Utrecht, May 8th 2018

Business Intelligence

Lecture 03 - Descriptive Analytics Part B
Data Warehousing, Business Performance Management

Georg Krempl

Algorithmic Data Analysis
Information and Computing Sciences
Utrecht University, The Netherlands

With particular thanks to

- Marco Spruits (previous BI lecturer)
- Armel Lefebvre (tutor, A.E.J.Lefebvre@uu.nl)
- Vincent Goris (student teaching assistant, vincent.goris93@gmail.com)
- Kristof Fellegi (student teaching assistant, k.fellegi@uu.nl)

Summary of the Previous Lecture(s)

- The Nature of Data
  See [Sharda et al., 2018, chapter 2.1], [Žliobaitė et al., 2016]
- Data Quality and Integrity
  See [Sharda et al., 2018, chapter 2.2–2.3]
- Data Preprocessing
  See [Sharda et al., 2018, chapter 2.4 & Application Case 2.1]

Figure:
[Sharda et al., 2018, p. 92 & 115]

- Statistics Repetition
  See [Hand et al., 2001, Appendix A.1]
- Statistical Modelling
  See [Sharda et al., 2018, chapter 2.5–2.7],
  [Cover and Thomas, 2006, chapter 2], [Hand, 2008], and others
- Business Reporting
  See [Sharda et al., 2018, chapter 2.7]
- Data Visualization
  See [Sharda et al., 2018, chapter 2.8–2.11] and [Segel and Heer, 2010]

Outline: Today's Lecture in the Context of Business Intelligence

Image not available due to copyright restrictions. Please refer to the source cited below.

Figure: Context of Data Warehousing within Business Intelligence (Source: [Sharda et al., 2018, page 84])

Descriptive Analytics Part B: Data Warehousing, Business Performance Management

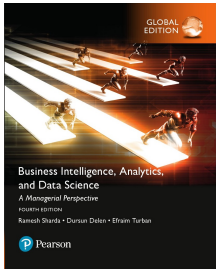


Figure: Textbook [Sharda et al., 2018, Chapter 3]
Sharda, Delen, Turban & King (2018). Business Intelligence, Analytics & Data Science: A Managerial Perspective 4th Global Edition, Pearson. ISBN-13: 9781292220567

Figure: Paper [Chaudhuri et al., 2011]
Chaudhuri, Dayal, Narasayya (2011)). An Overview of Business Intelligence Technology Communications of the ACM, vol. 54, no. 8 DOI:10.1145/1978542.1978562

Advanced Literature (Voluntary Further Reading):

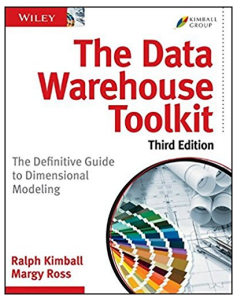


Figure: Textbook [Kimball and Ross, 2013]
Kimball, Ross (2013). The Data Warehouse Toolkit 3rd Edition, Wiley
ISBN-13: 978-1118530801

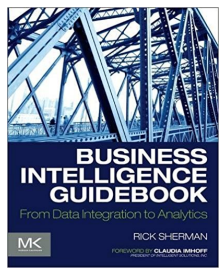


Figure: Textbook [Sherman, 2015, Part III-IV]
Sherman (2015). The Business Intelligence Guidebook 1st Edition, Morgan Kaufmann
ISBN-13: 978-0124114616

Outline and Summary[1]

- Definition & History of Data Warehouses
  See [Sharda et al., 2018, chapter 3.2]
- Data Warehousing Process & Architectures
  See [Sharda et al., 2018, chapter 3.3–3.4]
- Data Integration and ETL Processes
  See [Sharda et al., 2018, chapter 3.5]
- Data Warehouse Development
  See [Sharda et al., 2018, chapter 3.6]
- Data Modelling & Dimensional Modelling
  See [Sherman, 2015, **chapter 9** and parts of 8] and [Sharda et al., 2018, chapter 3.7–3.8]
- Issues, Considerations & Future Trends
  See [Sharda et al., 2018, chapter 3.7–3.8] and **[Chaudhuri et al., 2011]**
- Business Performance Management
  See [Sharda et al., 2018, chapter 3.9-3.11]

▸ Start ▸ Appendix

[1] Note: This lecture integrates knowledge from several further sources (cited where used, except if my own).

Outline and Summary[1]

- Definition & History of Data Warehouses
  See [Sharda et al., 2018, chapter 3.2]
- Data Warehousing Process & Architectures
  See [Sharda et al., 2018, chapter 3.3–3.4]
- Data Integration and ETL Processes
  See [Sharda et al., 2018, chapter 3.5]
- Data Warehouse Development
  See [Sharda et al., 2018, chapter 3.6]
- Data Modelling & Dimensional Modelling
  See [Sherman, 2015, **chapter 9** and parts of 8] and [Sharda et al., 2018, chapter 3.7–3.8]
- Issues, Considerations & Future Trends
  See [Sharda et al., 2018, chapter 3.7–3.8] and **[Chaudhuri et al., 2011]**
- Business Performance Management
  See [Sharda et al., 2018, chapter 3.9-3.11]

▸ Start ▸ Appendix

[1] Note: This lecture integrates knowledge from several further sources (cited where used, except if my own).

Data Warehouse: Motivation

> *The company being profiled within this case study is a privately held Dutch retailer with 2,800 stores across several European countries. This retailer has* **15 different brands** *that include products from toys to cookware,* **each** *brand* **having its own infrastructure**. *Each business entity is* **managed independently** *as one of 15 individual companies in the way they develop their processes, maintain their legacy systems, and make business decisions across finance, IT, supply chain, and general operations.*

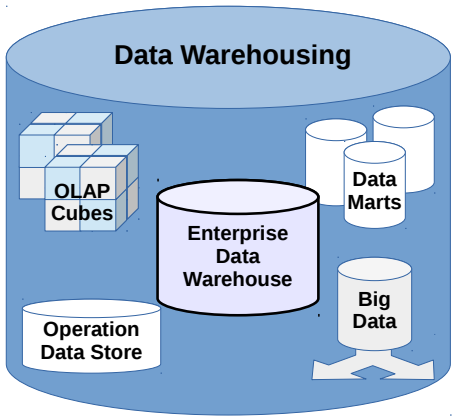
Application Case 3.3
[Sharda et al., 2018, p. 177–179]

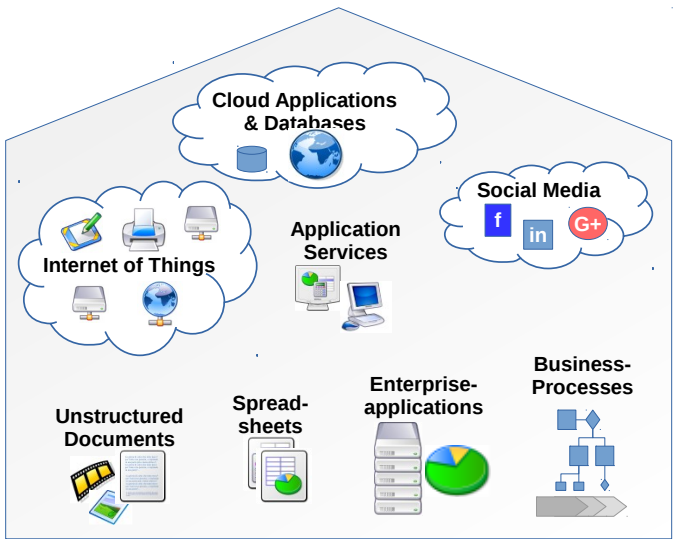
What are the challenges here?

