SIT103/SIT772 Database Fundamentals

Week 9

Business Intelligence

&

Big Data

Dr Iynkaran Natgunanathan

Last Week

- Procedural Language SQL (PL/SQL)
  - o Anonymous PL/SQL blocks
  - o Triggers
  - o Stored Procedures
  - o Cursors
  - o PL/SQL Functions

- Embedded SQL: A brief introduction

Last week's OnTrack Task

- No Pass Task
  - Students aiming for a 'P' are encouraged to play with ORACLE and practice PL/SQL in ORACLE

- 8.1C Online Quiz 2 (the same as 5.2C Online Quiz 1)
  - Do the online Quiz 2 in the CloudDeakin site
  - Submit the screenshot of your Quiz Result (80% or more)
  - Two attempts, 1 hours (60 mins) to complete once started
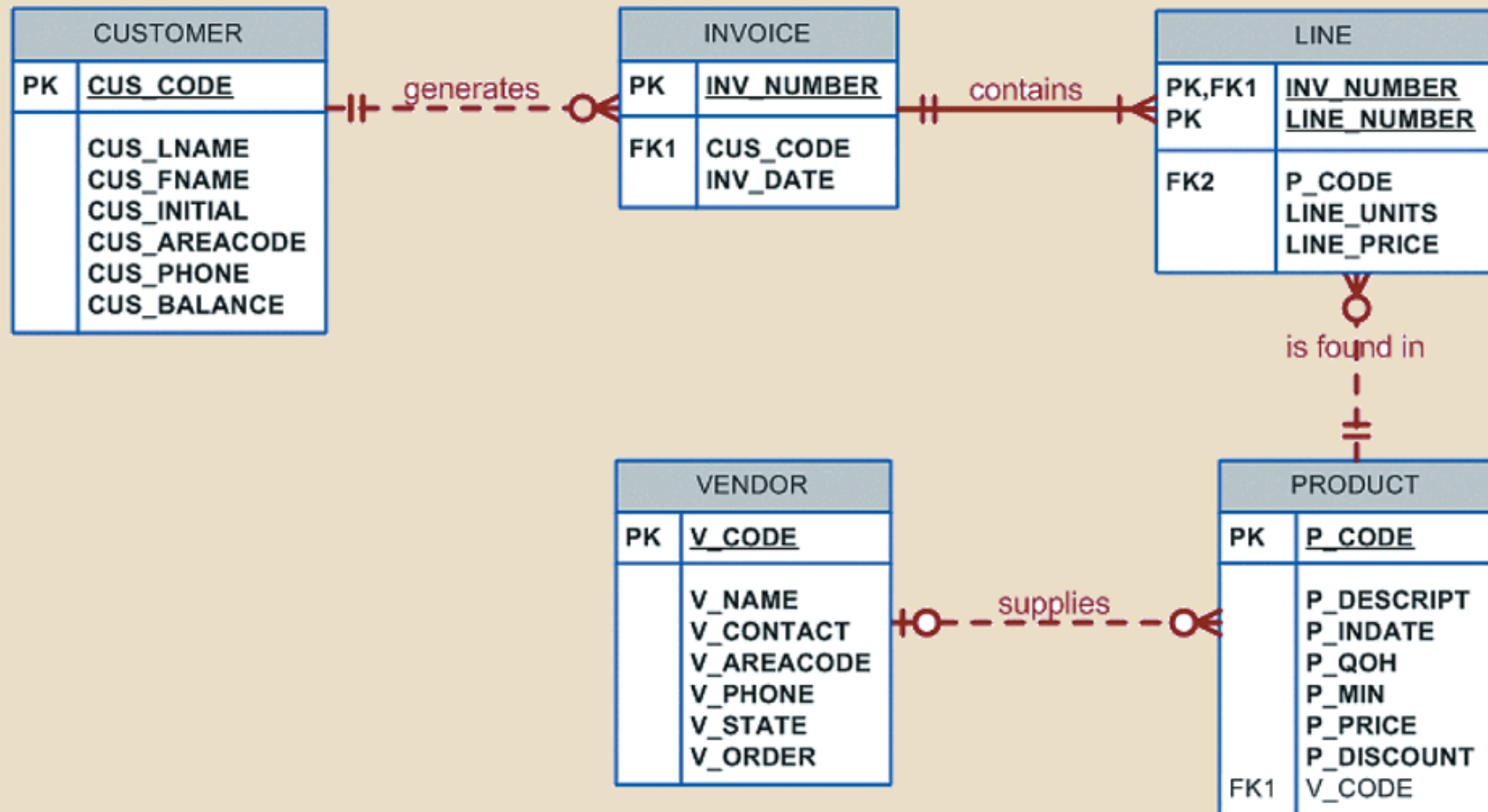
- 8.2D PL/SQL Exercise

Any Questions?

This week

- Operational and Decision Support Data

- Business Intelligence

- Data Warehouse and Data Marts

- Data Analytics and Data Mining

- Data Visualisation

- Big Data

- NoSQL Databases

Data storage & management discussed so far

FIGURE 8.1 DATABASE MODEL

Operational/Transactional DB

- So far, we discuss understanding, modelling and implementing data requirements for business operations of a company

  - **Operational database**

  - started from business operations/scenario

  - records each and every business transactions

    e.g., who bought what product on what day

  - stores a huge amount of data

    e.g., think about products bought by customers every day at all Coles stores

- How does Coles extract business insights from such massive volume of transactional data to support strategic decision making?

