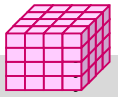
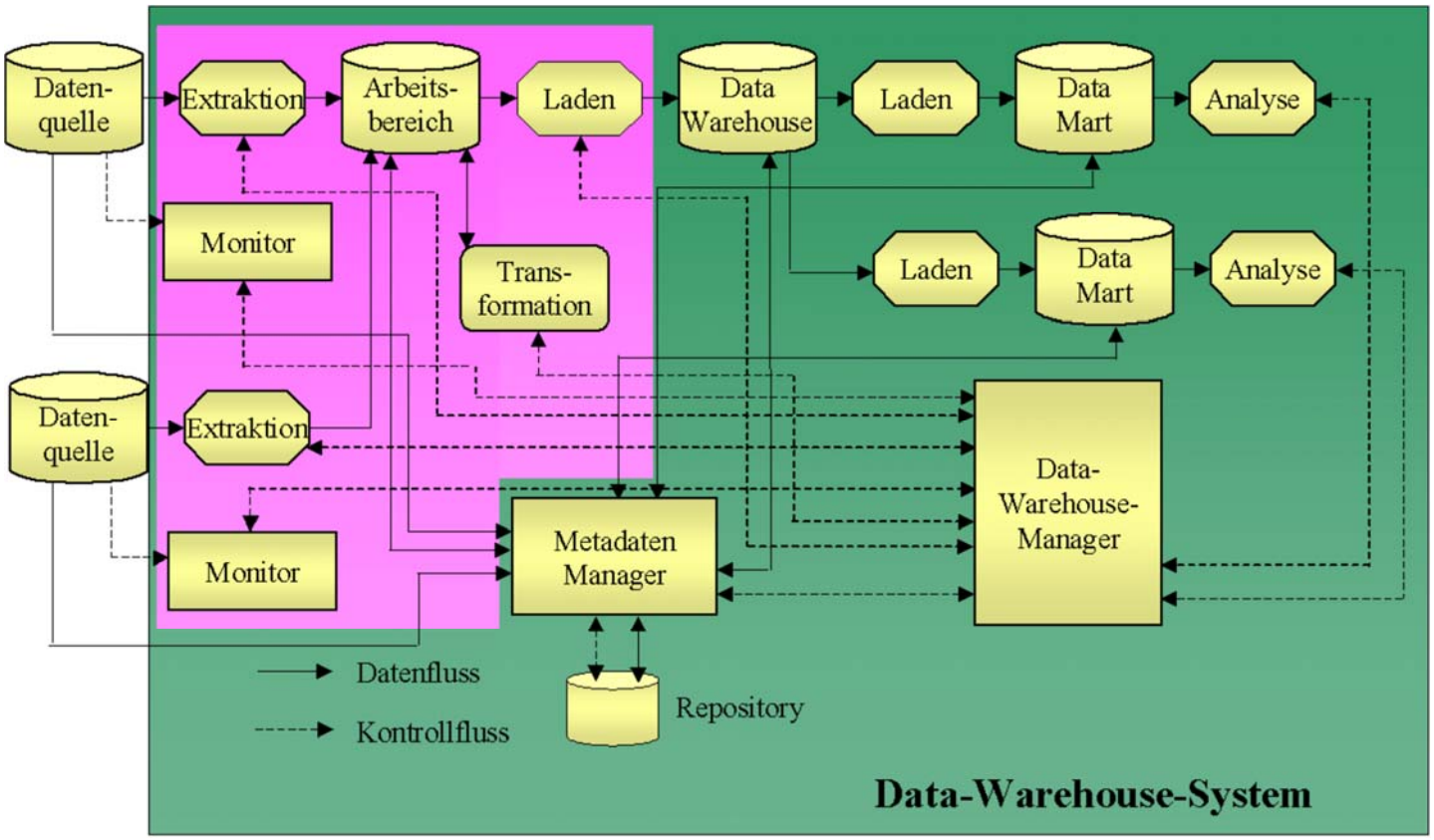
2. Architektur von Data Warehouse-Systemen

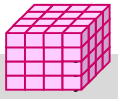
- Referenzarchitektur
  - Scheduler
  - Datenquellen
  - Datenextraktion
  - Transformation und Laden
- Abhängige vs. unabhängige Data Marts
- Operational Data Store (ODS)
- Metadatenverwaltung
  - Technische vs. fachliche Metadaten
  - CWM: Common Warehouse Model
- Data Warehouse Appliances
  - Bsp: In-MemoryWarehouse SAP HANA
- DWH und Big Data / Data Lake
- Master Data Management (MDM)



DW-Referenzarchitektur

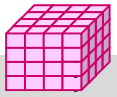


Quelle: Bauer/Günzel



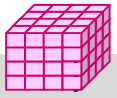
Phasen des Data Warehousing

1. Überwachung der Quellen auf Änderungen durch Monitore
2. Kopieren der relevanten Daten mittels Extraktion in temporären Arbeitsbereich
3. Transformation der Daten im Arbeitsbereich (Bereinigung, Integration)
4. Kopieren der Daten ins Data Warehouse (DW) als Grundlage für verschiedene Analysen
5. Laden der Daten in Data Marts (DM)
6. Analyse: Operationen auf Daten des DW oder DM



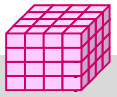
Datenquellen

- Lieferanten der Daten für das Data Warehouse (gehören nicht direkt zum DW)
- Merkmale
  - intern (Unternehmen) oder extern (z.B. Internet)
  - ggf. kostenpflichtig
  - i.a. autonom
  - i.a. heterogen bzgl. Struktur, Inhalt und Schnittstellen (Datenbanken, Dateien)
- Qualitätsforderungen:
  - Verfügbarkeit von Metadaten
  - Konsistenz (Widerspruchsfreiheit)
  - Korrektheit (Übereinstimmung mit Realität)
  - Vollständigkeit (z.B. keine fehlenden Werte oder Attribute)
  - Aktualität
  - Verständlichkeit
  - Verwendbarkeit
  - Relevanz



