Chapter 13

Business Intelligence and Data Warehouses

Learning Objectives (1 of 2)

- In this chapter, you will learn:
  - How business intelligence provides a comprehensive business decision support framework
  - About business intelligence architecture, its evolution, and reporting styles
  - About the relationship and differences between operational data and decision support data
  - What a data warehouse is and how to prepare data for one

Learning Objectives (2 of 2)

- In this chapter, you will learn:
  - What star schemas are and how they are constructed
  - About data analytics
  - About online analytical processing (OLAP)
  - How SQL extensions are used to support OLAP-type data manipulations

Business Intelligence (BI) (1 of 2)

- Comprehensive, cohesive, integrated set of tools and processes
  - Captures, collects, integrates, stores, and analyzes data
- Generates and presents information to support business decision making
- Allows a business to transform:
  - Data into information
  - Information into knowledge
  - Knowledge into wisdom

Business Intelligence (BI) (2 of 2)

- Concepts, practices, tools and techniques to help business
  - Understand its core capabilities
  - Provide snapshots of the company situation
  - Identify key opportunities to create a competitive advantage
- Provides a framework for
  - Collecting and storing operational data and aggregating it into decision support data
  - Analyzing decision support data and presenting generated information to end users to support business decisions
  - Making business decision which generates more data
  - Monitoring results to evaluate outcomes and predicting future outcomes with a high degree of accuracy

