Data Warehouse

Martin Clement
Teamleiter Analytics
Martin.Clement@atvantage.com

Overview of the lecture

1.	Introduction to Data Warehousing	16.10.2025
2.	DWH Definition, Architecture & Data Modeling	21.10.2025
3.	ETL & Metadata	30.10.2025
4.	OLAP & Multidimensional Models	06.11.2025
5.	DWH Architectures & Organization	13.11.2025
6.	Visualization	20.11.2025
7.	Data Mining	27.11.2025
8.	Technical Trends I	04.12.2025
9.	Technical Trends II	11.12.2025
10.	Prüfung	16.12.2025

Further reading Data Warehouse

- Data Warehouse: From Architecture to Implementation von Barry Devlin von Addison-Wesley Longman, Amsterdam , ISBN 0-201-96425-2
- Data Warehouse Systeme: Architektur, Entwicklung, Anwendung. Andreas Bauer, Holger Günzel dpunkt.verlag 2009, ISBN 3-89864-540-1
- Ralph Kimball, Mary Ross: *The Data Warehouse Toolkit. The Complete Guide to Dimensional Modeling.* Wiley, ISBN 0-471-20024-7
- Vom Data Warehouse zum Corporate Knowledge Center, Eitel von Maur, Robert Winter, Physica Verlag (Springer Gruppe), ISBN 3-7908-1536-5

- William H. Inmon : *Building the Data Warehouse.* John Wiley & Sons
- William H. Inmon, Richard D. Hackathorn: *Using the Data Warehouse.* John Wiley & Sons, ISBN 0-471-05966-8

What the hell is a Data Warehouse?

Whiteboard Session

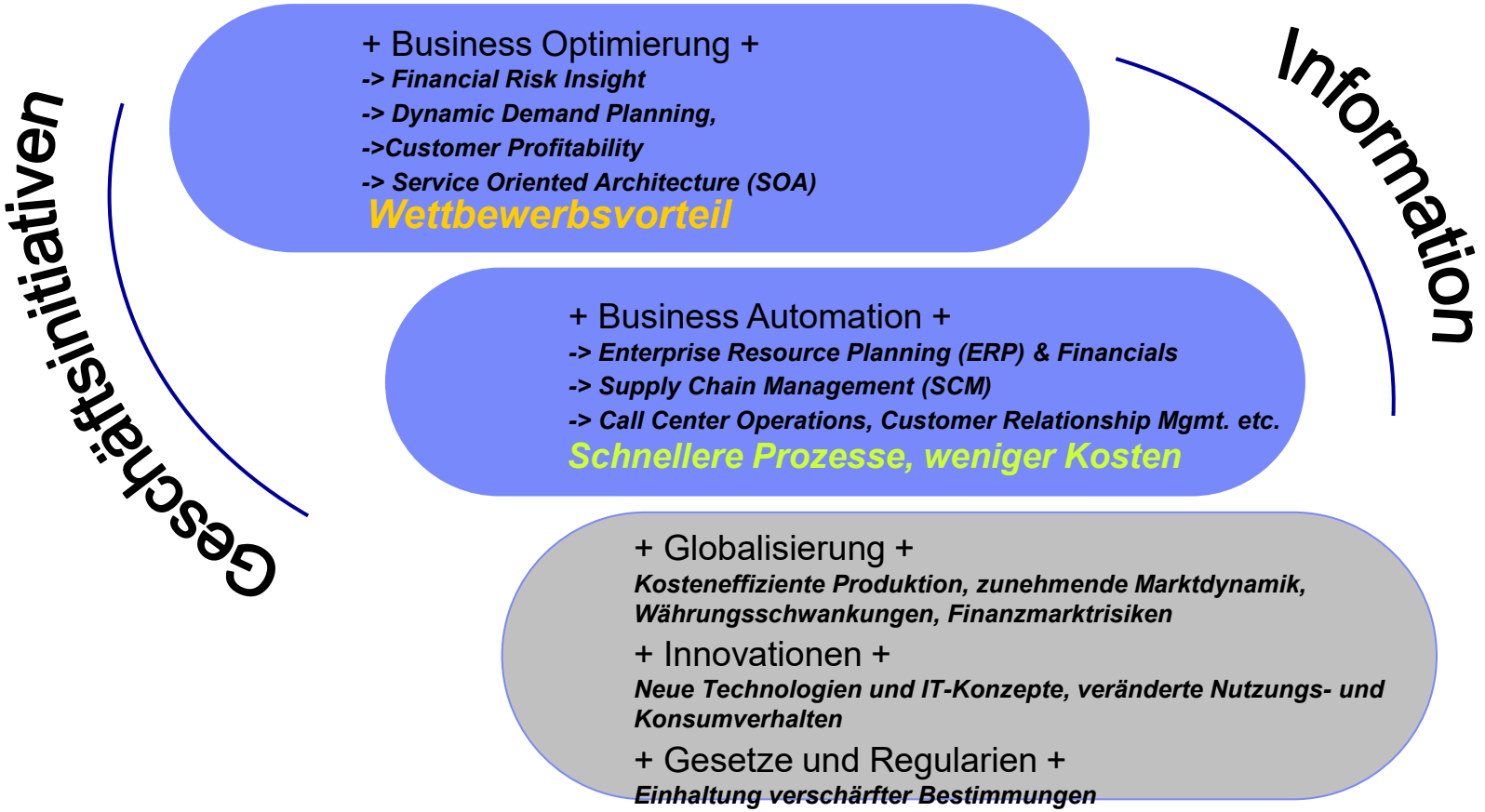
TIMETOACT GROUP
SOFTWARE & CONSULTING

Data Warehouse Background

?

