

# Information Engineering 2

## Data Warehousing Einführung

Prof. Dr. Kurt Stockinger

# Prof. Dr. Kurt Stockinger



- Dozent an der ZHAW seit 1. August 2013
- 2007-2013:  
Data Warehouse & Business Intelligence Architect bei **Credit Suisse**, Zürich & Forschungsprojekte mit ETH Zürich
- 2004-2007:  
Forschungstätigkeit am **Lawrence Berkeley National Laboratory**, Berkeley, Kalifornien (Scientific Data Management)
- 2002-2003:  
Forschungstätigkeit am **CERN** (Grid Computing)
- 1999-2001:  
Doktorat in Informatik am **CERN** (Zugriffsoptimierung für objektorientierte Datenbanken)  
Gastforscher am **California Institute of Technology**, Pasadena, Kalifornien

# Über diesen Kurs

- Zwei Themen: **Data Warehousing (DWH) & Big Daten:**
  - Wie man strukturierte Daten aufbereitet, modelliert und für die Analyse bereitstellt
  - Wie man skalierbare Analysesysteme mit Big Data Technologie aufbaut und nutzt.
- Unterlagen, Organisatorisches, Praktika, ... finden Sie auf Moodle  
URL: <https://moodle2.zhaw.ch>  
Name: **Information Engineering 2 - FS 2022**



# Semesterplan

SW	Datum	Vorlesungsthema	Praktikum
1	23.02.2022	Data Warehousing Einführung	Praktikum 1: KNIME Tutorial
2	02.03.2022	Dimensionale Datenmodellierung 1	Praktikum 1: KNIME Tutorial (Vertiefung)
3	09.03.2022	Dimensionale Datenmodellierung 2	Praktikum 2: Datenmodellierung
4	16.03.2022	Datenqualität und Data Matching	Praktikum 3: Star-Schema, Bonus: Praktikum 4: Slowly Changing Dimensions
5	23.03.2022	Big Data Einführung	DWH Projekt - Teil 1
6	30.03.2022	Spark - Data Frames	DWH Projekt - Teil 2 (Abgabe: 4.4.2022 23:59:59)
7	06.04.2022	Data Storage: Hadoop Distributed File System & Parquet	Praktikum 1: Data Frames
8	13.04.2022	Query Optimization	Praktikum 2: Data Storage
9	20.04.2022	Spark Best Practices & Applications	Praktikum 3: Query Optimization & Performance Analysis
10	27.04.2022	Machine Learning mit Spark 1	Praktikum 3: Query Optimization & Performance Analysis (Vertiefung)
11	04.05.2022	Machine Learning mit Spark 2 + Q&A	Praktikum 4: Machine Learning (Regression)
12	11.05.2022	NoSQL Systems	Big Data Projekt - Teil 1
13	18.05.2022	Keine Vorlesung (Arbeit am Projekt)	Big Data Projekt - Teil 2
14	25.05.2022	Keine Vorlesung (Arbeit am Projekt)	Big Data Projekt - Teil 3 (Abgabe: 30.5.2022 23:59:59)

# Kurslogistik

«Most of the things you *need* will *be brought* to you;  
most of the things you *want* you have to *go get*»

## Lektionen

Der Unterricht beginnt pünktlich, und endet pünktlich

Laptops etc. werden nur für den Unterricht benutzt



## Selbststudium

2+2 ECTS =      1½ h Vorlesung,  
                      1½ h Praktikum und  
                      3 h Selbststudium (ca.) pro Woche (+ Prüfungsvorbereitung)

## Modulnote

Semesterendprüfung (90 min., 1 A4-Blatt, d.h. 2 A4-Seiten handschriftlich)	70%
Praktika (DWH 10%, Big Data 20%)	30%

# Praktika

- Einteilung in Übungsgruppen siehe Stundenplan
- Bewertung: **Praktika + 2 Mini-Projekte:**
  - Data Warehousing:
    - Praktikum 2 und 3 jeweils 1 Punkt (Abgabe + kurze Vorstellung im Praktikum)
    - DWH Projekt (8P): Arbeit in 2er Gruppe
  - Big Data:
    - Praktikum 2,3 und 4 jeweils 2 Punkte (Abgabe + kurze Vorstellung im Praktikum)
    - Big Data Projekt (14P): Arbeit in 2er oder 3er Gruppe
- Bewertet werden **zwei Kurzberichte von max. je 10 Seiten** (Power Point Präsentation)

# ZHAW Datalab: Est. 2013

(zhaw.ch/datalab)

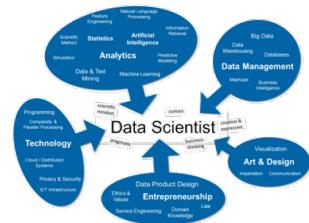
## Forerunner

- **One of the first** interdisciplinary data science initiatives in Europe
- One of the first interdisciplinary labs at ZHAW



## Foundation

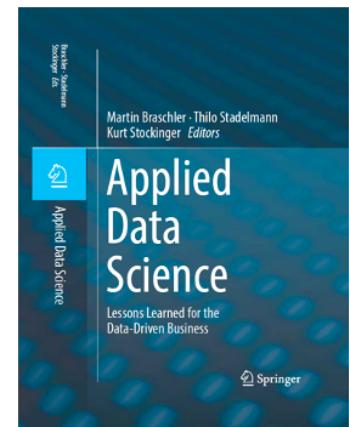
- **People:** ca. 130 researchers from 11 institutes and centers across 4 departments
- Vision: Nationally leading and internationally recognized center of excellence
- Mission: Generate projects through critical mass and mutual relationships
- Competency: Data product design with structured and unstructured data