Figure 13.1 - Business Intelligence Framework

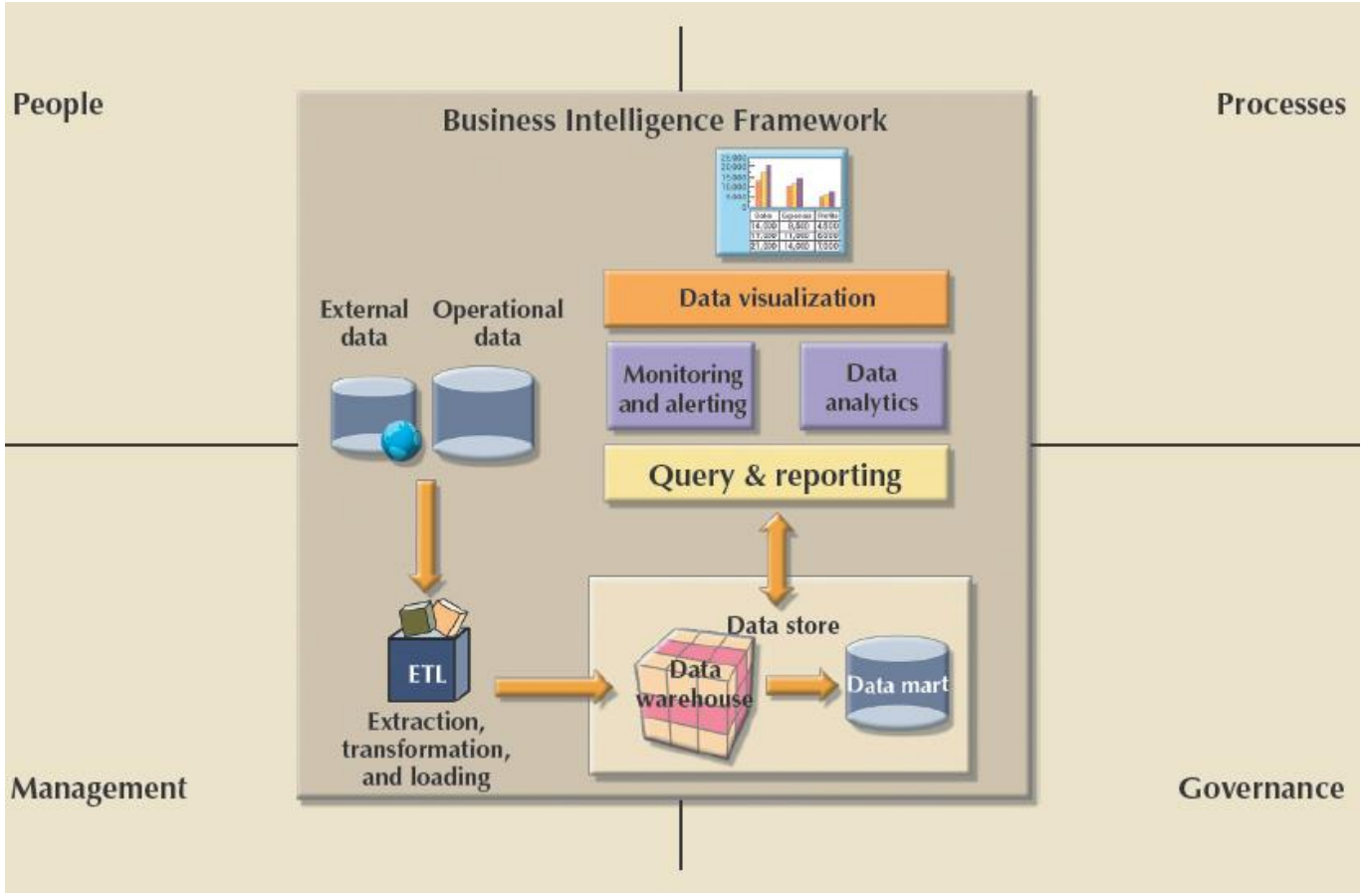


Table 13.3 – Sample of Business Intelligence Tools (1 of 2)

TOOL	DESCRIPTION	SAMPLE VENDORS
Dashboards and business activity monitoring	**Dashboards** use web-based technologies to present key business performance indicators or information in a single integrated view, generally using graphics that are clear, concise, and easy to understand.	Salesforce IBM/Cognos BusinessObjects Information Builders iDashboards
Portals	**Portals** provide a unified, single point of entry for information distribution. Portals are a web-based technology that use a web browser to integrate data from multiple sources into a single webpage. Many different types of BI functionality can be accessed through a portal.	Oracle Portal Actuate Microsoft SAP
Data analysis and reporting tools	These advanced tools are used to query multiple and diverse data sources to create integrated reports.	Microsoft Reporting Services MicroStrategy SAS WebReportStudio
Data-mining tools	These tools provide advanced statistical analysis to uncover problems and opportunities hidden within business data. Chapter 14 covers data mining in more detail.	SAP Teradata MicroStrategy MS Analytics Services

Table 13.3 – Sample of Business Intelligence Tools (2 of 2)

TOOL	DESCRIPTION	SAMPLE VENDORS
Data warehouses (DW)	The data warehouse is the foundation of a BI infrastructure. Data is captured from the production system and placed in the DW on a near real-time basis. BI provides company-wide integration of data and the capability to respond to business issues in a timely manner.	Microsoft Oracle IBM/Cognos Teradata
OLAP tools	Online analytical processing provides multidimensional data analysis.	IBM/Cognos BusinessObjects OracleMicrosoft
Data visualization	These tools provide advanced visual analysis and techniques to enhance understanding and create additional insight of business data and its true meaning.	Dundas Tableau QlikView Actuate

Practices to Manage Data (1 of 2)

- **Master data management (MDM)**: Collection of concepts, techniques, and processes for identification, definition, and management of data elements
- **Governance**: Method of government for controlling business health and for consistent decision making
- **Key performance indicators (KPI)**: Numeric or scale-based measurements that assess company's effectiveness in reaching its goals

Practices to Manage Data (2 of 2)

- **Data visualization**: Abstracting data to provide information in a visual format
  - Enhances the user's ability to efficiently comprehend the meaning of the data
  - Techniques:
    - Pie charts and bar charts
    - Line graphs
    - Scatter plots
    - Gantt charts
    - Heat maps

Reporting Styles of a Modern BI System

- Advanced reporting
- Monitoring and alerting
- Advanced data analytics

Business Intelligence Benefits

- Improved decision making
- Integrating architecture
- Common user interface for data reporting and analysis
- Common data repository fosters single version of company data
- Improved organizational performance