Business Intelligence (BI)

- Comprehensive, cohesive, integrated set of tools and processes used to

  - capture, collects, integrates, stores, and analyzes data

  - generate and present information to support **business decision making**

- Allows transformation

  - **Data into information**

  - Information into knowledge

Business Intelligence (2)

- BI is not a product by itself
- Concepts, practices, tools and techniques to help business
  - understand core capabilities
  - provide snapshots of the company situation
  - identify key opportunities to create a competitive advantage
- Provides a framework
  - 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 decisions which generate more data
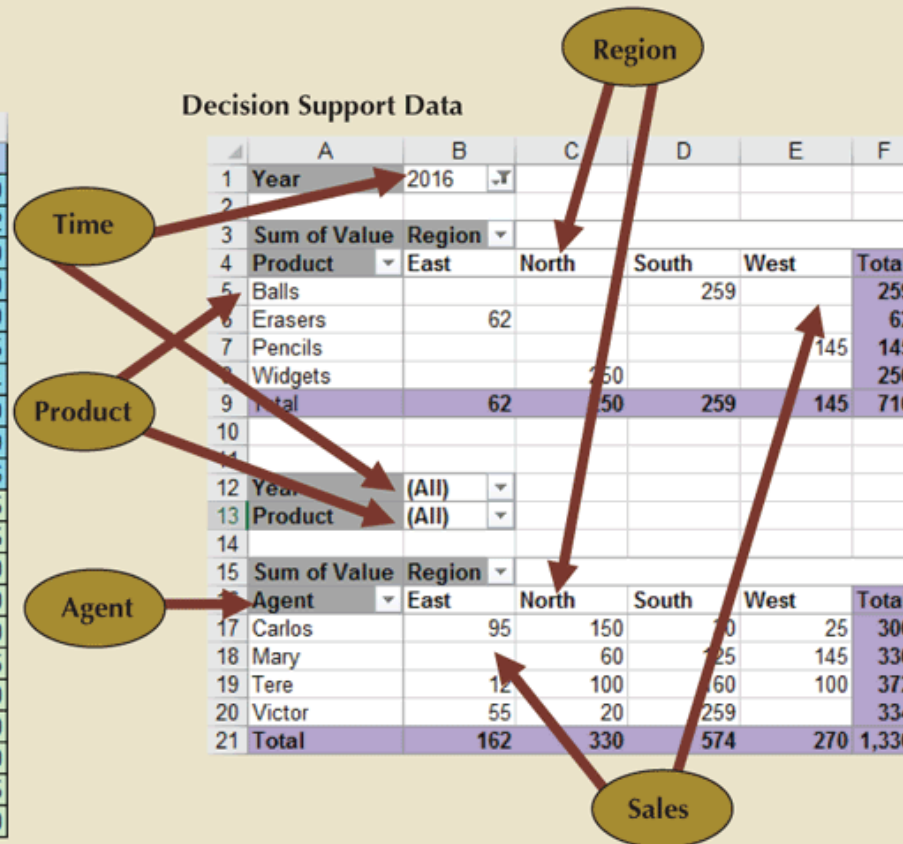 - monitoring results to evaluate outcomes and predicting future outcomes with a high degree of accuracy

Operational Data

	A	B	C	D	E
1	Year	Region	Agent	Product	Value
2	2016	East	Carlos	Erasers	50
3	2016	East	Tere	Erasers	12
4	2016	North	Carlos	Widgets	120
5	2016	North	Tere	Widgets	100
6	2016	North	Carlos	Widgets	30
7	2016	South	Victor	Balls	145
8	2016	South	Victor	Balls	34
9	2016	South	Victor	Balls	80
10	2016	West	Mary	Pencils	89
11	2016	West	Mary	Pencils	56
12	2017	East	Carlos	Pencils	45
13	2017	East	Victor	Balls	55
14	2017	North	Mary	Pencils	60
15	2017	North	Victor	Erasers	20
16	2017	South	Carlos	Widgets	30
17	2017	South	Mary	Widgets	75
18	2017	South	Mary	Widgets	50
19	2017	South	Tere	Balls	70
20	2017	South	Tere	Erasers	90
21	2017	West	Carlos	Widgets	25
22	2017	West	Tere	Balls	100

Decision Support Data

	A	B	C	D	E	F
1	Year	2016				
2						
3	Sum of Value	Region				
4	Product	East	North	South	West	Total
5	Balls				259	259
6	Erasers	62				62
7	Pencils				145	145
8	Widgets		250			250
9	Total	62	250	259	145	716
10						
12	Year	(All)				
13	Product	(All)				
14						
15	Sum of Value	Region				
16	Agent	East	North	South	West	Total
17	Carlos	95	150	30	25	300
18	Mary		60	125	145	330
19	Tere	12	100	160	100	372
20	Victor	55	20	259		334
21	Total	162	330	574	270	1,336

Operational data has a narrow time span, low granularity, and single focus. Such data is usually represented in tabular format, in which each row represents a single transaction. This format often makes it difficult to derive useful information.

Decision support system (DSS) data focuses on a broader time span, tends to have high levels of granularity, and can be examined in multiple dimensions. For example, note these possible aggregations:

- Sales by product, region, agent, and so on
- Sales for all years or only a few selected years
- Sales for all products or only a few selected products