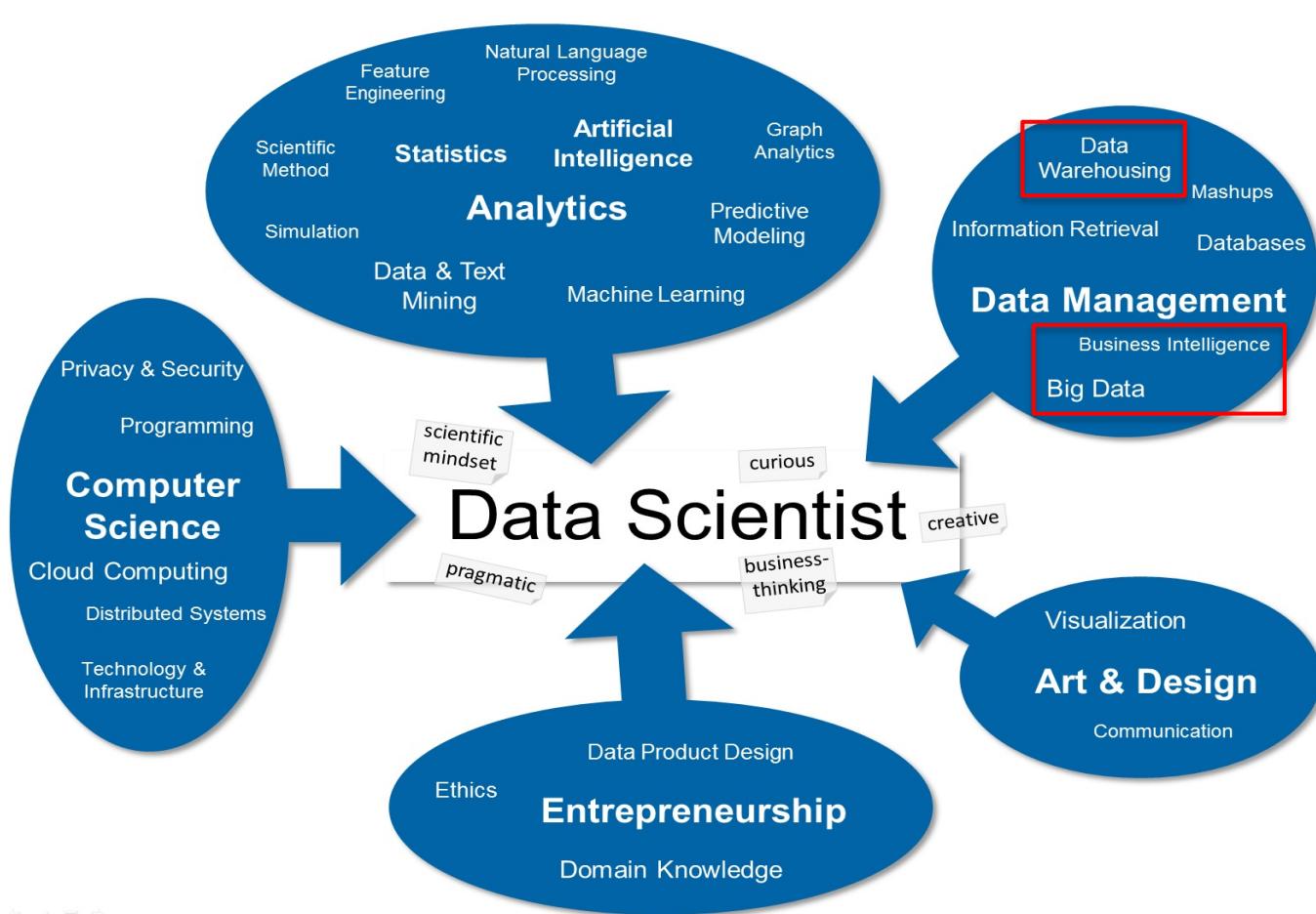


## Success factors

- **Lean** organization and operation → geared towards projects
- Years of successful **pre-Datalab collaboration**
- Founder of Swiss Conference on Data Science

<https://www.sds2022.ch/>



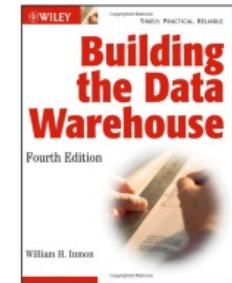
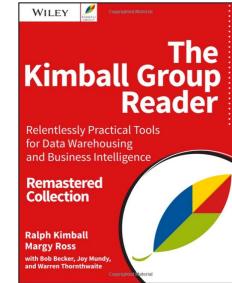


T. Stadelmann, K. Stockinger, M. Braschler, M. Cieliebak, G. Baudinot, O. Dürr, A. Ruckstuhl,  
**Applied Data Science in Europe.** In: European Computer Science Summit. ECSS 2013. Amsterdam, The Netherlands: IEEE.  
<http://pd.zhaw.ch/publikation/upload/204718.pdf>