- many organizational units (brands, divisions, . . .)
- managed independently
- heterogenous IS landscape

- *Decision support* require integrating data from many units & systems
- Operational systems and databases (OLTP) are designed independently and with different objectives in mind (*efficiency* and *consistency*)

Data Warehouse: Definition





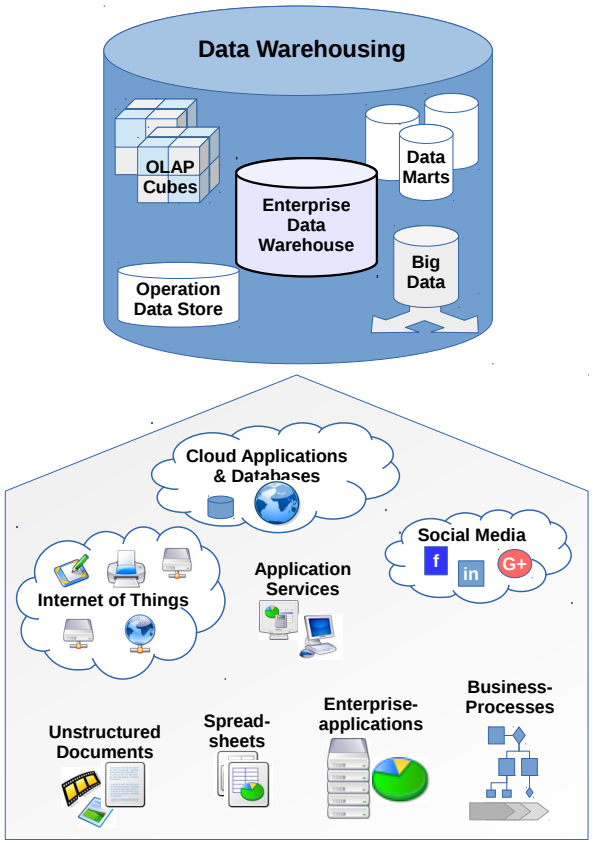
Data Warehouse

- *Central physical* repository
- for *decision support*, providing
- *enterprise-wide*, *cleansed* data
- in a *standardized format*,
- mostly *structured* and *analytics-ready*

Data Warehousing

- subject-oriented
- integrated
- time-variant
- nonvolatile
- collection of data (& metadata)
- often not normalized
- to support decision-making

Data Warehousing: Definition



subject-oriented

- organized by detailed subject
- reduced to decision-relevant info
- *not product-oriented as in operational DBs*

integrated

- fetch data from different sources
- transform into consistent format

time-variant

- temporal quality: time-stamped data
- time as *the* important dimension

nonvolatile

- once stored, data is not to be changed
- obsolete data is discarded, not updated

Data Warehousing: Further (Potential) Characteristics

Web-Based

- designed for web-based applications

Relational / Multidimensional

- either relational structured, or multidimensional structured

Client/Server

- for easy access by end users

Real Time

- real-time, active, data-access and analytics capabilities

Inclusion of Metadata

- metadata describing how data is organized and (to be) used

Data Warehouse: Main Types of DW

Data Marts

- smaller and focused on a particular subject/business unit
- **dependent** of a centralized enterprise data warehouse
  follows a single entreprise-wide data model
- **independent data mart** designed for one unit

Operational Data Stores

- consolidation of data from multiple sources
- near-real-time, integrated view of volatile current data
- used for **short-term decisions**, **short-term memory**
- often used as **interim staging area** for DW

Enterprise Data Warehouse

- **large-scale** data warehouse
- used **accross the enterprise**
- consolidation of data from multiple sources
- standard format for BI/DSS

Data Warehouse: Historical Development

Image not available due to copyright restrictions. Please refer to the source cited below.

Figure: Timeline of Historical Events in DW Development (Source: [Sharda et al., 2018, page 158])

Data Warehouse: Historical Development

Image not available due to copyright restrictions. Please refer to the source cited below.

Figure: Timeline of Historical Events in DW Development (Source: [Sharda et al., 2018, page 158])

Milestones

1970s	Mainframe computers Data mostly hidden
1980s	Personal Computers Islands of data Distributed DBMS
1990s	Centralized Data Warehouse Data copied & integrated
2000s	DM & Predictive Analytics
2000+	(Big) Data Analytics

Outline and Summary[1]

- Definition & History of Data Warehouses
  See [Sharda et al., 2018, chapter 3.2]
- Data Warehousing Process & Architectures
  See [Sharda et al., 2018, chapter 3.3–3.4]
- Data Integration and ETL Processes
  See [Sharda et al., 2018, chapter 3.5]
- Data Warehouse Development
  See [Sharda et al., 2018, chapter 3.6]
- Data Modelling & Dimensional Modelling
  See [Sherman, 2015, **chapter 9** and parts of 8] and [Sharda et al., 2018, chapter 3.7–3.8]
- Issues, Considerations & Future Trends
  See [Sharda et al., 2018, chapter 3.7–3.8] and **[Chaudhuri et al., 2011]**
- Business Performance Management
  See [Sharda et al., 2018, chapter 3.9-3.11]

▸ Start ▸ Appendix

[1] Note: This lecture integrates knowledge from several further sources (cited where used, except if my own).

Data Warehousing Process

Data Warehousing

The process

- from data sources
- to data extraction, transformation, loading
- into a comprehensive database
- including managing and retrieving metadata

Image not available due to copyright restrictions. Please refer to the source cited below.

Figure: Data Warehouse Framework and Views (Source: [Sherda et al., 2018, page 164])

Data Warehousing Architecture: Components

Data Sources

- indep. operational "legacy" systems
- internal OLTP or ERP systems
- external data providers
- web data (logs)