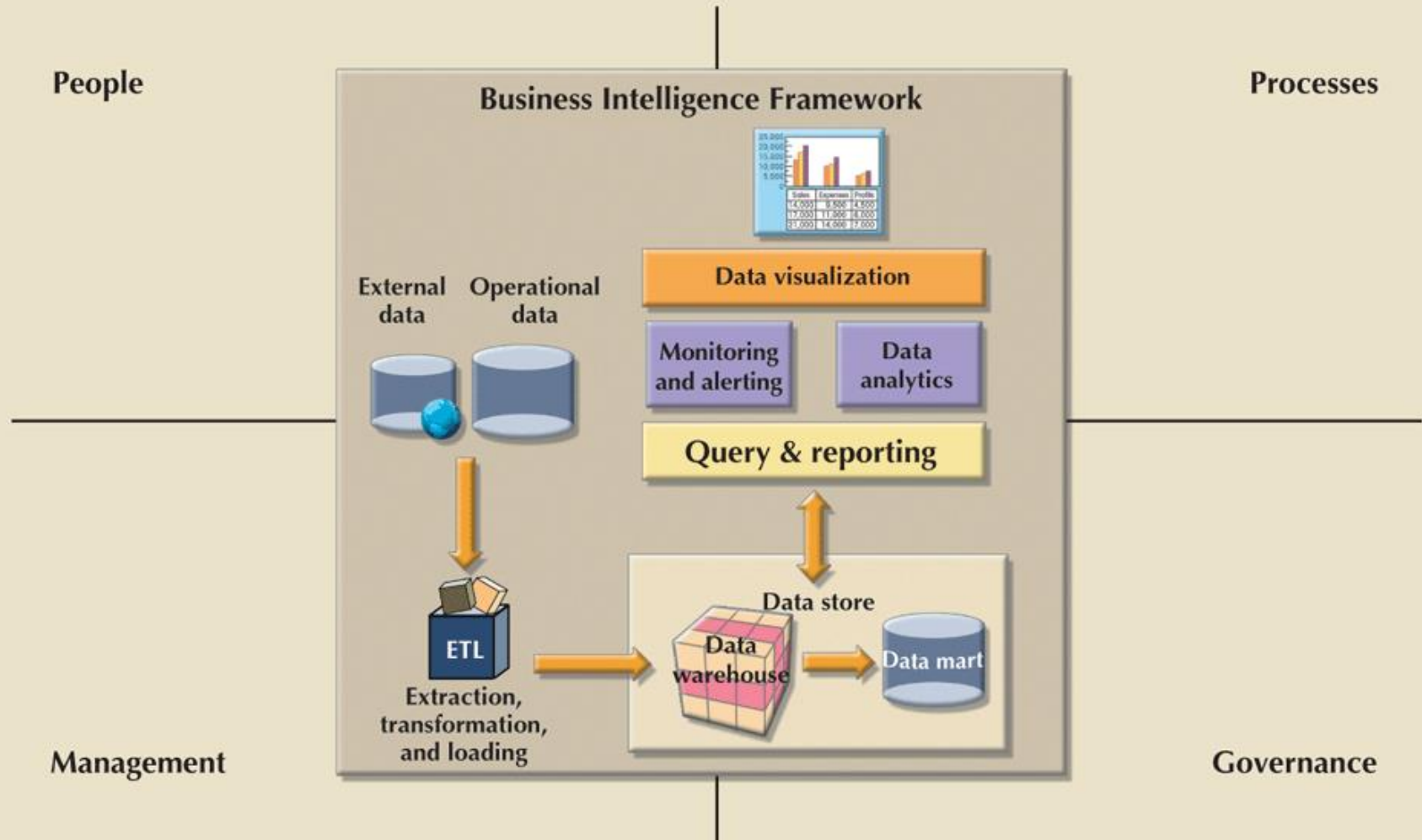# Operational vs Decision Support Data (2)

