Tamkang University
淡江大學

商業智慧實務

Tamkang University

Practices of Business Intelligence

描述性分析 II：

商業智慧與資料倉儲

(Descriptive Analytics II: Business Intelligence and Data Warehousing)

1071BI05
MI4 (M2084) (2888)
Wed, 7, 8 (14:10-16:00) (B217)

Min-Yuh Day
戴敏育
Assistant Professor
專任助理教授
Dept. of Information Management, Tamkang University
淡江大學 資訊管理學系
http://mail. tku.edu.tw/myday/

2018-10-17

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容(Subject/Topics)
1	2018/09/12	商業智慧實務課程介紹 (Course Orientation for Practices of Business Intelligence)
2	2018/09/19	商業智慧、分析與資料科學 (Business Intelligence, Analytics, and Data Science)
3	2018/09/26	人工智慧、大數據與雲端運算 (ABC: AI, Big Data, and Cloud Computing)
4	2018/10/03	描述性分析I：數據的性質、統計模型與可視化 (Descriptive Analytics I: Nature of Data, Statistical Modeling, and Visualization)
5	2018/10/10	國慶紀念日 (放假一天) (National Day) (Day off)
6	2018/10/17	描述性分析II：商業智慧與資料倉儲 (Descriptive Analytics II: Business Intelligence and Data Warehousing)

課程大綱 (Syllabus)

週次 (Week)　日期 (Date)　內容(Subject/Topics)

7　2018/10/24　預測性分析I：資料探勘流程、方法與演算法
(Predictive Analytics I: Data Mining Process, Methods, and Algorithms)

8　2018/10/31　預測性分析II：文本、網路與社群媒體分析
(Predictive Analytics II: Text, Web, and Social Media Analytics)

9　2018/11/07　期中報告 (Midterm Project Report)

10　2018/11/14　期中考試 (Midterm Exam)

11　2018/11/21　處方性分析：最佳化與模擬
(Prescriptive Analytics: Optimization and Simulation)

12　2018/11/28　社會網絡分析
(Social Network Analysis)

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容(Subject/Topics)
13	2018/12/05	機器學習與深度學習 (Machine Learning and Deep Learning)
14	2018/12/12	自然語言處理 (Natural Language Processing)
15	2018/12/19	AI交談機器人與對話式商務 (AI Chatbots and Conversational Commerce)
16	2018/12/26	商業分析的未來趨勢、隱私與管理考量 (Future Trends, Privacy and Managerial Considerations in Analytics)
17	2019/01/02	期末報告 (Final Project Presentation)
18	2019/01/09	期末考試 (Final Exam)

Business Intelligence (BI)

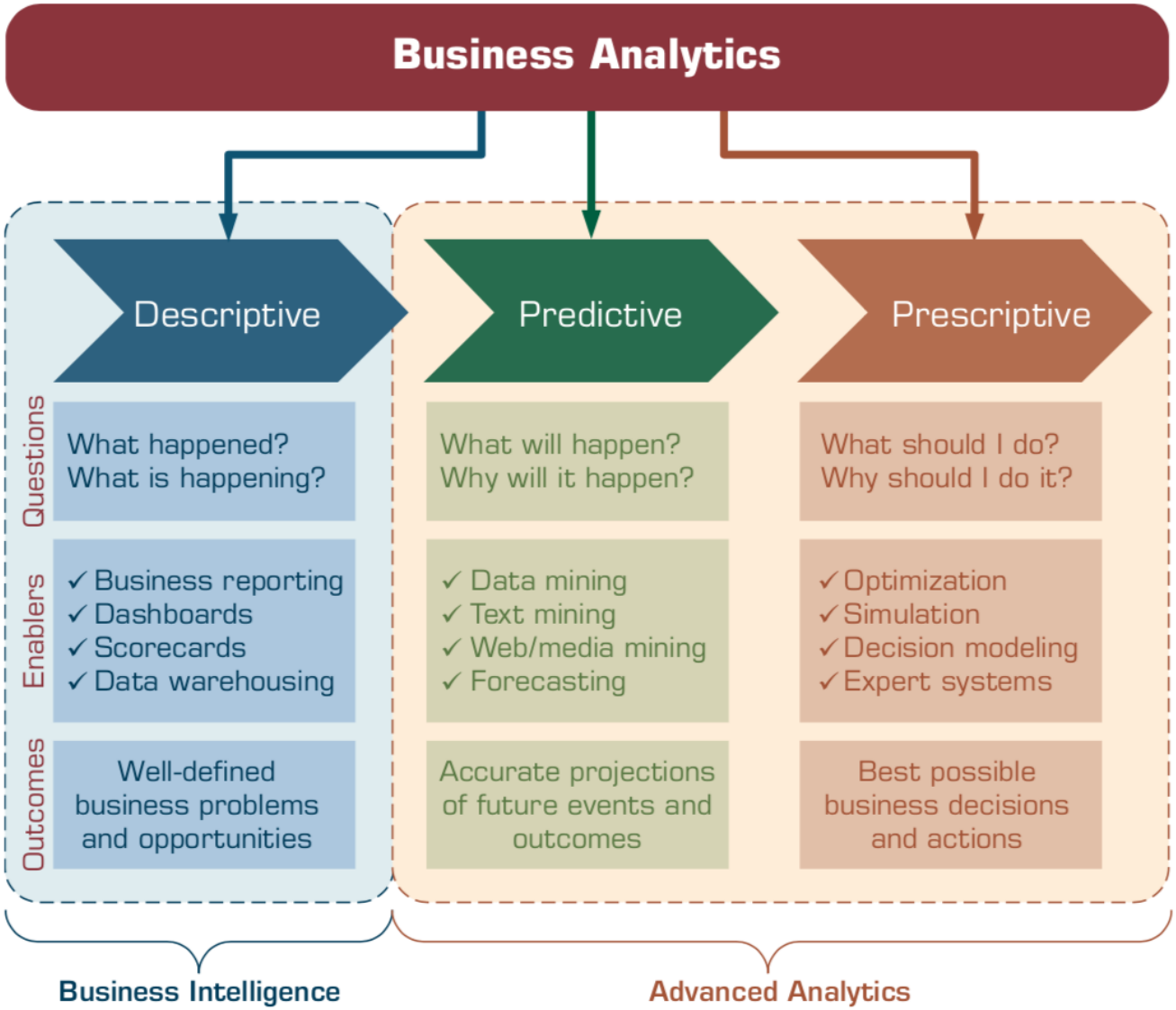
1. Introduction to BI and Data Science
2. Descriptive Analytics
3. Predictive Analytics
4. Prescriptive Analytics
5. Big Data Analytics
6. Future Trends

