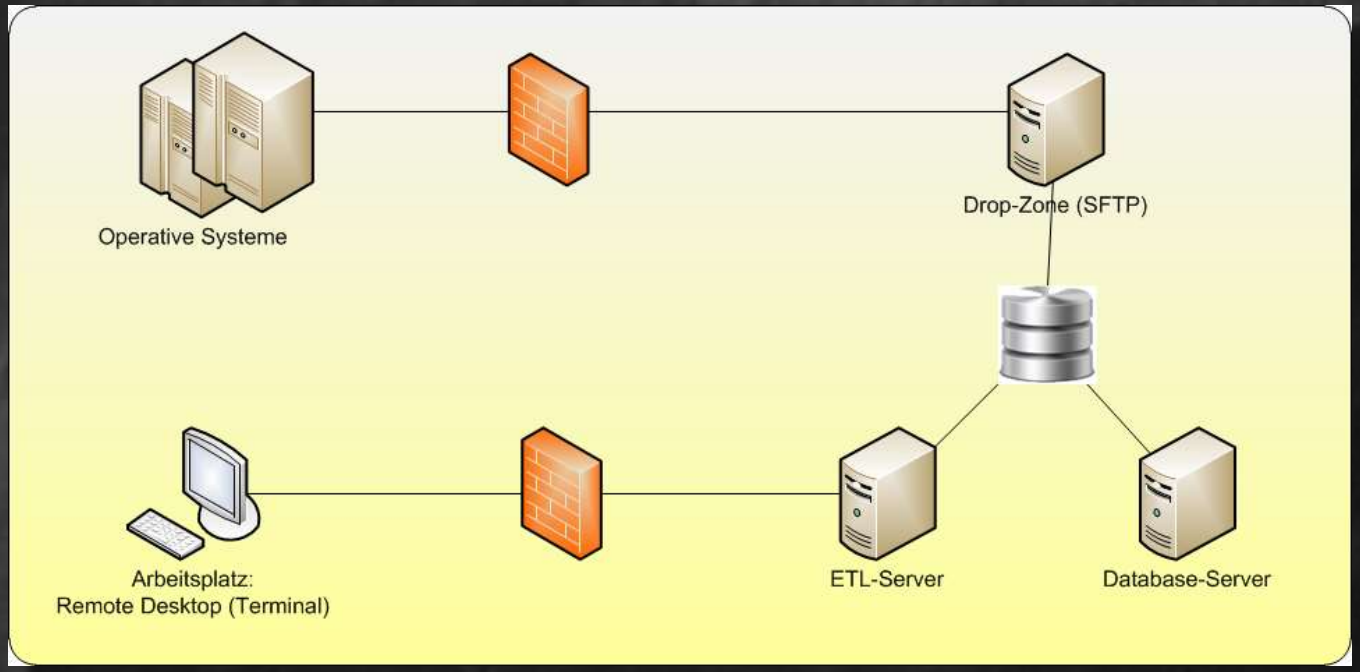
Horizontale Skalierung



- Viele verteilte Rechner teilen sich die Arbeit
- Datenverteilung & Ergebniseinsammlung zeitintensiv (Netzwerk)
- Betriebskosten

