
Big Data

Neue Erkenntnisse aus Daten gewinnen

Thomas Klughardt

Senior Systems Consultant



Dell Software Lösungsbereiche

Transform



Data center and cloud management

- Foglight APM, Virtualization & Database
- KACE 1000/2000/3000
- Migration Manager, Recovery Manager, ActiveRoles Server & Change Auditor
- Dell Cloud Manager

Inform



Information management

- Boomi AtomSphere, Boomi MDM
- Toad Business Intelligence Suite
- Toad for Oracle, Toad for SQL Server, Toad for Cloud Databases
- SharePlex

Connect



Mobile workforce management

- SonicWALL Next-Generation Firewalls
- SonicWALL Mobile Connect
- KACE 1000/2000/3000
- Dell Workspaces - Mobile & Desktops

Protect

Security



- SonicWALL email security and anti-spam
- SonicWALL next-generation firewalls
- Dell One Identity Manager/Password Mgr
- SonicWALL Secure Remote Access

Data protection

- AppAssure/DL4000
- NetVault Backup
- Deduplication appliance: DR4100
- Email Archive Manager and Message One



Agenda

- Das Ziel
- Was bedeutet Big Data?
- Plattformen
 - NoSQL Systeme
 - Die Mischung macht's
- Herausforderungen
- Fazit



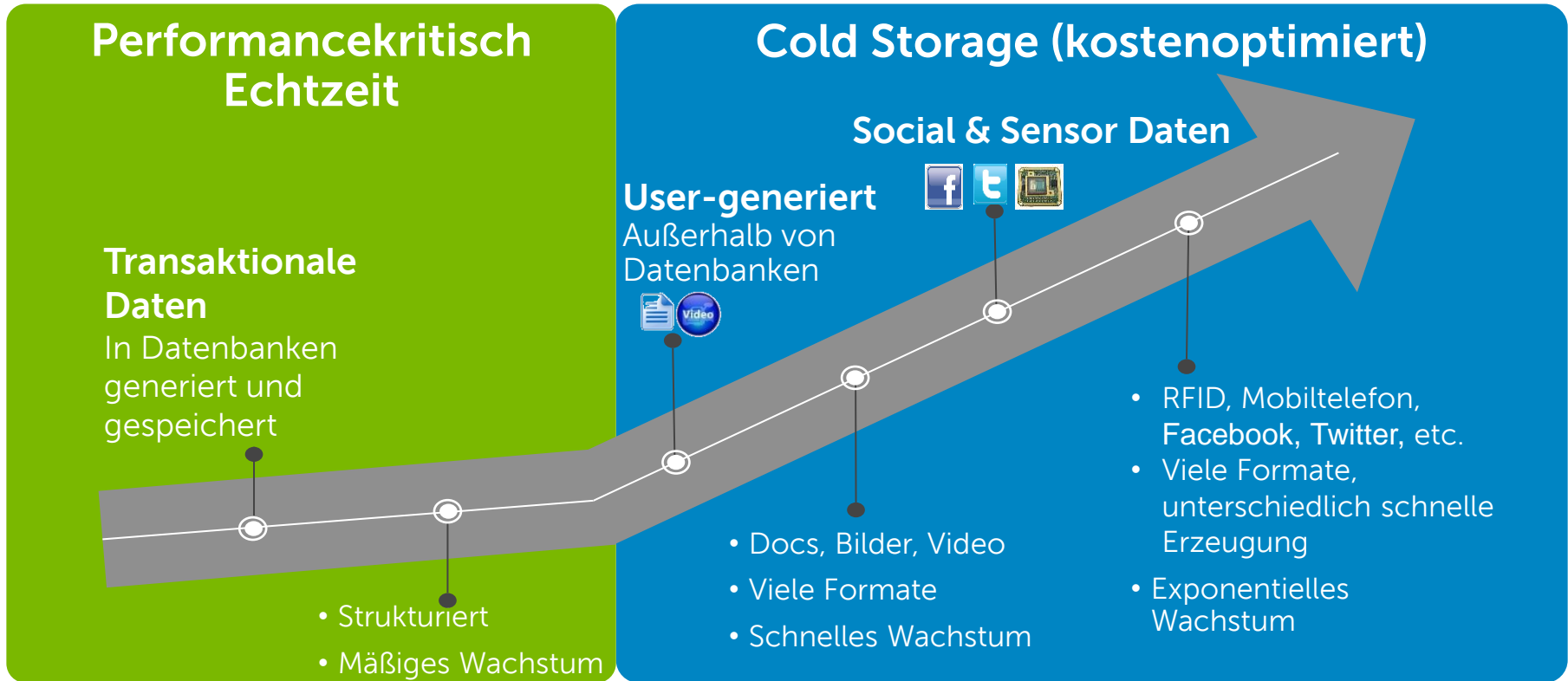
Neue Erkenntnisse – wo möchten wir hin?