Descriptive Analytics II: Business Intelligence and Data Warehousing

Outline

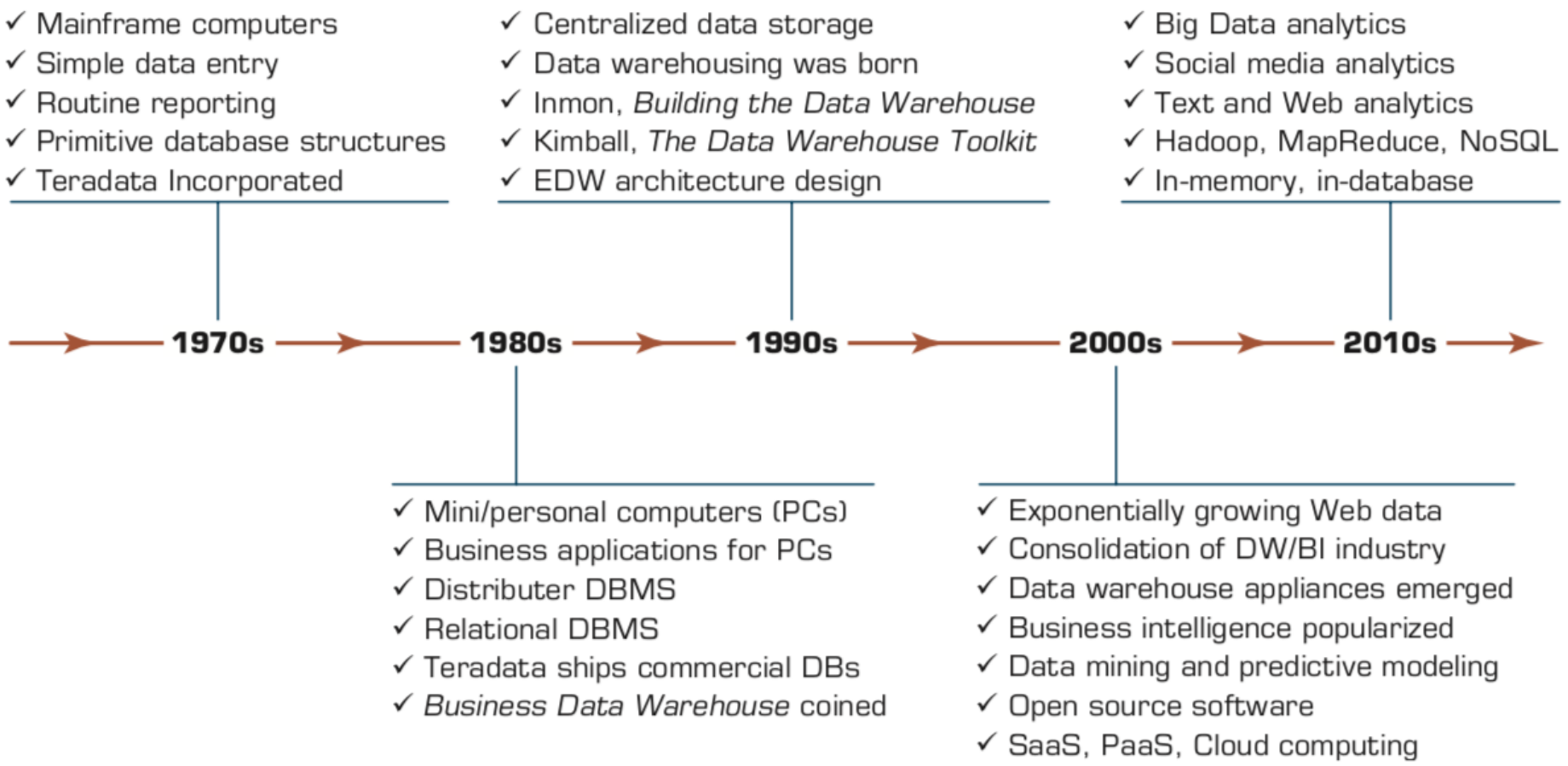
- Descriptive Analytics II
- Business Intelligence
- Data Warehousing
- Data Integration and the Extraction, Transformation, and Load (ETL) Processes
- Business Performance Management (BPM)
- Performance Measurement
  - Balanced Scorecards
  - Six Sigma

Relationship between Business Analytics and BI, and BI and Data Warehousing



Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

A List of Events That Led to Data Warehousing Development

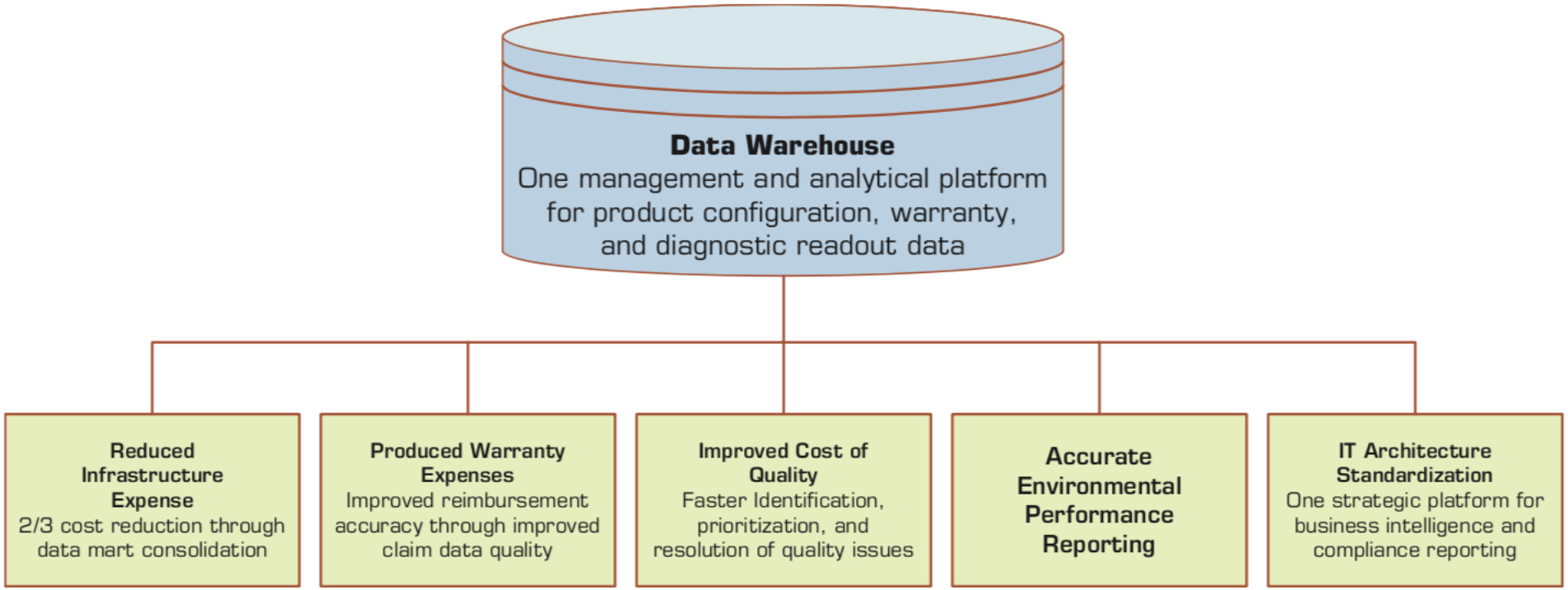


Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Characteristics of Data Warehousing