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

BI Architecture

FIGURE 13.1   BUSINESS INTELLIGENCE FRAMEWORK

People

Processes

Business Intelligence Framework

External data   Operational data

Data visualization

Monitoring and alerting

Data analytics

Query & reporting

ETL

Extraction, transformation, and loading

Data store

Data warehouse

Data mart

Management

Governance

BI Components

Business Intelligence Framework

TABLE 13.2

BASIC BI ARCHITECTURAL COMPONENTS

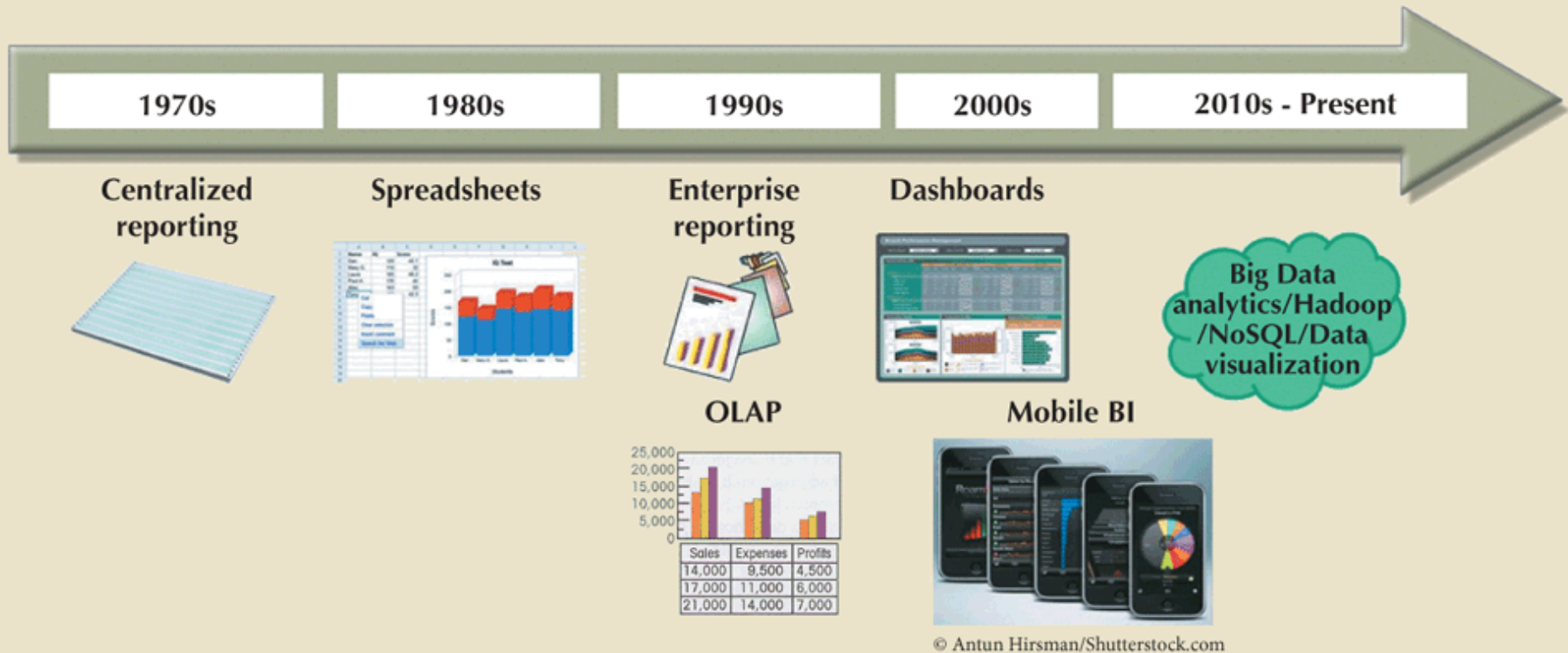| COMPONENT	DESCRIPTION
ETL tools	Data **extraction, transformation, and loading (ETL)** tools collect, filter, integrate, and aggregate internal and external data to be saved into a data store optimized for decision support.
Data store	The data store is optimized for decision support and is generally represented by a *data warehouse* or a *data mart*. The data is stored in structures that are optimized for data analysis and query speed.
Query and reporting	This component performs data selection and retrieval, and it is used by the data analyst to create queries that access the database and create the required reports.
Data visualization	This component presents data to the end user in a variety of meaningful and innovative ways. This tool helps the end user select the most appropriate presentation format, such as summary reports, maps, pie or bar graphs, mixed graphs, and static or interactive dashboards.