Warum sind Informationen wichtig?

Unternehmen auf der Suche nach Wettbewerbsvorteilen.



Das Ziel neuer Geschäftsinitiativen ist die Optimierung des Business. Der Erfolg ist abhängig von den vorhanden Informationen; eine effizientere Nutzung der Informationen ist notwendig.

?

Aber, wo liegt das Problem bei Informationen?

Ein Paradoxon wird zur Herausforderung.

Rohstoff: Wasser

- Häufig vorkommender Rohstoff, der sogar durch Erderwärmung wächst.
- Circa ¾ der Erde sind mit Wasser bedeckt.

› **Trotzdem verdursten Menschen.**

Rohstoff: Daten

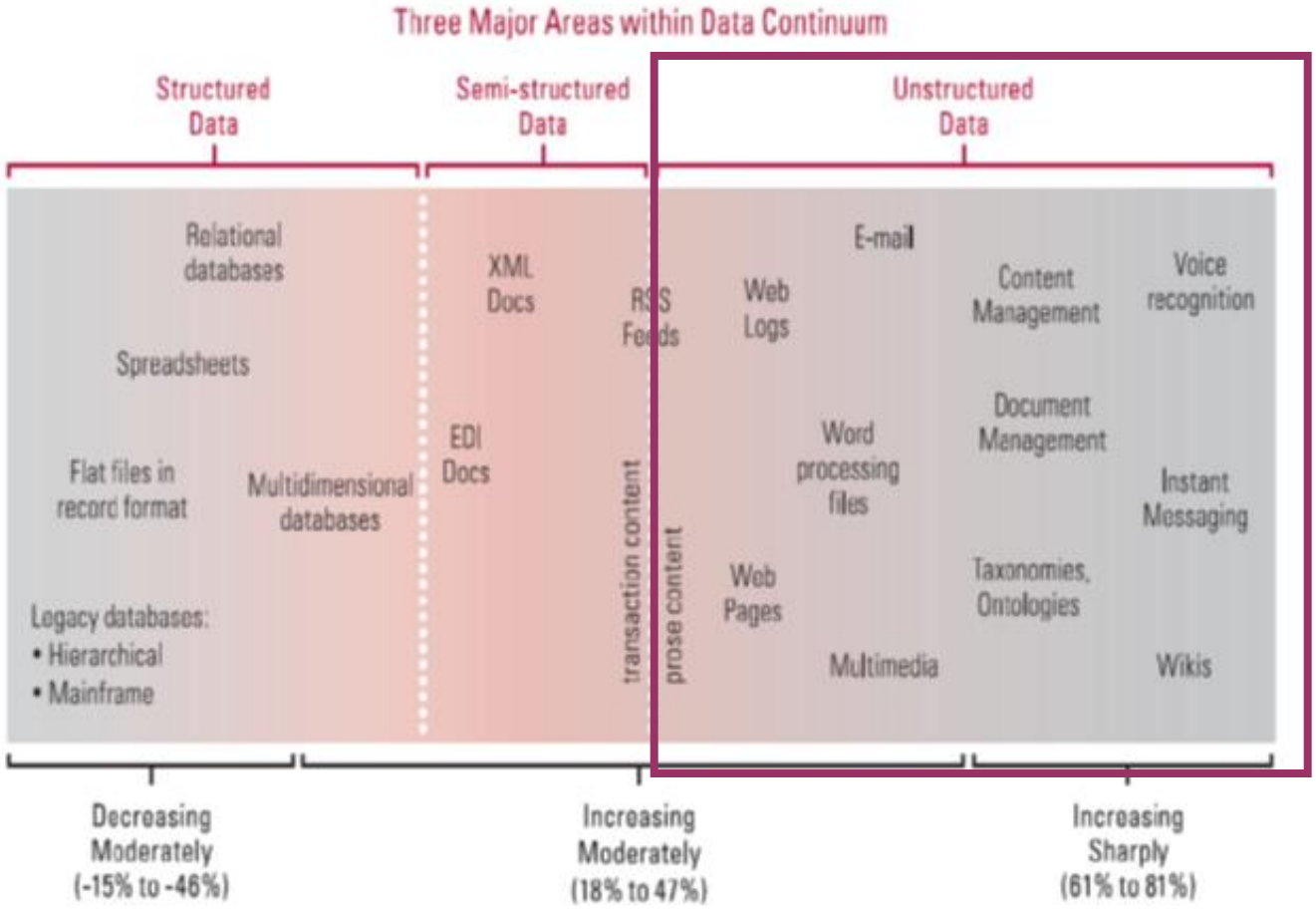
- Überall und immer mehr. Datenflut!
- Verzehnfachung der Daten von 2016 bis 2025 (163 Zetabytes!)*

› **Trotzdem fehlen Unternehmen Informationen.**

› Die bloße quantitative Existenz eines Rohstoffs bringt noch keinen Vorteil ...und wird eventuell sogar zum Problem und unkalkulierbarem Kostenfaktor!