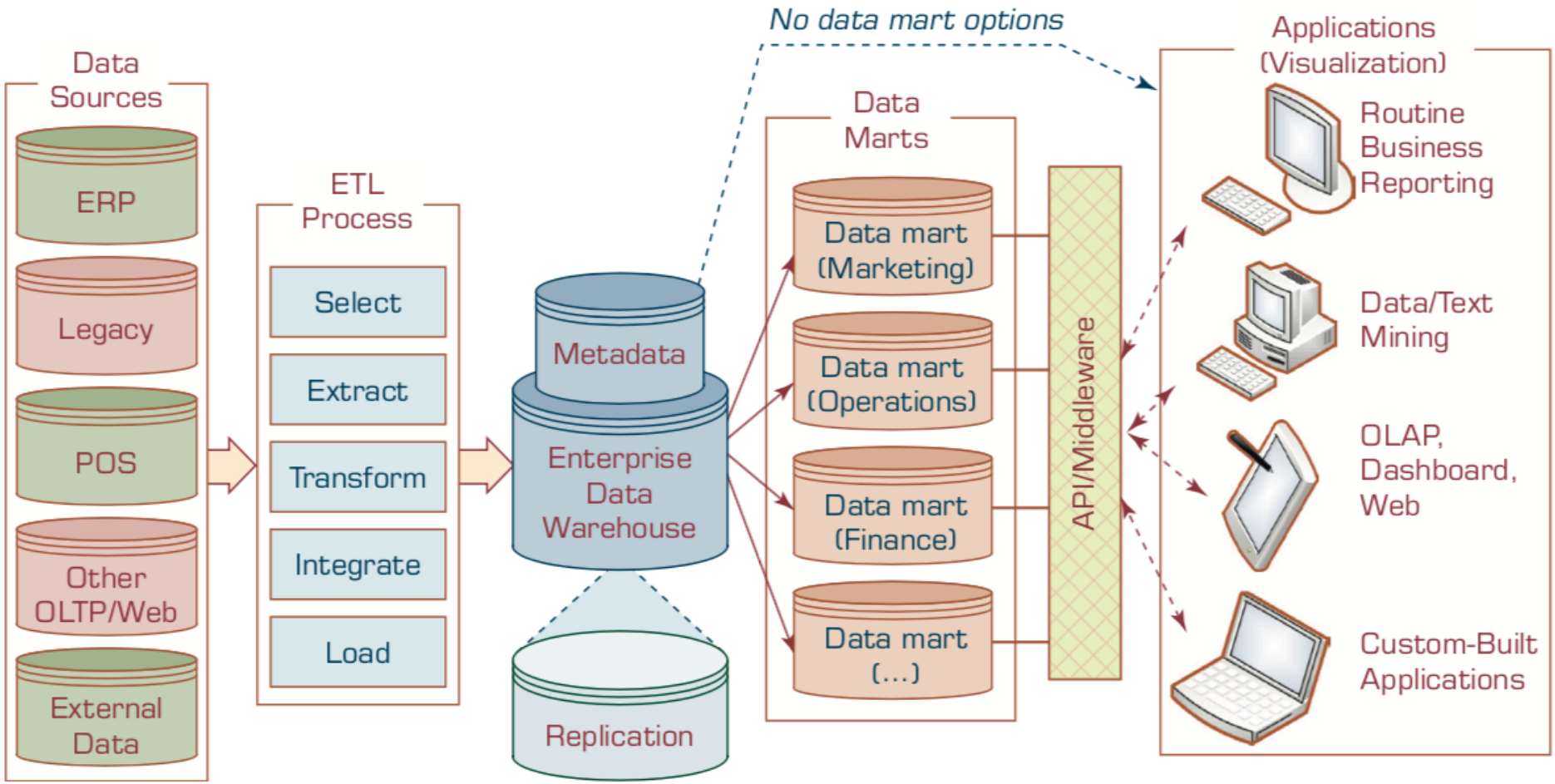
- Subject oriented
  - Data are organized by detailed subject, such as sales, products, or customers, containing only information relevant for decision support.
- Integrated
  - Integration is closely related to subject orientation.
- Time variant (time series)
  - A warehouse maintains historical data.
- Nonvolatile
  - After data are entered into a data warehouse, users cannot change or update the data.

Data-Driven Decision Making—Business Benefits of the Data Warehouse



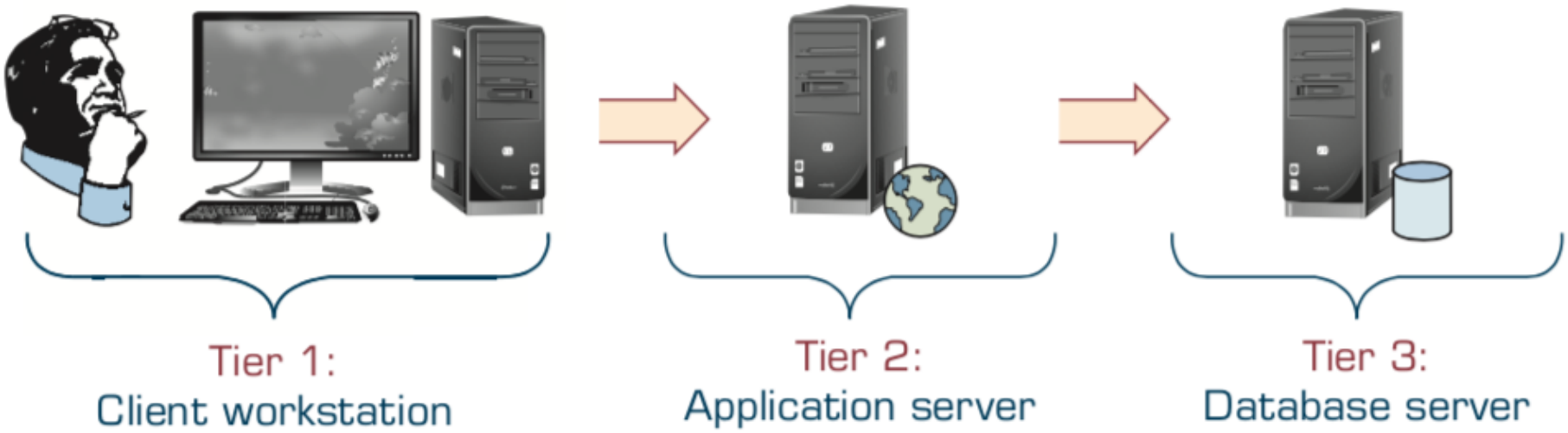
Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

A Data Warehouse Framework and Views



Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Architecture of a Three-Tier Data Warehouse



Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

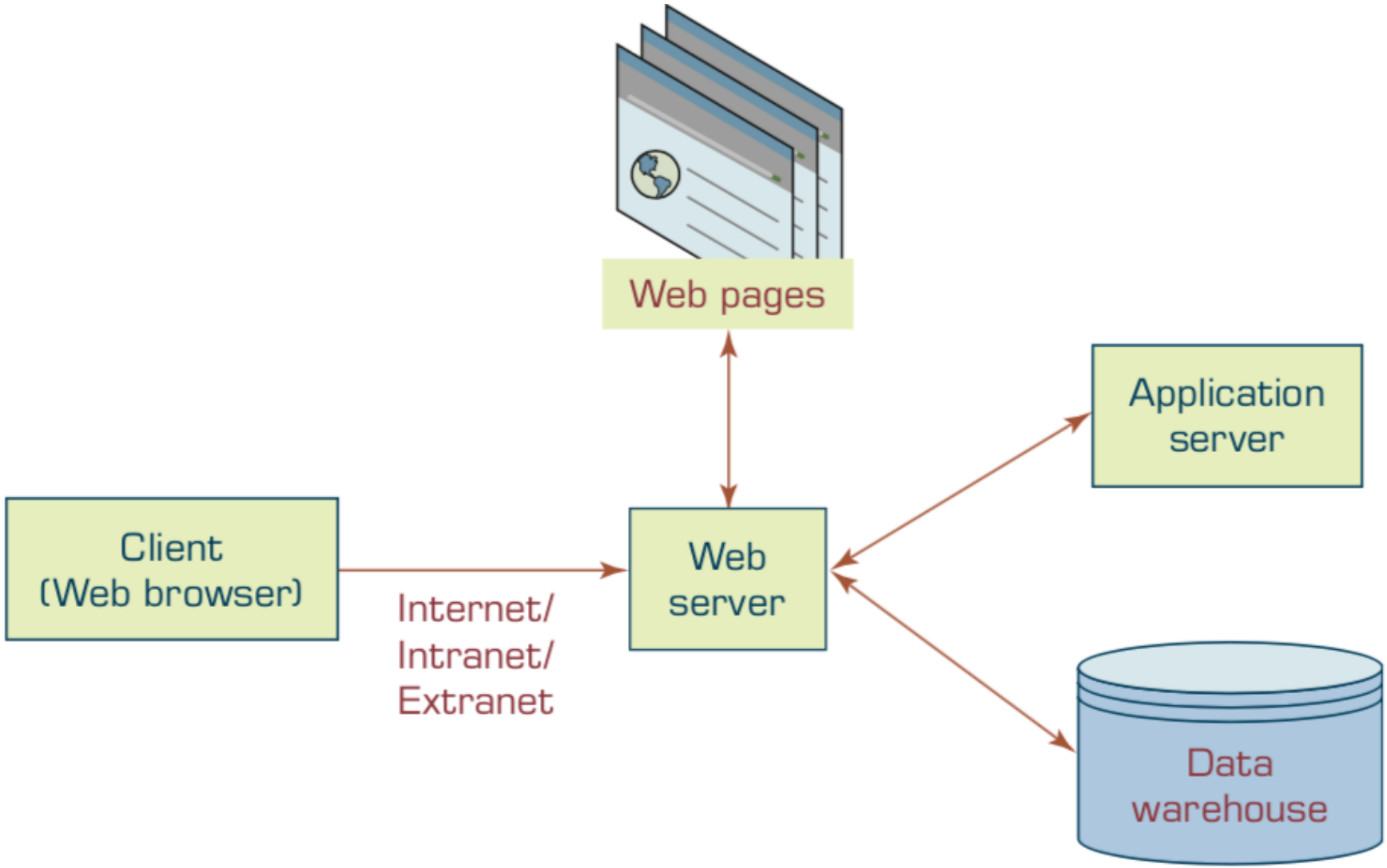
Architecture of a Two-Tier Data Warehouse