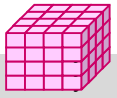
Data-Warehouse-Manager/Scheduler

- Ablaufsteuerung: Initiierung, Steuerung und Überwachung der einzelnen Prozesse
- Initiierung des Datenbeschaffungsprozesses und Übertragung der Daten in Arbeitsbereich
  - in regelmäßigen Zeitabständen (jede Nacht, am Wochenende etc.)
  - bei Änderung einer Quelle: Start der entsprechenden Extraktionskomponente
  - auf explizites Verlangen durch Administrator
- Fehlerfall: Dokumentation von Fehlern, Wiederanlaufmechanismen
- Zugriff auf Metadaten aus dem Repository
  - Steuerung des Ablaufs
  - Parameter der Komponenten



Datenextraktion

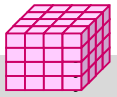
- Monitore: Entdeckung von Änderungen in einer Datenquelle
  - interne Datenquellen: aktive Mechanismen
  - externe Datenquellen: Polling / periodische Abfragen
- Extraktionskomponenten: Übertragung von Daten aus Quellen in Arbeitsbereich
  - periodisch
  - auf Anfrage
  - ereignisgesteuert (z.B. bei Erreichen einer definierten Anzahl von Änderungen)
  - sofortige Extraktion
- Performance-Probleme bei großen Datenmengen
- Autonomie der Quellsysteme ist zu wahren
- unterschiedliche Funktionalität der Quellsysteme
  - Nutzung von Standardschnittstellen (z.B. ODBC) oder Eigenentwicklung
  - Nutzung spezieller Funktionalität, z.B. von DBS-Quellen



Datenextraktion: Strategien

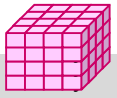
- Snapshots: periodisches Kopieren des Datenbestandes in Datei
- Trigger
  - Auslösen von Triggern bei Datenänderungen und Kopieren der geänderten Tupel
- Log-basiert
  - Analyse von Transaktions-Log-Dateien der DBMS zur Erkennung von Änderungen
- Nutzung von DBMS-Replikationsmechanismen

	Autonomie	Performanz	Nutzbarkeit
Snapshot			
Log			
Trigger			
Replikation			



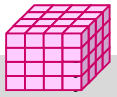
Datentransformation und Laden

- *Arbeitsbereich* (engl.: *Staging Area*)
  - Temporärer Zwischenspeicher zur Integration und Bereinigung
  - Laden der Daten ins DW erst nach erfolgreichem Abschluss der Transformation
  - Keine Beeinflussung der Quellen oder des DW
  - Keine Weitergabe fehlerbehafteter Daten
- *Transformationskomponente*: Vorbereitung der Daten für Laden
  - Data Auditing: Datenüberprüfung, Aufspüren von Abweichungen (z.B. mit Data Mining)
  - Vereinheitlich von Datentypen, Datumsangaben, Maßeinheiten, Kodierungen etc.
  - **Data Cleaning** / Scrubbing: Beseitigung von Verunreinigungen, fehlerhafte oder fehlende Werte, Redundanzen, veralteten Werte
- *Ladekomponente*: Übertragung der bereinigten und aufbereiteten (z.B. aggregierten) Daten in DW
  - Nutzung spezieller Ladewerkzeuge (z.B. Bulk Loader)
  - Historisierung: zusätzliches Abspeichern geänderter Daten anstatt Überschreiben
  - Offline vs. Online-Laden (Verfügbarkeit des DW während des Ladens)



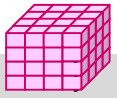
Data Warehouse

- Relationale, mehrdimensionale oder kombinierte Speicherung der Daten (ROLAP, MOLAP, HOLAP)
- oft Trennung zwischen
  - relationalen Basis-DB (Warehouse) mit Detaildaten und
  - mehreren abgeleiteten Datenwürfel (Cubes) mit aggregierten Daten
- Änderungen im (Basis-) Data Warehouse nach Laden müssen auf Cubes angewandt werden



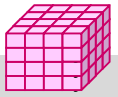
Data Marts

- Was ist eine Data Mart?
  - eine Teilmenge des Data Warehouse
  - inhaltliche Beschränkung auf bestimmten Themenkomplex oder Geschäftsbereich
- führt zu verteilter DW-Lösung
- Gründe für Data Marts
  - Performance: schnellere Anfragen, weniger Benutzer, Lastverteilung
  - Eigenständigkeit, Datenschutz
  - ggf. schnellere Realisierung
- Probleme
  - zusätzliche Redundanz
  - zusätzlicher Transformationsaufwand
  - erhöhte Konsistenzprobleme
- Varianten
  - Abhängige Data Marts
  - Unabhängige Data Marts