Wo? - DQ-Infrastruktur

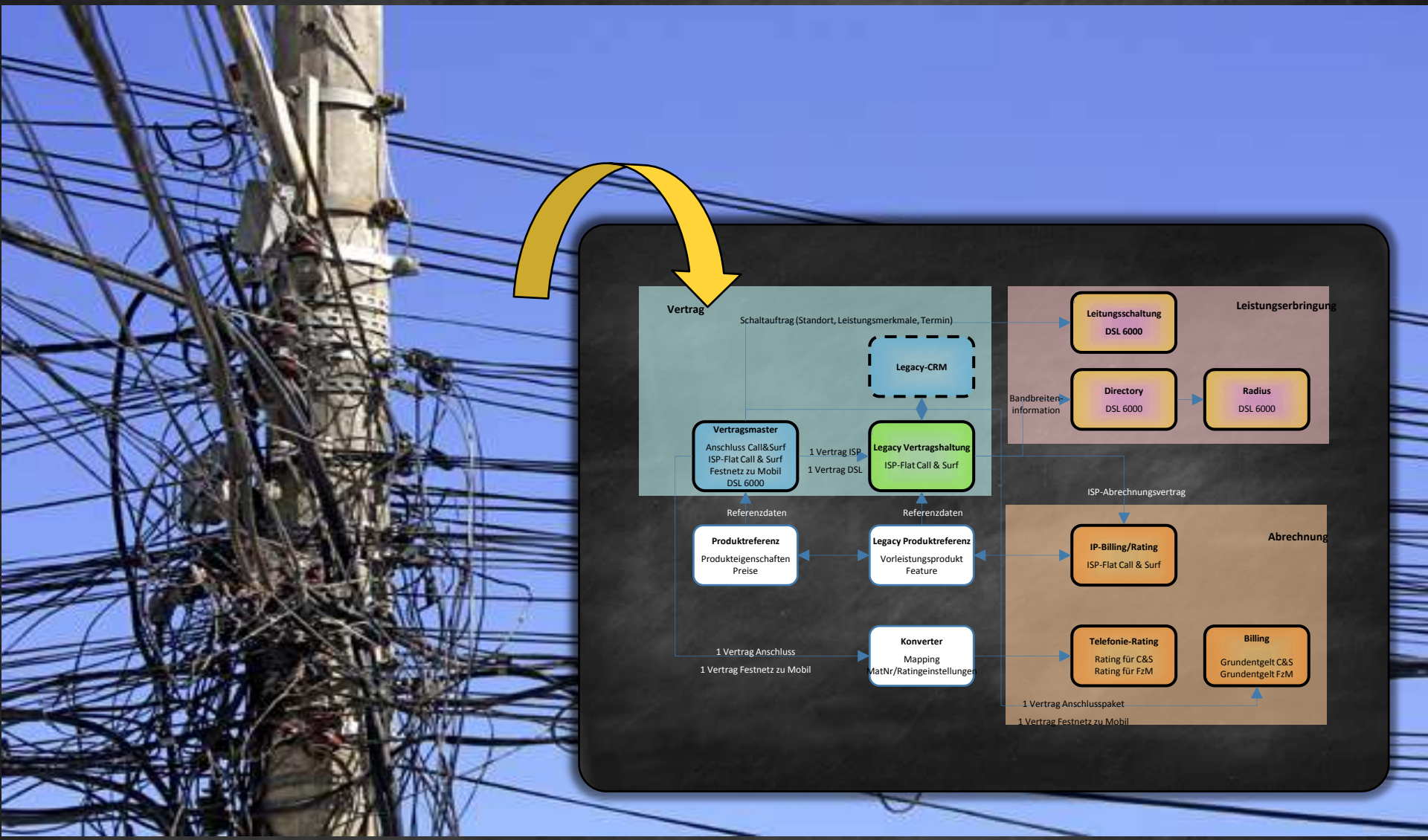
Vertikale Skalierung



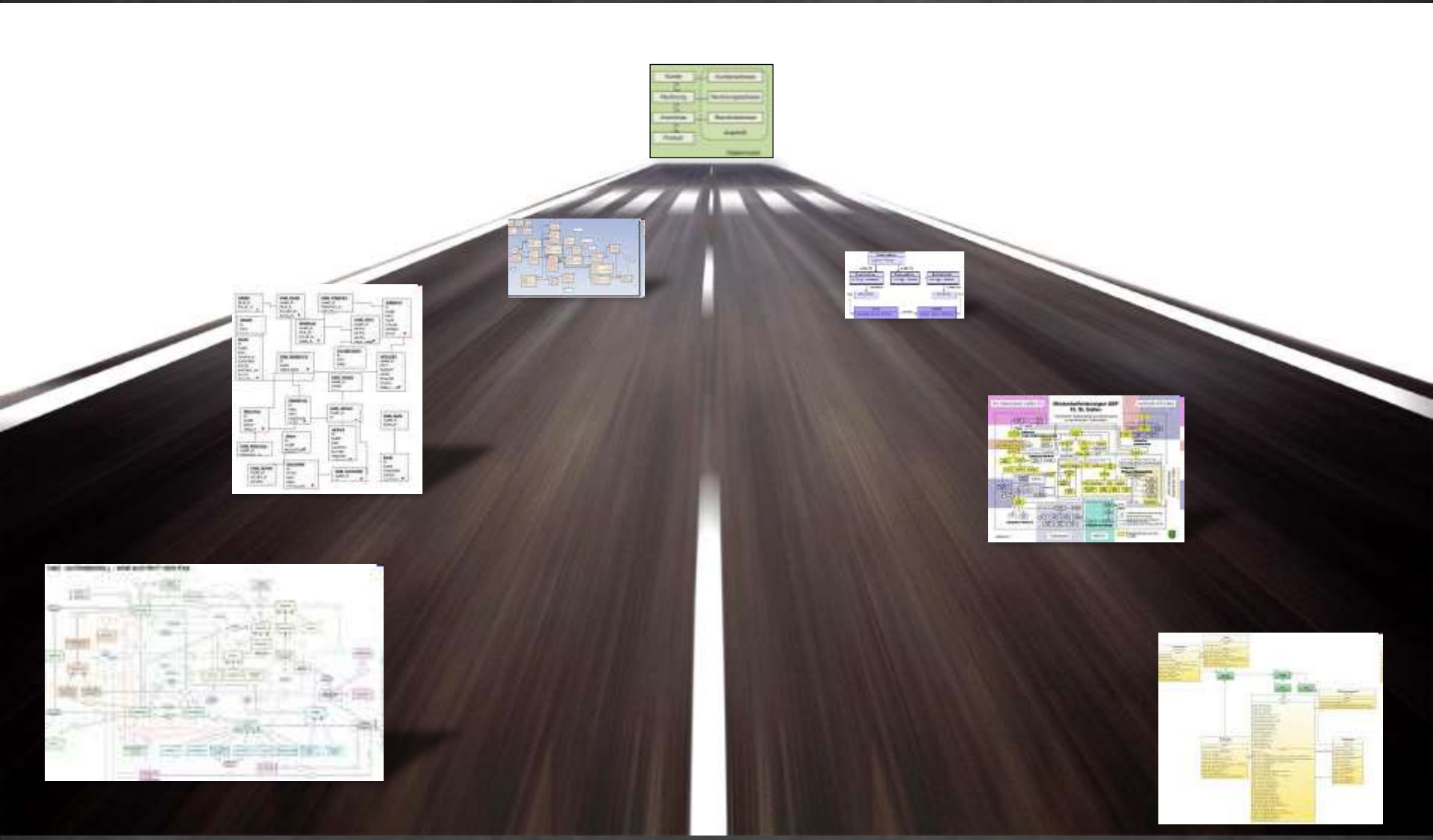
- **sehr viel RAM**
- **leistungsfähige Mehrprozessorsysteme**
- gute Storage I/O-Anbinung (SAN)
- kurze Wege für die Daten
- In-Memory Datenverarbeitung, z.B. mit spaltenorientierter Datenbank

Was? - Prozesse und Systeme analysieren.

System-Archäologie



Was? – Datenmodelle konsolidieren



Modelle rein illustrativ!

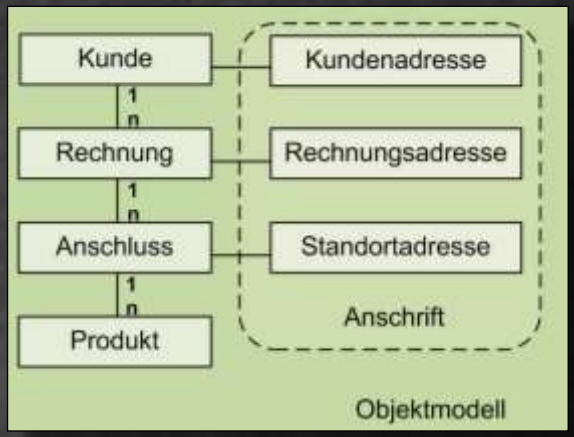
Bildnachweis: © iStock.com/carloscastilla