Tier 1:
Client workstation

Tier 2:
Application & database server

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Architecture of Web-Based Data Warehousing



Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

5 Alternative Data Warehouse Architectures

a. Independent data marts.

b. Data mart bus architecture

c. Hub-and-spoke architecture

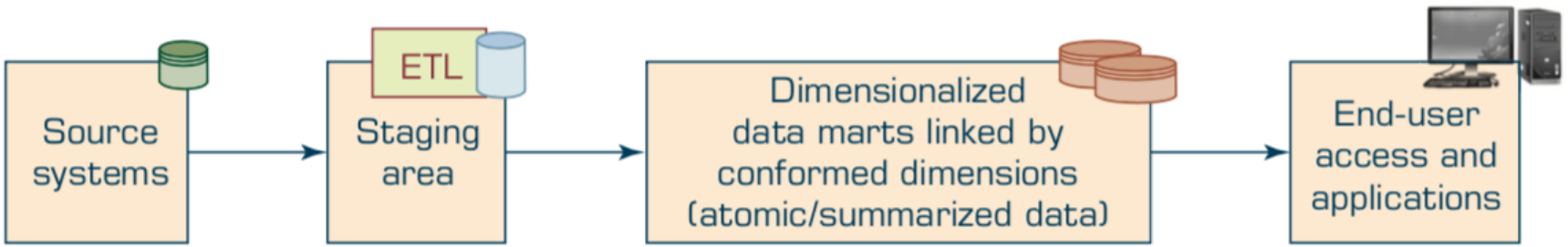
d. Centralized data warehouse

e. Federated data warehouse

5 Alternative Data Warehouse Architectures

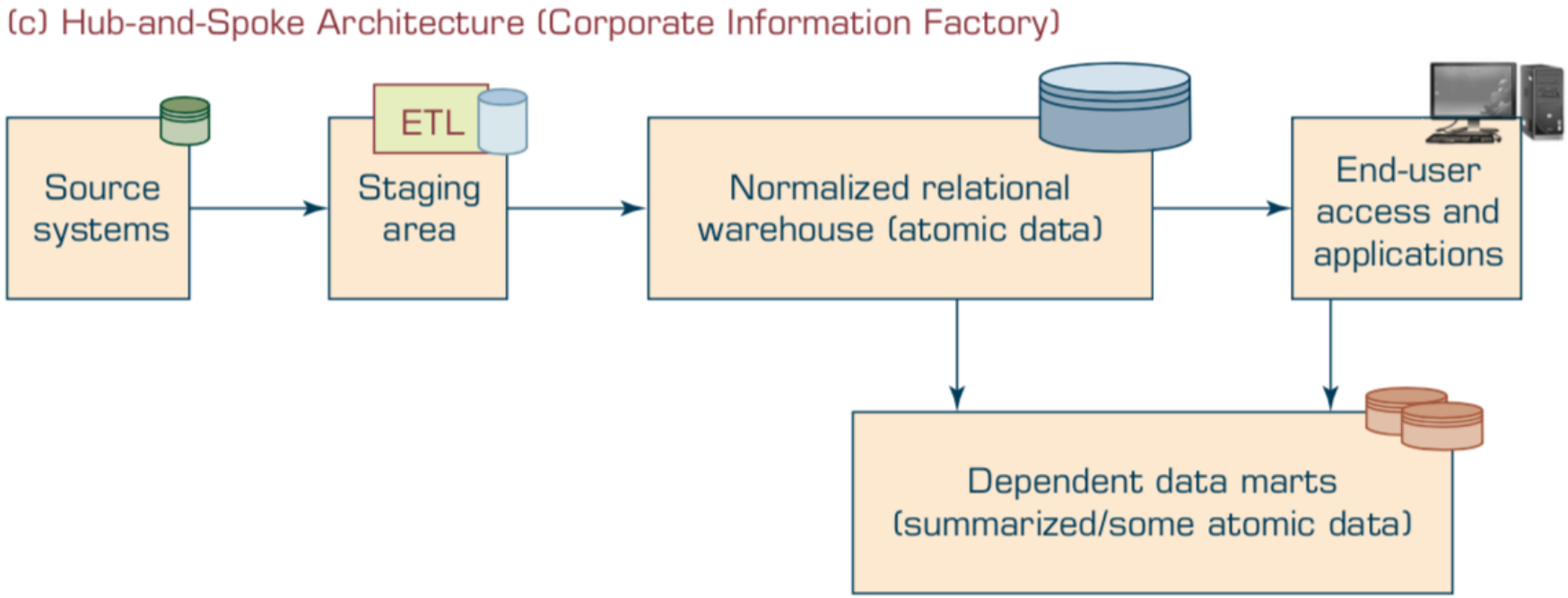
(a) Independent Data Mart Architectures

(b) Data Mart Bus Architecture with Linked Dimensional Data Marts



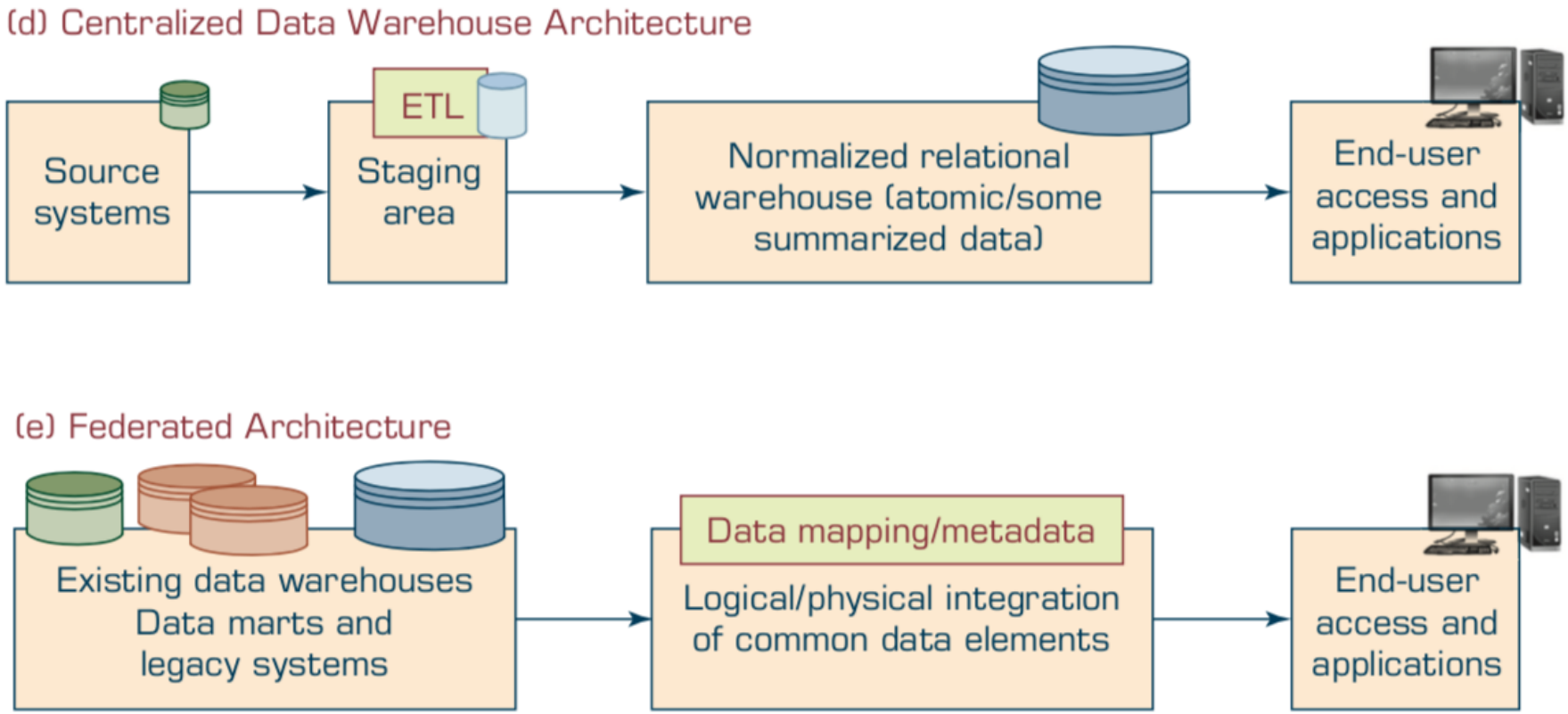
Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