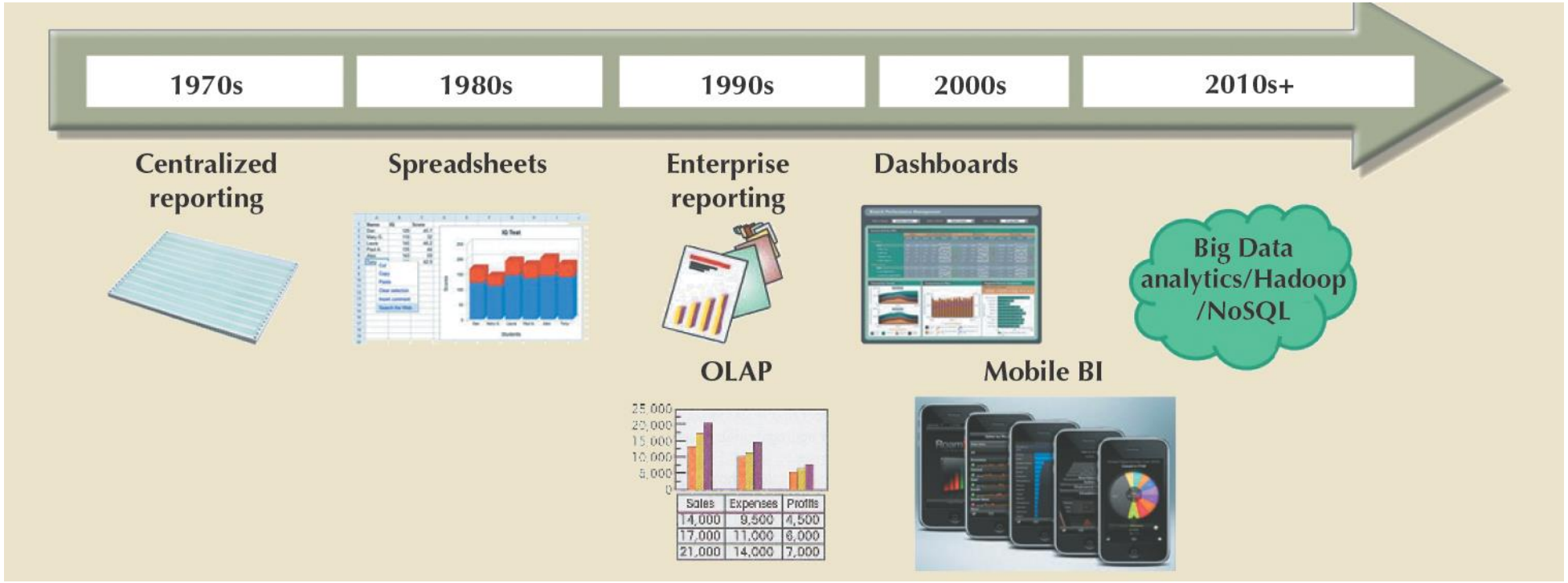
Table 13.4 - Business Intelligence Evolution (1 of 2)

SYSTEM TYPE	DATA SOURCE	DATA EXTRACTION/ INTEGRATION PROCESS	DATA STORE	END-USER QUERY TOOL	END USER PRESENTATION TOOL
Traditional mainframe based online transaction processing (OLTP)	Operational data	None Reports read and Summarized data Directly from operational data	None Temporary files used for reporting purposes	Very basic Predefined reporting formats Basic sorting, totaling, and averaging	Very basic Menu-driven, predefined reports, text and numbers only
Managerial information system (MIS)	Operational data	Basic extraction and aggregation Read, filter, and summarize operational data into intermediate data store	Lightly aggregated data in RDBMS	Same as above, in addition to some ad hoc reporting using SQL	Same as above, in addition to some ad hoc columnar report definitions
First-generation departmental decision support system (DSS)	Operational data External data	Data extraction and integration process populates DSS data store Run periodically	First DSS database generation Usually RDBMS	Query tool with some analytical capabilities and predefined Reports	Spreadsheet style Advanced presentation tools with plotting and graphics capabilities

Table 13.4 - Business Intelligence Evolution (2 of 2)

SYSTEM TYPE	DATA SOURCE	DATA EXTRACTION/ INTEGRATION PROCESS	DATA STORE	END-USER QUERY TOOL	END USER PRESENTATION TOOL
First-generation BI	Operational data External data	Advanced data extraction and integration Access diverse data sources, filters, aggregations, classifications, scheduling, and conflict Resolution	Data warehouse RDBMS technology Optimized for query purposes Star schema Model	Same as above	Same as above, in addition to multidimensional presentation tools with drill-down Capabilities
Second- generation BI Online analytical processing (OLAP)	Same as above	Same as above	Data warehouse stores data in MDBMS Cubes with multiple Dimensions	Adds support for end-user based data analytics	Same as above, but uses cubes and multidimensional matrixes; limited by cube size Dashboards Scorecards Portals
Third-generation Mobile, cloud based, and Big Data	Same as above Includes social media and Machine generated Data	Same as above Cloud-based	Same as above Cloud-based Hadoop and No SQL Databases	Advanced analytics Limited ad hoc Interactions	Mobile devices: smartphones and tablets