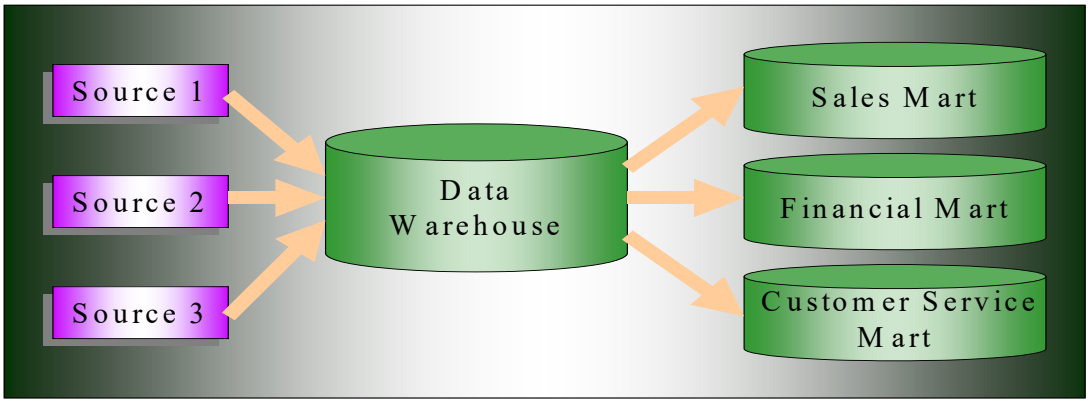


Data Warehouse vs. Data Mart

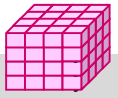
	Data Warehouse	Data Mart
Philosophie	anwendungsneutral	anwendungsbezogen
Adressat der Datenbereitstellung	Unternehmen	
Datenmenge / Detaillierungsgrad		
Umfang historischer Daten		
Optimierungsziel	Datenmenge	
Anzahl	eins (wenige)	
Typische DB-Technologie	relational	



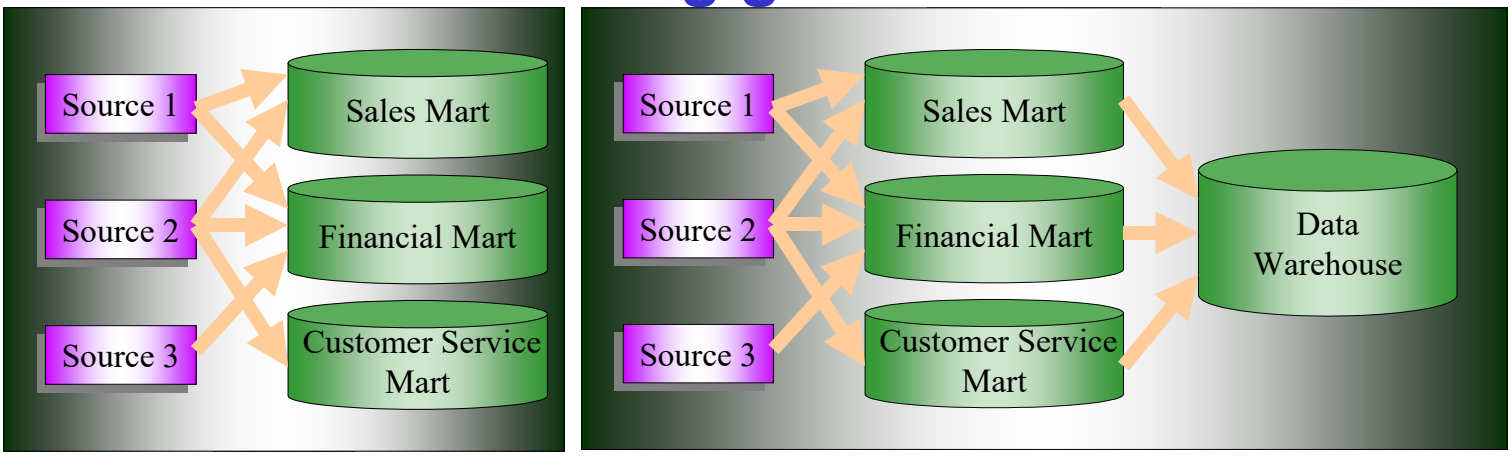
Abhängige Data Marts



- „Nabe- und Speiche“-Architektur (hub and spoke)
- Data Marts sind Extrakte aus dem zentralen Warehouse
  - strukturelle Ausschnitte (Teilschema, z.B. nur bestimmte Kennzahlen)
  - inhaltliche Extrakte (z.B. nur bestimmter Zeitraum, bestimmte Filialen ...)
  - Aggregierung (geringere Granularität), z.B. nur Monatssummen
- Vorteile:
  - relativ einfach ableitbar (Replikationsmechanismen des Warehouse-DBS)
  - Analysen auf Data Marts sind konsistent mit Analysen auf Warehouse
- Nachteil: Entwicklungsdauer (Unternehmens-DW zunächst zu erstellen)