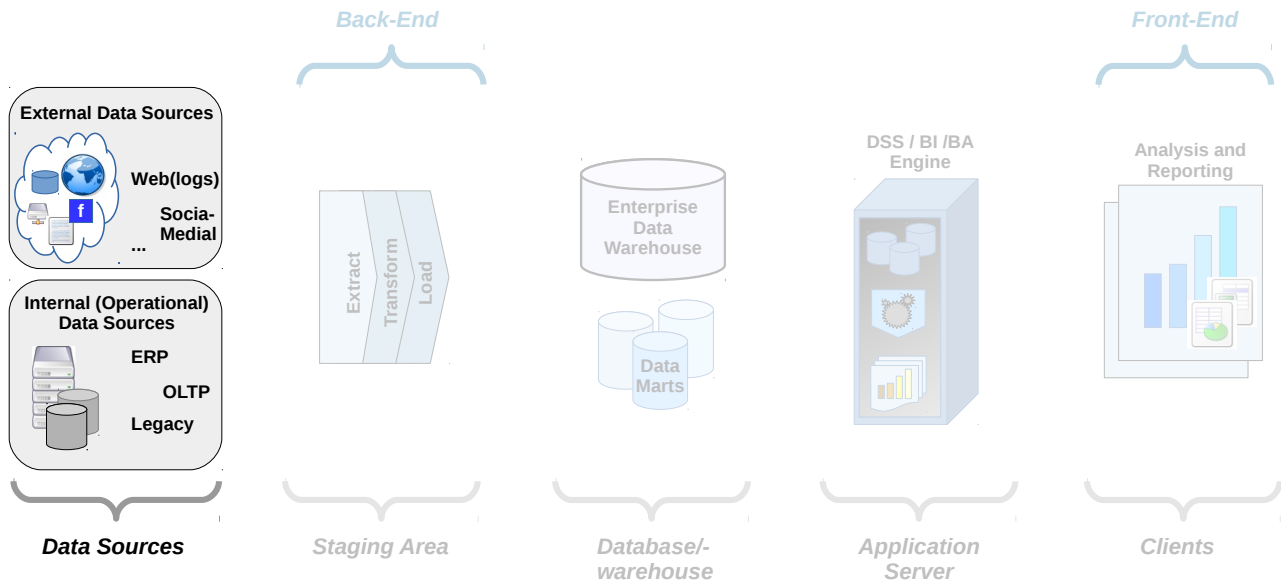


Figure: Data Warehouse Architecture Components

Data Warehousing Architecture: Components

Data Extraction, Transformation

- Subprocess for selection, extraction, transformation of data

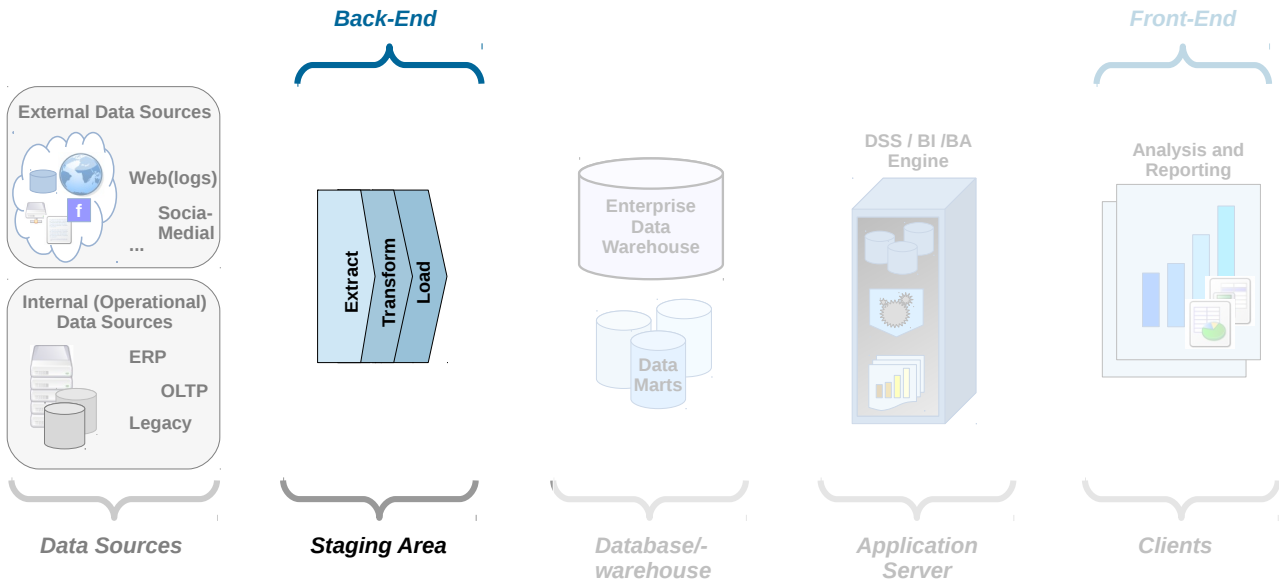


Figure: Data Warehouse Architecture Components

Data Warehousing Architecture: Components

Data Loading

- Loading into a staging area for transformation and cleansing
- Loading into data warehouse

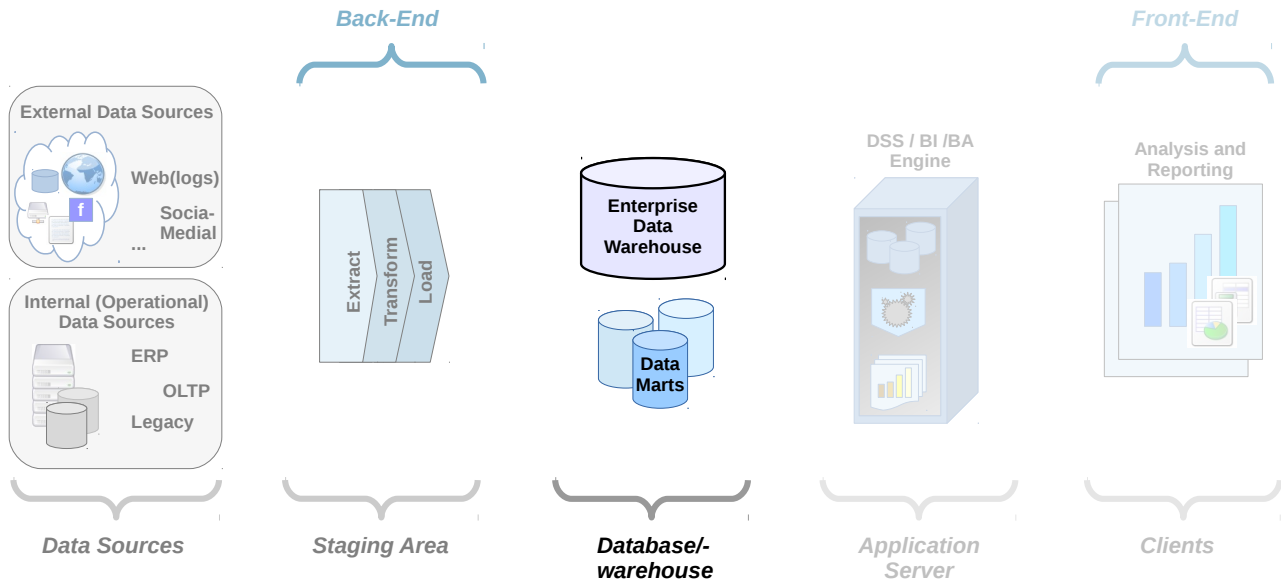


Figure: Data Warehouse Architecture Components

Data Warehousing Architecture: Components

DSS/BI/BA Engine

- Analytics and DSS Software
- Typically an application server

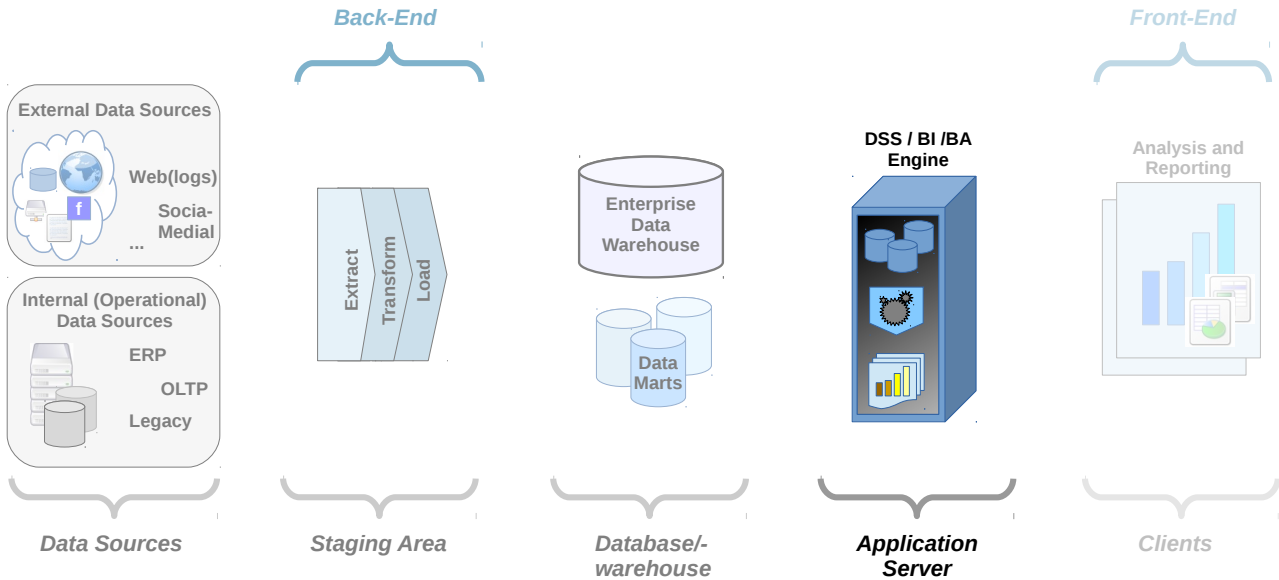


Figure: Data Warehouse Architecture Components

Data Warehousing Architecture: Components

DSS/BI/BA Engine

- Analytics and DSS Software
- Typically an application server
- Sometimes combined with the DB server (two-tier architecture)