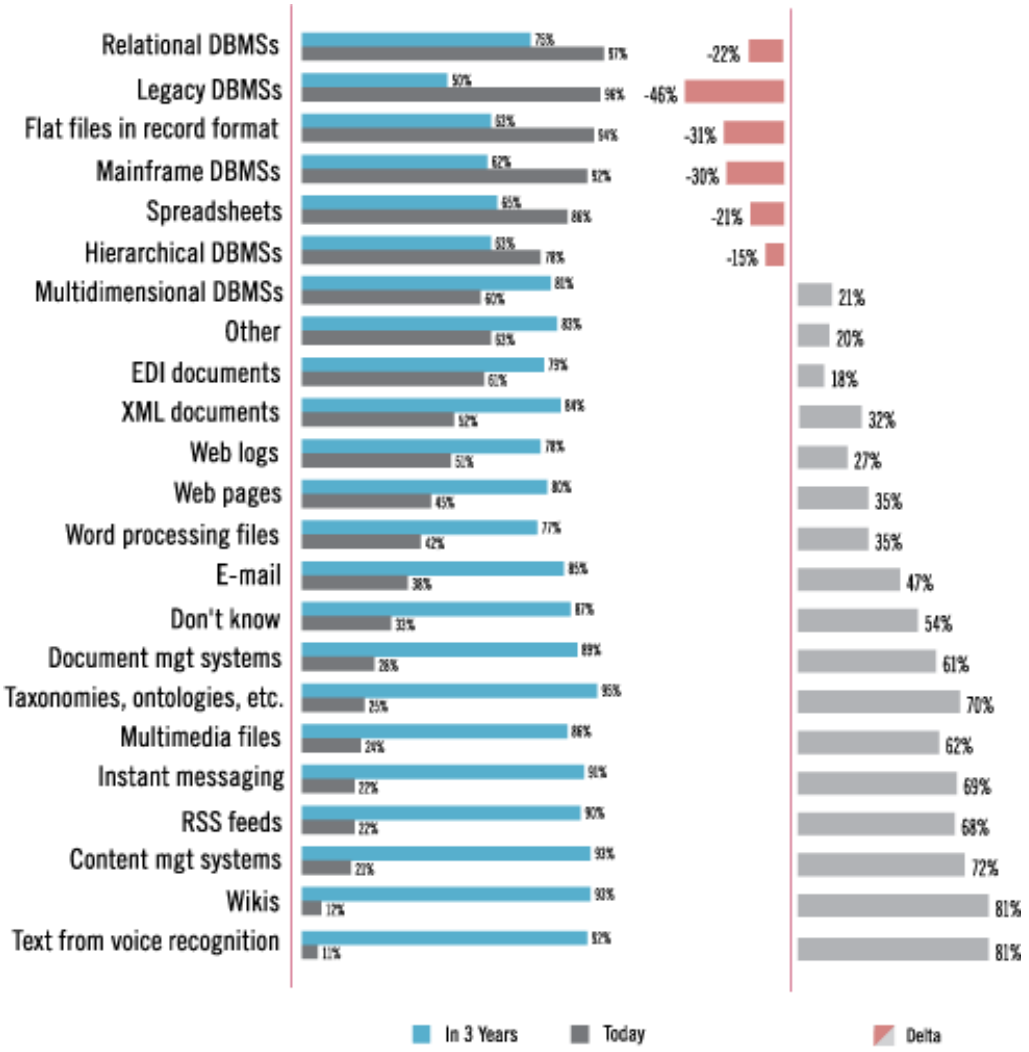
* Quelle: Statista 2016

Die Datenflut nimmt zu.