Figure 13.3 - Evolution of BI Information Dissemination Formats



Business Intelligence Technology Trends

- Data storage improvements
- Business intelligence appliances
- Business intelligence as a service
- Big Data analytics
- Personal analytics

Decision Support Data (1 of 2)

- Effectiveness of BI depends on quality of data gathered at operational level
- Operational data
  - Seldom well-suited for decision support tasks
  - Stored in relational database with highly normalized structures
  - Optimized to support transactions representing daily operations

Decision Support Data (2 of 2)

- Differ from operational data in:
  - Time span
  - Granularity
    - **Drill down**: Decomposing a data to a lower level
    - **Roll up**: Aggregating a data into a higher level
  - Dimensionality

Table 13.5 - Contrasting Operational and Decision Support Data Characteristics

CHARACTERISTIC	OPERATIONAL DATA	DECISION SUPPORT DATA
Data currency	Current operations Real-time data	Historic data Snapshot of company data Time component (week/month/year)
Granularity	Atomic-detailed data	Summarized data
Summarization level	Low; some aggregate yields	High; many aggregation levels
Data model	Highly normalized Mostly relational DBMSs	Non-normalized Complex structures Some relational, but mostly multidimensional DBMSs
Transaction type	Mostly updates	Mostly query
Transaction volumes	High-update volumes	Periodic loads and summary calculations
Transaction speed	Updates are critical	Retrievals are critical
Query activity	Low to medium	High
Query scope	Narrow range	Broad range
Query complexity	Simple to medium	Very complex
Data volumes	Hundreds of gigabytes	Terabytes to petabytes

Decision Support Database Requirements (1 of 2)

- Database schema
  - Must support complex, non-normalized data representations
  - Data must be aggregated and summarized
  - Queries must be able to extract multidimensional time slices

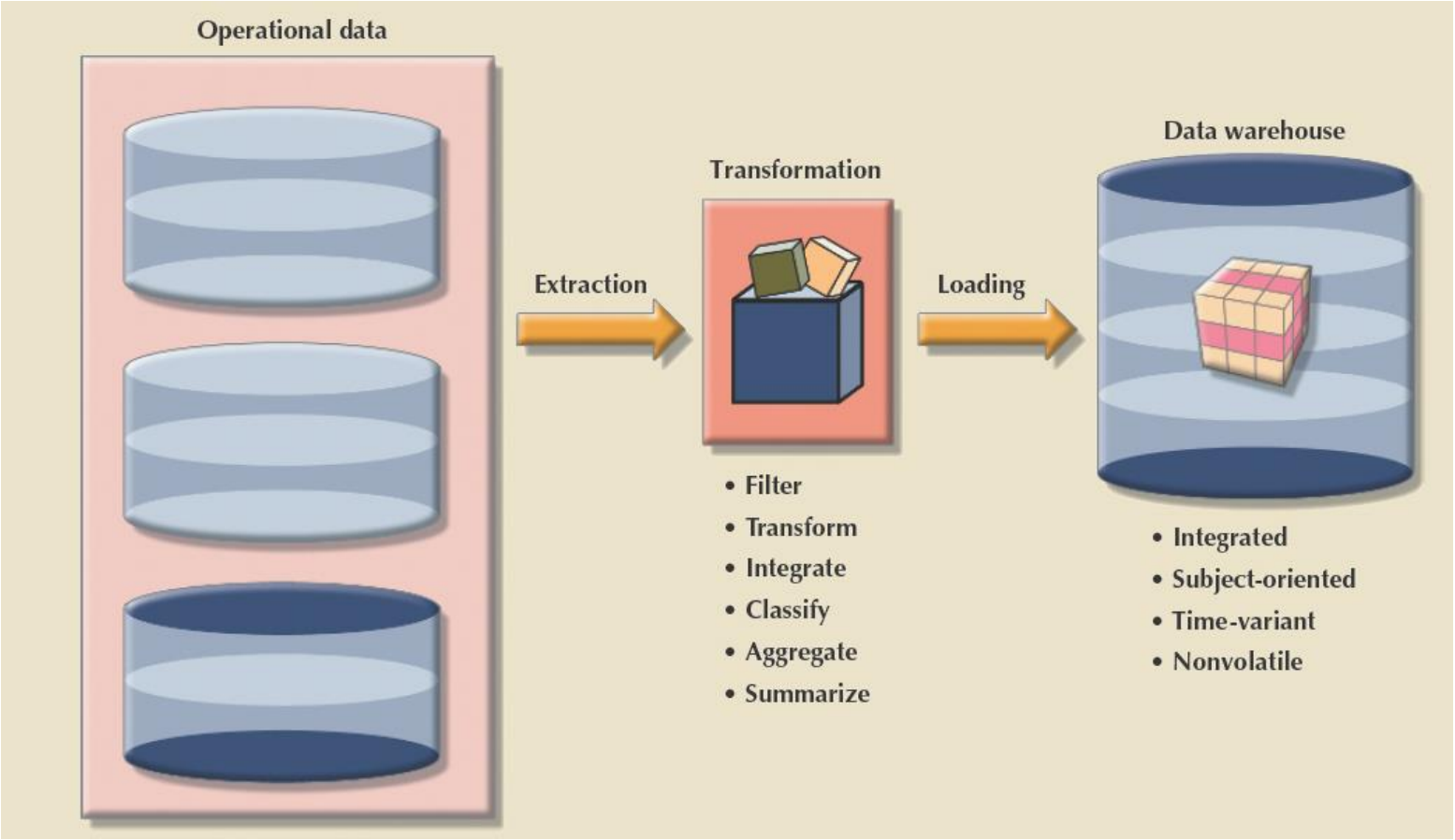
Decision Support Database Requirements (2 of 2)