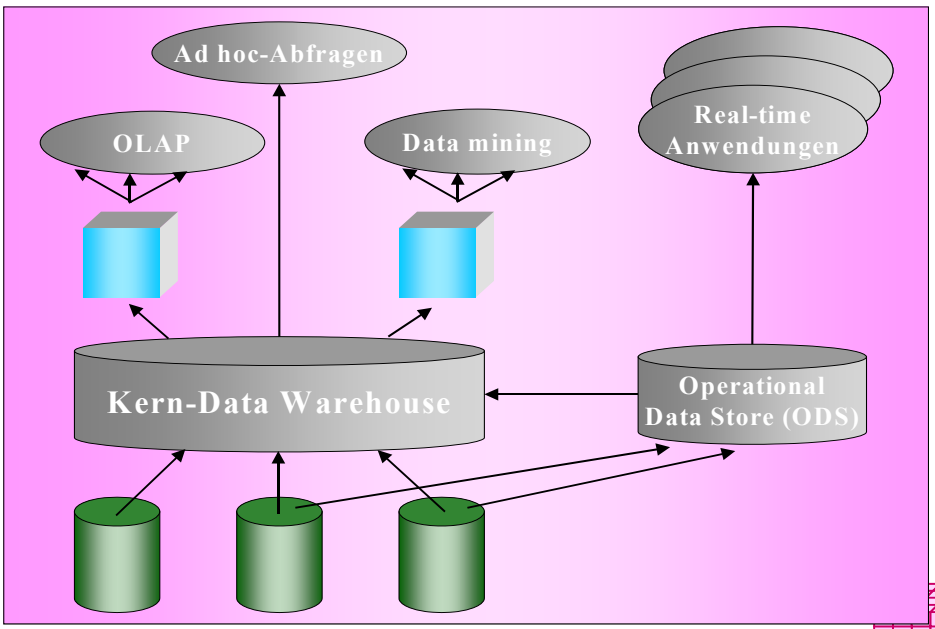
Unabhängige Data Marts



- Variante 1: kein zentrales, unternehmensweites DW
  - wesentlich einfachere und schnellere Erstellung der DM verglichen mit DW
  - Datenduplizierung zwischen Data Marts, Gefahr von Konsistenzproblemen
  - Aufwand wächst proportional zur Anzahl der DM
  - schwierigere Erweiterbarkeit
  - keine unternehmensweite Analysemöglichkeit
- Variante 2: unabhängige DM + Ableitung eines DW aus DM
- Variante 3: unabhängige DM + Verwendung gemeinsamer Dimensionen

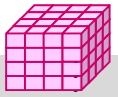
Operational Data Store (ODS)

- optionale Komponente einer DW-Architektur zur Unterstützung operativer (Realzeit-) Anwendungen auf integrierten Daten
  - grössere Datenaktualität als Warehouse
  - direkte Änderbarkeit der Daten
  - geringere Verdichtung/Aggregation, da keine primäre Ausrichtung auf Analysezwecke
- Probleme
  - weitere Erhöhung der Redundanz
  - geänderte Daten im ODS

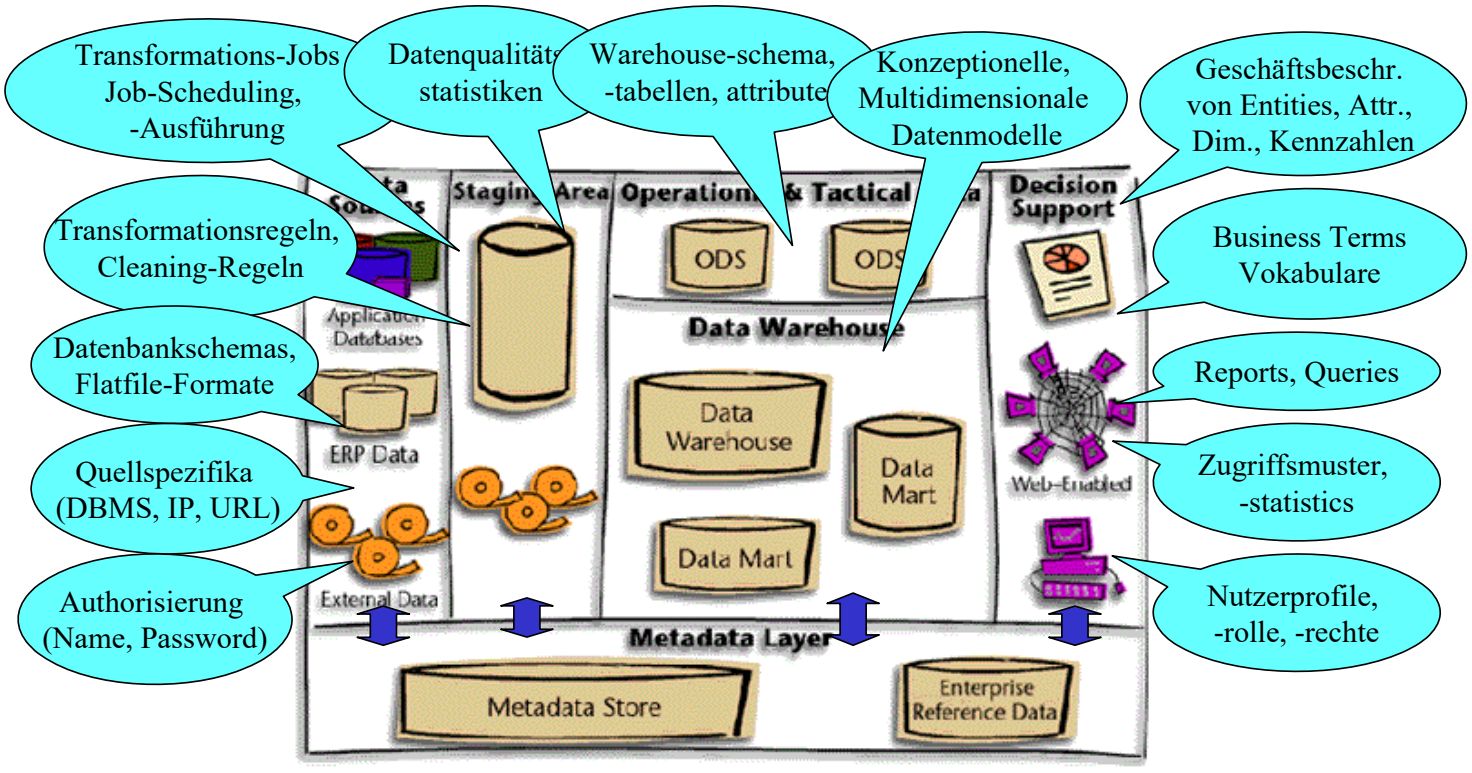


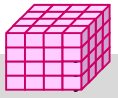
Metadaten-Verwaltung

- Anforderungen an Metadaten-Verwaltung / Repository
  - Bereitstellung aller relevanten Metadaten auf aktuellem Stand
  - flexible Zugriffsmöglichkeiten (DB-basiert) über mächtige Schnittstellen
  - Versions- und Konfigurationsverwaltung
- Unterstützung für technische und fachliche Aufgaben und Nutzer
  - Technische Metadaten vs. Business-Metadaten
- Realisierungsformen
  - werkzeugspezifisch: fester Teil von Werkzeugen
  - allgemein einsetzbar: generisches und erweiterbares Repository-Schema (Metadaten-Modell)
- zahlreiche proprietäre Metadaten-Modelle
- Standardisierungsbemühungen mit begrenztem Erfolg
  - Open Information Model (OIM) - wurde 2000 eingestellt
  - Common Warehouse Metamodel (CWM) der OMG (Object Management Group)
- häufig Integration von bzw. Austausch zwischen dezentralen Metadaten-Verwaltungssystemen notwendig