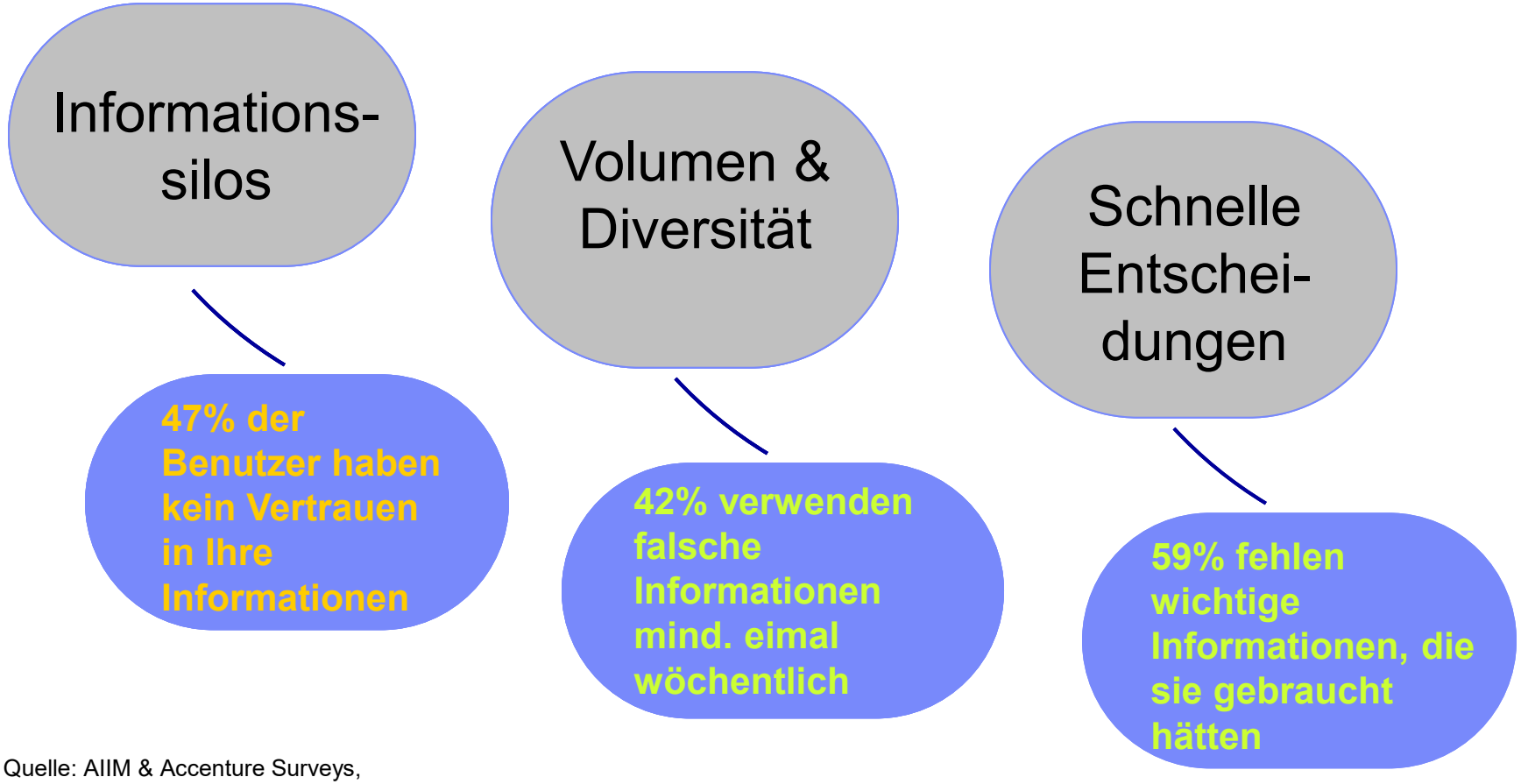
Besonders in unstrukturierten Daten stecken immer mehr Informationen. Sie wachsen durch neue Technologien überproportional an.

Die Entwicklung der Datenflut



QUELLE: TDWI

Unternehmen sehen ihr Informationsmanagement kritisch.



Quelle: AIIM & Accenture Surveys,

Die Herausforderung ist die richtige Information zum richtigen Zeitpunkt am richtigen Ort bereitzustellen.

? Wie bekommt man die Datenflut in den Griff und setzt Informationen in Wettbewerbsvorteile um?

Ein Rohstoff wird zum Wertstoff.

So wie Wasser letztlich durch den Einsatz von Hilfsmitteln und Verfahren für den Menschen nutzbar und wertvoll wird …

Aus Daten werden wertvolle Informationen.

› ... so benötigen Daten verschiedene Technologien und Prozesse, um für Unternehmen zur wertvollen Information zu werden.

Mehrstufiger Aufbereitungsprozeß im DWH

Business Optimierung, bessere Geschäftsergebnisse

Customer & Product Profitability
Financial Risk Insight
Workforce Optimization
Dynamic Supply Chain
Multi-Channel Marketing

End-to-End Lösungen

Business Intelligence and Performance Manangement

Analysieren, Planen, Verstehen und Steuern

Information Integration, Warehousing and Master Data Mgmt.

Bereitstellung zuverlässiger Informationen

Enterprise Data Management

Enterprise Content Management

Verwaltung von Daten und Inhalten während ihres gesamten Lebenszyklus ...