Data monitoring and alerting	This component allows real-time monitoring of business activities. The BI system will present concise information in a single integrated view. This integrated view could include specific metrics about the system performance or activities, such as number of orders placed in the last four hours, number of customer complaints by product by month, and total revenue by region. Alerts can be placed on a given metric; once the value of a metric goes below or above a certain baseline, the system will perform a given action, such as emailing shop floor managers, presenting visual alerts, or starting an application.
Data analytics	This component performs data analysis and data-mining tasks using the data in the data store. This tool advises the user as to which data analysis tool to select and how to build a reliable business data model. Business models are generated by special algorithms that identify and enhance the understanding of business situations and problems.
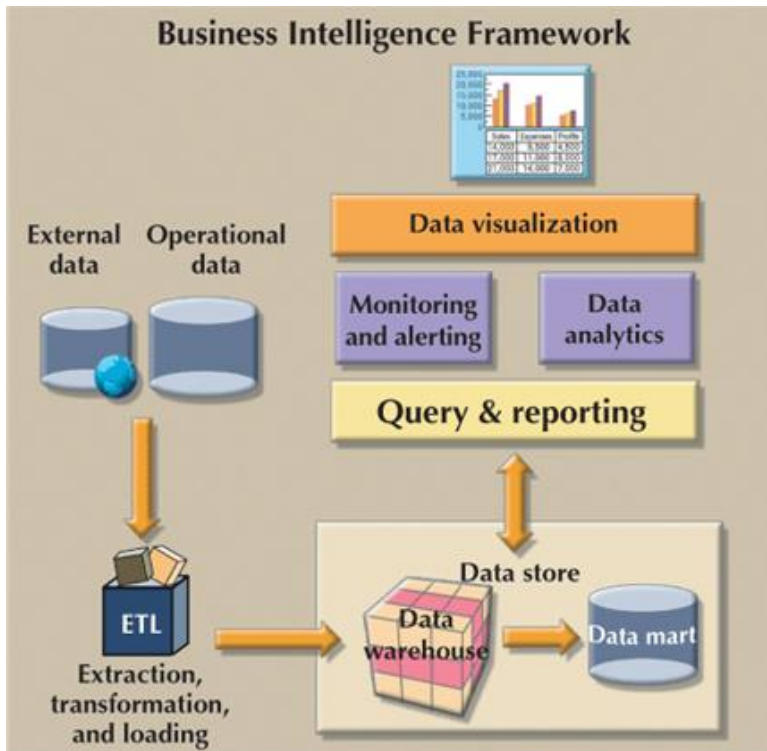
BI Tools

TABLE 13.3

SAMPLE OF BUSINESS INTELLIGENCE TOOLS

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 Tableau
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.	SAP Teradata MicroStrategy MS Analytics Services
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 Oracle Microsoft
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 Microsoft PowerBI

BI Benefits

- Improved decision making for **competitive advantage**

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

- Achieving all these benefits takes a lot of human, financial, technological resources, and time
  - BI benefits are not achieved overnight; are the result of a focused company-wide effort that could take a long time
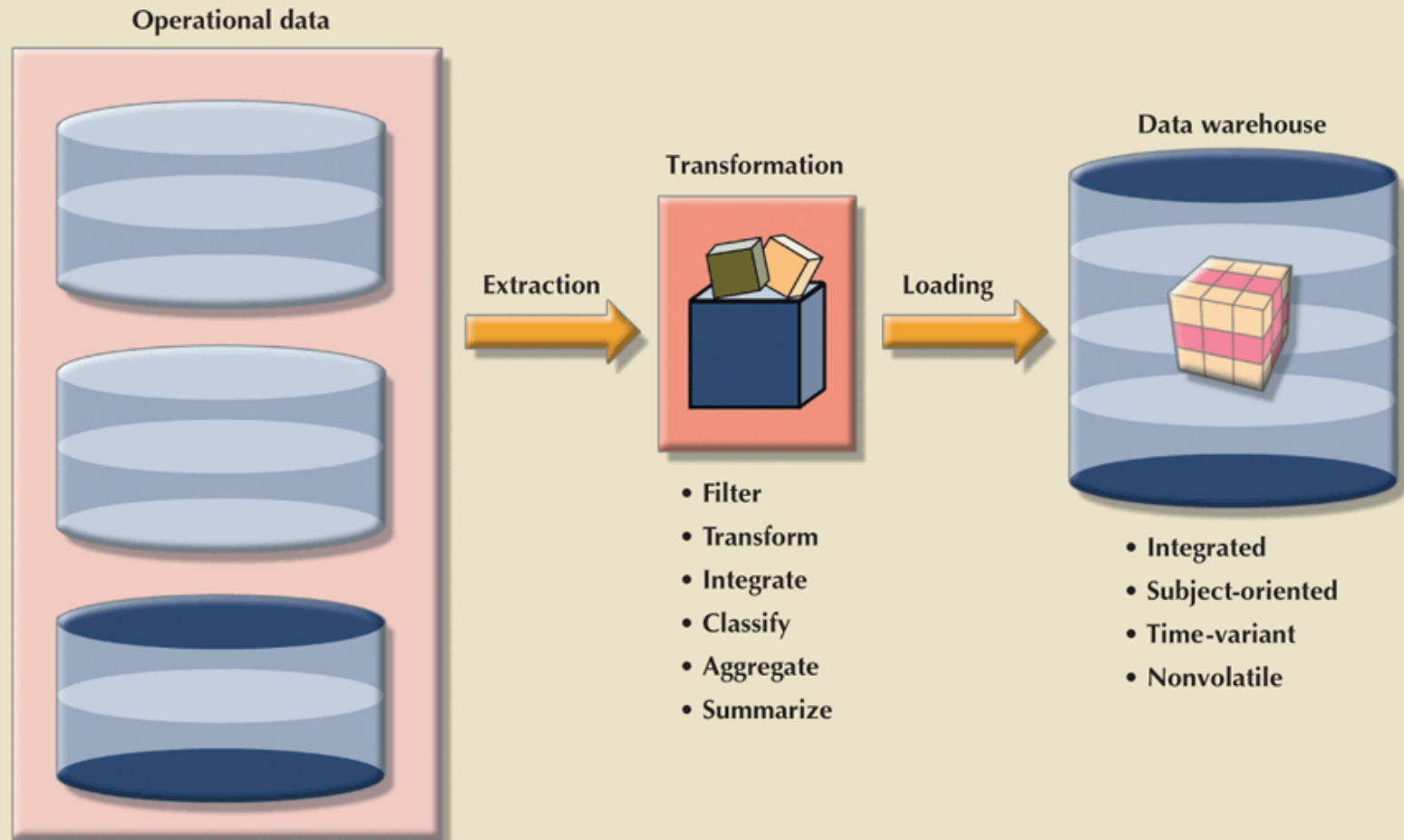
Evolution of BI