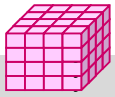
Metadaten im Data Warehouse-Kontext





Technische vs. Fachliche Metadaten

- Technische Metadaten
  - Quell-, Ziel-Systeme
    - technische Charakteristika für Zugriff (IP, Protokoll, Benutzername und Passwort, etc.)
  - Warehouse-Administration (Datenaktualisierung, -archivierung, Optimierung)
    - Systemstatistiken (Usage Patterns, nutzer-/gruppenspezifische CPU-/ IO-Nutzung, ...)
    - Logging-Information, Job-Ausführungsstatus …
  - Schemata: Datenbank-Schemata, Dateiformate
  - Datenabhängigkeiten: (technische) Mappings
    - Operationale Systeme <-> Data Warehouse/Data Marts: Datentransformations-Regeln
    - Data Warehouse/Data Marts <-> Datenzugriff-Tools: technische Beschreibung von Queries, Cubes …
- Fachliche Metadaten (Buasiness-Metadaten)
  - Informationsmodelle, konzeptuelle Datenmodelle
  - Unternehmens-/Branchen-spezifische Vokabulare (Business terms), Terminologien
  - Abbildungen Business Terms <->Warehouse/Data Mart-Elementen (Dimensionen, Attribute, Fakten)
  - Zuordnung Nutzer zu Rollen, Interessensgebieten … (Personalisierung)
  - Geschäftsbeschreibung von Kennzahlen (KPI), Queries, Reports
  - Datenqualität
    - Herkunft (lineage): aus welchen Quellen stammen die Daten? Besitzer?
    - Richtigkeit (accuracy): welche Transformation wurden angewendet?
    - Aktualität (timeliness): wann war der letzte Aktualisierungsvorgang?



Business Metadaten: Beispiel

- Business Terms für Versicherungsindustrie

Liability Insurance:

Insurance covering the legal liability of the insured resulting from injuries to a third party to their body or damage to their property.

Life Insurance:

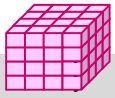
Insurance providing payment of a specified amount on the insured's death, either to his or her estate or to a designated beneficiary.

Liquor Liability Insurance:

Provides protection for the owners of an establishment that sells alcoholic beverages against liability arising out of accidents caused by intoxicated customers.

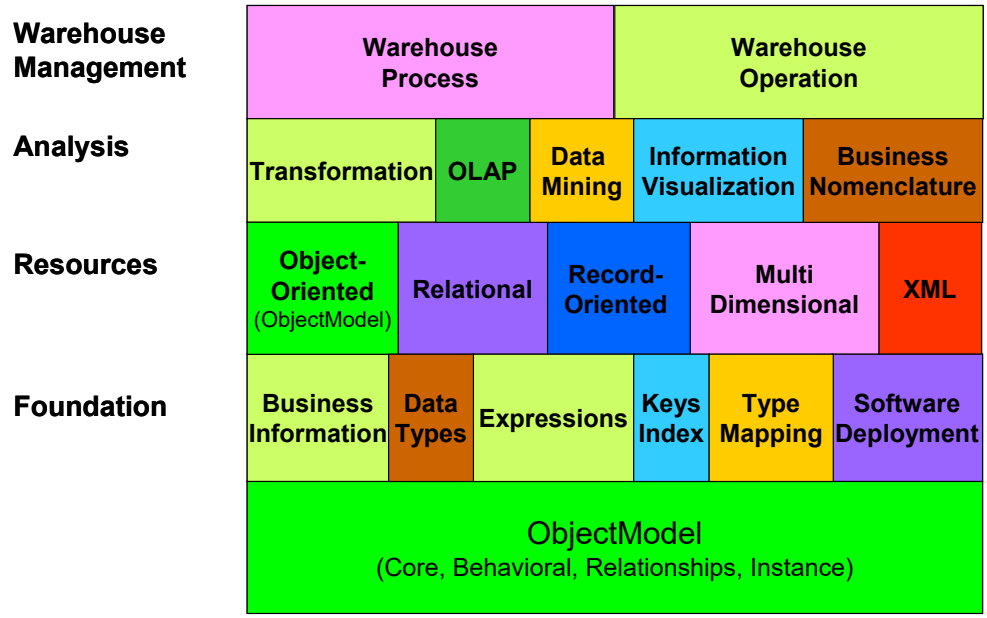
Long-Term Disability Insurance:

Insurance to provide a reasonable replacement of a portion of an employee's earned income lost through serious illness or injury during the normal work career