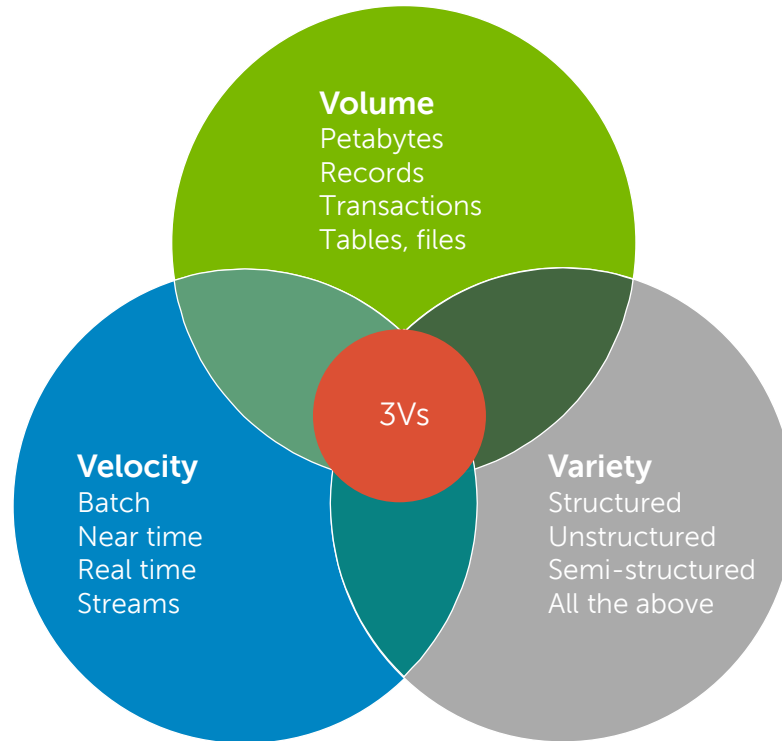
Was bedeutet Big Data?



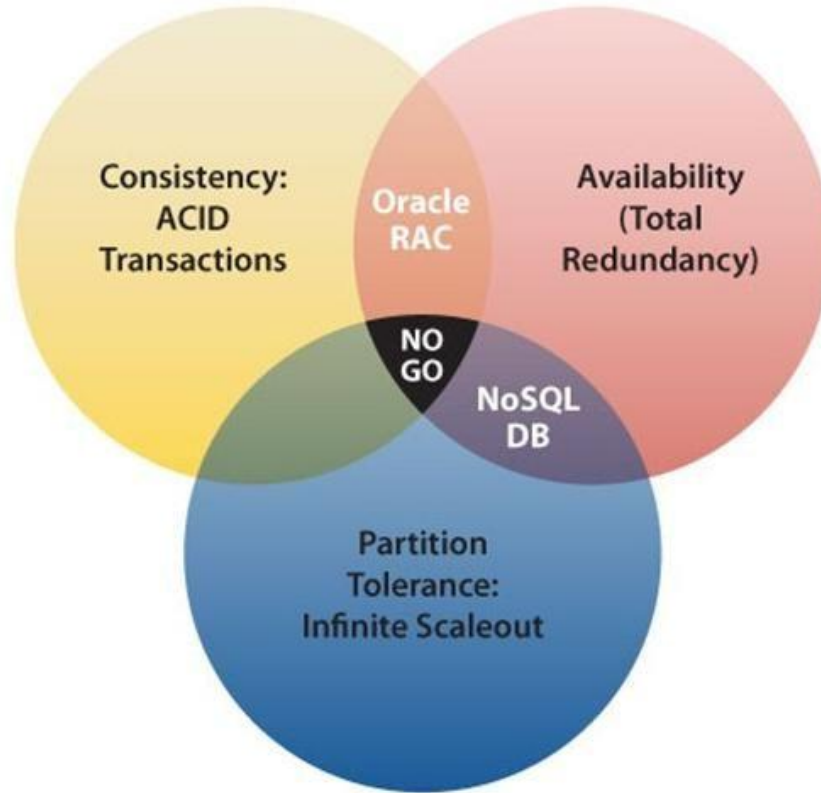
Was bedeutet Big Data?