FIGURE 13.2 EVOLUTION OF BI INFORMATION DISSEMINATION FORMATS

ETL Process and Data Warehouse

FIGURE 13.4 THE ETL PROCESS

Operational data

Transformation

Extraction

Loading

Data warehouse

- Filter
- Transform
- Integrate
- Classify
- Aggregate
- Summarize

- Integrated
- Subject-oriented
- Time-variant
- Nonvolatile

Data in Data Warehouse

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.

Data Marts

- A Data Mart is a small single-subject subset of Data Warehouse
  - provides decision support to a small group of people with faster data access

- Data Marts and Data Warehouse co-exist within a BI environment
  - they are different in size and scope of the problem being solved
  - problem definition and data requirements are essentially the same

- Data Marts provide people at different management levels access to required data with summarization, aggregation and presentation needed for them to make decisions at their levels

Data Analytics

- Subset of business intelligence (BI) functionality that encompasses a wide range of mathematical, statistical, and modeling techniques with the purpose of extracting knowledge from data

- Data analytics is a "shared" service in the BI framework
  - used in all levels with the BI framework, including queries and reporting, monitoring and alerting, and visualization
  - Discovers characteristics, relationships, dependencies, or trends in the organization's data and explains the discoveries and predicts future event

Extracting Knowledge from Data

FIGURE 13.18 EXTRACTING KNOWLEDGE FROM DATA
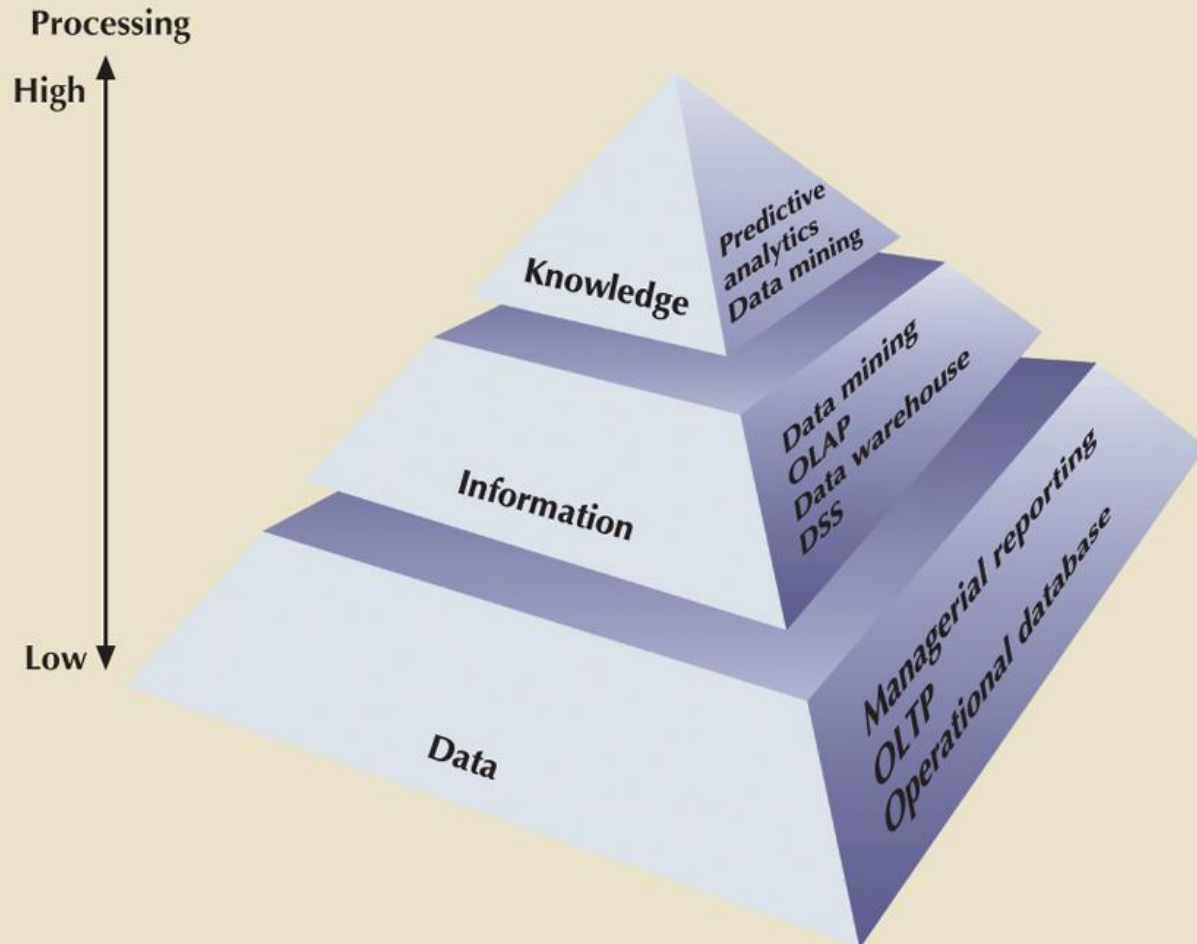
Explanatory vs Predictive Analytics
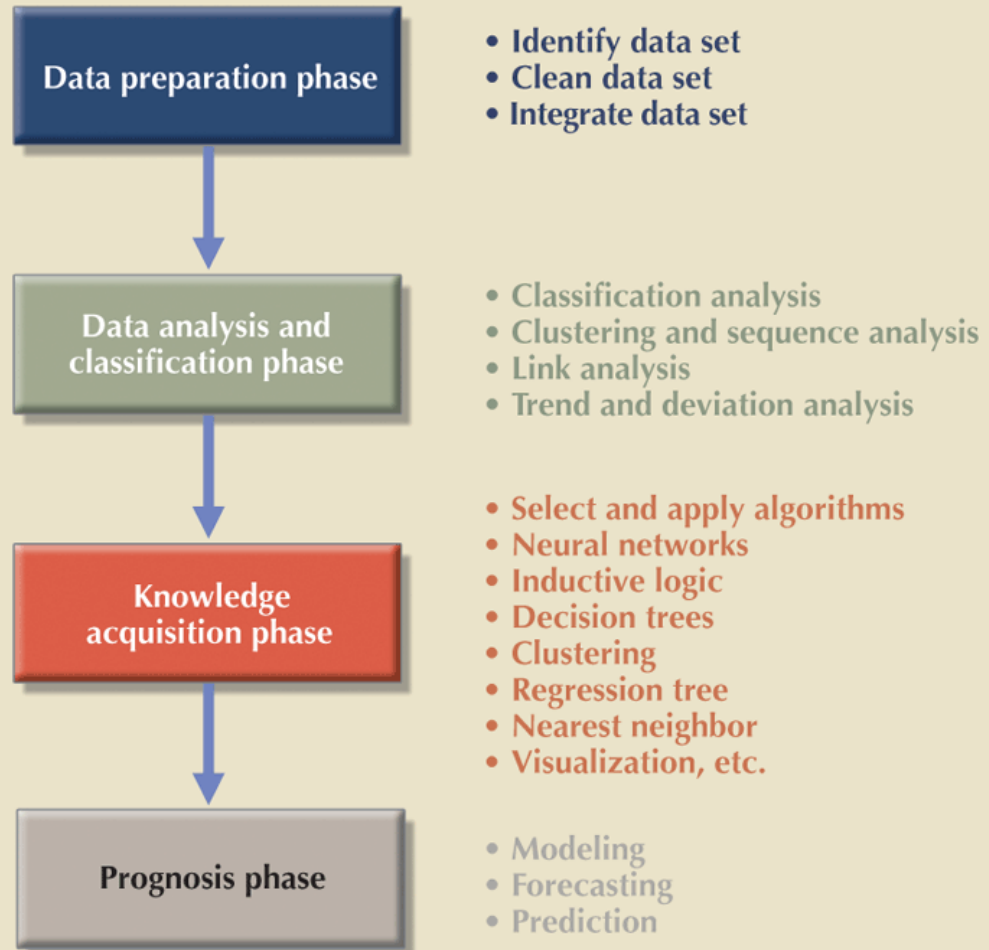
- **Explanatory analytics**: focuses on discovering and explaining data characteristics and relationships based on existing data
    - what products customers buy together?
    - customer/marketing segmentation

- **Predictive analytics**: focuses on predicting future data outcomes with a high degree of accuracy
    - sales forecast for next week/month
    - credit risk prediction

Data Mining

Analyzing massive amounts of data to uncover hidden trends, patterns, and relationships; to form computer models to simulate and explain the findings; and to use such models to support business decision making

FIGURE 13.19   DATA-MINING PHASES

Data preparation phase
- Identify data set
- Clean data set
- Integrate data set

Data analysis and classification phase
- Classification analysis
- Clustering and sequence analysis
- Link analysis
- Trend and deviation analysis

Knowledge acquisition phase
- Select and apply algorithms
- Neural networks
- Inductive logic
- Decision trees
- Clustering
- Regression tree
- Nearest neighbor
- Visualization, etc.

Prognosis phase
- Modeling
- Forecasting
- Prediction

Data Visualisation

- Visual representation of data that enhances the user's ability to comprehend the meaning of the data
  - Goal is to allow the user to quickly and efficiently see the data's big picture by identifying trends, patterns, and relationships

Visualisation makes it easier to understand data, in particular, large amount of data