- Data extraction and loading
  - Allow batch and scheduled data extraction
  - Support different data sources and check for inconsistent data or data validation rules
  - Support advanced integration, aggregation, and classification
- Database size should support
  - **Very large databases (VLDBs)**
  - Advanced storage technologies
  - Multiple-processor technologies

Table 13.8 - Characteristics of Data Warehouse Data and Operational Database Data

CHARACTERISTIC	OPERATIONAL DATABASE DATA	DATA WAREHOUSE DATA
Integrated	Similar data can have different representations or meanings. For example, Social Security numbers may be stored as ###-##-#### or as #########, and a given condition may be labeled as T/F or 0/1 or Y/N. A sales value may be shown in thousands or in millions.	Provide a unified view of all data elements with a common definition and representation for all business units.
Subject-oriented	Data is stored with a functional, or process, orientation. For example, data may be stored for invoices, payments, and credit amounts.	Data is stored with a subject orientation that Facilitates multiple views of the data and decision making. For example, sales may be recorded by product, division, manager, or region.
Time-variant	Data is recorded as current transactions. For example, the sales data may be the sale of a product on a given date, such as $342.78 on 12-MAY-2016.	Data is recorded with a historical perspective in mind. Therefore, a time dimension is added to facilitate data analysis and various time comparisons.
Nonvolatile	Data updates are frequent and common. For example, an inventory amount changes with each sale. Therefore, the data environment is fluid.	Data cannot be changed. Data is added only periodically from historical systems. Once the data is properly stored, no changes are allowed. Therefore, the data environment is relatively static.

Figure 13.5 - The ETL Process



Data Marts

- Small, single-subject data warehouse subset
- Provide decision support to a small group of people
- Benefits over data warehouses
  - Lower cost and shorter implementation time
  - Technologically advanced
  - Inevitable people issues

Table 13.9 - Twelve Rules for a Data Warehouse

RULE NO.	DESCRIPTION
1	The data warehouse and operational environments are separated.
2	The data warehouse data is integrated.
3	The data warehouse contains historical data over a long time.
4	The data warehouse data is snapshot data captured at a given point in time.
5	The data warehouse data is subject oriented.
6	The data warehouse data is mainly read-only with periodic batch updates from operational data. No online updates are allowed.
7	The data warehouse development life cycle differs from classical systems development. Data warehouse development is data-driven; the classical approach is process-driven.
8	The data warehouse contains data with several levels of detail: current detail data, old detail data, lightly summarized data, and highly summarized data.
9	The data warehouse environment is characterized by read-only transactions to very large data sets. The operational environment is characterized by numerous update transactions to a few data entities at a time.
10	The data warehouse environment has a system that traces data sources, transformations, and storage.
11	The data warehouse's metadata is a critical component of this environment. The metadata identifies and defines all data elements. The metadata provides the source, transformation, integration, storage, usage, relationships, and history of each data element.
12	The data warehouse contains a chargeback mechanism for resource usage that enforces optimal use of the data by end users.