# Data Warehousing

# Literatur

- Kimball, R., & Ross, M. (2015). *The kimball group reader: Relentlessly practical tools for data warehousing and business intelligence remastered collection.* John Wiley & Sons.
- Inmon, W. H. (2005). *Building the data warehouse.* John Wiley & Sons.
- Ehrenmann, Pieringer, R., Stockinger, K. (2012). Is there a Cure-All for Business Analytics? Case Studies of Exemplary Businesses in Banking, Telecommunications, and Retail, *Business Intelligence Journal*, Vol 17, No.3, pp. 28-39



# Use Case einer Grossbank

- **Filialen** in Zürich, New York, Singapur
- **Aufgaben:**
  - Monatsendabrechnung:
    - Net New Assets (Neugeldzufluss)
    - Assets under Management (verwaltete Bankprodukte)
  - Performance Analyse:
    - Welche Produkte sind am rentabelsten?
  - Marketing:
    - Welchen Kunden soll man das neue Produkt „High-Interest-Yielder“ anbieten?
- **Ausgangslage:**
  - Alle Filialen haben Daten in mehreren Excel-Sheets gespeichert
  - Daten kommen von unterschiedlichen Systemen mit unterschiedlichem Aktualitätsgrad
  - Die Berechnungen laufen zu unterschiedlichen Zeiten
  - Buchhaltungssysteme gemäss lokalen Vorschriften



# Herausforderungen

Welche Probleme können bei diesen Berechnungen auftreten?

# ... und Lösungen

- Unterschiedliche Excel-Sheets geraten **ausser Synchronisation**:
  - Keine gemeinsame goldene Quellen (z. B. Produkte)
- Unterschiedliche **Berechnungsmethoden**
- Erkennt man einen **Berechnungsfehler**, so müssen alle Excel-Sheets ausgebessert werden
- Unterschiedliche **Zeitzonen** führen zu unterschiedlichen Ausgangslagen für Berechnungen

→ Abhilfe schafft ein **Data Warehouse**

# Ziel: Data Warehouse (DWH)

- Integration unterschiedlicher Daten in ein System (goldene Quelle)
- Daten werden einmal pro Tag in das DWH geladen
- Alle Applikationen haben eine gemeinsame Datenbasis mit zentraler Berechnung



# Begriffsdefinition

Wir lernen folgende Begriffe kennen:

- Decision Support System
- Data Warehouse
- Business Intelligence
- Expert System

# Definition: Decision Support System

Entscheidungsunterstützungssysteme sind Softwaresysteme, die für menschliche **Entscheidungsträger** für **operative und strategische Aufgaben** relevante Informationen ermitteln, aufbereiten, übersichtlich zusammenstellen und bei der Auswertung helfen.

[Wikipedia](#)

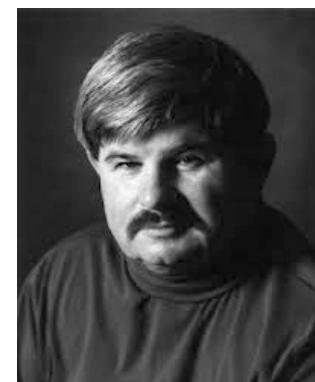
System	Einführung	Fokus
Decision Support Systems (DSS)	Anfang 1970er	Modellfokus
Executive Information Systems (EIS)	Ende 1980er	Präsentationsfokus
Data Warehouse (DWH)	Anfang 1990er	Datenfokus
OLAP (Online Analytical Processing)	Anfang 1990er	Modellfokus
Business Intelligence	Anfang 1990er	Präsentationsfokus
Advanced Analytics	Ab ca. 2012	Analysefokus

# Definition: Data Warehouse

A data warehouse is a

- subject-oriented, -- getrennt nach Fachbereichen (sachorientiert)
- integrated, -- integriert mit einheitlichem Datenmodell
- time-variant, -- Datenzustände zu unterschiedl. Zeitpunkten
- nonvolatile -- dauerhaft gespeichert und nicht mehr geändert

collection of data in support of management's decision-making process.

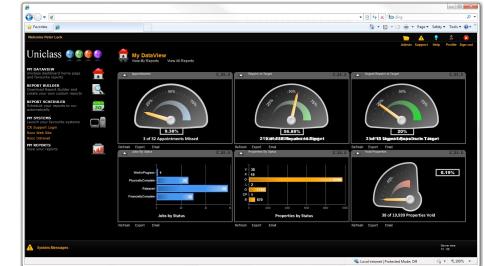


Inmon

- Begriff in 1980ern durch IBM geprägt

# Definition: Business Intelligence

Verfahren und Prozesse zur **systematischen Analyse** (Sammlung, Auswertung und Darstellung) von Daten in elektronischer Form.



Ziel ist die **Gewinnung von Erkenntnissen**, die in Hinsicht auf die Unternehmensziele **bessere** operative oder strategische **Entscheidungen** ermöglichen.