Was bedeutet Big Data?



Das CAP Theorem



© DBPeditas.com

Plattformen

NoSQL
Systeme



Arten von NoSQL Systemen (Auszug)

- Wide Column Store / Column Families
- Document Store
- Key Value / Tuple Store
- Graph Databases
- Multimodel Databases
- Object Databases
- XML Databases
- Grid & Cloud Database Solutions
- Multidimensional Databases
- Multivalued Databases
- Event Sourcing
- Andere
 - z.B. Lotus Notes Domino

Weiterführende Informationen: <http://nosql-databases.org/>

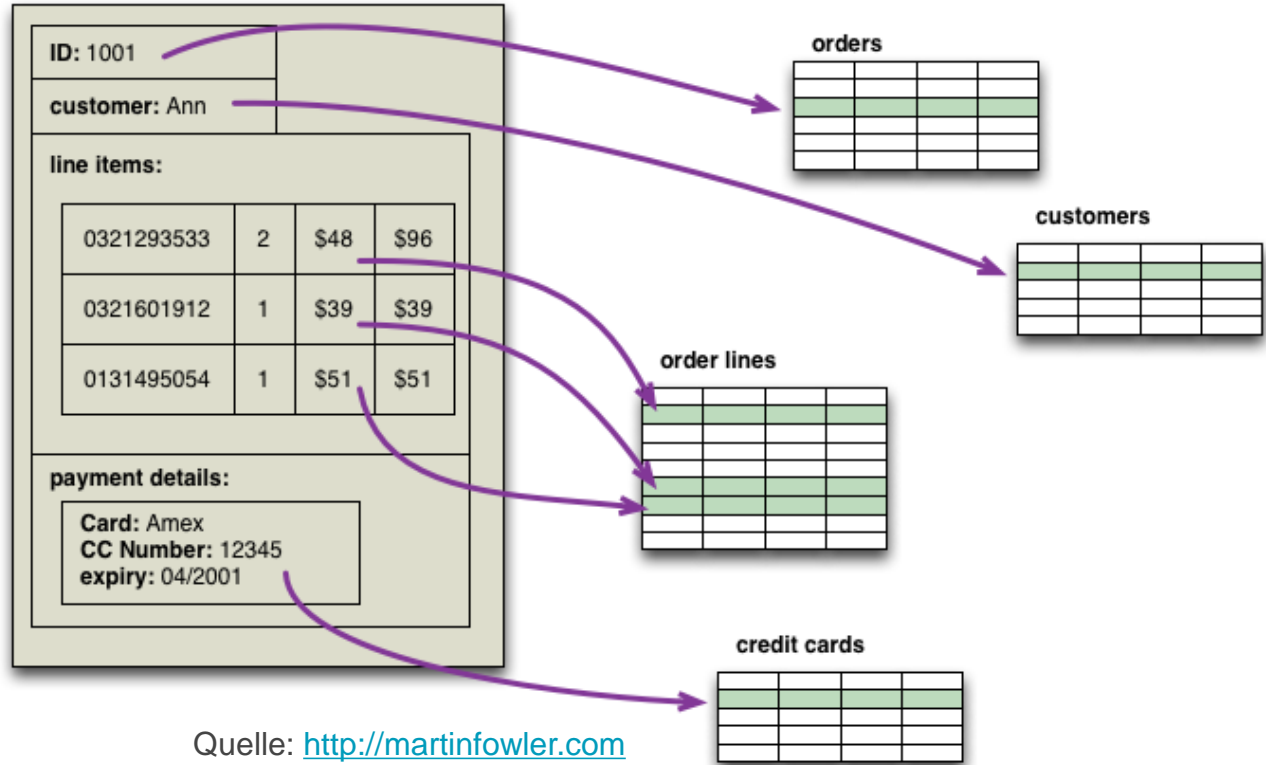


Aggregatororientierte Datenbanken

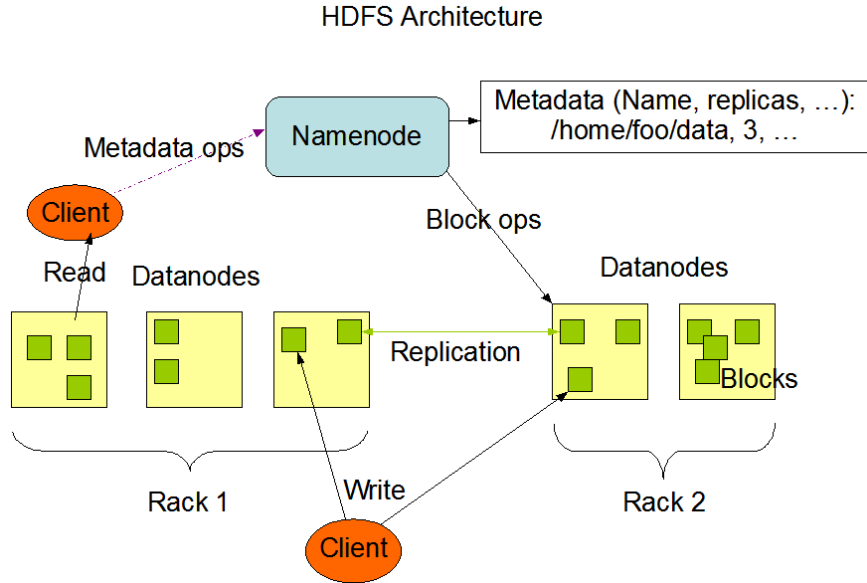
- Column Stores
- Document Stores
- Key Value Stores

- Denormalisiert
- Schnell und skalierbar

- Daten sind Aggregate



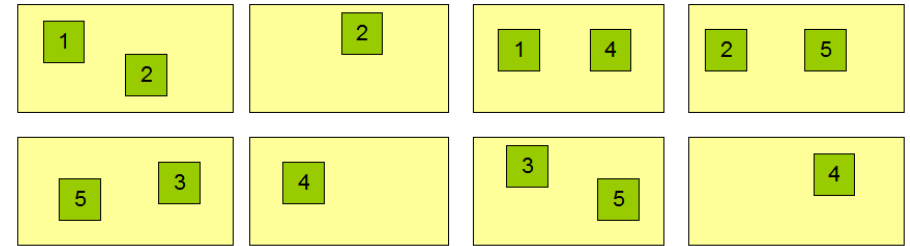
Hadoop ist erst mal nur ein Dateisystem...