5 Alternative Data Warehouse Architectures



Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

5 Alternative Data Warehouse Architectures



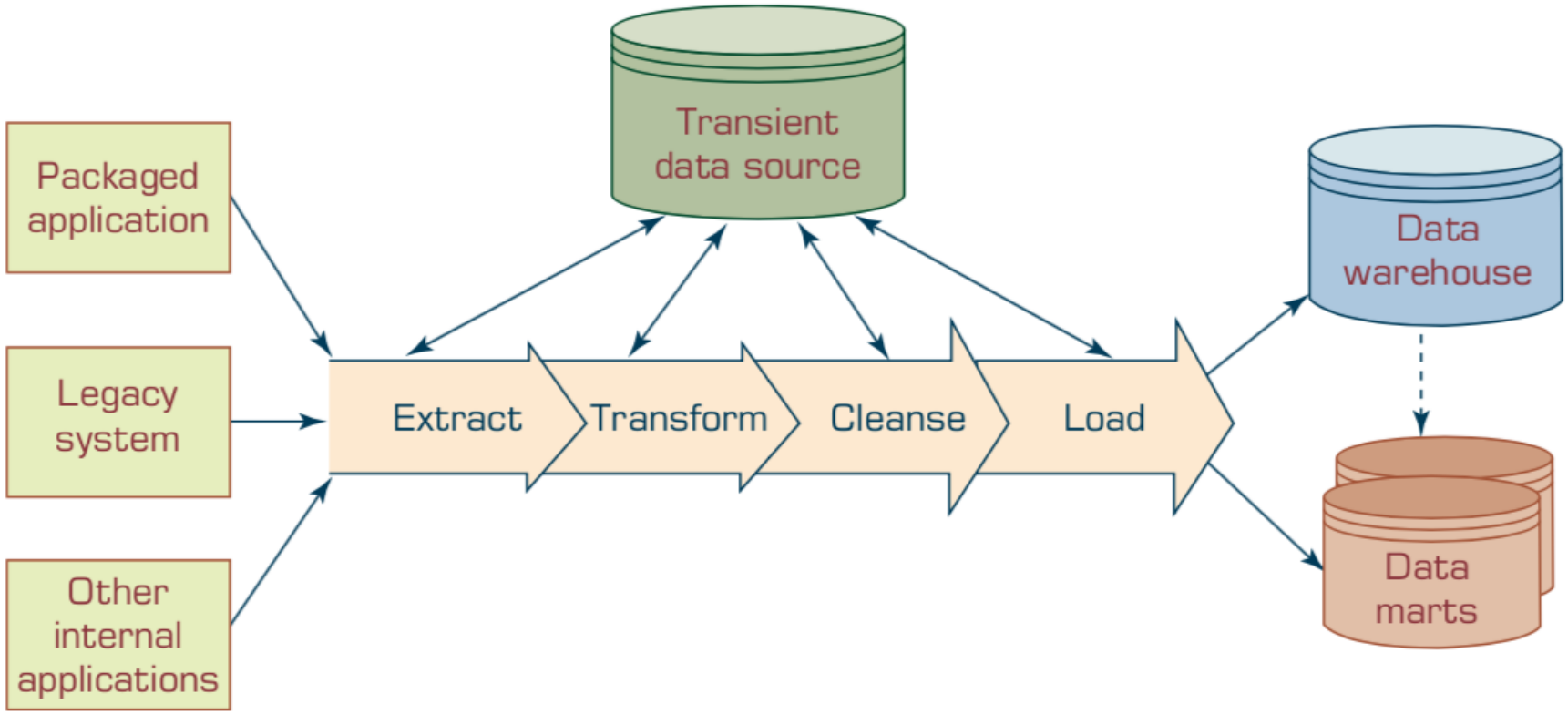
Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Average Assessment Scores for the Success of the DW Architectures

	Independent DMs	Bus Architecture	Hub-and-Spoke Architecture	Centralized Architecture (No Dependent DMs)	Federated Architecture
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
Organizational Impacts	4.66	5.34	5.24	5.30	4.77

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

The ETL Process



Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Sample List of Data Warehousing Vendors

Vendor	Product Offerings
Business Objects (businessobjects.com)	A comprehensive set of BI and data visualization software (now owned by SAP)
Computer Associates (cai.com)	Comprehensive set of data warehouse (DW) tools and products
DataMirror (datamirror.com)	DW administration, management, and performance products
Data Advantage Group (dataadvantagegroup.com)	Metadata software
Dell (dell.com)	DW servers
Embarcadero Technologies (embarcadero.com)	DW administration, management, and performance products
Greenplum (greenplum.com)	Data warehousing and data appliance solution provider (now owned by EMC)
Harte-Hanks (harte-hanks.com)	Customer relationship management (CRM) products and services
HP (hp.com)	DW servers
Hummingbird Ltd. (hummingbird.com)	DW engines and exploration warehouses

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Sample List of Data Warehousing Vendors

Vendor	Product Offerings
Hyperion Solutions (hyperion.com)	Comprehensive set of DW tools, products, and applications
IBM InfoSphere (www-01.ibm.com/software/data/infosphere)	Data integration, DW, master data management, Big Data products
Informatica (informatica.com)	DW administration, management, and performance products
Microsoft (microsoft.com)	DW tools and products
Netezza	DW software and hardware (DW appliance) provider (now owned by IBM)
Oracle (including PeopleSoft and Siebel; oracle.com)	DW, ERP, and CRM tools, products, and applications
SAS Institute (sas.com)	DW tools, products, and applications
Siemens (siemens.com)	DW servers
Sybase (sybase.com)	Comprehensive set of DW tools and applications
Teradata (teradata.com)	DW tools, DW appliances, DW consultancy, and applications

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Contrasts between the DM and EDW Development Approaches