[Wikipedia](#)

- Erstmalige Begriffsverwendung in IBM Journal 1958
- Popularität durch Gartner 1989
- Heute wird oft der Begriff „Analytics“ verwendet

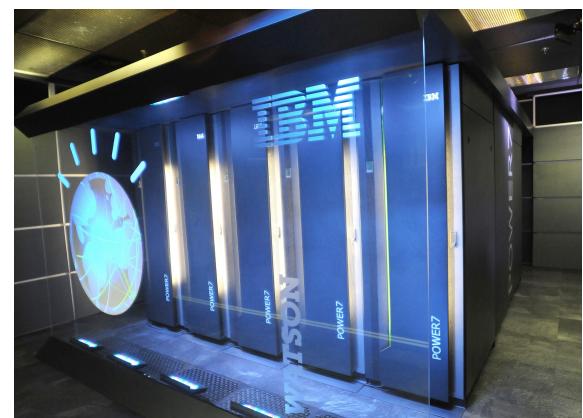
# Abgrenzung zu Expertensystemen

Ein **Expertensystem** ist ein Computerprogramm, das Menschen bei der **Lösung von komplexeren Problemen** wie ein Experte unterstützen kann, indem es Handlungsempfehlungen aus einer Wissensbasis ableitet.

Expertensysteme sind ein Teilbereich der **künstlichen Intelligenz**. Beispiele sind Systeme zur Unterstützung **medizinischer Diagnosen** oder zur Analyse wissenschaftlicher Daten.

[Wikipedia](#)

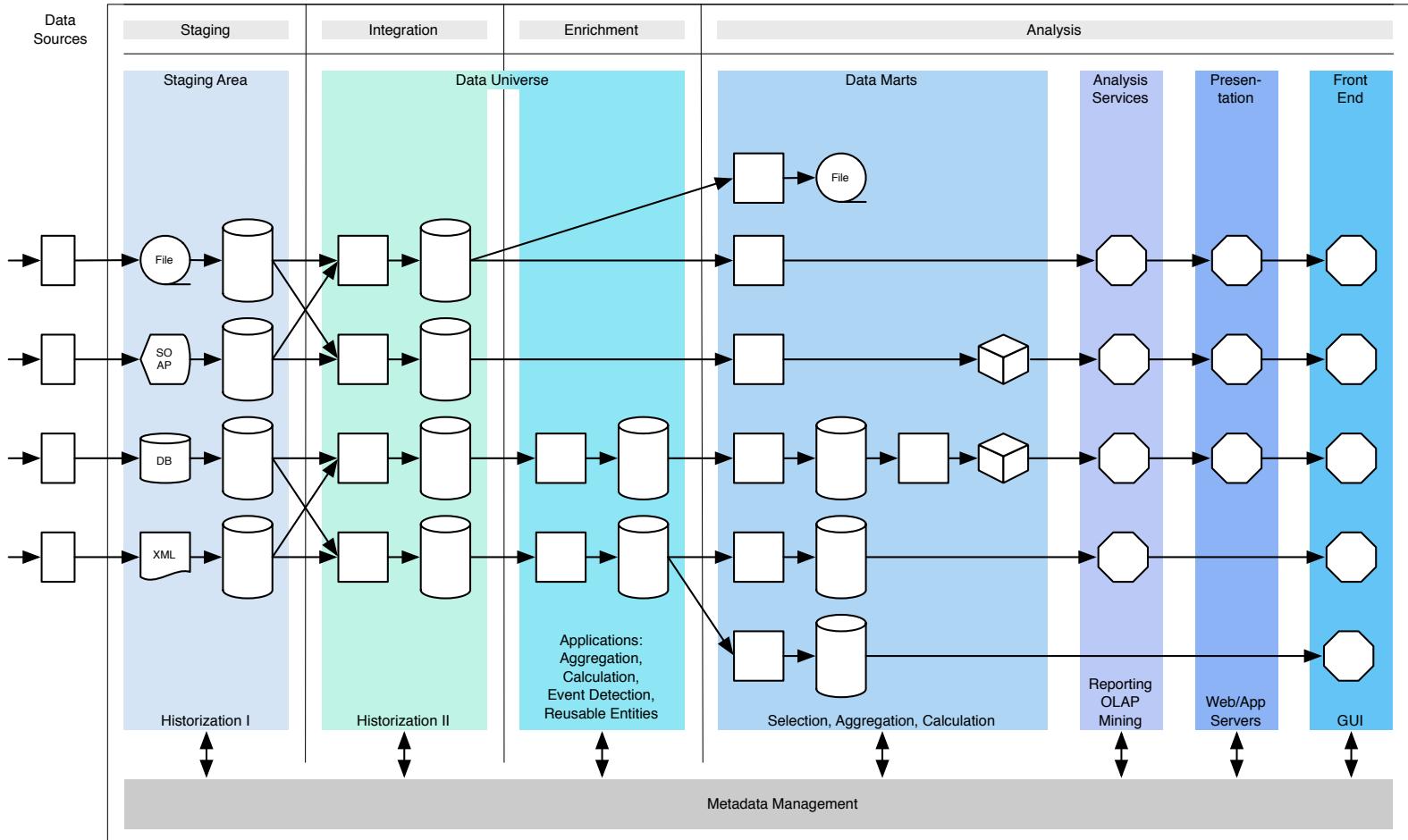
- Erste Arbeiten stammen aus den 1960ern
- Seit den 1980ern in kommerziellem Einsatz