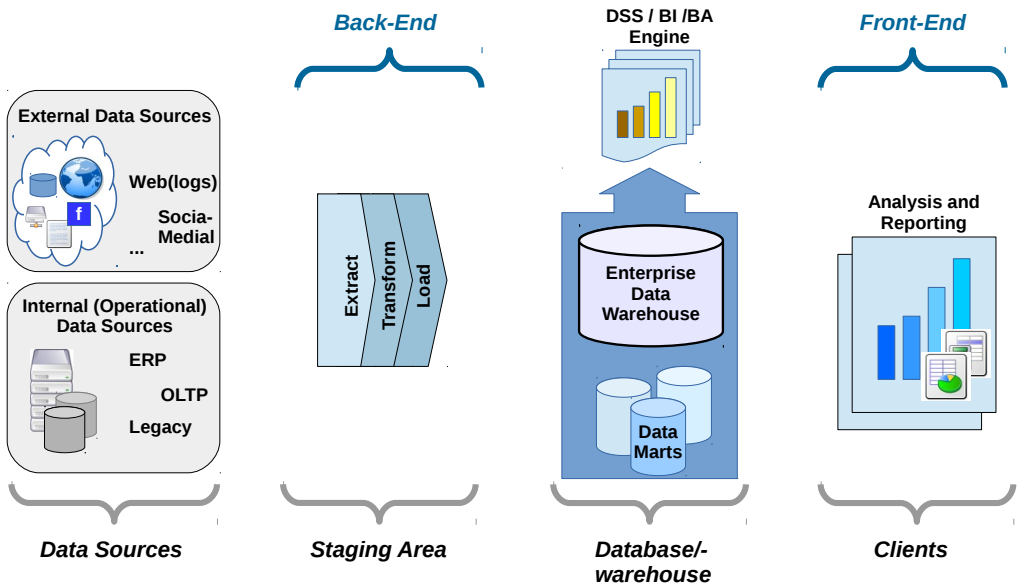


Figure: Data Warehouse Architecture Components

Data Warehousing Architecture: Components

Client / GUI

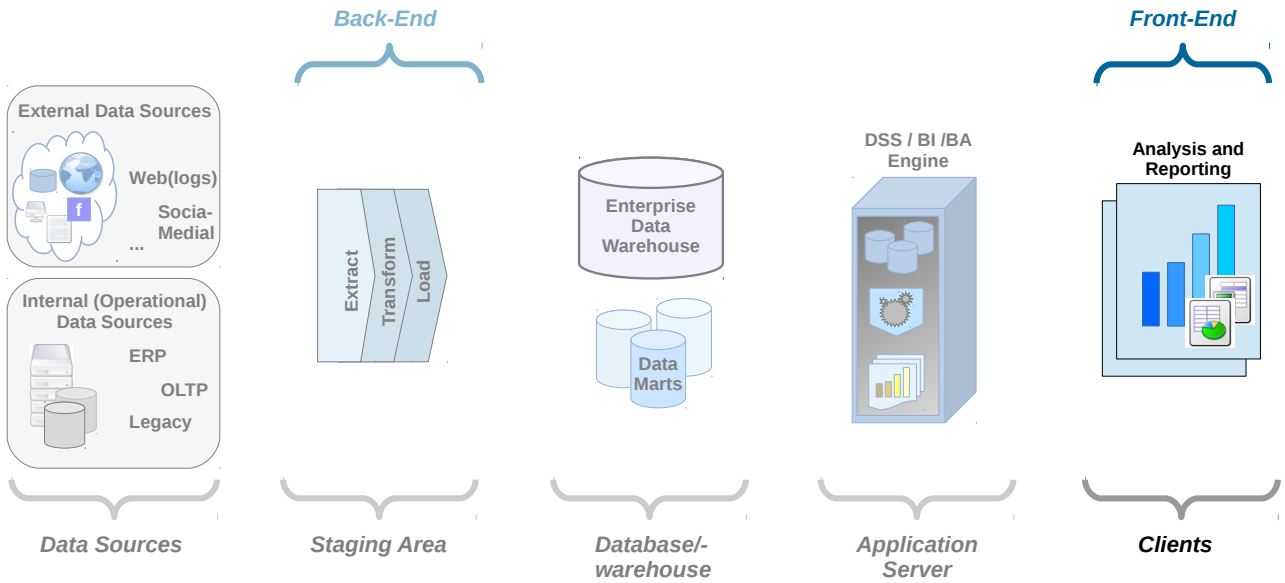


Figure: Data Warehouse Architecture Components

Web-Based Data Warehouse Architecture

- Dedicated Application, Database, and Web-Server

Image not available due to copyright restrictions. Please refer to the source cited below.

Figure: Web Data Warehouse (Source: [Sharda et al., 2018, page 167])

(Further) Data Warehousing Architectures

Indepedent Data Mart

- Data Marts that are independent of a central Data Warehouse

Image not available due to copyright restrictions. Please refer to the source cited below.

Figure: Independent Data Mart (Source: [Sharda et al., 2018, page 169])

(Further) Data Warehousing Architectures

Data Mart Bus

- Data Marts are linked by conformed dimensions

Image not available due to copyright restrictions. Please refer to the source cited below.

Figure: Data Mart Bus (Source: [Sharda et al., 2018, page 169])

(Further) Data Warehousing Architectures

Hub-and-Spoke Architecture

- Normalized Relational Warehouse with
- Dependent Data Marts

Image not available due to copyright restrictions. Please refer to the source cited below.

Figure: Hub-and-Spoke Architecture (Source: [Sharda et al., 2018, page 169])

(Further) Data Warehousing Architectures

Centralized Data Warehouse

- Normalized Relational Warehouse

Image not available due to copyright restrictions. Please refer to the source cited below.

Figure: Normalized Relational Warehouse (Source: [Sharda et al., 2018, page 169])

(Further) Data Warehousing Architectures

Federated Architecture

- Several Data Warehouses, Data Marts, and Legacy Systems
- Data Mapping between them

Image not available due to copyright restrictions. Please refer to the source cited below.

Figure: The Federated Architecture (Source: [Sharda et al., 2018, page 169])

Data Warehousing Architectures: Comparision[2]

	Indep. D.-Marts	Bus Arch.	Hub-&Spoke Arch.	Centr. Arch.	Fed. Arch.
Popularity	12%	26%	**39%**	17%	4%
Information Quality	4.42	5.16	**5.35**	5.23	4.73
System Quality	4.59	**5.60**	5.56	5.41	4.69
Individual Impacts	5.08	**5.80**	5.62	5.64	5.15
Organiz. Imp.	4.66	**5.34**	5.24	5.30	4.77

Table: Results of an empirical study by Ariyachandra and Watson (2006)

Discussion

- Predominantly used (in 2006) was hub-and-spoke architecture (39%), followed by bus, centralized, and independent architectures
- Hub-and-Spoke: most expensive and time-consuming to implement, mostly with entreprise-wise implementations & larger warehouses
- Bus, Hub- & Spoke, and Centralized architecture perform very similar

[2] Based on Ariyachandra and Watson (2006), cited in [Sharda et al., 2018, p. 170].