Effort	DM Approach	EDW Approach
Scope	One subject area	Several subject areas
Development time	Months	Years
Development cost	$10,000 to $100,000+	$1,000,000+
Development difficulty	Low to medium	High
Data prerequisite for sharing	Common (within business area)	Common (across enterprise)
Sources	Only some operational and external systems	Many operational and external systems
Size	Megabytes to several gigabytes	Gigabytes to petabytes
Time horizon	Near-current and historical data	Historical data
Data transformations	Low to medium	High
Update frequency	Hourly, daily, weekly	Weekly, monthly
Technology		
Hardware	Workstations and departmental servers	Enterprise servers and mainframe computers
Operating system	Windows and Linux	Unix, Z/OS, OS/390
Databases	Workgroup or standard database servers	Enterprise database servers
Usage		
Number of simultaneous users	10s	100s to 1,000s
User types	Business area analysts and managers	Enterprise analysts and senior executives
Business spotlight	Optimizing activities within the business area	Cross-functional optimization and decision making

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

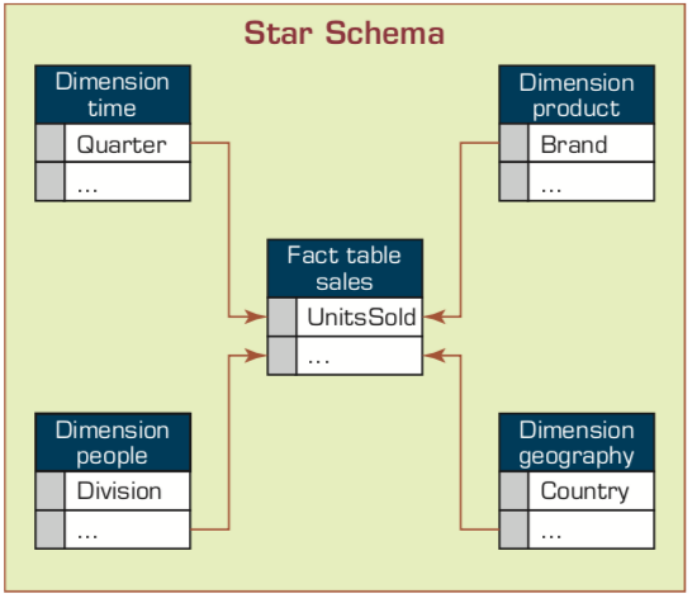
Essential Differences between Inmon's and Kimball's Approaches

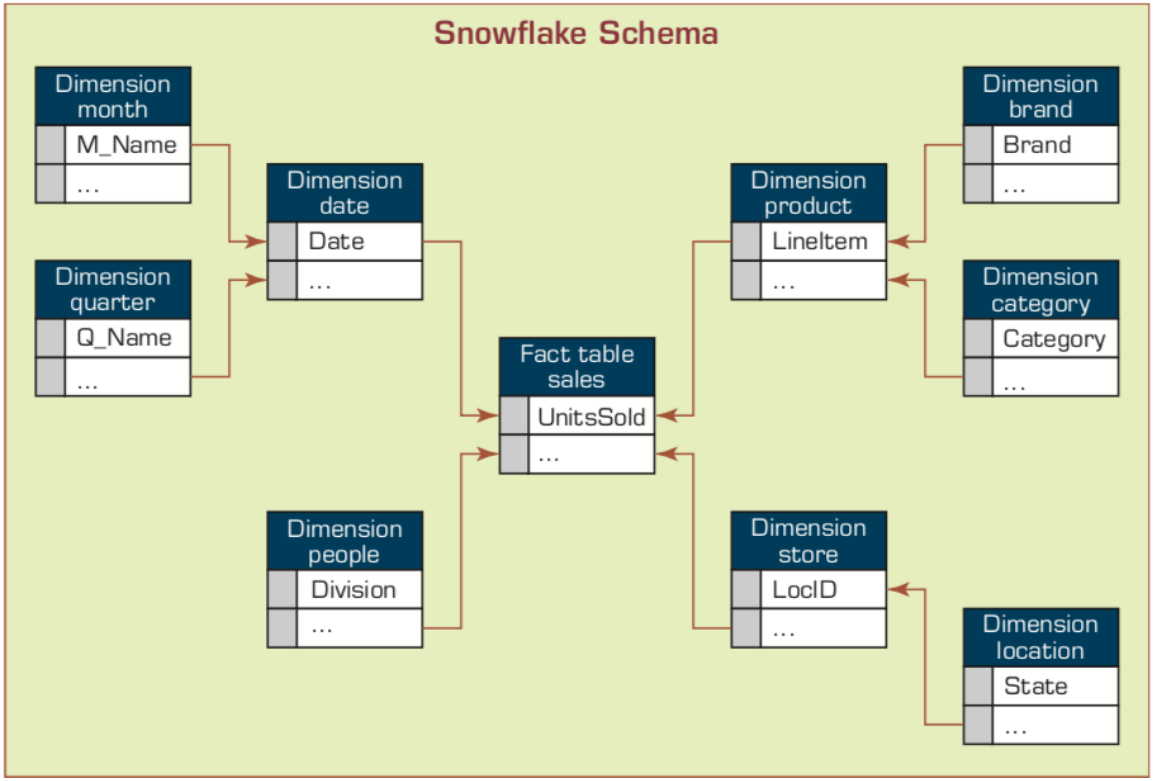
Characteristic	Inmon	Kimball
Methodology and Architecture		
Overall approach	Top-down	Bottom-up
Architecture structure	Enterprise-wide (atomic) data warehouse "feeds" departmental databases	DMs model a single business process, and enterprise consistency is achieved through a data bus and conformed dimensions
Complexity of the method	Quite complex	Fairly simple
Comparison with established development methodologies	Derived from the spiral methodology	Four-step process; a departure from RDBMS methods
Discussion of physical design	Fairly thorough	Fairly light
Data Modeling		
Data orientation	Subject or data driven	Process oriented
Tools	Traditional (entity-relationship diagrams [ERD], data flow diagrams [DFD])	Dimensional modeling; a departure from relational modeling
End-user accessibility	Low	High
Philosophy		
Primary audience	IT professionals	End users
Place in the organization	Integral part of the corporate information factory	Transformer and retainer of operational data
Objective	Deliver a sound technical solution based on proven database methods and technologies	Deliver a solution that makes it easy for end users to directly query the data and still get reasonable response times

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Representation of Data in Data Warehouse