IBM Watson

19

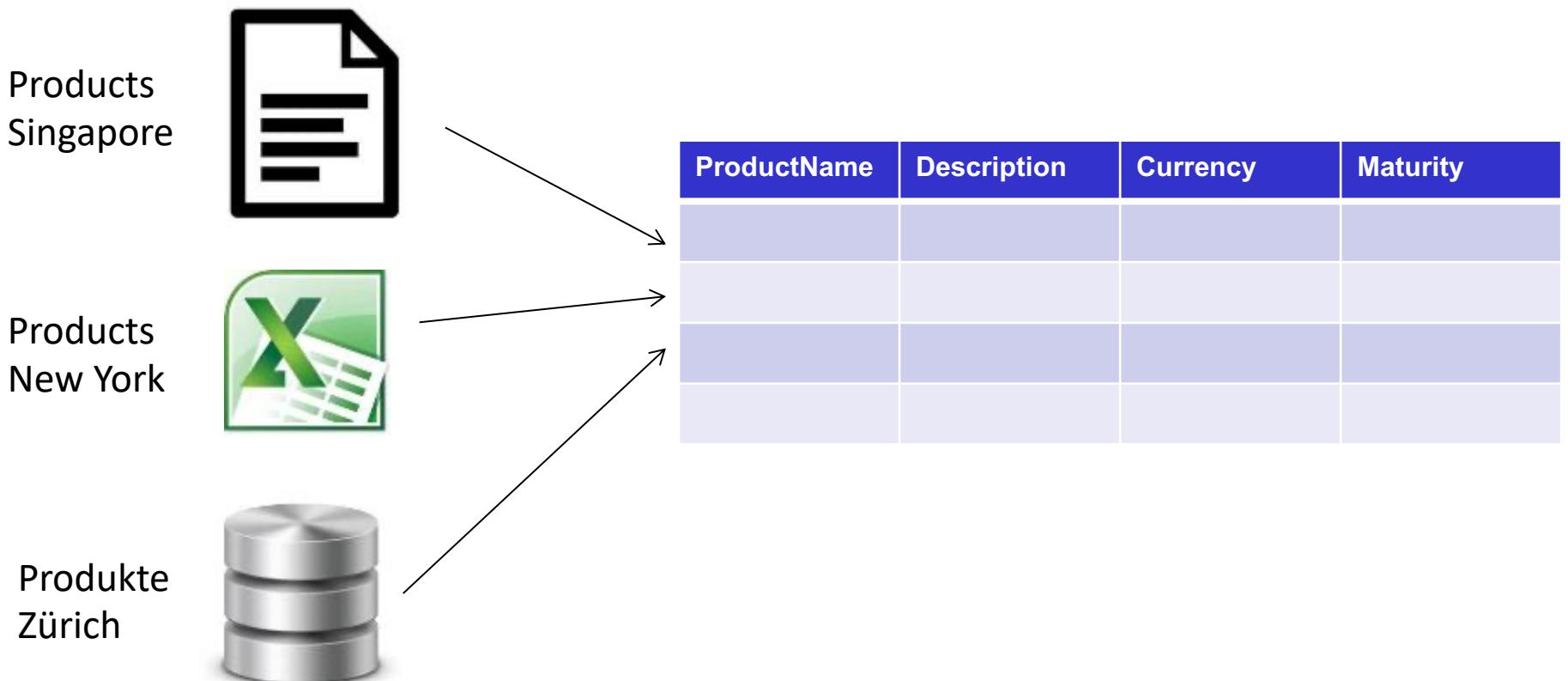
# Referenzarchitektur eines DWHs



# Ein einfaches DWH Beispiel

# Staging Layer #1

- Zusammenfügen unterschiedlicher Datenquellen im DWH:
  - Jede Datenquelle kann unterschiedliches Format haben



# Staging Layer #2

Products  
Singapore



Products  
New York



Produkte  
Zürich



ProductName	Description	Currency	Maturity
		877	March 3, 2014
Bond which yields 3% interest rate	Bond-3	US Dollars	7/5/2009
Anleihe42	Anleihe mit 4.2% Zinsen	CHF	31-12-2017

Welche Fehler sind aufgetreten und wie können sie behoben werden?

# Integration Layer

- Integrieren der Daten aus Staging Area in einheitliches Datenmodell

Personen-DB New York:

Name	First Name	Address
Page	Larry	CA 94740, Benvenue Ave 2449, Berkeley

Personen-DB Zürich:

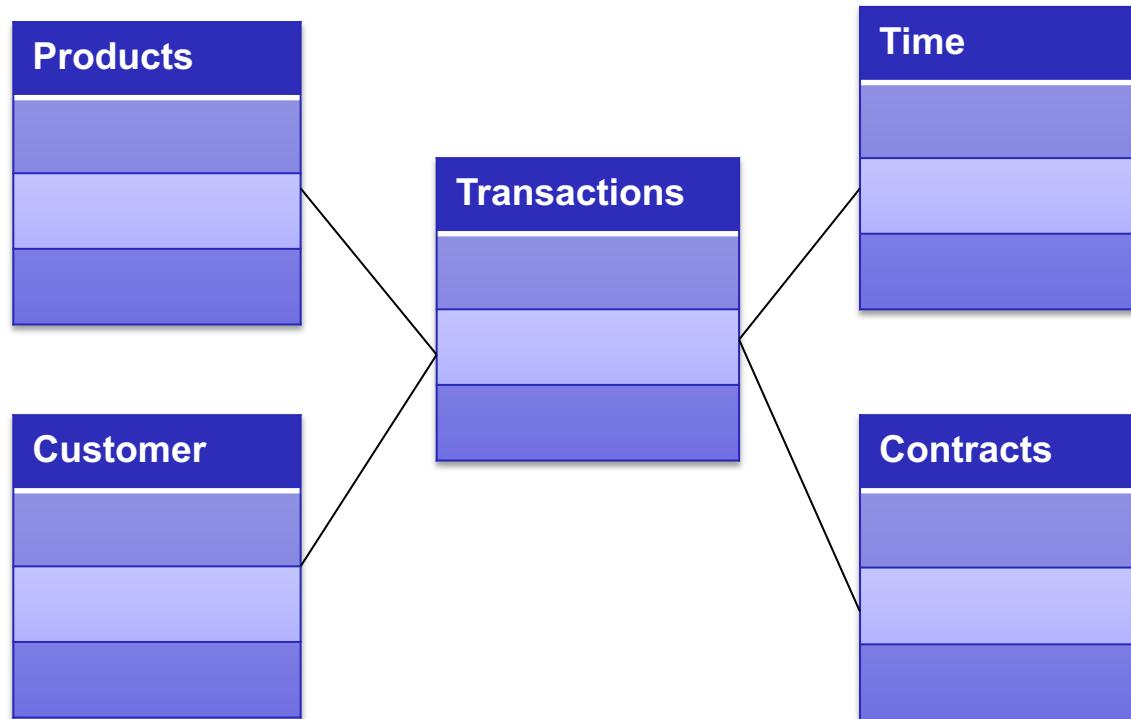
Vorname	Nachname	PLZ	Stadt	Strasse
Peter	Müller	8001	Zürich	Bahnhofstrasse 15



FirstName	LastName	PO_Box	City	Street
Peter	Müller	CH 8001	Zürich	Bahnhofstrasse 15
Larry	Page	CA 94740	Berkeley	Benvenue Ave 2449

# Analysis Layer: Data Marts

- Datenmodell ist für Abfragen optimiert
- Denormalisiert
- Star-Schema

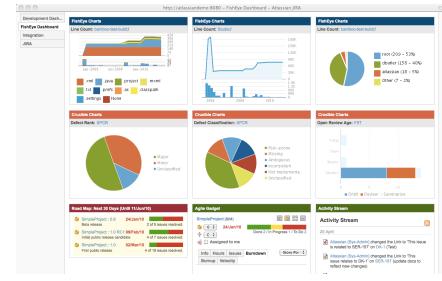


# Business Intelligence / Analytics

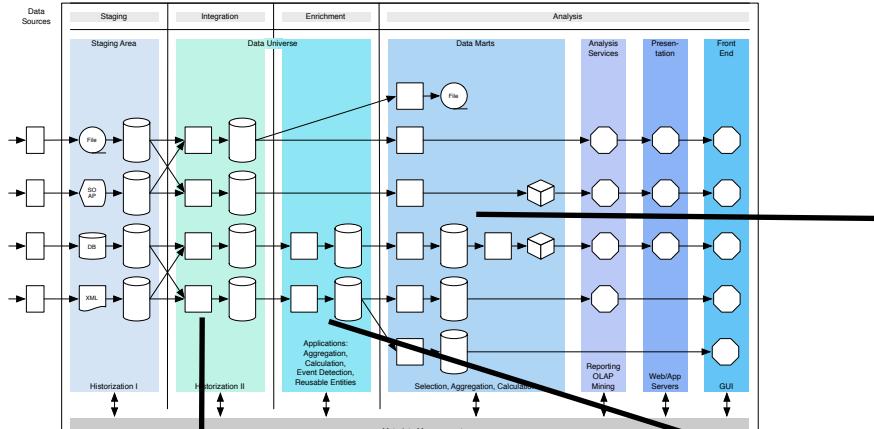
Auswertungen basieren auf den Data Marts

Typische Fragestellungen einer Bank:

- Entsprechen die Monatsberichte den **Basel II / III** Anforderungen?
- Welche **Kunden** haben die letzten X Monate in die strukturierten **Produkte Y** investiert?
- Welches sind die **Risiken aller Vermögen**, die von Zürich aus verwaltet werden?
- Wie hoch ist der **Neugeldzufluss** im 3. Quartal?
- Was sind die wichtigsten **Kennzahlen** des Unternehmens und wie ist deren **Entwicklung** der letzten 4 Quartale?