Data Warehousing Architectures: Issues to Consider

Which Database Management System (DBMS)?

- Mostly relational DBMS, like Oracle, Microsoft SQL,
- Open Source variantly like PostgreSQL, ...

Will parallel processing / partitioning be used?

- Parallel processing: Enables simultaneous processing of queries
- Partitioning (splitting into smaller subtables) for access efficiency

Use of Data Migration Tools to Load the Warehouse

- Moving data is tedious and costly (see application case study 3.3)
- Tools ease this process (in particular on standard sources)

Tools for Data Retrieval and Analysis

Data Warehousing Architectures: Influence Factors[3]

1. Information interdependence between organizational units
2. Upper managements information needs
3. Urgency of need for a data warehouse
4. Nature of end-user tasks
5. Constraints on resources
6. Strategic view of the data warehouse prior to implementation
7. Compatibility with existing systems
8. Perceived ability of the in-house IT staff
9. Technical issues
10. Social/political factors

[3] Based on Ariyachandra and Watson (2005). Cited from [Sharda et al., 2018, p. 170].

Outline and Summary[1]

- Definition & History of Data Warehouses
  See [Sharda et al., 2018, chapter 3.2]
- Data Warehousing Process & Architectures
  See [Sharda et al., 2018, chapter 3.3–3.4]
- Data Integration and ETL Processes
  See [Sharda et al., 2018, chapter 3.5]
- Data Warehouse Development
  See [Sharda et al., 2018, chapter 3.6]
- Data Modelling & Dimensional Modelling
  See [Sherman, 2015, **chapter 9** and parts of 8] and [Sharda et al., 2018, chapter 3.7–3.8]
- Issues, Considerations & Future Trends
  See [Sharda et al., 2018, chapter 3.7–3.8] and **[Chaudhuri et al., 2011]**
- Business Performance Management
  See [Sharda et al., 2018, chapter 3.9-3.11]

▸ Start ▸ Appendix

[1] Note: This lecture integrates knowledge from several further sources (cited where used, except if my own).

Data Integration: Preparation for Lecture

Reading assignments

- Application Case Study 3.3
  [Sharda et al., 2018, pages 177–179]
- Paper [Chaudhuri et al., 2011]

Questions?

Data Warehouse: Motivation

> *The company being profiled within this case study is a privately held Dutch retailer with 2,800 stores across several European countries. This retailer has* **15 different brands** *that include products from toys to cookware,* **each** *brand* **having its own infrastructure**. *Each business entity is* **managed independently** *as one of 15 individual companies in the way they develop their processes, maintain their legacy systems, and make business decisions across finance, IT, supply chain, and general operations.*

Application Case 3.3
[Sharda et al., 2018, p. 177–179]

What are the challenges here?

- many organizational units (brands, divisions, . . .)
- managed independently
- heterogenous IS landscape

- *Decision support* require integrating data from many units & systems
- Operational systems and databases (OLTP) are designed independently and with different objectives in mind (*efficiency* and *consistency*)

Application Case Study 3.3[4]: Discussion

Data Integration: Issues & Lessons Learned

- 6 months for (unsuccessful) integration of SAP ERP
- Estimation of required integration time: 400 days for development of loading from one (out of 15!) SAP ERP system
- Approach using automated system: 5 days for initial SAP ERP loading
  Further 45 days for customising and preparing data for consumption
- Study available (standard) technologies/solutions, focus own development efforts on tasks that require customised/individual solutions

From Prototype to Rollout for Integrating all Brands

- Aim: Integrate all brands (on the long run)
- Approach: Develop a framework for integration (using ETL as ELT tool) on a subset of brands, later apply it to all remaining brands

Further Learning Points

- Keep system design as simple as possible
- Develop standard data governance approach (data integrity beyond implementation!)
- Identify user's needs, e.g., on acceptable latency

[4] See [Sharda et al., 2018, pages 177–179].

Data Integration Capabilities

1. Data Access (from any source)
2. Data Federation (Integration across multiple sources)
3. Change Capture (of changes at the enterprise data source)

Data Integration Capabilities

- Enterprise application integration
  and Service-Oriented Architecture
- Enterprise information integration
- Extraction, Transformation, Load - Process

Data Integration: Enterprise Application Integration (EAI) and Enterprise Information Integration (EII)

Enterprise Application Integration (EAI)

- Integration of application functionality
- Focus on *sharing functionality* rather than data
- Often accomplished by using SOA (Service-Oriented Architecture) Services
- Usable for fast populating of a data warehouse (near real-time)

Enterprise Information Integration (EII)

- Allows virtual data integration by generating views
- Uses predefined Metadata (e.g., XML-based)

Data Integration: Extraction, Transformation, Load - Process

Extraction, Transformation, Load (ETL)

- Extraction: Reading from one/many DBs
- Transformation: Converting to the form needed for the DW or another DB
- Load: Writing the data into the DW

- Often input is written to staging tables to facilitate load
- ETL tools are not only used for populating a Data Warehouse, also for transport between databases
- Automated metadata management (e.g., creation of metadata on how data changes during transport)
- Administration of all runtime processess and operations (e.g., scheduling, error managements, logging)

Image not available due to copyright restrictions. Please refer to the source cited below.

Data Integration: Extraction, Transformation, Load - Process

Issues to Consider when Selecting ETL Tools

- OLAP and Data Mining tools rely no the data transformation!
- ETL tools should support an unlimited number of data source architectures
- Metadata should be automatically handled (capturing, delivering)
- Tools should be conform to open standards
- Interface should be easy to use for developer and functional users
  Complexity in Learning Data Transformation Tools!
- Success only becomes measureable after tools have been learned and the process is completed

Outline and Summary[1]

- Definition & History of Data Warehouses
  See [Sharda et al., 2018, chapter 3.2]
- Data Warehousing Process & Architectures
  See [Sharda et al., 2018, chapter 3.3–3.4]
- Data Integration and ETL Processes
  See [Sharda et al., 2018, chapter 3.5]
- Data Warehouse Development
  See [Sharda et al., 2018, chapter 3.6]
- Data Modelling & Dimensional Modelling
  See [Sherman, 2015, **chapter 9** and parts of 8] and [Sharda et al., 2018, chapter 3.7–3.8]
- Issues, Considerations & Future Trends
  See [Sharda et al., 2018, chapter 3.7–3.8] and **[Chaudhuri et al., 2011]**
- Business Performance Management
  See [Sharda et al., 2018, chapter 3.9-3.11]

▸ Start ▸ Appendix

[1] Note: This lecture integrates knowledge from several further sources (cited where used, except if my own).

Data Warehouse Development: Main Philosophies

Enterprise Data Warehouse (EDW)

- Propagated by Bill Inmon
- Top-Down approach

Data Mart Approach (DM)

- Propagated by Ralph Kimball
- Bottom-Up approach

Hosted Data Warehouse (HDW)

- Outsourcing DW to external provider ("cloud")
- For small enterprises cheaper
- Privacy and security concerns

Data Warehouse Development: (Dis)Advantages of Hosted Data Warehouses[5]

Possible Advantages

- Minimal investment (in infrastructure) required
- Focus on core competencies/business
- Free-up of money / capacity on in-house systems
- Allows powerful solutions at lower cost
- Scalability/Adaptable to varying needs (e.g., growth)
- Often better integrated & optimised solutions (hard- & software)

Possible Disadvantages

- Loss of competencies & control
- For large enterprises in-house solution might be cheapter on the long-run
- Outsourcing sensible applications and data:
  Security & privacy concerns

[5] See also [Sharda et al., 2018, Technology Insight 3.1].