... und deren Nutzung für Geschäftsprozesse

Introduction to Data Warehousing

TIMETOACT GROUP
SOFTWARE & CONSULTING

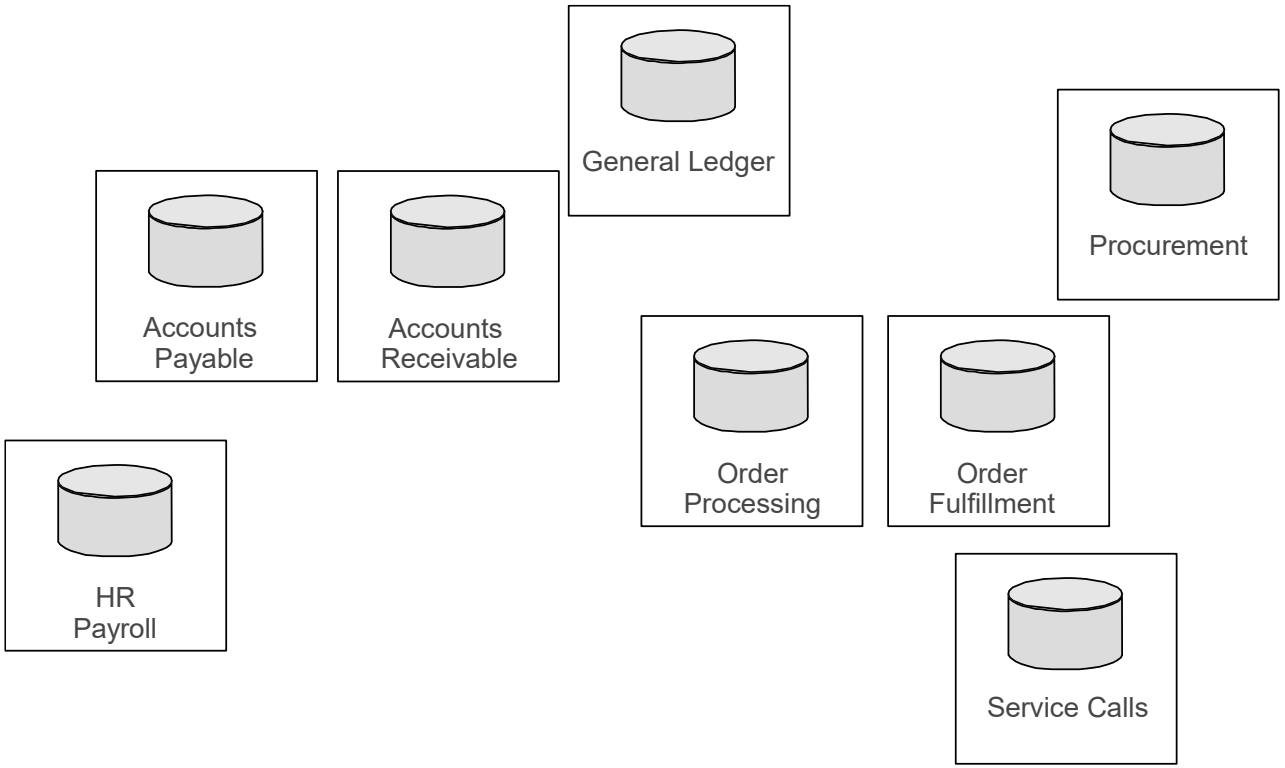
Information Technology (1960‘s – 80‘s)

- Support of few business processes in separated applications
  - applications separated
  - target: replace manual and time consuming activities
  - support daily transactions,
    - often batch processing
  - data embedded in process / process-specific application

    →process-orientation

Information Technology (1960‘s – 80‘s)

- Example: systems throughout a company



Information Technology (1960‘s – 80‘s)

- Storage of data
  - mostly on tape
  - very little space, therefore data encoded and compressed
  - one separate tape for each process / application

    →data spread
  - later introduction of databases –
    however, no consolidation of data throughout the company

Information Technology (1960‘s – 80‘s)

- Databases connected through common unique keys, numbers, ...
- More and more applications access the different databases
- Complex structure of systems and databases, but no overall view of company possible

Decison support in the 60‘s – 80‘s

- „Management Information Systems“ (MIS)
  - since late 60es
  - did not really work
- “Unplanned decision support"
  - Management needs reports / combined data from different systems to do make decisions for company
    → complex relationships between business processes
  - Reports all manually written by IT people
    → extract, combine, cache, accumulate data
    → can take several days
  - Error prone
    - relevant information may be forgotten or combined in a wrong way
  - No ad-hoc questions possible

Major problems for effective decision support

- Distributed data
- Different data structures
- Historic data
- System workload
- Inadequate technology

Distributed data

- Data resides on
  - different systems
  - different applications
- Has to be accumulated on one system for further analysis

Different data structures

- Systems developed independently from each other
  - different data types
    - E.g.; zip-code as integer or character string
  - different encodings
    - E.g.; m-f, m-w for gender
  - different field lengths
    - E.g.; address field

Issues with historic data

- Usually: Data archived after max. 3 months
  - daily transactions produce lots of data
  - limited size of storage

  → high amounts of data fill up systems
- For reports, already archived data may have to be accessed

Issues with system workload

- Systems constructed for daily transaction business
  - handle a constant (large) amount of transactions
  - handle transactions concurrently
  - each transaction accesses small amount of data
  - very simple
- Reports:
  - few reports at one time
  - access lots of data
  - not on a regular basis
  - complex arithmetic operations
- Consequences:
  - systems stressed by additional load (due to reports)
  - not optimized for this kind of workload
  - performance of daily transaction business jeopardized
  - may possibly lead to system failure!

Inadequate technology

- Capacity and cost of hard disk drives
  - E.g. 1979: 8″-Winchester hard disk drive of IBM
    - Capacity 5 MB
    - Cost 10 000 DM.
- Performance and cost of processors
- No graphical front-ends
- Network technology in its beginnings
  - Ethernet 1973

Conclusion

- "operative systems" not suitable for analytical evaluations
- Need for a new, separated system
  - fast answers, ad-hoc questions possible
  - no interference with daily transaction business

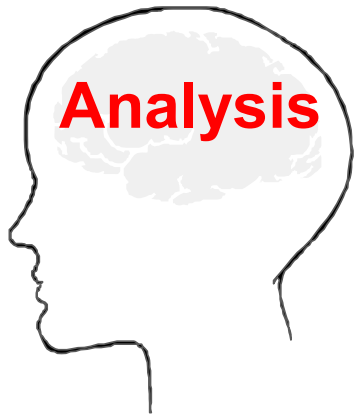
→Data Warehouse

Group task

- Assume that you are the CIO of a bank, a retail company, a telco provider or an online store.
- From what you have learned so far
  - List possible (functional and non-functional) requirements for a data warehouse
    Think of the deficiencies of transactional systems like
    - Distributed data
    - Different data structures
    - Problem with historic data
    - Problem with system workload
    - Inadequate technology
  - What are requirements from an end-user perspective.