# Detaillierung Referenzarchitektur



## Analytische Schicht:

- End User Interface für Fachseite
- Aggregationen nach Business-Sicht (performance-optimierte Strukturen) in Data Marts
- Modellierung: relational denormalisiert, Sternschemata, flache Strukturen
- Zugriffsschicht für Reporting Applikationen (Standard und ad hoc)

## Integrationsschicht:

- Enterprise Data Warehouse Modell
- Zentrale Plattform für Informationsversorgung (single version of truth)
- Themen- und Businessorientierte Interpretation der Daten für analytische Nutzung
- Normalisierte Modellierung
- Basis für die analytische Schicht

## Anreicherungsschicht:

- Zentrale Erzeugung wiederverwendbarer Daten. Beispiele:
  - Kunden, Verträge, ...
  - Berechnete KPIs (Key Performance Indicators)
  - Klassifikationen
  - Segmente
  - Simulationsdaten
  - Conformed Dimensions

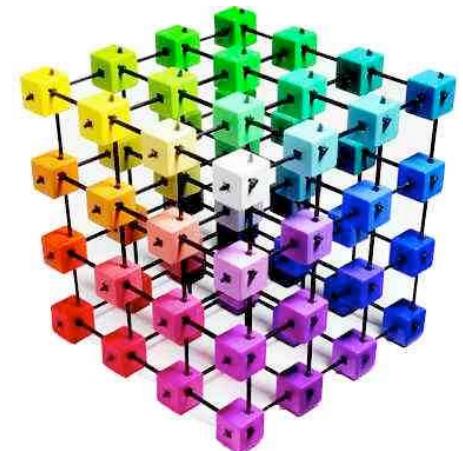
# Definition: OLTP vs. OLAP #1

- **OLTP = Online Transaction Processing:**
  - Viele kleine Transaktionen: updates oder inserts
  - Normalisiertes Datenschema: keine Redundanz in den Daten
  - Konsistente Zugriffe auf aktuelle Daten
- OLTP-Beispiele:
  - Hotelreservierungen
  - Bankomatabbuchungen
  - ERP-Systeme (Enterprise Resource Planning, Warenwirtschaft)
- Ziel: So viele Transaktionen wie möglich prozessieren

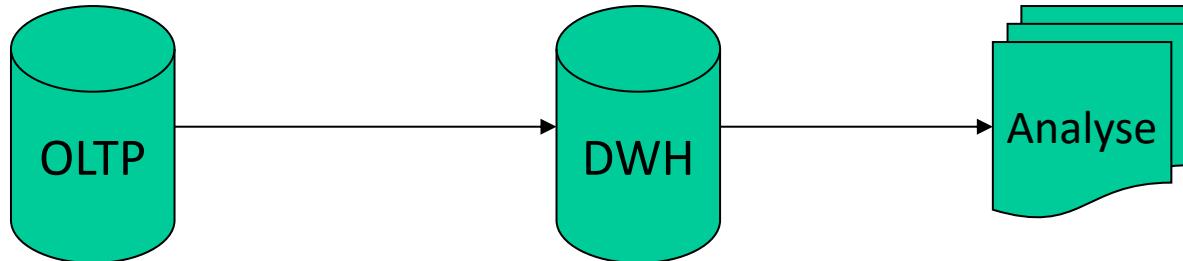


# Definition: OLTP vs. OLAP #2

- **OLAP = Online Analytical Processing:**
  - Grosse Queries mit vielen Joins: keine Updates
  - Redundanz in den Daten, um Abfragen zu optimieren:
    - Materialized views (Ergebnisse werden persistent gespeichert)
    - Spezielle Indizes (Bitmap Indizes)
    - De-normalisierung
  - Periodisches Neuladen (z.B. täglich oder monatlich)
- **OLAP-Beispiele:**
  - Management Information System von Grossfirmen
  - Wissenschaftliche Datenbanken (CERN, Bioinformatik)
  - Wetterdaten
- **Ziel: Queries mit sehr niedrigen Antwortzeiten (Sekunden bis Minuten)**