Science of Data Visualisation
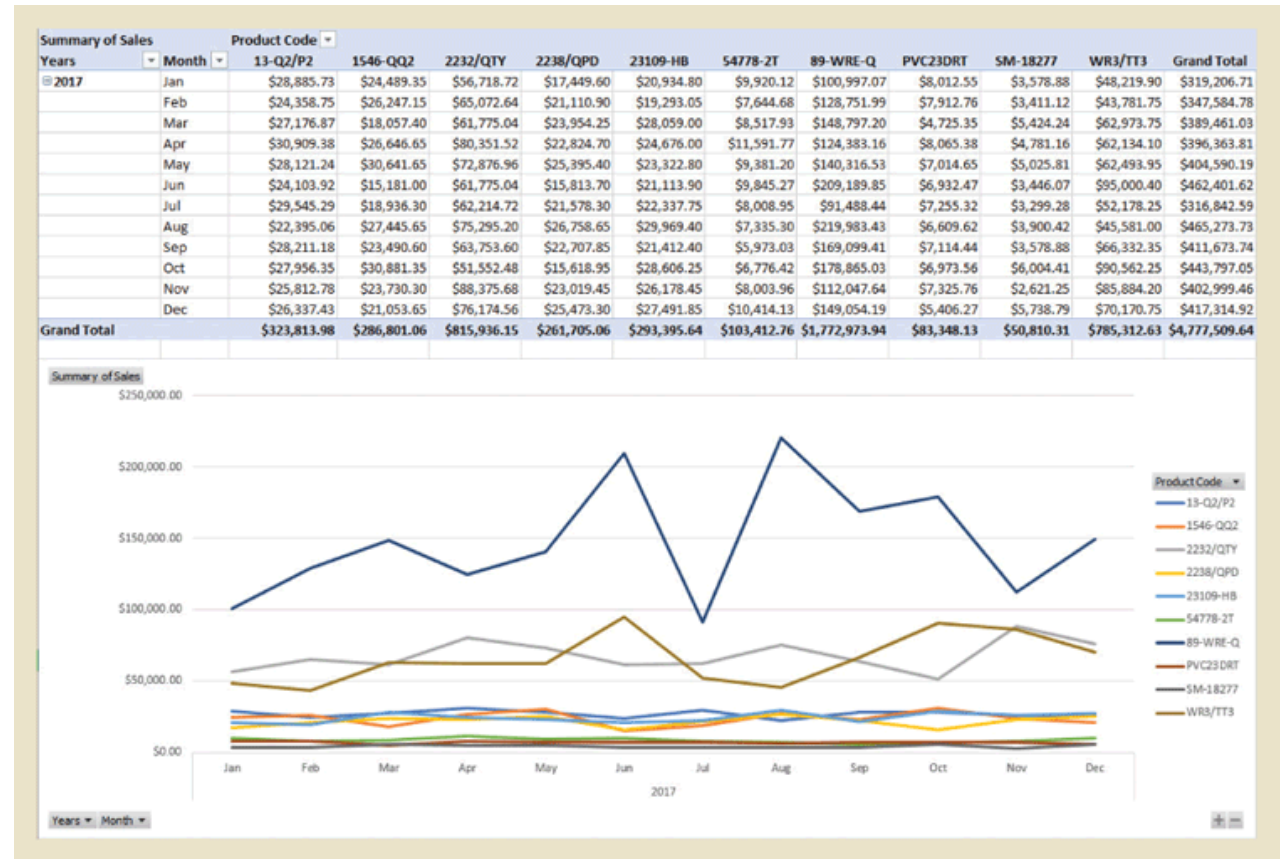
- Roots in cognitive sciences: how the human brain receives, interprets, organizes, and processes information
  - **Pattern recognition**: visually identifying trends, distributions, and relationships
  - **Spatial awareness**: use of size and orientation to compare and relate data
  - **Aesthetics**: use of shapes and color to highlight and contrast data composition and relationships

FIGURE 13.28 THE POWER OF VISUAL COMMUNICATION
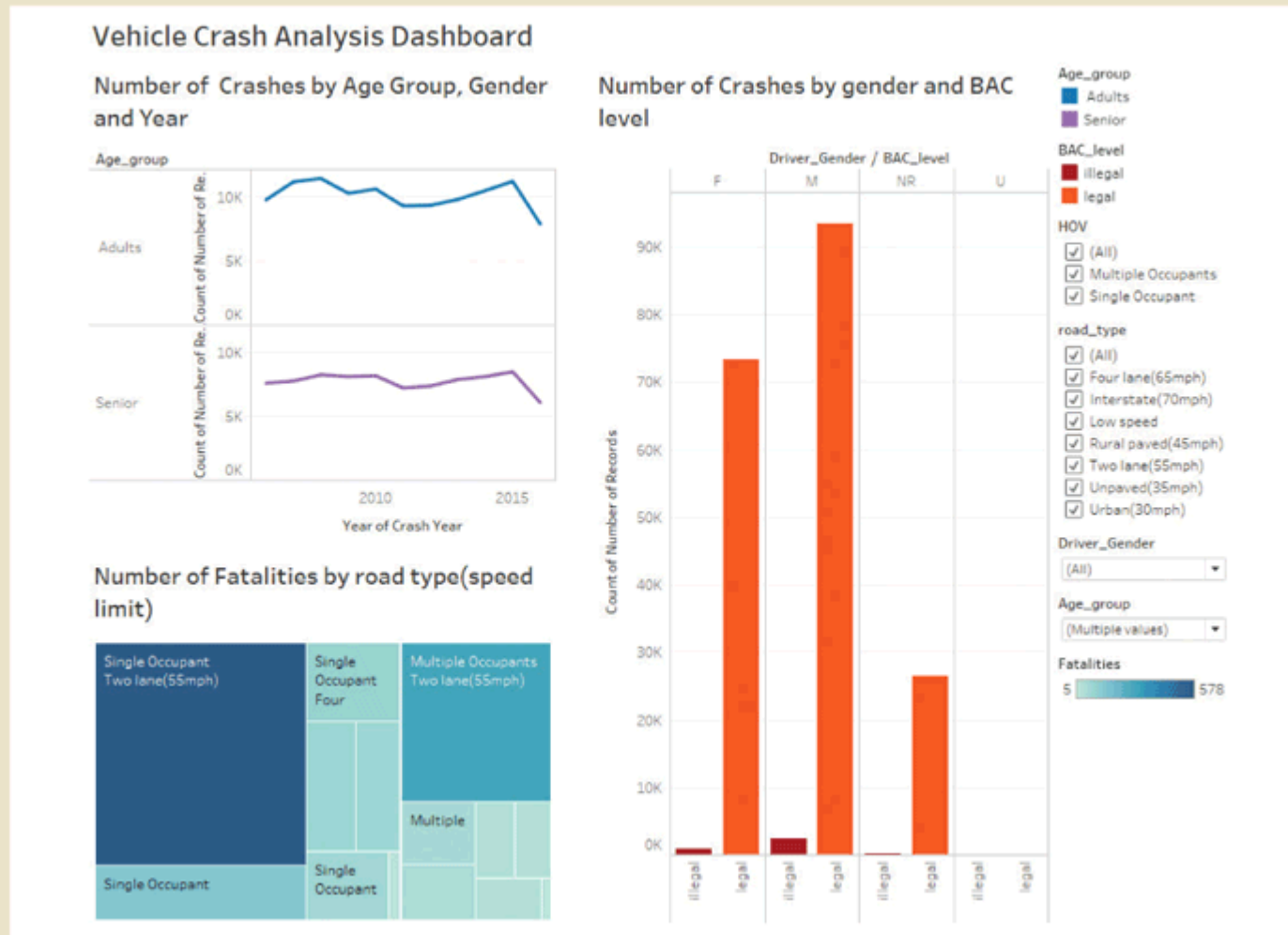
How many soccer balls are there?