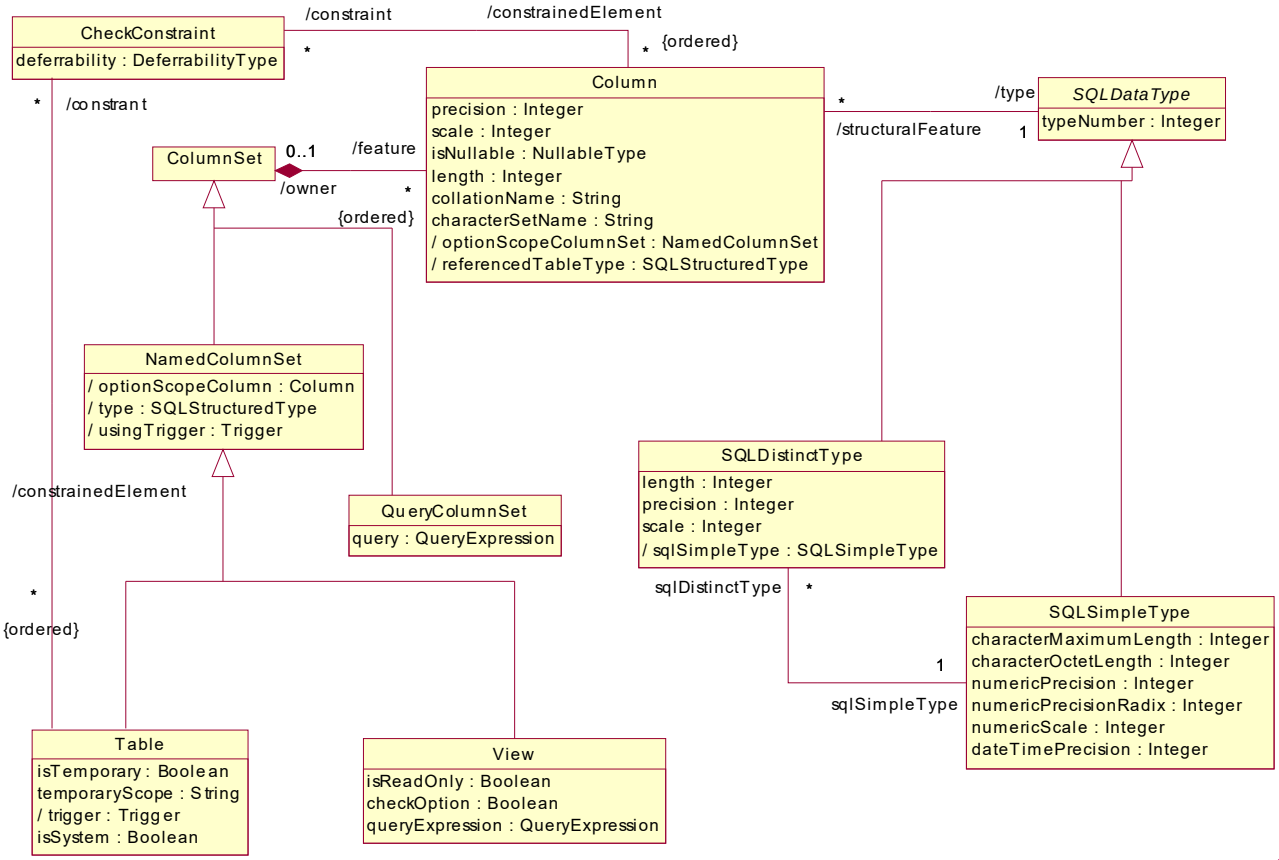
Common Warehouse Metamodel (CWM)

- umfassende UML-basierte Metadaten-Modelle für Data Warehousing
- OMG-Standard
  - CWM 1.0: 2001
  - CWM 1.1: 2002



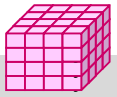
- Web-Infos: *www.omg.org* , *www.cwmforum.org*
- geringe Produktunterstützung