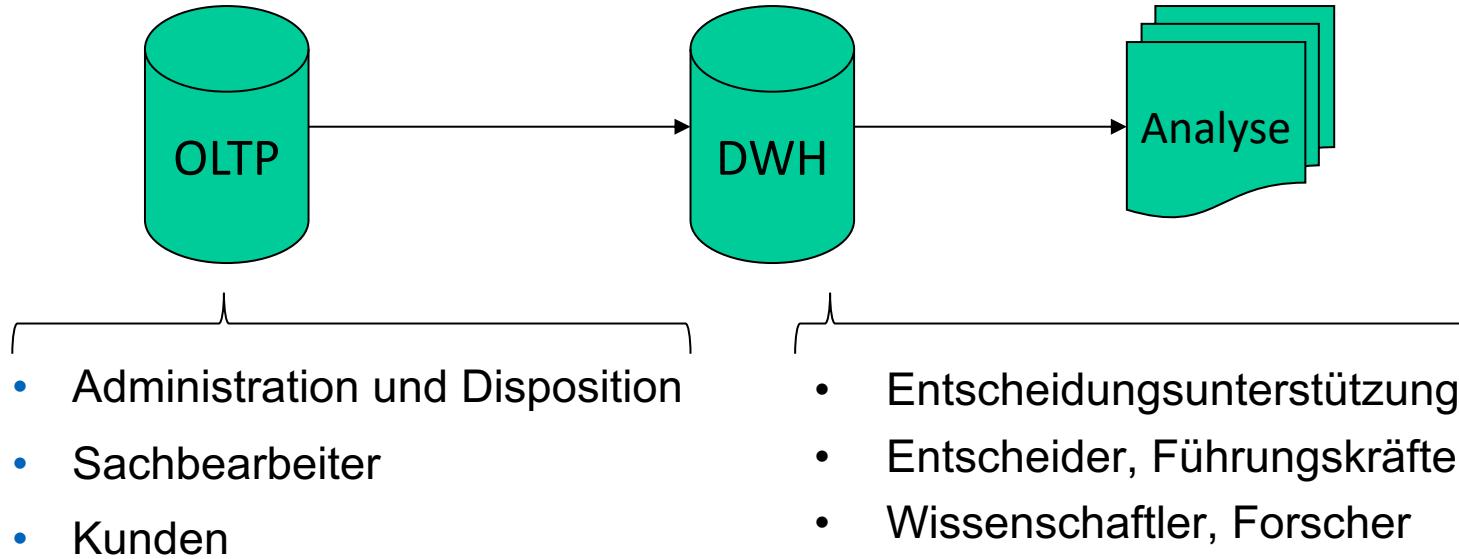


# Unterschiede in der Verarbeitung



Merkmal	<b>OLTP: Online Transaction Processing</b>	<b>DWH: Data Warehouse</b>
Verwendung für	Lesen, Ändern, Hinzufügen, Löschen	Lesen und Hinzufügen
Zugriff	vorhersehbar, repetitiv	ad hoc, heuristisch, periodisch
Antwortzeit	kurz ( $x < \text{Sekunde}$ )	mittel bis lang Sekunden bis Minuten
Datenstruktur	Einfach (Views oder Tables)	Komplex (Unions, Joins etc.)
Transaktionsvolumen	Wenige Rows	Teile oder ganze Tabellen
Transaktionsdauer	Kurze Lese- und Schreibvorgänge	Lange Lesevorgänge

# Unterschiede in der Anwendung



# Geschichte von Data Warehouses

- Erste industrielle Data Warehouses seit 1995
- Zu Beginn sehr hohe Fehlerquote (80%)
- Warum?
  - Datenintegration und –bereinigung schwierig
  - Datenmodelle haben wirkliches Business nur schlecht abgebildet
- DWH sind sehr teuer
- Erste grosse Erfolgsgeschichte:
  - WalMart reduziert Kosten bis zu 20% wegen DWH

# Tools

- Data Warehouse:
  - Commercial: DB2, Microsoft, Oracle, SAS, Teradata,...
  - Open source: Pentaho, JasperSoft,...
- Business Intelligence:
  - Commercial: SAP BusinessObjects, MS BI, Oracle, SAS,...
  - OpenSource: QlikView, Pentaho,...
- ETL (Extract, Transfer, Load)
  - Commercial: PowerCenter, Warehouse Builder, SAS,...
  - OpenSource: Pentaho, JasperSoft,...



# ZHAW-Forschung im Bereich DWH

**Applied Data Science** pp 333-351 | [Cite as](#)

## Data Warehousing and Exploratory Analysis for Market Monitoring

[Authors](#) [Authors and affiliations](#)

Melanie Geiger, Kurt Stockinger [✉](#)

Chapter  
First Online: 14 June 2019

2.4k  
Downloads

### Abstract

With the growing trend of digitalization, many companies plan to use machine learning to improve their business processes or to provide new data-driven services. These companies often collect data from different locations with sometimes conflicting context. However, before machine learning can be applied, heterogeneous datasets often need to be integrated, harmonized, and cleaned. In other words, a data warehouse is often the foundation for subsequent analytics tasks.

In this chapter, we first provide an overview on best practices of building a data warehouse. In particular, we describe the advantages and disadvantage of the major types of data warehouse architectures based on Inmon and Kimball. Afterward, we describe a use case on building an e-commerce application where the users of this platform are provided with information about healthy products as well as products with sustainable production. Unlike traditional e-commerce applications, where users need to log into the system and thus leave personalized traces when they search for specific products or even buy them afterward, our application allows full anonymity of the users in case they do not want to log into the system. However, analyzing anonymous user interactions is a much harder problem than analyzing named users. The idea is to apply modern data warehousing, big data technologies, as well as machine learning algorithms to discover patterns in the user behavior and to make recommendations for designing new products.

[https://link.springer.com/chapter/10.1007/978-3-030-11821-1\\_18](https://link.springer.com/chapter/10.1007/978-3-030-11821-1_18)

# Building Data Systems with Academia and Industry

- SODA – Search Over Data Warehouse:
  - ("Future ZHAW employee" + Credit Suisse + ETH Zurich)
  - Accessing **business data warehouses** in natural language
- Bio-SODA:
  - (ZHAW + Swiss Institute of Bioinformatics)
  - Accessing **bioinformatics databases** in natural language
- NQuest - Natural Language Query Exploration System:
  - (ZHAW + Zurich Startup Veezoo)
  - Accessing **databases and (partially) machine learning** in natural language
- GraphQueryML – Using Machine Learning to Optimize Queries in Graph Databases
  - (ZHAW + University of Konstanz)
  - Using machine learning for **query optimization**
- INODE – Intelligent Open Data Exploration System
  - (ZHAW + 8 partners in Europe)
  - Exploring **structured and unstructured data** in natural language

# Zusammenfassung

- Data Warehousing ist ein **wichtiges Thema** für viele Unternehmen
- **Ganzheitliche Sicht** auf die Unternehmensdaten
- Kombiniert **Datenbankthemen mit Datenanalyse** (Machine Learning)
- Data Warehousing ist **nicht** einfach nur ein **Tool**
- **Datenmodellierung, Datenintegration und Performanceaspekte** sind wesentlich