Star Schema

- Data-modeling technique
- Maps multidimensional decision support data into a relational database
- Creates the near equivalent of multidimensional database schema from existing relational database
- Yields an easily implemented model for multidimensional data analysis

Components of Star Schemas

Facts

- Numeric values that represent a specific business aspect

Dimensions

- Qualifying characteristics that provide additional perspectives to a given fact

Attributes

- Used to search, filter, and classify facts
- **Slice and dice**: Ability to focus on slices of the data cube for more detailed analysis

Attribute hierarchies

- Provides a top-down data organization

Star Schema Representation

- Facts and dimensions represented by physical tables in data warehouse database
- Many-to-one (M:1) relationship between fact table and each dimension table
- Fact and dimension tables
  - Related by foreign keys
  - Subject to primary and foreign key constraints
- Primary key of a fact table
  - Is a composite primary key because the fact table is related to many dimension tables
  - Always formed by combining the foreign keys pointing to the related dimension tables

Performance-Improving Techniques for the Star Schema (1 of 2)

- Normalizing dimensional tables
  - **Snowflake schema**: Dimension tables can have their own dimension tables
- Maintaining multiple fact tables to represent different aggregation levels
- Denormalizing fact tables

Performance-Improving Techniques for the Star Schema (2 of 2)

- Partitioning and replicating tables
  - **Partitioning**: Splits tables into subsets of rows or columns and places them close to customer location
  - **Replication**: Makes copy of table and places it in a different location
  - **Periodicity**: Provides information about the time span of the data stored in the table

Online Analytical Processing (OLAP)

- Advanced data analysis environment that supports decision making, business modeling, and operations research
- Characteristics:
  - Multidimensional data analysis techniques
  - Advanced database support
  - Easy-to-use end-user interfaces

Multidimensional Data Analysis Techniques

- Data are processed and viewed as part of a multidimensional structure
- Augmented by the following functions:
  - Advanced data presentation functions
  - Advanced data aggregation, consolidation, and classification functions
  - Advanced computational functions
  - Advanced data-modeling functions

Advanced Database Support

- OLAP tools must have the following features to deliver efficient decision support:
  - Access to many different kinds of DBMSs, flat files, and internal and external data sources
  - Access to aggregated data warehouse data and operational database detail data
  - Advanced data navigation features
  - Rapid and consistent query response times
  - Ability to map end-user requests
  - Support for very large databases

Easy-to-Use End-User Interface

- Proper implementation leads to simple navigation and accelerated decision making or data analysis
- Advanced OLAP features are more useful when access is kept simple
- Many interface features are borrowed from previous generations of data analysis tools