CWM: Relationales Teilmodell

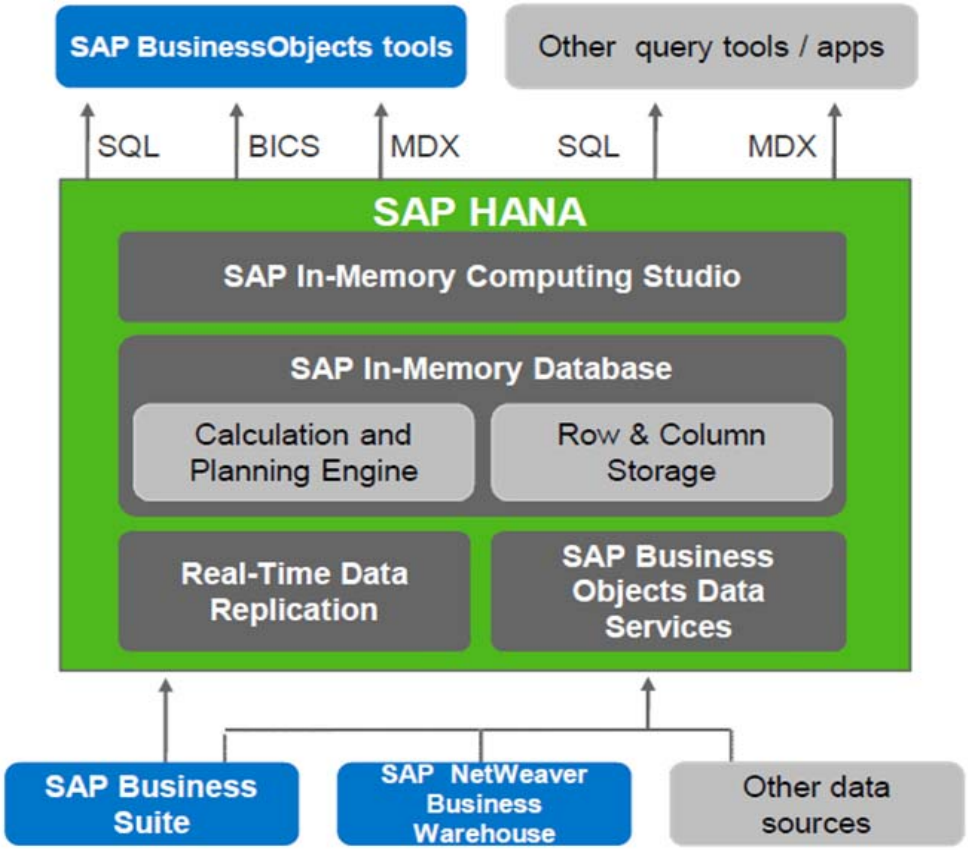


Data Warehouse Appliances

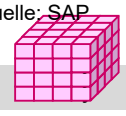
- vorkonfigurierte, komplette Data Warehouse-Installation
  - mehrere Server, Externspeicher, Software, etc.
  - Pricing nach Datenumfang (nicht Hardware)
- Pioniere mit Spezial-Hardware: Teradata, Netezza (jetzt IBM)
- allgemeine Optimierungen im Rahmen von „NewSQL“-Systemen
  (z.B. SAP HANA, EMC Greenplum, MS SQL Server PDW (Parallel Data Warehouse), HP Vertica, EXASOL …)
  - Nutzung riesiger Hauptspeicher-Datenbanken (In-Memory Data Warehouses)
  - Column-Store-Techniken mit Datenkompression
  - Parallelverarbeitung
  - Nutzung von Flash-Speichern (Solid-State Disks) statt Magnetplatten
- Vorteile von Data Warehouse Appliances
  - hohe Leistung / Skalierbarkeit
  - hohe Verfügbarkeit (durch eingebaute Fehlerbehandlungsmechanismen)
  - geringer Administrationsaufwand
  - schnelle DWH-Realisierung/Nutzung



In-Memory-Datenbanktechnologie

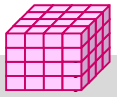


Quelle: SAP



SAP HANA - Merkmale

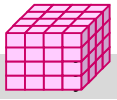
- dramatische Beschleunigung der DB-Verarbeitung
  - Vermeidung langsamer Plattenzugriffe
  - neue auf In-Memory-Verarbeitung zugeschnittene Datenstrukturen und Algorithmen
  - Vermeidung von Indexstrukturen, Cubes etc.
- gleichzeitige Unterstützung von OLTP + OLAP
  - Record Store und Column Store
  - hohe Datenaktualität
- Einschränkungen
  - ETL/Datenintegrationsaufgaben für Daten außerhalb der In-Memory-DB bleiben bestehen
  - geschlossene Umgebung
  - hohe Kosten