(1) Star Schema (2) Snowflake Schema





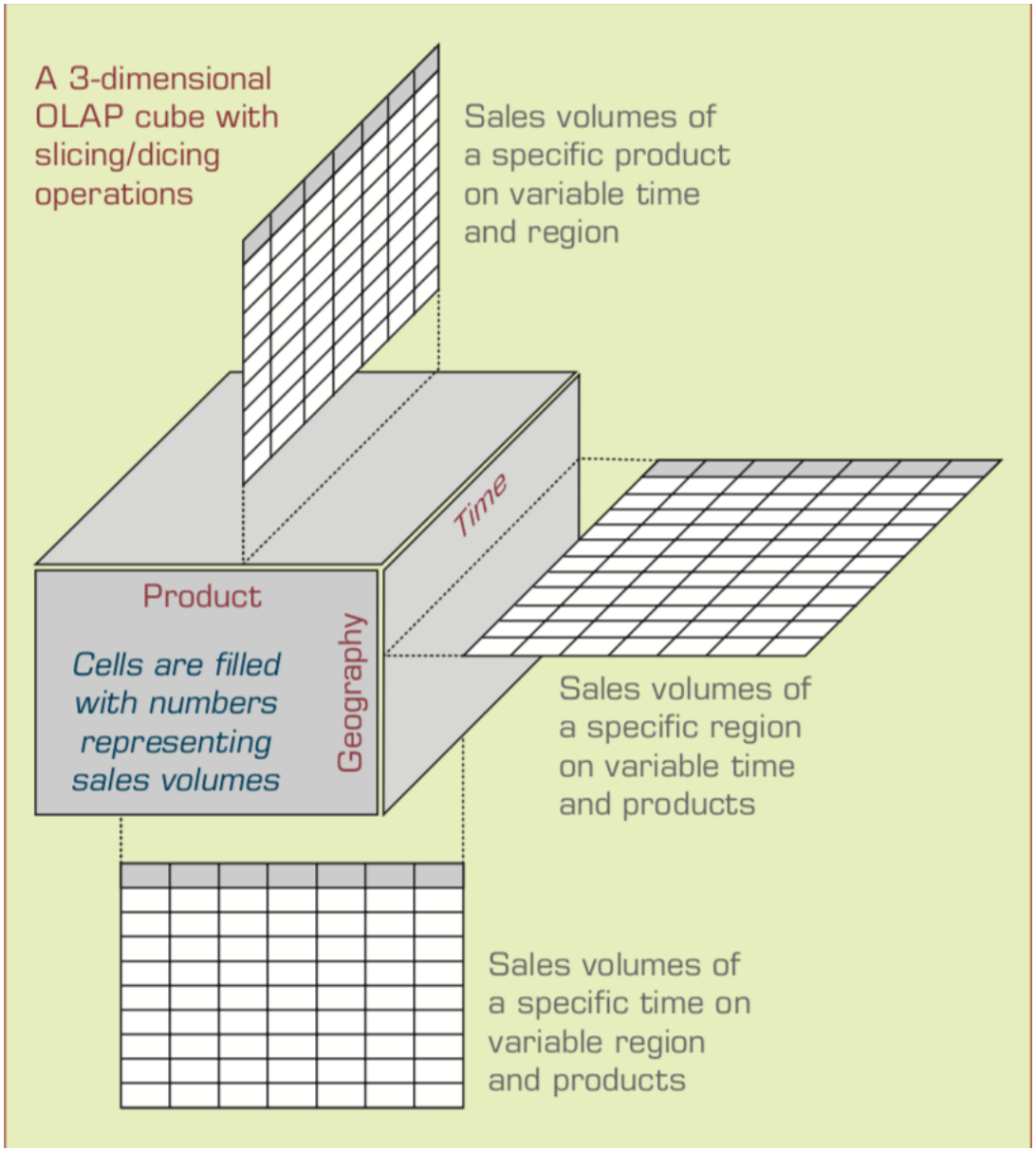
Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

A Comparison between OLTP and OLAP

Criteria	OLTP	OLAP
Purpose	To carry out day-to-day business functions	To support decision making and provide answers to business and management queries
Data source	Transaction database (a normalized data repository primarily focused on efficiency and consistency)	Data warehouse or DM (a nonnormalized data repository primarily focused on accuracy and completeness)
Reporting	Routine, periodic, narrowly focused reports	Ad hoc, multidimensional, broadly focused reports and queries
Resource requirements	Ordinary relational databases	Multiprocessor, large-capacity, specialized databases
Execution speed	Fast (recording of business trans-actions and routine reports)	Slow (resource intensive, complex, large-scale queries)

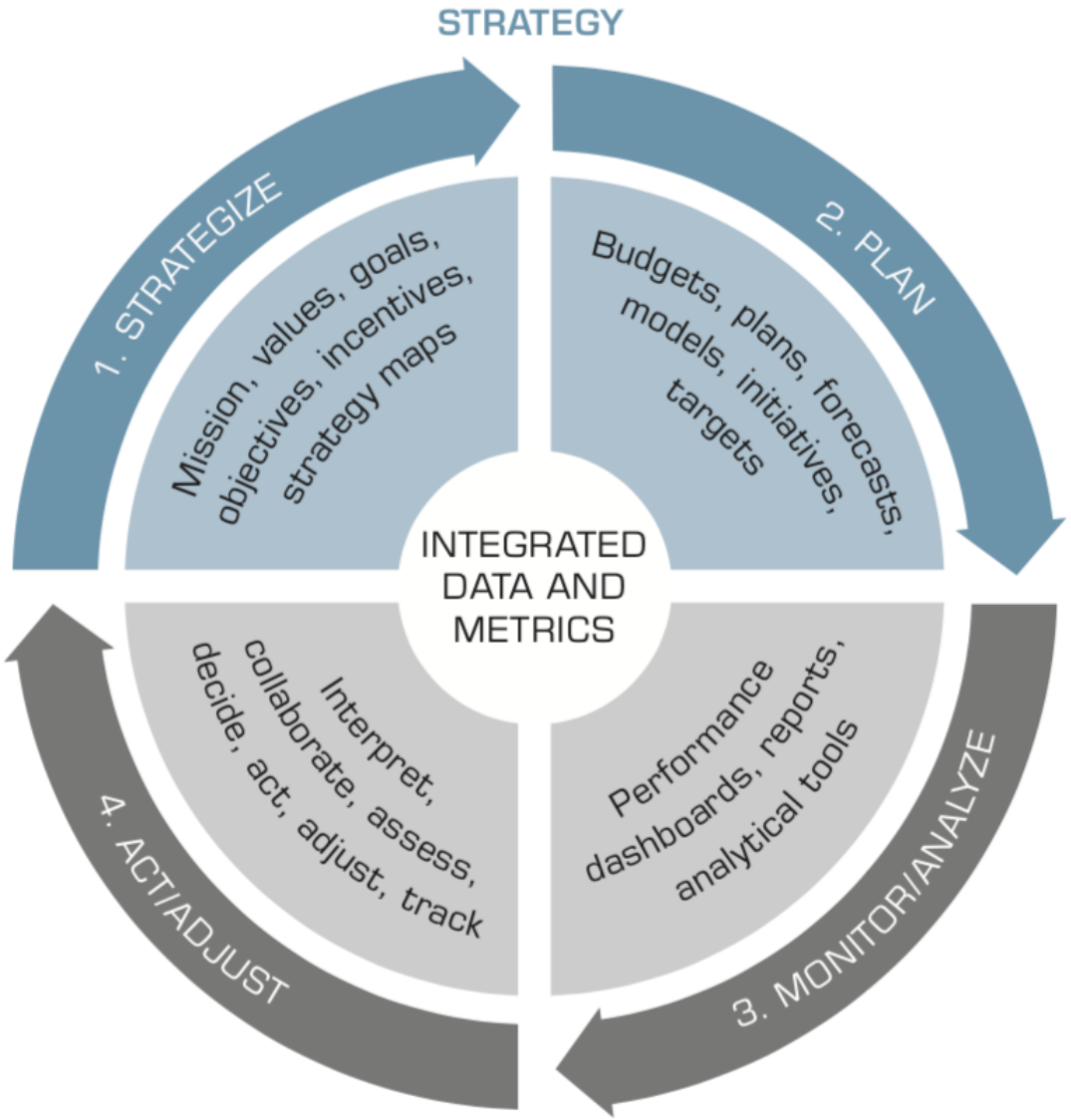
Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Slicing Operations on a Simple Three-Dimensional Data Cube



Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Business Performance Management (BPM) Closed-Loop BPM Cycle