What is Business Intelligence / Data Warehouse ?

In a word

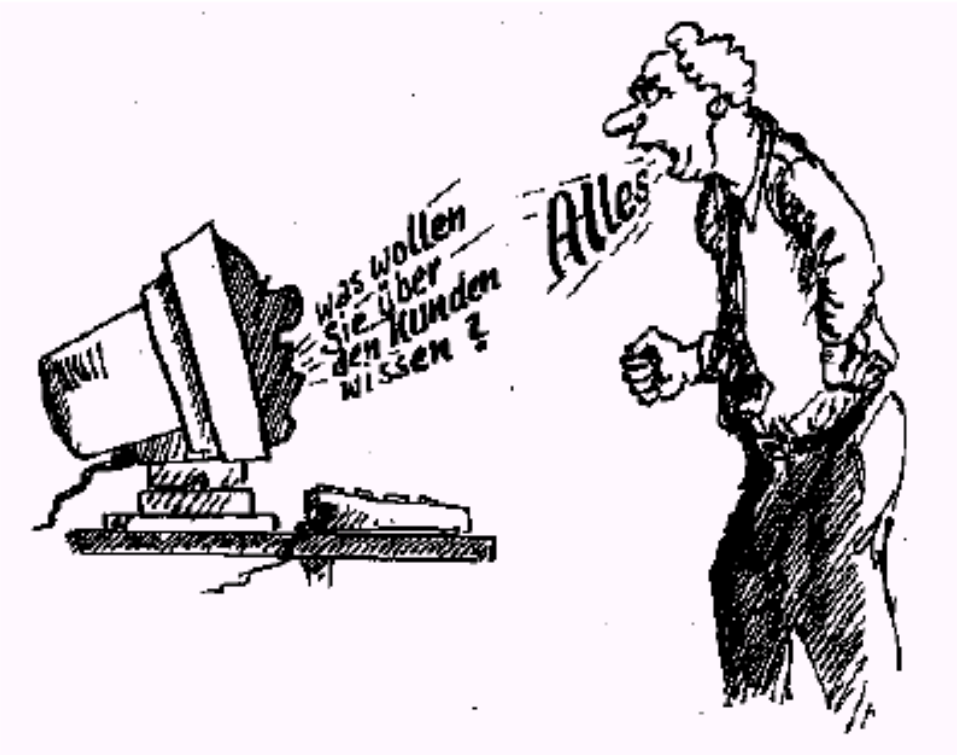


Business Intelligence / Data Warehousing is the process of gathering, consolidating, and analyzing data from multiple sources for strategic decision making

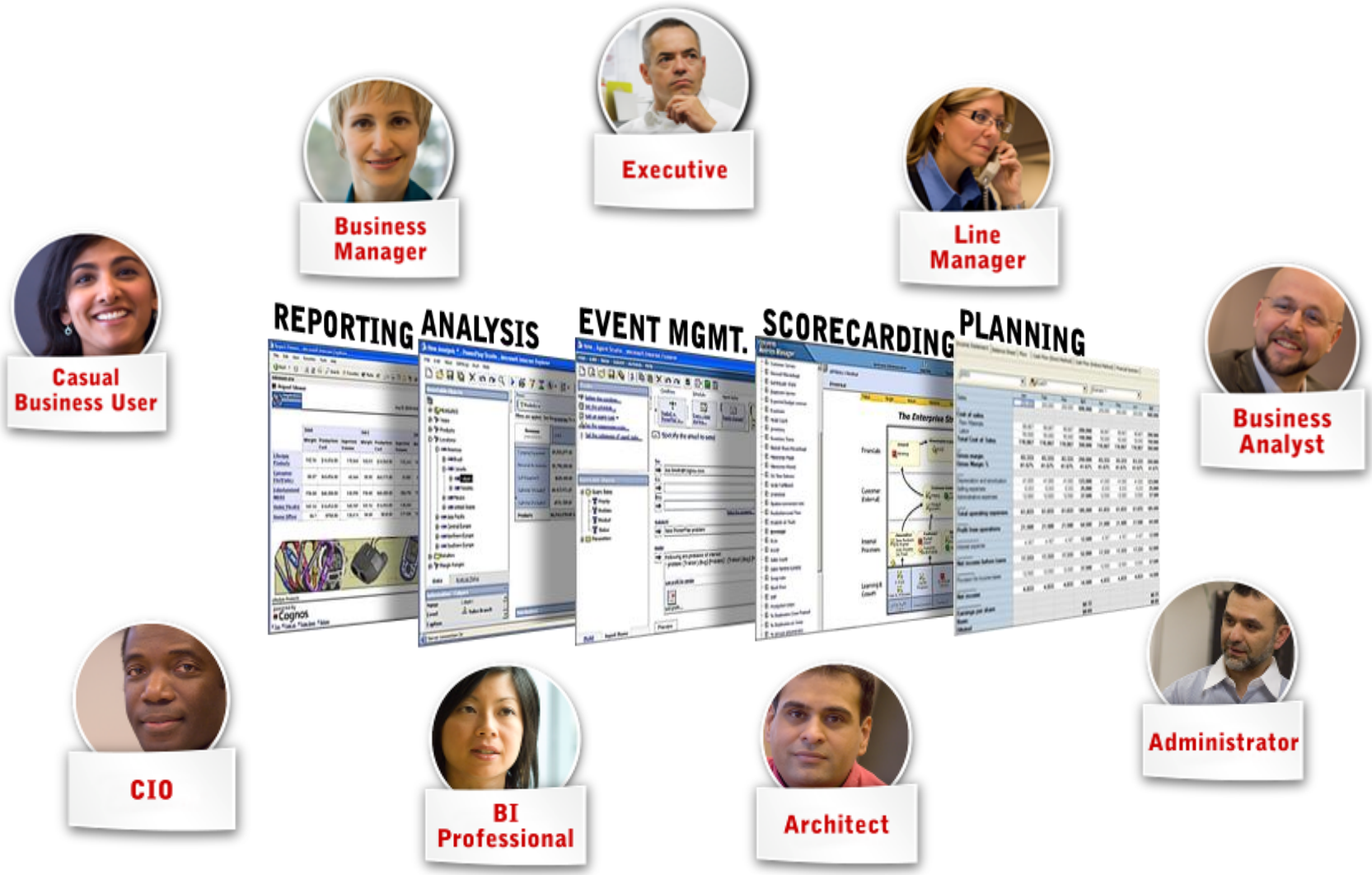
- BI derives new value from your transactional data
- BI supports strategic planning, monitoring, and efficiency
- BI delivers *knowledge* of the customer, suppliers, and channels
- BI unifies the enterprise with a single version of the truth