Block Replication

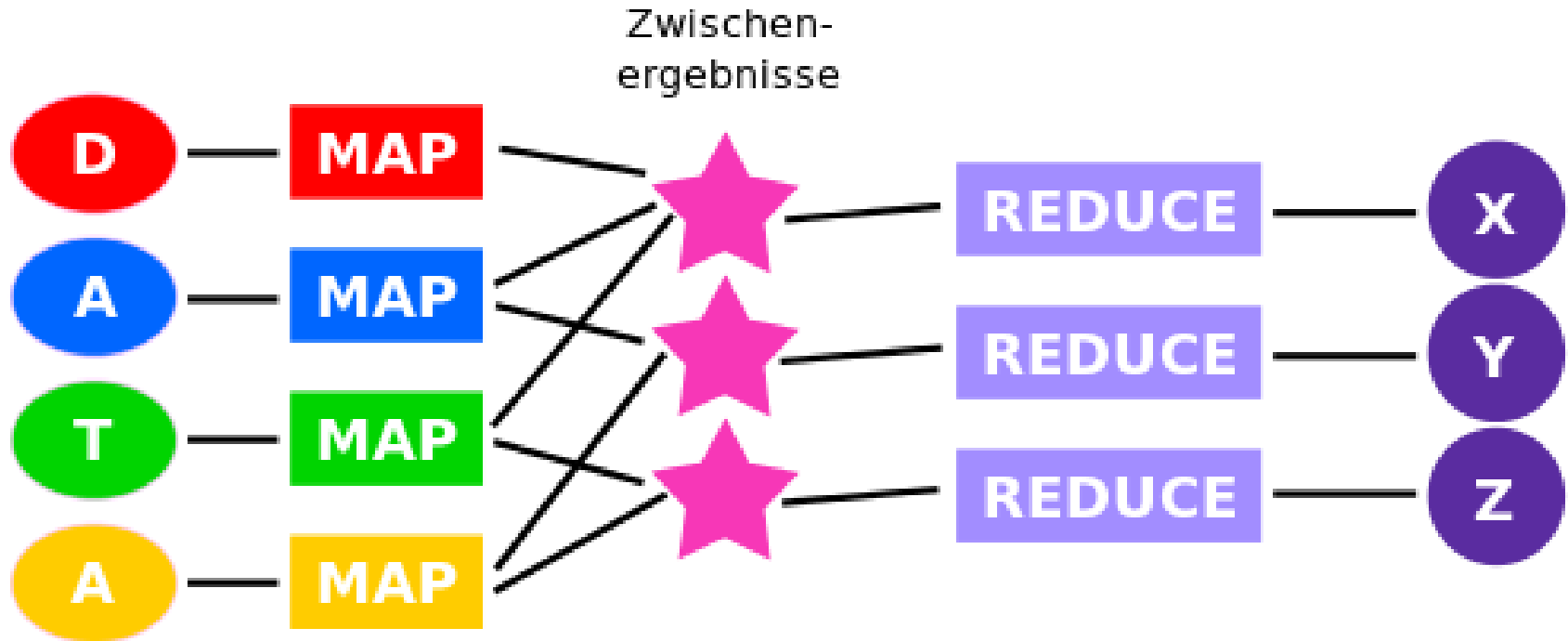
Namenode (Filename, numReplicas, block-ids, ...)
/users/sameerp/data/part-0, r:2, {1,3}, ...
/users/sameerp/data/part-1, r:3, {2,4,5}, ...