Data Warehouse Development: Comparison EDW and DM

Characteristics	**EDW Approach**	**DM Approach**
Author	Bill Inmon	Ralph Kimball
Overall	top-down	bottom-up
Architecture	a single enterprise-wide (atomic) Datawarehouse feeds departmental DBs	single business process is modelled; consistency by data bus and conforming dimensions
Complexity	high	simple
Tools	ER-diagrams,	dimensional
Tools	data flow diagrams,	modelling

Table: Comparison between Enterprise Data Warehouse and Data Marts.
Source: Adaptation by [Sharda et al., 2018, p.182] from M. Breslin (2004)

Data Warehouse Development: Comparison EDW and DM

Effort	EDW Approach	DM Approach
Scope	several subject areas	one subject area
Development	years for implementation	months for implementation
	$1,000,000+$ Euros	$10,000 - 100,000+$ Euros
	high difficulty	low-medium difficulty
Data prerequisite	common accross enterprise	common within unit
Sources	many operational & external	only few operational & ext.
Size	Giga- to Petabytes	Mega- to Gigabytes
Time Horizon	Historical Data	Near-Current Hist. Data
Update Freq.	weekly, monthly	weekly, daily, hourly
Hardware	Enterprise Servers, mainframe comp.	Workgroup / standard DB Servers
Simult. Users	100 to 1000+	10*s*
Use	Cross-functional optim. & decision making	Optimization within business area

Table: Comparison between Enterprise Data Warehouse and Data Marts.
Source: Adaptation by [Sharda et al., 2018, p.181] from Van den Hoven, J. (2003)

Outline and Summary[1]

- Definition & History of Data Warehouses
  See [Sharda et al., 2018, chapter 3.2]
- Data Warehousing Process & Architectures
  See [Sharda et al., 2018, chapter 3.3–3.4]
- Data Integration and ETL Processes
  See [Sharda et al., 2018, chapter 3.5]
- Data Warehouse Development
  See [Sharda et al., 2018, chapter 3.6]
- Data Modelling & Dimensional Modelling
  See [Sherman, 2015, **chapter 9** and parts of 8] and [Sharda et al., 2018, chapter 3.7–3.8]
- Issues, Considerations & Future Trends
  See [Sharda et al., 2018, chapter 3.7–3.8] and **[Chaudhuri et al., 2011]**
- Business Performance Management
  See [Sharda et al., 2018, chapter 3.9-3.11]

▸ Start ▸ Appendix

[1] Note: This lecture integrates knowledge from several further sources (cited where used, except if my own).

Data Modelling[6]

Levels of Data Models

Three levels with increasing complexity and level of detail:

Conceptual Data Model: Business View

Logical Data Model: Architect View

Pysical Data Model: Developer View

Image not available due to copyright restrictions. Please refer to the source cited below.

Data Modelling: Recap ER-Diagram and Normalisation[7]

- ER-modelling
- Normalisation
- primary key: used to identify an instance of an entity
- foreign key: reference to an instance of another entity (based on their primary key)
- alternate key: could be used in place of a primary key
- surrogate key: primary key generated by DBMS (e.g., based on counter)

Image not available due to copyright restrictions. Please refer to the source cited below.

Data Modelling[8]

Conceptual Data Model

- Overall view of the structure of data in business context
- Independent of any database or physical storage structure
- May contain objects that are never being implemented in physical DB but needed for understanding

Tool for business and IT to define

- Data requirements scope
- Enterprise-wide business terms & measures
- Names, data types, characteristics of entities and their attributes

[8] See [Sherman, 2015, chapter 8].

Data Modelling (2)

Logical Data Model

- Still independent of any database or physical storage structure
- First step in designing architecture of the application
- Specifies the entities & attributes to be implemented
- Identifies relationships between these entities & attributes, and business rules
- Defines primary keys, foreign keys, alternate keys[9]

Physical Data Model

- Represents the logical data model in a DB schema, DBMS specific
- Defines physical objects such as tables and columns
- Specifies referential integrity rules,
  including foreign keys, constraints, event triggers
- Contains DBMS-specific performance and optimisation entities

[9] See [Sherman, 2015, chapter 8, section Entity-Relationship Modelling] for explanations of ER-modelling, primary / foreign / alternate keys, and normalisation.

Data Modelling Workflow[10]

1. Gather business requirements
   - Analyse the data needed by business requirements
   - Identify the data relationships
2. Create the three data models top-down:
   2.1 Conceptual
   2.2 Logical
   2.3 Physical
3. Support the application development
   - Create application specifications
   - Develop applications
   - Deploy applications

Image not available due to copyright restrictions. Please refer to the source cited below.

Dimensional Modelling[11]: Multidimensionality

Definition of Multidimensionality

> *"The ability to organize, present, and analyze data by several dimensions, such as sales by region, by product, by salesperson, and by time (four dimensions)" [Sharda et al., 2018]*

Multidimensional Presentation

Dimensions: products, salespeople, market segments, business units, geographical locations, distribution channels, country, or industry

Measures: money, sales volume, head count, inventory profit, actual versus forecast

Time: daily, weekly, monthly, quarterly, or yearly

[11] Recommended reading: [Sherman, 2015, chapter 9].

Dimensional Modelling: Facts and Fact Tables

- (Numerical) measurement of business activity (descriptive dec. analysis attrib.)
- Examples: Sales, expenses, inventory levels
- Ideally stored at the most granular level, leaving aggregation to analytics
  E.g., Salesprice of every unit in euros and eurocents (flexibility vs. storage)
- Primary key: typically combination of foreign keys, or surrogate key

Types of Facts

Is summarisation possible across dimensions?

additive: sum can be calculated across **all** dimensions
Example: Quantity of sold items (sum over time, stores, . . .)

semiadditive: sum can be calculated across **some** dimensions
Example: Bank account balance (sum over accounts but not months)

non-additive: sum **can not** be calculated over **any** dimension
Example: Unit price, ratios, temperature

Note that other aggregation operations (e.g., averaging) might be reasonable.

Fact Tables

Contain columns of two types:

1. Measures of business activity
2. (Foreign) Keys pointing to dimension tables

Dimensional Modelling: Fact Table Example

Image not available due to copyright restrictions. Please refer to the source cited below.

Figure: Fact Table Example (Source: [Sherman, 2015, page 199])

Dimensional Modelling: Dimensions and Dimension Tables

- Define the Who What Where Why
- Establish business context of the measures
- Used for filtering and aggregating content in fact table
  (e.g., the dimensions used to slice & dice the values in the fact table)
  based on one-to-many relationships with rows in the central fact table
- Descriptive in nature (although some are numeric)
- Often, there is a **dimension hierarchy**
  Example: Geographical dimension (continent:country:province:city)
- Primary key: ideally surrogate key, or source system's primary key

Dimensions

- descriptive, i.e. understandable for business people and BI application designer
- complete, **no missing values! Why?**
- unique, to ensure values are uniquely identifiable
- valid, to be useful for business decisions

Dimensional Modelling: Example

Image not available due to copyright restrictions. Please refer to the source cited below.

Figure: Dimensional Modelling Example (Source: [Sherman, 2015, page 235])

Dimensional Modelling

- A retrieval-based system supporting high-volume query access

Star Schema

- **Star (join) schema** is the most common and simplest dimensional modelling style
- Comprises a **fact table** surrounded by and connected to several **dimension tables**
- Design aim: fast query response time, simplicity, ease of maintenance for read-only DB
- Consequences of optimisation for analytics rather than storage:
  - to avoid redundancy in fact table, it uses foreign keys pointing to dimension tables
  - dimensions are de-normalised, i.e. a single table per dimension combining attributes that spead across multiple tables in the source system

Snowflake Schema

- Extends the star schema
- Usually only one central fact table, connected to
- multiple dimensions that are normalised into multiple related tables

Dimensional Modelling: Star Schema Example

Image not available due to copyright restrictions. Please refer to the source cited below.