Data Warehouse Definitions

What do we expect from the Data Warehouse?



Die “Single version of the truth” für alle Benutzer



Data Warehouse User

- Wants to access and analyze all important data from desktop
- Uses data analysis as facility for daily work
- Wants data to be fast and dynamically available
- Identifies data correlations by browsing through the data
  - does possibly not know what (s)he is looking for

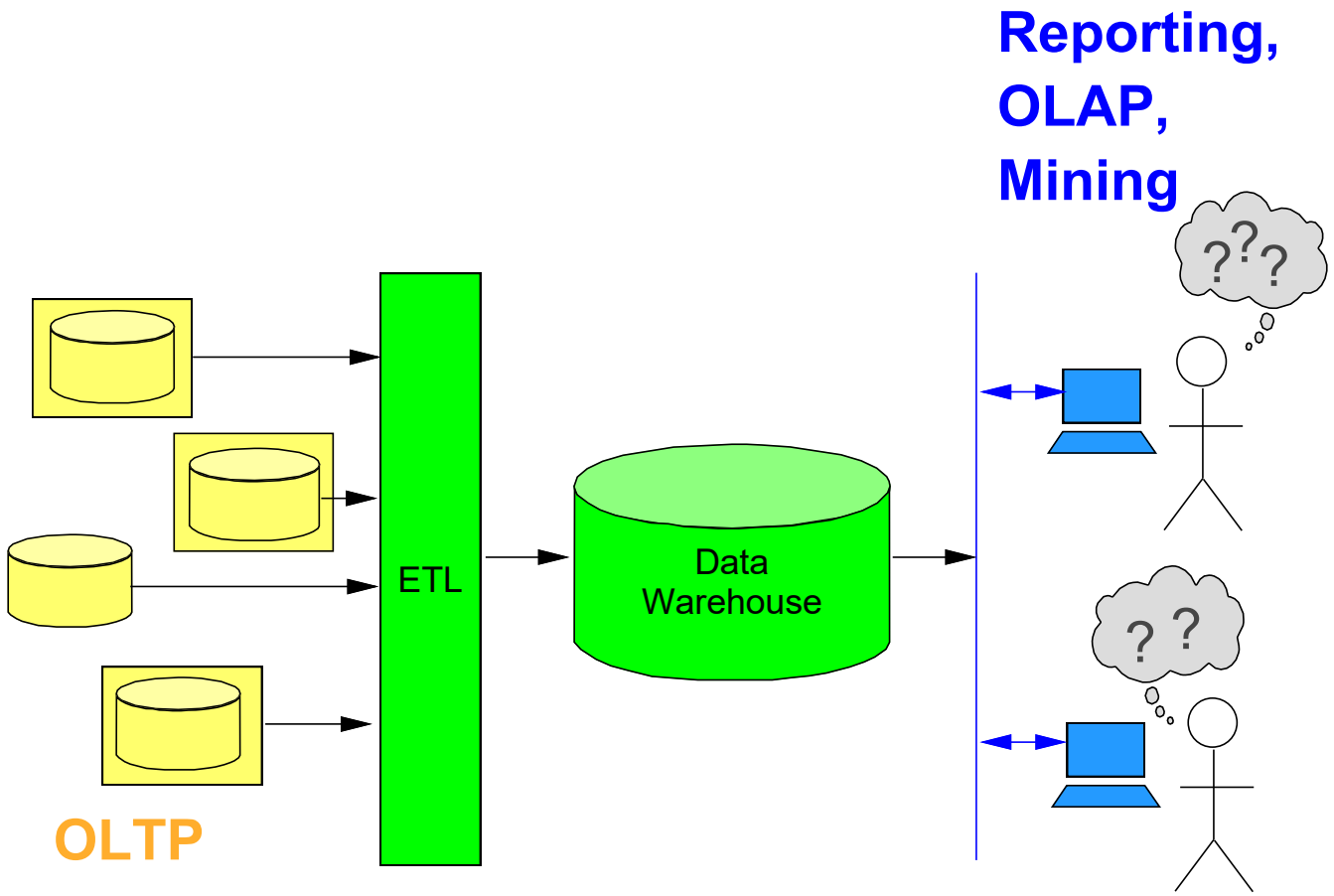
Data Warehouse

- Answer business-critical questions
- Decision support for management
- Overcomes difficulties when using existing transaction systems for those tasks
- Not a product, but a overall concept

Data Warehouse

- Contains data from different systems
- Imports data from different systems on a regular basis
  - summarize data
  - provide historic data
  - generate metadata
- Completely new system

Basic Data Warehouse architecture



Data Warehouse Definition (Inmon)

A data warehouse is a:

- Subject oriented,
- Integrated,
- Time variant,
- Non volatile

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

Subject Orientation

- A data warehouse is organized around around the major subjects of the enterprise like
  - customer,
  - vendor,
  - product.
- In contrast to the process/functional orientation of applications such a
  - loans,
  - savings,
  - bankcard.

Integration

- Data contained in the warehouse are integrated.
- Aspects of integration
  - consistent naming conventions,
  - consistent measurement of variables,
  - consistent encoding structures,
  - consistent physical attributes of data

→„One version of truth“

Time-variance

- All data in the data warehouse is accurate as of some moment in time.
  - Has to be associated with a time stamp.
- In the operational environment data is accurate as of the moment of access.
- Once data is correctly recorded in the data warehouse, it cannot be updated.
  - Data warehouse data is, for all practical purposes, a long series of snapshots.
- Operational data, being accurate as of the moment of access, can be updated as the need arises.

Non-volatile

- Operations in operational environment
  - insert
  - delete
  - update
  - read
- Operations in a data warehouse
  - the initial and additional loading of data,
  - the access of data.

Data warehouse characteristics

- subject-oriented
  - not process-oriented as old systems were
- integrated
  - data is integrated (no problems any more with different representations of same data, example with gender)
- time-variant
  - data is constantly added, data warehouse data is actual
- non-volatile
  - after insertion of data in warehouse: no update allowed: old data is "preserved"

Characteristics of operative databases

- **O**n**l**ine **T**ransaction **P**rocessing (OLTP)