Figure 13.17 – OLAP Architecture

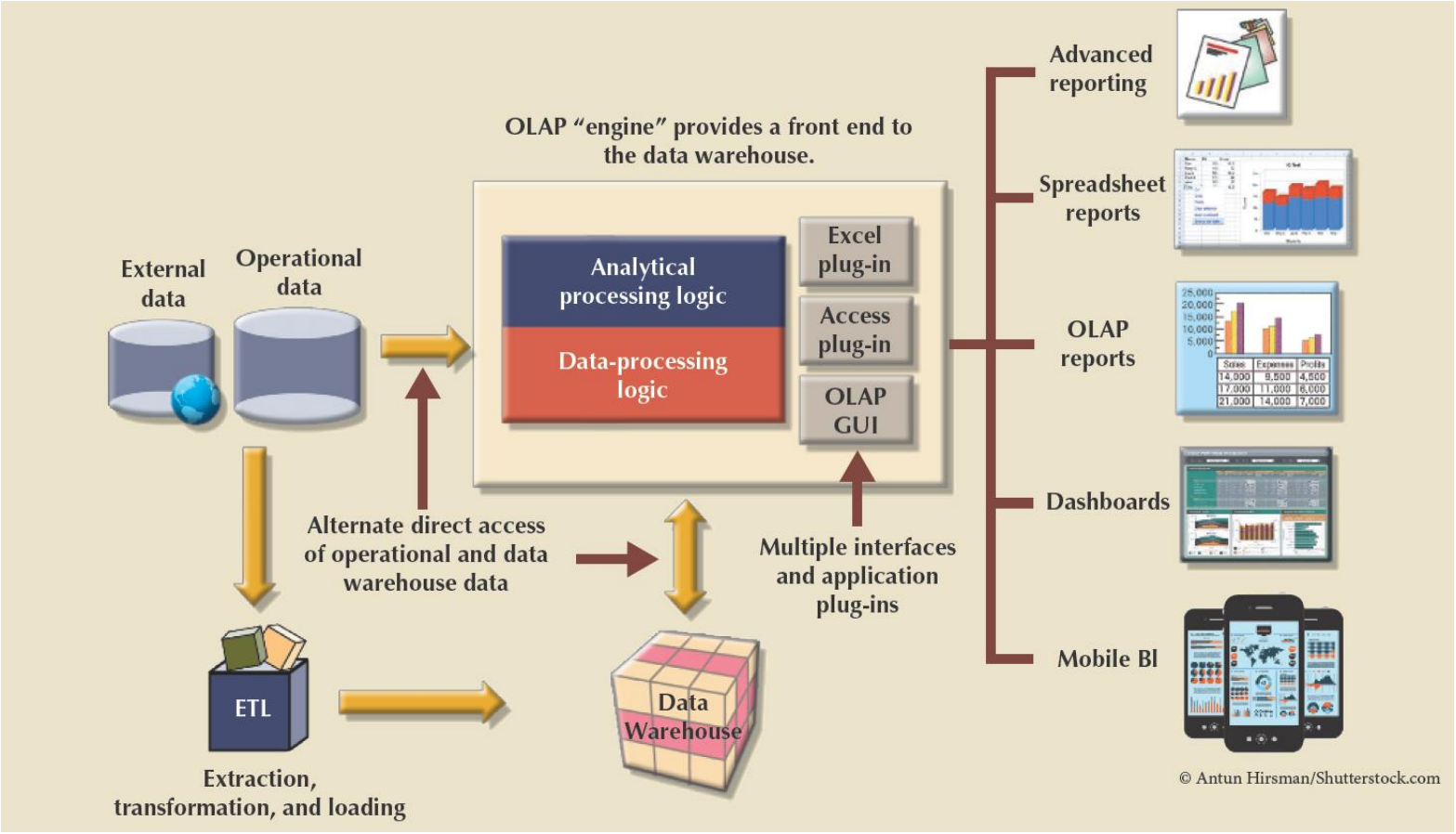
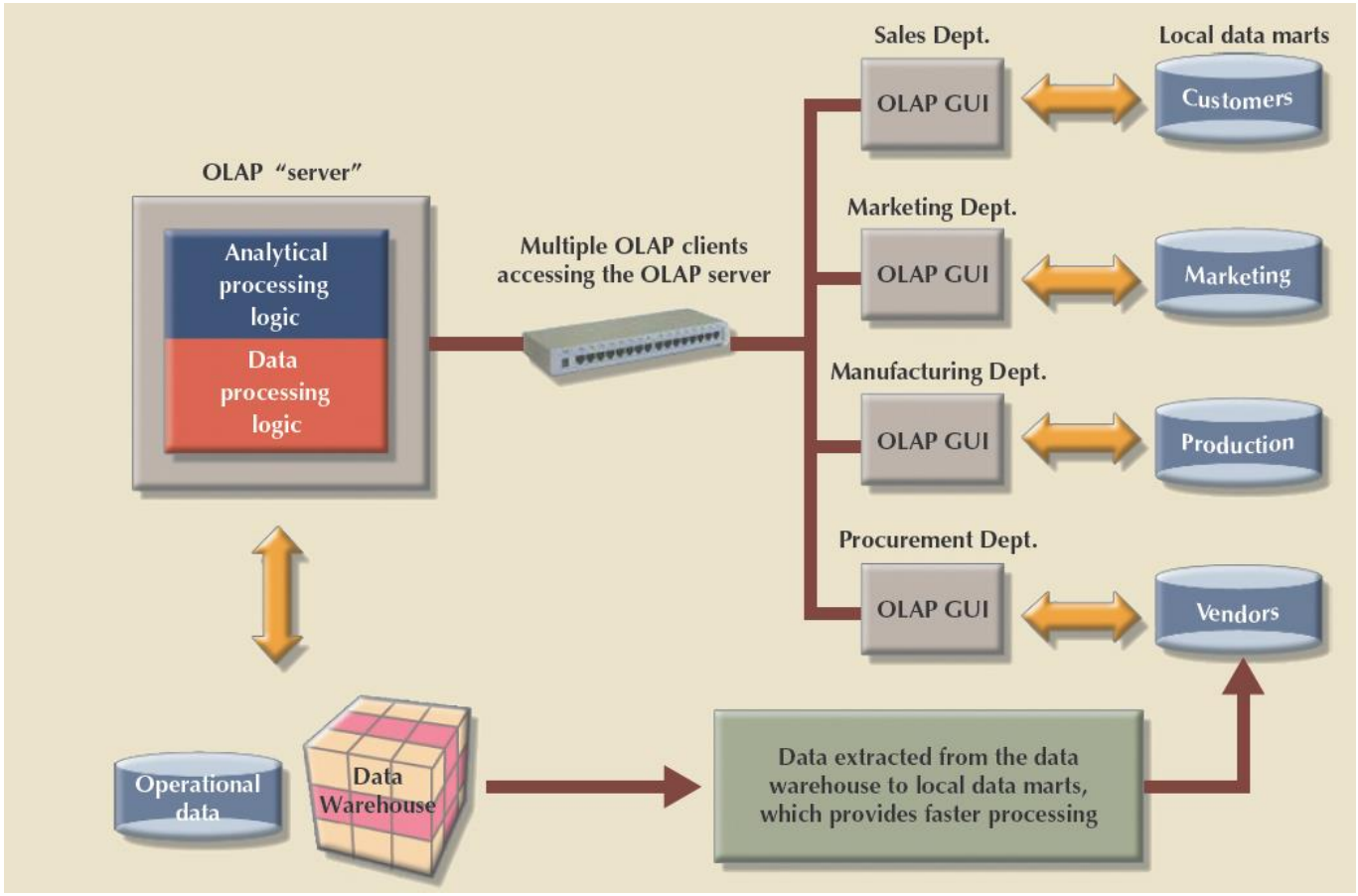