Figure: Star Schema Example (Source: [Sherman, 2015, page 209])

Dimensional Modelling: Snowflake Schema Example

Image not available due to copyright restrictions. Please refer to the source cited below.

Figure: Snowflake Schema Example (Source: [Sherman, 2015, page 210])

Dimensional Modelling: Star & Snowflake Schema Examples

Image not available due to copyright restrictions. Please refer to the source cited below.

Figure: The Star & Snowflake Schema (Source: [Sharda et al., 2018, page 184])

Dimensional Modelling: Multifact Star Schema Example

Image not available due to copyright restrictions. Please refer to the source cited below.

Figure: Multifact Star Schema Example (Source: [Sherman, 2015, page 212])

- In practice, multiple facts are required to be considered together
- They share dimensions, called **conformed dimensions**
- For a discussion on conforming dimensions and Master Data Management, see also [Sherman, 2015, p. 220–221]

Dimensional Modelling: Mapping to Business Report

Image not available due to copyright restrictions. Please refer to the source cited below.

Figure: Mapping from Dimensional Model to Business Report (Source: [Sherman, 2015, page 217])

Dimensional Modelling: ER vs. Dimensional Model

Image not available due to copyright restrictions. Please refer to the source cited below.

Figure: ER vs. Dimensional Model (Source: [Sherman, 2015, page 216])

Note: For differences between ER and Dimensional Modelling see also [Sherman, 2015, Table 9.1, p. 215]

Analysis of Data in Data Warehouse: OLAP

Online Analytical Processing (OLAP)

- Approach for answering ad-hoc questions (reports)
- by executing multidimensional analytical queries against Data Warehouses (or other organisational data repositories)
- Convert data into information for decision support
- Conducting statistical and other analyses
- Most common data anylsis technique for DW
- Data cubes, drill-down / rollup, slice & dice, ...
- Compare to Online Transaction Processing (OLTP):
  Capturing and storing data from ERP, CRM, POS, ... (see next page)
  The main focus there is on efficiency of routine tasks

Multidimensional Schema: Online Analytical Processing (OLAP)

Image not available due to copyright restrictions. Please refer to the source cited below.

Figure: Example of a 3-Dimensional Data Cube (Source: [Sharda et al., 2018, page 186])

Data Cube

- multidimensional array (data structure)
- supports fast analysis of data
- overcomes limitations of traditional RDBMS (those focus on row-wise adding/deleting/updating capabilities)

Slice

- A 2D-subset of a multidim. array
- Example: Sales of a specific product along time & region
- Generalisation beyond 2D: **Dice**

Online Analytical Processing (OLAP)

Image not available due to copyright restrictions. Please refer to the source cited below.

Figure: Example of a 3-Dimensional Data Cube (Source: [Sharda et al., 2018, page 186])

Drill Down/Up

- Navigating through levels of aggregation/summarisation
- E.g.: Going from total sales to sales of a specific product at a time & region

Roll-Up

- Computing all of the data relationships for one or more dimensions

Pivot

- Changing dimensional orientation of a report / query page display

Outline and Summary[1]

- Definition & History of Data Warehouses
  See [Sharda et al., 2018, chapter 3.2]
- Data Warehousing Process & Architectures
  See [Sharda et al., 2018, chapter 3.3–3.4]
- Data Integration and ETL Processes
  See [Sharda et al., 2018, chapter 3.5]
- Data Warehouse Development
  See [Sharda et al., 2018, chapter 3.6]
- Data Modelling & Dimensional Modelling
  See [Sherman, 2015, **chapter 9** and parts of 8] and [Sharda et al., 2018, chapter 3.7–3.8]
- Issues, Considerations & Future Trends
  See [Sharda et al., 2018, chapter 3.7–3.8] and **[Chaudhuri et al., 2011]**
- Business Performance Management
  See [Sharda et al., 2018, chapter 3.9-3.11]

▸ Start ▸ Appendix

[1] Note: This lecture integrates knowledge from several further sources (cited where used, except if my own).

Issues and Considerations

- **Sponsorship Chain:** Support from top-level executives
  That is, a resouce-providing "sponsor" and a decisive but flexible "project driver"
  Why: DW needs architectural framework for decision analysis throughout enterprise
- **Expectation Management:** Realistic expectations, plan capacity & speed reserves
- **Data Landfill:** Filling Data Warehouse with irrelevant data
- **Awareness of DW DB Design Particularities:**
  redundant, not normalized, multidimensional
  (vs. transactional DB design: non-redunant, normalized, relational)
- **User before Technology:** DW must serve the user
  Focus on user- rather than technology-orientation
- **Moving from Reports to Alerts:** Periodical reports are a 1st step, but event-triggered as-soon-as-possible alerts are the aim
- **Clear and Simple:** Avoid conflicting data definitions and formats

Securing a Data Warehouse[12]

1. Establish Corporate & Security Policies & Procedures
   Communicate them to employees on all levels
2. Develop logical security procedures & access restriction techniques
   This comprises authentication, access control, encryption
3. Limit physical access to data center
4. Establish an internal control review process
   Should focus on security & privacy

[12] Based on [**?**], cited in [Sharda et al., 2018, p.190–191].

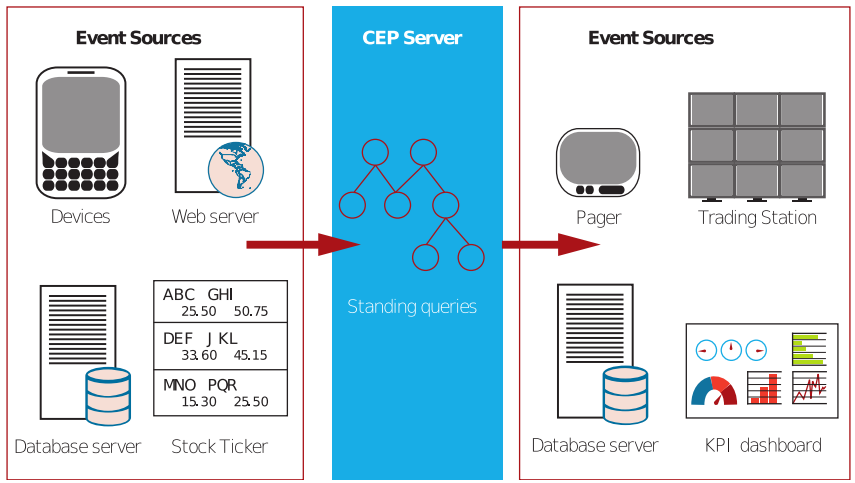
Current Trends & Developments

Big Data

- Recap: Velocity, Volume, and Variety
- Data from many sources (Web, Social Media, Internet of Things)
- Data of many kind (structured, semi- and unstructured)
- Already discussed: Streaming/Online processing

Near Real-Time Streaming BI

- Complex Event Processing (CEP) Engines
- Search for **events** over streaming operational data: pre-specified patterns or temporal trends
- Do not require data being loaded into warehouse
- Define so-called **standing** (or continuous) queries that use operations such as filtering, windowing, aggregation, unions, joins



Current Trends & Developments: Data Lakes

Data Lakes

- Data Warehouse: focus on cleaned, structured & tabular data
  Schema-on-write: Shape & structure data when writing it into DW

- **Data Lake:** Large storage for all types of data
  Loosely defined, stored in native/raw format
  Schema-on-read: Shape & structure data when reading it from DL
  Flexible, but sophisticated (requires specialists as users)
  Low cost storage, but limitations in speed

Current Trends & Developments: DB Technologies

In-Database Processing

- "Put the algorithm where the data is"
- Integrate analytics algorithms into DW
- Advantage: Performance improvements, high-throughput
- Disadvantage: Complexity, requires specialised tools

Database Technologies

- **Columnar:** column-oriented processing
  Efficient when doing operations over most rows but (few) columns
  Allows better compression (due to homogenity of data within each column)
- **In-Memory Storing:** (RAM rather than hard disk, fast but costly)

Current Trends & Developments: Distributed Systems

Distributed Systems

- Platforms based on distributed file systems
- Highly parallel processing using **MapReduce**
- Open Source Apache **Hadoop** Ecosystem
- Tools for translating SQL-like queries
  (also available for R!)