Datanodes



Quelle: Apache Commons

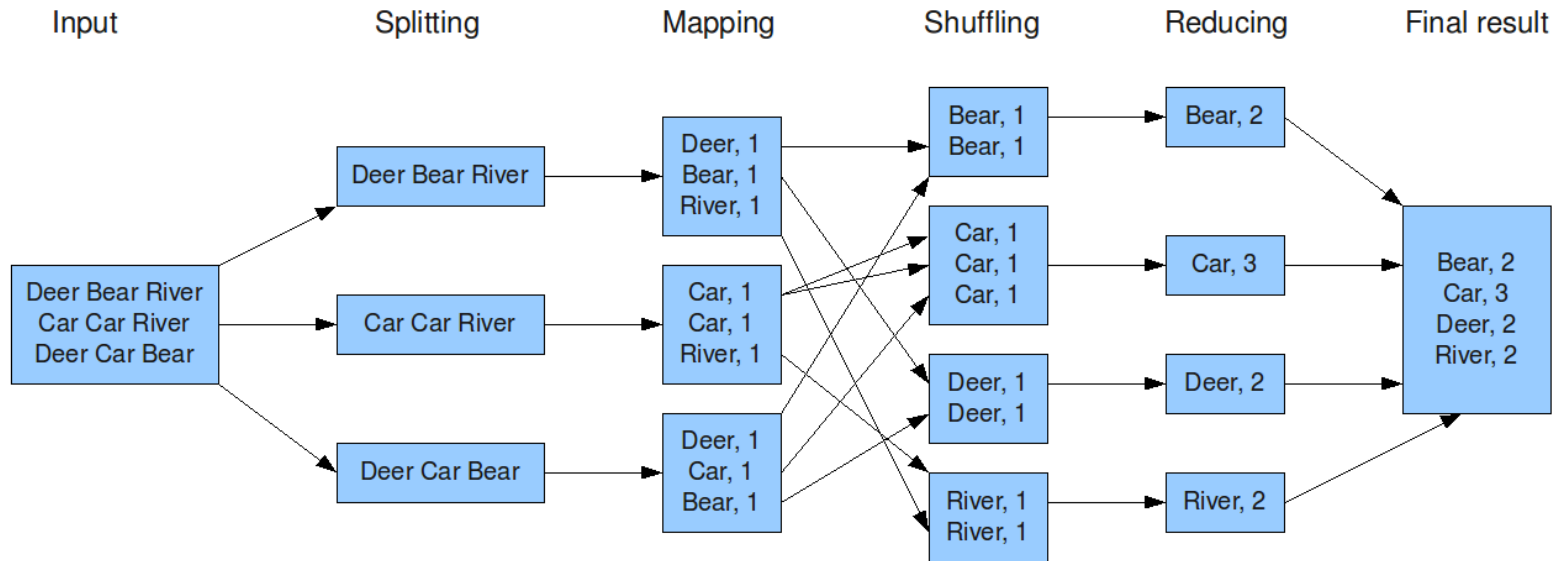
... mit einer Map-Reduce Implementierung



Quelle: Wikipedia

Ein Beispiel – WordCount

The overall MapReduce word count process



Quelle: <http://blog.trifork.com/>

Die Mischung macht's



Was ist mit Map Reduce abbildbar?

- Gut: Statistische Funktionen
 - Count, Min, Max, Average, Pivot Element, etc.
- Gut: Sortierungen (z.B. Terasort)
- Gut: Konvertierungen/Transformationen von Streams
 - MPEG -> AVI
 - WAV -> MP3
- Schlecht: Daten, die voneinander Abhängig sind (Joins)
 - Zuerst relevante Informationen extrahieren und zusammen ablegen.
 - Dann zusammenhängende Daten verarbeiten
- Schlecht: Echtzeitabfragen
 - Map Reduce ist ein Batch Processing Framework



Verschiedene Plattformen für verschiedene Dinge

- Relationale Datenbank
 - Auftragsverwaltung
 - ERP System
- Hadoop Cluster
 - Sensordaten
 - Datenhalde und Rechencluster
- Aggregatororientierte NoSQL Datenbank
 - CRM
 - Webanwendungen
- Graph Datenbanken und andere spezielle Datenbanken
 - Koordinaten, Beziehungen, Entfernungen, Kosten, etc.
 - Spezialanwendungen



Traditioneller Ansatz vs. Big Data Architektur

- Relationale Datenbank
 - Strukturiertes Schema; normalisierte Daten
 - Schema on Write
 - Verknüpfbare Daten
 - Konsistentes Modell
- Big Data Architektur
 - Mischung aus relationalen und nicht-relationalen Datenbanken
 - Erfassung und Speicherung von unstrukturierten und strukturierten Daten
 - Direkte Auswertung oder Aggregation in relationale Daten
 - Schema on Read; nach Aggregation meist Schema on Write
- Big Data \neq NoSQL
 - NoSQL Systeme normalerweise nur ein Bestandteil einer Big Data Lösung.