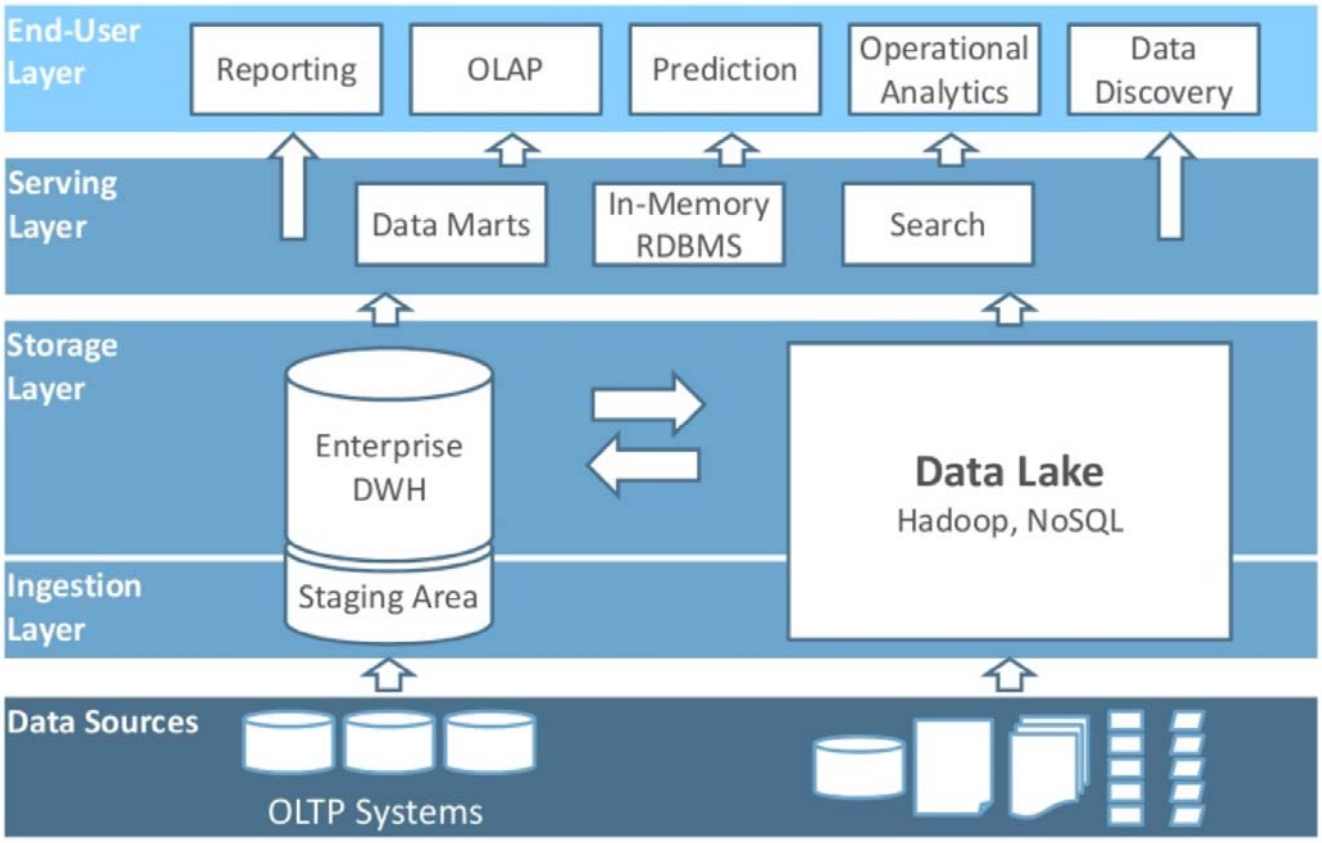
Data Warehouse vs Big Data

- Data Warehouse
  - Analyse strukturierter Daten auf Basis eines Datenbankschemas
  - Schema first / Schema on Write
- Big-Data-Technologien (z.B. Hadoop/Spark) komplementär nutzbar
  - Beschleunigung der ETL-Prozesse
  - Erfassung und Analyse unstrukturierter und hoch-dynamischer Daten (Texte, Bilder, Datenströme …)
  - Erfassung zunächst in schemalosem Data Lake
  - Datenaufbereitung bei Bedarf („Schema on Read“)
  - Analyse mit Frameworks wie Apache Spark/Flink, Textsuche etc.
  - Datenaustausch Data Lake – Data Warehouse möglich

www.pmone.com



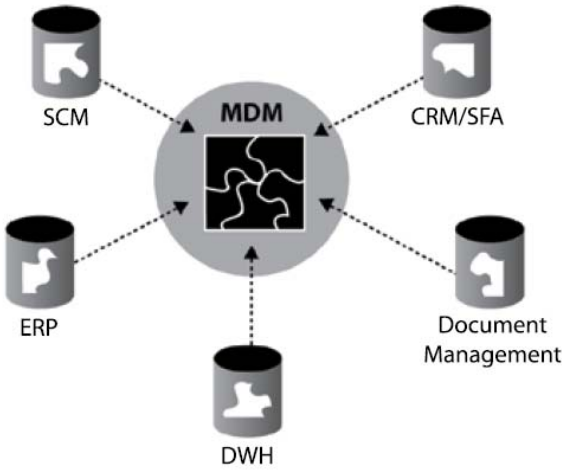
BI-Architektur mit DWH und Data Lake



Quelle: J. Albrecht

Master Data Management (MDM)

- Bereitstellung integrierter Stammdaten (Masterdaten, Referenzdaten) nicht nur für Analyse, sondern auch für operative Anwendungen und Geschäftsprozesse
  - Kundendaten, Produktdaten, Mitarbeiterdaten, …
  - Abgrenzung zu transaktionalen Daten (Bewegungsdaten)
- Ziel: „Single View of Truth“
- ähnliche Probleme wie DWH, z.B. Duplikatbehandlung
- spezialisierte Lösungsansätze
  - CDI (Customer Data Integration)
  - PIM (Product Information Management): Produkte, Hersteller, Preise
- MDM Teil von Anwendungs-architekturen (SOA)
  - z.B. von SAP, IBM , Oracle, Microsoft ….
  - Replikation/Caching von Masterdaten in Anwendungen mit Änderungsmöglichkeit

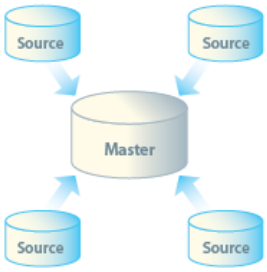


Quelle: IBM

MDM-Architektur

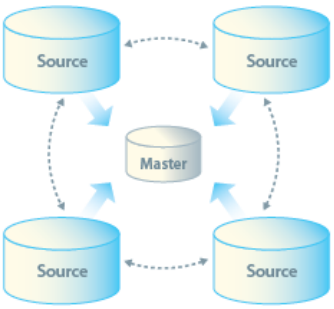
- Materialisierte oder virtuelle Realisierung eines MDM-Hubs

Consolidation



- Master is single version of truth
- Data quality management at master
- Updates occur at sources
- Updates propagated to master

Registry

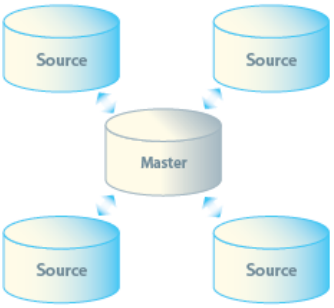


- Multiple versions of truth
- Data quality management is ongoing
- Updates occur at sources
- Keys and metadata updated in registry
- Updates propagated to other sources (optional)

Transaction

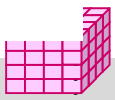
- Master is single version of truth
- Data quality management at master
- Updates occur at master
- Updates propagated to sources

Coexistence



- Master is single version of truth
- Data quality management is ongoing
- Updates occur at sources or master
- Updates propagated to other sources

Quelle: Information Builders, White Paper 2010



Zusammenfassung

- wesentliche Komponenten der Referenzarchitektur
  - ETL-Komponenten inklusive Monitoring und Scheduling
  - Arbeitsbereich (Staging Area)
  - Data Warehouse/Cubes und Data Marts
  - Metadaten-Verwaltung
- Extraktionsansätze: Snapshot, Trigger, Log-Transfer, DBS-Replikationsverfahren
- abhängige vs. unabhängige Data Marts
- Unterstützung operativer Anwendungen auf integrierten Daten
  - ODS: Online Data Store
  - MDM: Master Data Management
- zunehmender Einsatz voroptimierter DWH Appliances mit In-Memory-Technologie, Column Stores, …
- Ko-Existenz DWH und Big-Data-Technologien / Data Lakes
  - flexible Unterstützung unstrukturierter und hoch-dynamischer Daten
  - parallele ETL-Verarbeitung und Datenanalyse