Further Trends

- Open Source Software
- Software as a Service, extending the Application Service Provider model
- Virtualisation & Cloud Computing

Outline and Summary[1]

- Definition & History of Data Warehouses
  See [Sharda et al., 2018, chapter 3.2]
- Data Warehousing Process & Architectures
  See [Sharda et al., 2018, chapter 3.3–3.4]
- Data Integration and ETL Processes
  See [Sharda et al., 2018, chapter 3.5]
- Data Warehouse Development
  See [Sharda et al., 2018, chapter 3.6]
- Data Modelling & Dimensional Modelling
  See [Sherman, 2015, **chapter 9** and parts of 8] and [Sharda et al., 2018, chapter 3.7–3.8]
- Issues, Considerations & Future Trends
  See [Sharda et al., 2018, chapter 3.7–3.8] and **[Chaudhuri et al., 2011]**
- Business Performance Management
  See [Sharda et al., 2018, chapter 3.9-3.11]

▸ Start ▸ Appendix

[1] Note: This lecture integrates knowledge from several further sources (cited where used, except if my own).

Data Warehousing: Aligning it with Business Strategy

Expedias **problem was not lack of data**. *The customer satisfaction group at Expedia knew that it had lots of data. In all, there were 20 disparate databases with 20 different owners. [. . .]*

The business analyst spent 2 to 3 weeks every month pulling and aggregating the data, leaving virtually no time for analysis. Eventually, the group realized that **it wasnt enough to aggregate the data**. *The* **data needed to be viewed in the context of strategic goals**, *and individuals had to take ownership of the results.*

Application Case 3.6
[Sharda et al., 2018, p. 208–209]

Key Messages:

- data alone is worthless: "data landfill"
- view it in context of strategy!

Approach:

1. Analyse fundamental performance drivers, and the link between performance & costs
2. Decide how to measure satisfaction
   Basis for scorecards & KPIs
3. Set right performance targets
4. Put data into context
   Link to ongoing actions

Data Warehousing: Aligning it with Business Strategy

Image not available due to copyright restrictions. Please refer to the source cited below.

Figure: The closed-loop BPM cycle (Source: [Sharda et al., 2018, page 197])

Strategy:

- Where do we want to go?
- Long-time perspective (several years)
- Identify & state organisation's mission, vision, objectives
- Develop according plans (strategic, tactical, operational)

Plan:

- How do we get there?
- Detailed **operational plan**
  1. Start with revenue-generating tactics
  2. Next, focus on their associated costs
  3. Finally, overhead & financing costs

Data Warehousing: Aligning it with Business Strategy (2)

Image not available due to copyright restrictions. Please refer to the source cited below.

Figure: The closed-loop BPM cycle (Source: [Sharda et al., 2018, page 197])

Monitor/Analyse:

1. How are we doing?
2. Identify indicators or measures
3. Develop measuring strategy
4. Examples of Measures:
   Key performance indicators (KPIs)
   See next slides ...

Act and Adjust:

1. What needs to be done differently?

Performance Measurement: Key Performance Indicator (KPI)

Key Performance Indicator

- A **strategically aligned** measure
  embodies a strategic objective
- Measures **performance against a target**
  E.g., achievment target
- with a **performance range** (above/on/below target)
- **Encoding** that alows visual display (e.g., percentage)
- Assigned to **time frame**

Performance Measurement: Key Performance Indicator (KPI) (2)

Categories of KPIs

- Outcome KPIs or lagging indicators
  measure output of past activity
- Driver KPIs or leading indicators
  measure activities that will impact outcome KPIs (later)

Examples

- Customer performance: Customer satisfaction/retention, ...
- Service performance: Metrics for service-call resolution rates, ...
- Sales operations: New pipeline accounts, sales meetings secured, ...
- Sales plan/forecast: Metrics for price-to-purchase accuracy, ...

Business Performance Management

- A set of integrated, closed-loop management & analytic processes
- Tools for defining strategic goals, operationalising (making them measurable) & evaluating achievment of them
- Methods & tools for monitoring key performance indicators (KPIs), linked to organisational strategy

BPM Instruments Overview

Balanced Scorecard

Indicators along 4 dimensions

- customer
- financial
- business process
- learning & growth

Dashboard

Graphical presentation of several performance indicators
in a single page using dials/gauges

Key Performance Indicator (KPI)

A strategically aligned measure of performance against a target,
suited for visual display

Outline and Summary[1]

- Definition & History of Data Warehouses
  See [Sharda et al., 2018, chapter 3.2]
- Data Warehousing Process & Architectures
  See [Sharda et al., 2018, chapter 3.3–3.4]
- Data Integration and ETL Processes
  See [Sharda et al., 2018, chapter 3.5]
- Data Warehouse Development
  See [Sharda et al., 2018, chapter 3.6]
- Data Modelling & Dimensional Modelling
  See [Sherman, 2015, **chapter 9** and parts of 8] and [Sharda et al., 2018, chapter 3.7–3.8]
- Issues, Considerations & Future Trends
  See [Sharda et al., 2018, chapter 3.7–3.8] and **[Chaudhuri et al., 2011]**
- Business Performance Management
  See [Sharda et al., 2018, chapter 3.9-3.11]

▸ Start ▸ Appendix

[1] Note: This lecture integrates knowledge from several further sources (cited where used, except if my own).

Outlook: Predictive Analytics

Next Lecture

- Predictive Analytics: Data Mining: [Sharda et al., 2018, chapter 4–5]

Any More Questions?

Thank you!

Appendix

Bibliography I


Chaudhuri, S., Dayal, U., and Narasayya, V. (2011).
An overview of business intelligence technology.
Communications of the ACM, 54(8):88–98.


Cover, T. M. and Thomas, J. A. (2006).
Elements of Information Theory.
Wiley-Interscience, 2 edition.


Hand, D. J. (2008).
Statistics: A very short introduction.
Oxford University Press.


Hand, D. J., Mannila, H., and Smyth, P. (2001).
Principles of Data Mining.
Adaptive Computation and Machine Learning. The MIT Press.


Kimball, R. and Ross, M. (2013).
The Data Warehouse Toolkit – The Definitive Guide to Dimensional Modelling.
Wiley, 3 edition.


Segel, E. and Heer, J. (2010).
Narrative visualization: Telling stories with data.
IEEE Transactions on Visualization and Computer Graphics, 16(6):1139–1148.

Bibliography II


Sharda, R., Delen, D., and Turban, E. (2018).
Business Intelligence, Analytics, and Data Science: A Managerial Perspective.
Pearson, 4 edition.


Sherman, R. (2015).
Business Intelligence Guidebook: From Data Integration to Analytics.
Morgan Kaufmann.


Winston, W. L. (1997).
Operations Research: Applications and Algorithms.
Wadsworth Publishing Company, 3rd edition edition.


Žliobaitė, I., Pechenizkiy, M., and Gama, J. (2016).
An overview of concept drift applications.
In Japkowicz, N. and Stefanowski, J., editors, *Big Data Analysis: New Algorithms for a New Society*, pages 91–114. Springer, Cham.