  - Small transactions, i.e. accounting
  - Mostly read, write, update operations
  - Optimized for data input
  - Operate on the most recent, valid state of the data
  - Operate on limited amount of data
  - Data model: Entity-Relationship Model

Characteristics of dispositive/analytical databases

- **O**n**l**ine **A**nalytical **P**rocessing (OLAP)
  - DSS: Decision Support System
  - MIS: Management Information System
  - Decision Support
  - Optimized for data analysis and extraction
  - Work typically with large amounts of data
  - Data Warehouse Front end

Operative vs. dispositive data

	Operative DBs	Dispositive DBs
Purpose	processing of daily business transactions	infomation for management (decision support)
Content	detailed, complete, most recent data	historic, stable and summarized, data
Data Amount	small amount of data per transaction	large amount of data for load, and often per query
Data Structure	complex, suitable for operational calculations	simpler, suitable for business analyses
Transactions	very short read/write transactions	long load operations, longer read transactions
Requirements	high performance for write- and update operations, low redundancy	high performance for read-operations

Operative vs. dispositive data

	Operative data	Dispositive data
Handling	structured, parallel processes with short and isolated ("atomic") transactions	complex and changing questions, unstructured analytical processes
Modeling	process- and function-oriented, individual for each application	subject-oriented and standardized
User Types	In-/output by business specialist	Analysis by manager, controllers, business analysts
# of Users	a lot	few(er)
System return time	Milliseconds	Seconds to minutes (even hours)

Data Warehousing Usage Types

- Information oriented
  - Mostly fixed reports „standard reporting“
  - 70 – 80 % of all users of a data warehouse
- Analysis oriented
  - Dynamic, „Ad hoc“ queries
  - Purpose analysis of complex problems
- Planning oriented
  - Model the business,
  - Future views based on historical data
- Campaign oriented
  - Executed as specific projects

Exercise

For one of the following companies

- Retail Bank
- Telecommunication compagny
- Online bookstore (like Amazon.com)
- Discount furniture store (like IKEA)
- Supermarket
- Stock Exchange

1. Outline the operational systems
   - characterize which operations are performed by them
     - mostly triggered by actions of a customer
   - which information is stored by these systems
   - how this information is manipulated
   - what questions can be answered by these systems
2. For a data warehouse
   - describe what information can be stored in it
   - what questions can/should be answered with this information