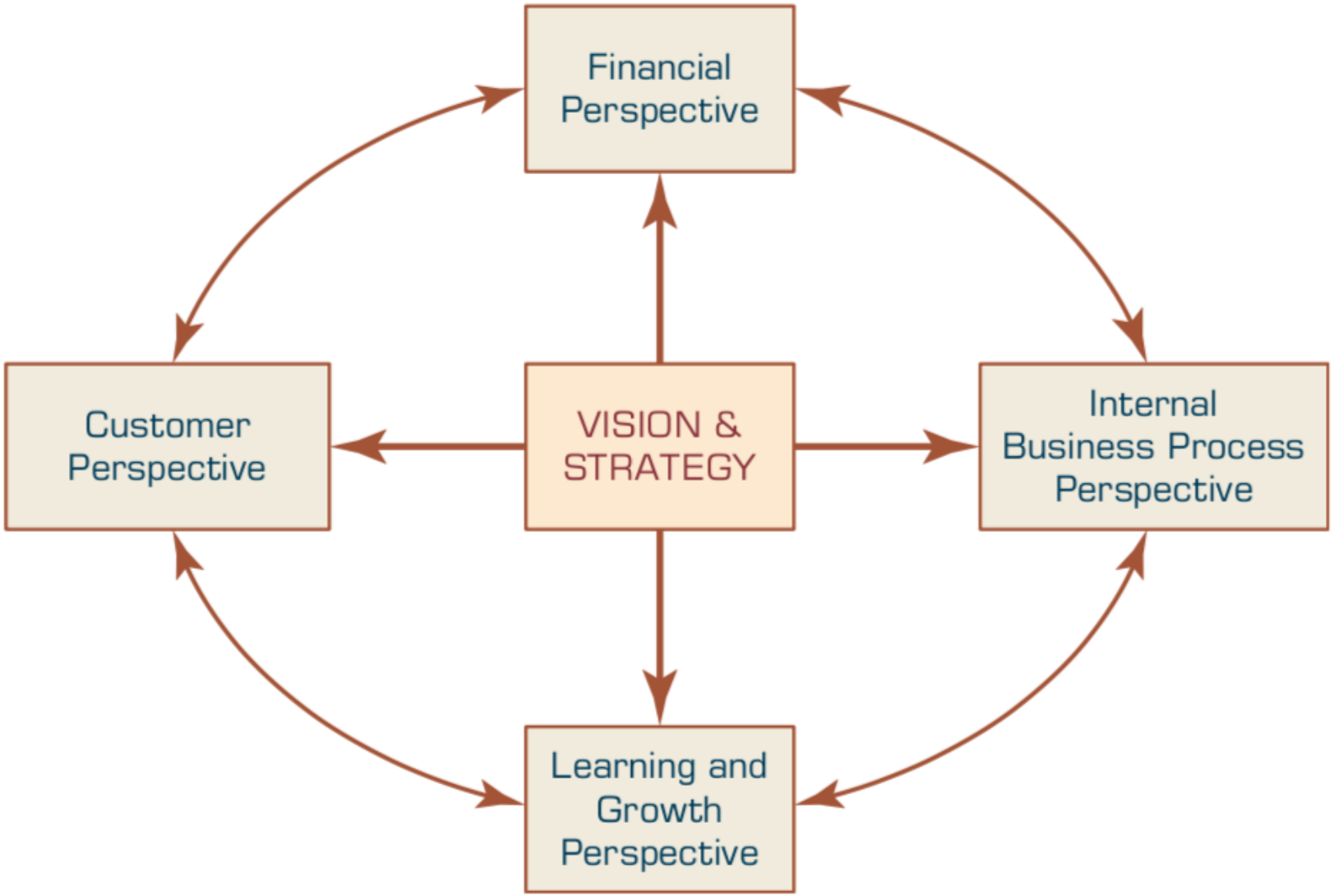


Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Business Performance Management (BPM) Closed-Loop BPM Cycle

1. ***Strategize***
   - ***Where do we want to go?***
2. ***Plan***
   - ***How do we get there?***
3. ***Monitor/Analyze***
   - ***How are we doing?***
4. ***Act and Adjust***
   - ***What do we need to do differently?***

Four Perspectives in Balanced Scorecard Methodology



Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Comparison of the Balanced Scorecard and Six Sigma

Balanced Scorecard	Six Sigma
Strategic management system	Performance measurement system
Relates to the longer-term view of the business	Provides snapshot of business's performance and identifies measures that drive performance toward profitability
Designed to develop a balanced set of measures	Designed to identify a set of measurements that impact profitability
Identifies measurements around vision and values	Establishes accountability for leadership for wellness and profitability
Critical management processes are to clarify vision/ strategy, communicate, plan, set targets, align strategic initiatives, and enhance feedback	Includes all business processes—management and operational
Balances customer and internal operations without a clearly defined leadership role	Balances management and employees' roles; balances costs and revenue of heavy processes
Emphasizes targets for each measurement	Emphasizes aggressive rate of improvement for each measurement, irrespective of target
Emphasizes learning of executives based on feedback	Emphasizes learning and innovation at all levels based on process feedback; enlists all employees' participation
Focuses on growth	Focuses on maximizing profitability
Heavy on strategic content	Heavy on execution for profitability
Management system consisting of measures	Measurement system based on process management

Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Six Sigma
The DMAIC Performance Model

- **D**efine
- **M**easure
- **A**nalyze
- **I**mprove
- **C**ontrol

The Joy of Stats:
200 Countries, 200 Years, 4 Minutes

https://www.youtube.com/watch?v=jbkSRLYSojo

Hans Rosling's 200 Countries, 200 Years, 4 Minutes - The Joy of Stats - BBC Four

Python Data Science Handbook in Google Colab

https://colab.research.google.com/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/Index.ipynb

Index.ipynb

File Edit View Insert Runtime Tools Help

SHARE

CODE TEXT CELL CELL COPY TO DRIVE CONNECT EDITING

Python Data Science Handbook

Jake VanderPlas

powered by jupyter

Jake VanderPlas

https://colab.research.google.com/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/Index.ipynb

Python Data Science Handbook in Google Colab

Table of Contents

https://colab.research.google.com/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/Index.ipynb

Python Data Science Handbook in Google Colab

2. Introduction to NumPy

- Understanding Data Types in Python
- The Basics of NumPy Arrays
- Computation on NumPy Arrays: Universal Functions
- Aggregations: Min, Max, and Everything In Between
- Computation on Arrays: Broadcasting
- Comparisons, Masks, and Boolean Logic
- Fancy Indexing
- Sorting Arrays
- Structured Data: NumPy's Structured Arrays

https://colab.research.google.com/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/Index.ipynb

Python Data Science Handbook in Google Colab

3. Data Manipulation with Pandas

- Introducing Pandas Objects
- Data Indexing and Selection
- Operating on Data in Pandas
- Handling Missing Data
- Hierarchical Indexing
- Combining Datasets: Concat and Append
- Combining Datasets: Merge and Join
- Aggregation and Grouping
- Pivot Tables
- Vectorized String Operations
- Working with Time Series
- High-Performance Pandas: eval() and query()
- Further Resources

https://colab.research.google.com/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/Index.ipynb

Python Data Science Handbook in Google Colab

4. Visualization with Matplotlib

- Simple Line Plots
- Simple Scatter Plots
- Visualizing Errors
- Density and Contour Plots
- Histograms, Binnings, and Density
- Customizing Plot Legends
- Customizing Colorbars
- Multiple Subplots
- Text and Annotation
- Customizing Ticks
- Customizing Matplotlib: Configurations and Stylesheets
- Three-Dimensional Plotting in Matplotlib
- Geographic Data with Basemap
- Visualization with Seaborn
- Further Resources

https://colab.research.google.com/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/Index.ipynb

Python Data Science Handbook in Google Colab

5. Machine Learning

- What Is Machine Learning?
- Introducing Scikit-Learn
- Hyperparameters and Model Validation
- Feature Engineering
- In Depth: Naive Bayes Classification
- In Depth: Linear Regression
- In-Depth: Support Vector Machines
- In-Depth: Decision Trees and Random Forests
- In Depth: Principal Component Analysis
- In-Depth: Manifold Learning
- In Depth: k-Means Clustering
- In Depth: Gaussian Mixture Models
- In-Depth: Kernel Density Estimation
- Application: A Face Detection Pipeline
- Further Machine Learning Resources

https://colab.research.google.com/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/Index.ipynb

Summary

- Descriptive Analytics II
- Business Intelligence
- Data Warehousing
- Data Integration and the Extraction, Transformation, and Load (ETL) Processes
- Business Performance Management (BPM)
- Performance Measurement
  - Balanced Scorecards
  - Six Sigma

References

- Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson.
- Jake VanderPlas (2016), Python Data Science Handbook: Essential Tools for Working with Data, O'Reilly Media.