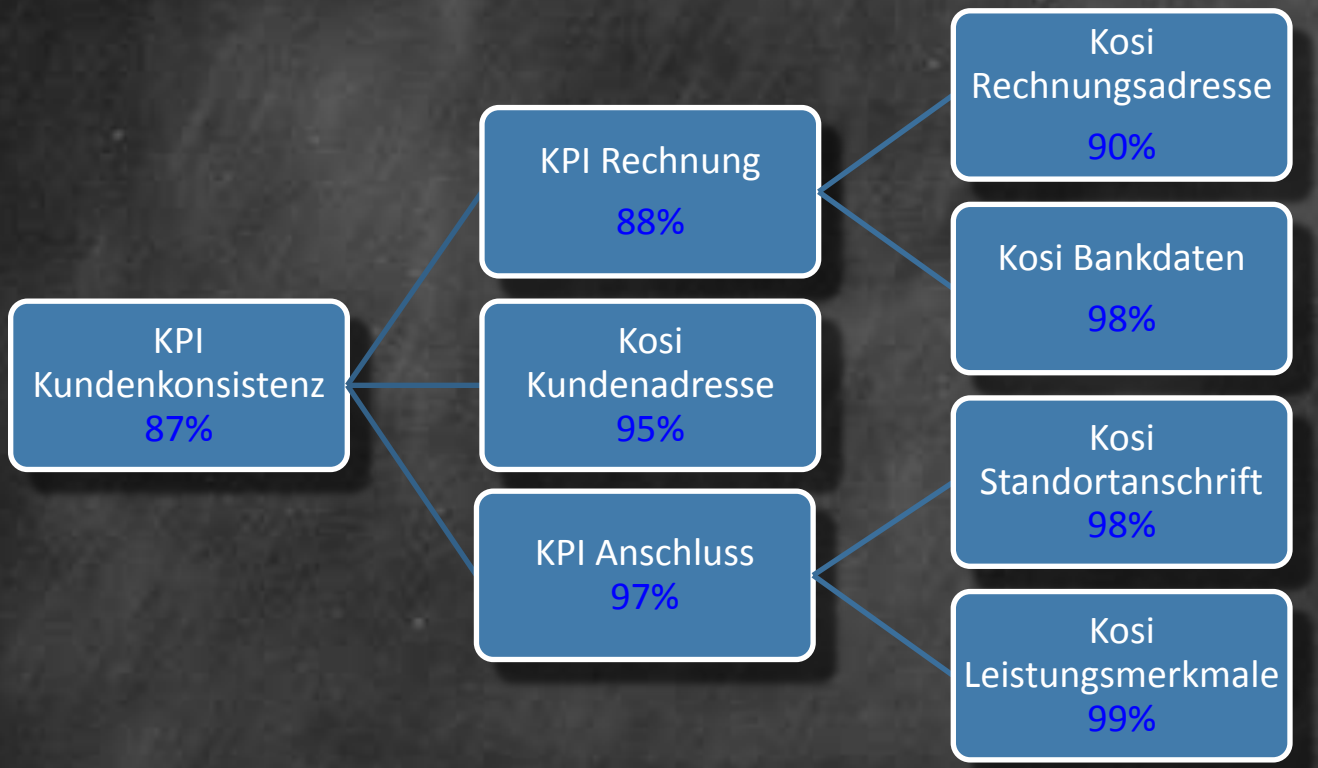
Regeln und Modelle dokumentieren

```
/**
      TOI-DSL müssen in CRMT eine TO-Nr zugeordnet haben.
*/
Regel Vertrag_11
{
      nr: "R77442101"
      system: CRMT
      short: {* Vertrag[TOI-DSL]: Hat genau eine TO-Nr in CRMT *}

      /**
            Für jeden TOI-DSL Vertrag muss es in der Tabelle
            Vertragsattribute (s_asset_xa) einen Eintrag
            mit der Attribut-Id 564 (TO-Nr) geben, der einen
            numerischen Wert enthält.
      */
      condition:
      {
            Für alle dsl : Vertrag«TOIDSL»  gilt:
                  Es existiert genau ein va : Vertragsattribut
                  mit:
                              va.fkVertrag = dsl.row_id      // join
                        und   va.attributId = "564"          // TO-Nr
                        und   matches("[0-9]+", va.wert)     // numerischer Wert
      }

      note:
      {*
            Ein TOI-DSL-Vertrag muss genau einem IspAccount zugeordnet werden.
      *}
}
```