Figure 13.18 – OLAP Server with Local Mini Data Marts



Relational Online Analytical Processing (ROLAP)

- Provides OLAP functionality using relational databases and familiar relational tools to store and analyze multidimensional data
- Extensions added to traditional RDBMS technology
  - Multidimensional data schema support within the RDBMS
  - Data access language and query performance optimized for multidimensional data
  - Support for very large databases (VLDBs)

Multidimensional Online Analytical Processing (MOLAP)

- Extends OLAP functionality to multidimensional database management systems (MDBMSs)
  - **MDBMS**: Uses proprietary techniques store data in matrix-like n-dimensional arrays
  - End users visualize stored data as a 3D **data cube**
    - Grow to n dimensions, becoming hypercubes
    - Held in memory in a **cube cache** to speed access
- **Sparsity**: Measures the density of the data held in the data cube

Table 13.12 - Relational versus Multidimensional OLAP

CHARACTERISTIC	ROLAP	MOLAP
Schema	Uses star schema Additional dimensions can be added dynamically	Uses data cubes Multidimensional arrays, row stores, column stores Additional dimensions require re-creation of the data cube
Database size	Medium to large	Large
Architecture	Client/server Standards-based	Client/server Open or proprietary, depending on vendor
Access	Supports ad hoc requests Unlimited dimensions	Limited to predefined dimensions Proprietary access languages
Speed	Good with small data sets; average for medium-sized to large data sets	Faster for large data sets with predefined dimensions

SQL Extensions for OLAP

The ROLLUP extension

- Used with GROUP BY clause to generate aggregates by different dimensions
- Enables subtotal for each column listed except for the last one, which gets a grand total
- Order of column list important

The CUBE extension

- Used with GROUP BY clause to generate aggregates by the listed columns
- Includes the last column

Materialized View

- Dynamic table that contains SQL query command to generate rows and stores the actual rows
- Created the first time query is run
  - Summary rows are stored in the table
- Automatically updated when base tables are updated
- Requires specified privileges