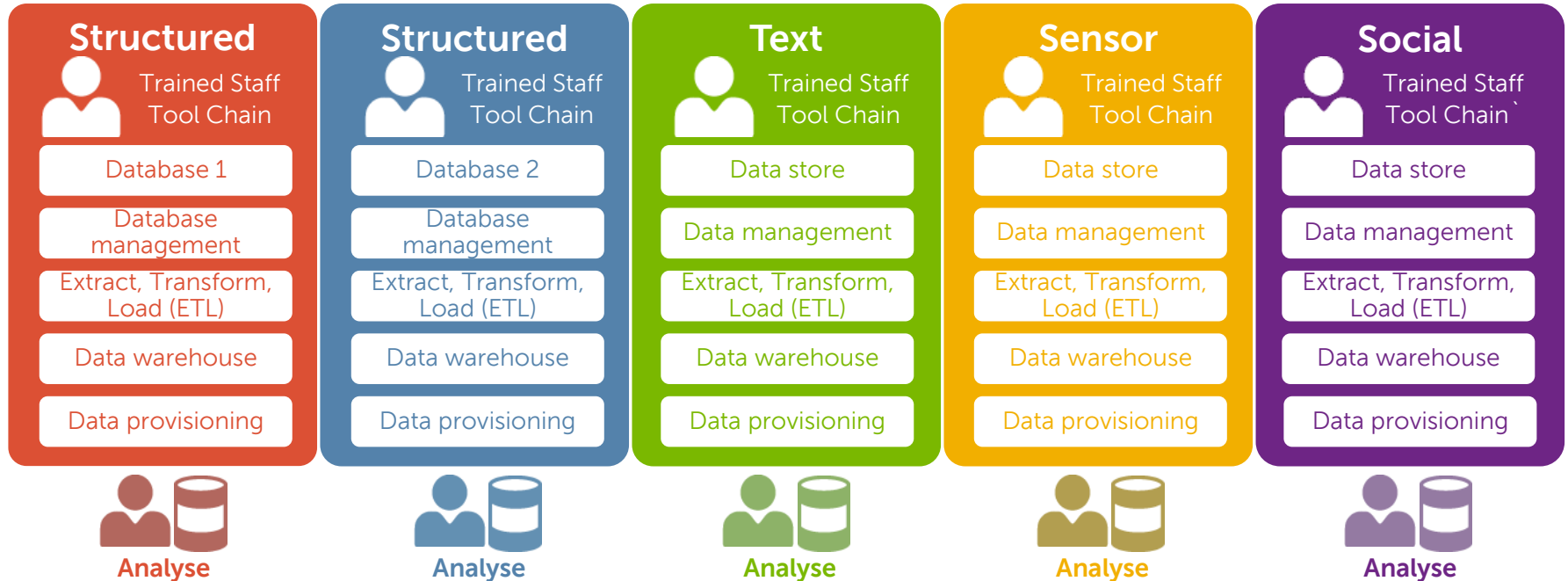
Herausforderungen

Das Ziel



Silos müssen überwunden werden.

Anwendungs- und Datenintegration



Neue Technologien und Werkzeuge

Management

Datenbank- management



Sichern,
Wiederherstellen,
Hochverfügbarkeit,
Zugriffskontrolle,
Performance

Integration

Datenintegration über Grenzen



On-Premise, Public
und Private Cloud,
Strukturiert,
Unstrukturiert,
Domänen, Systeme

Analyse

Analysen in Echtzeit Batch



Abfragen, Berichte,
Dashboards, KPIs,
Benchmarks,
Vorhersagen,
Simulationen

Fazit

Fazit

- Es gibt keine eierlegende Wollmilchsau
 - Vermeintliche Allheilmittel werden schnell entzaubert
- Big Data Plattformen erfordern zusätzliches Wissen
 - Es ist ein weiter Weg bis zur kompletten Plattform
- Die Anforderungen sind schon da und werden weiter kommen.
 - Besser verknüpfte Daten sind ein Wettbewerbsvorteil.
 - Deshalb auch besser jetzt schon damit beschäftigen.



Welche **Fragen** haben Sie?