Metriken festlegen

Key-Performance Indicators (KPI).





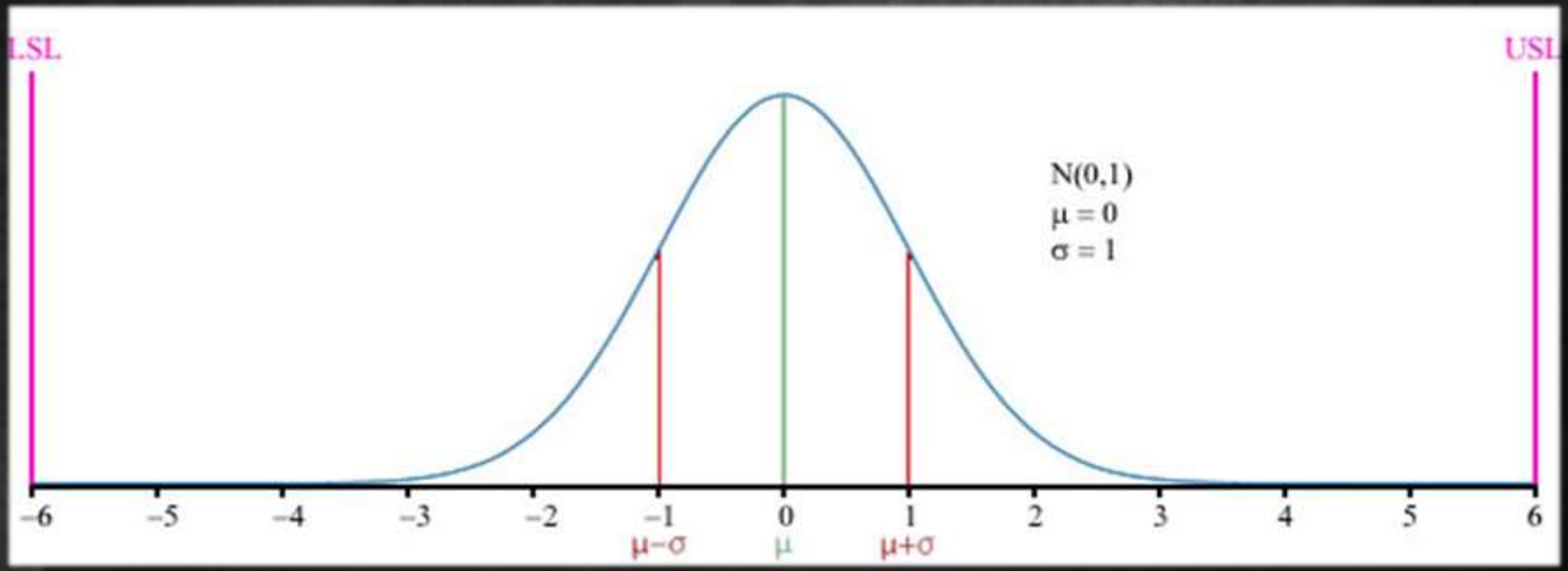
Veröffentlichen der Messergebnisse

- Dashboard
- OLAP-Würfel (Cubes)
- Listen

Darstellung aus einem Dashboard-Prototypen (Zahlen nur illustrativ).

Backup

Normalverteilung



Quelle: http://de.wikipedia.org/wiki/Six_Sigma#/media/File:6_Sigma_Normal_distribution.jpg