Panel A

Panel B

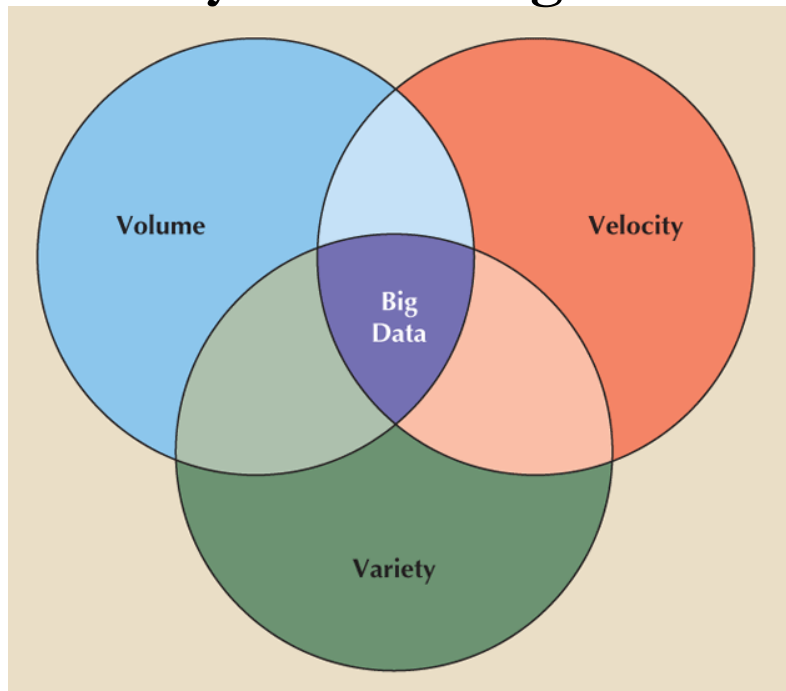# Data Visualisation Example

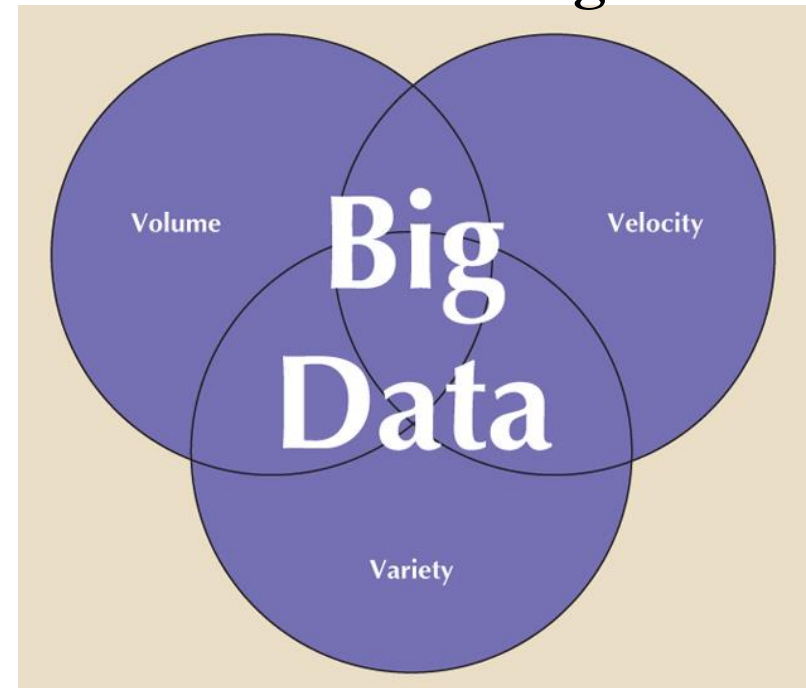FIGURE 13.29 VEHICLE CRASH ANALYSIS

Big Data

- Big Data generally refers to a set of data that displays the characteristics of volume, velocity, and variety (the **3V**s) to an extent that makes the data unsuitable for management by a relational database management system

Early view of Big Data

Current view of Big Data

Big Data (1)

- **Volume**: quantity of data to be stored
  - Scaling up: keeping the same number of systems but migrating each one to a larger system
  - Scaling out: when the workload exceeds server capacity, it is spread out across a number of servers

- **Velocity**: speed at which data is entered into system and must be processed
  - Stream processing: focuses on input processing and requires analysis of data stream as it enters the system
  - Feedback loop processing: analysis of data to produce actionable results

- **Variety**: variations in the structure of data to be stored
  - Structured data: fits into a predefined data model (e.g., transactional data)
  - Unstructured data: does not fit into a predefined model (e.g., images, text, audio, video)

Other Characteristics of Big Data

- **Variability**: changes in meaning of data based on context
  - Sentimental analysis: attempts to determine if a statement conveys a positive, negative, or neutral attitude about a topic (sarcasm to express negative sentiment)
- **Veracity**: trustworthiness of data
- **Value**: degree data can be analyzed for meaningful insight
- **Visualization**: ability to graphically resent data to make it understandable

- Relational databases may not necessarily be the best for storing and managing all organizational data
  - **Polyglot persistence**: coexistence of a variety of data storage and management technologies within an organization's infrastructure

Feedback Loop Processing

- Analysis of data to produce actionable results
- The process of capturing the data, processing it into usable information, and then acting on that information is a feedback loop

  E.g., Recommendations based on user clicks – news, tweets, FB feeds, products, books, songs, videos recommendation

FIGURE 14.3 FEEDBACK LOOP PROCESSING

Data is analyzed to determine other books and products the user may like

List of recommended items added to the user request

Data is captured about the user and about the book requested

Information requested by user plus information on recommendations are returned

User clicks on a link for a book
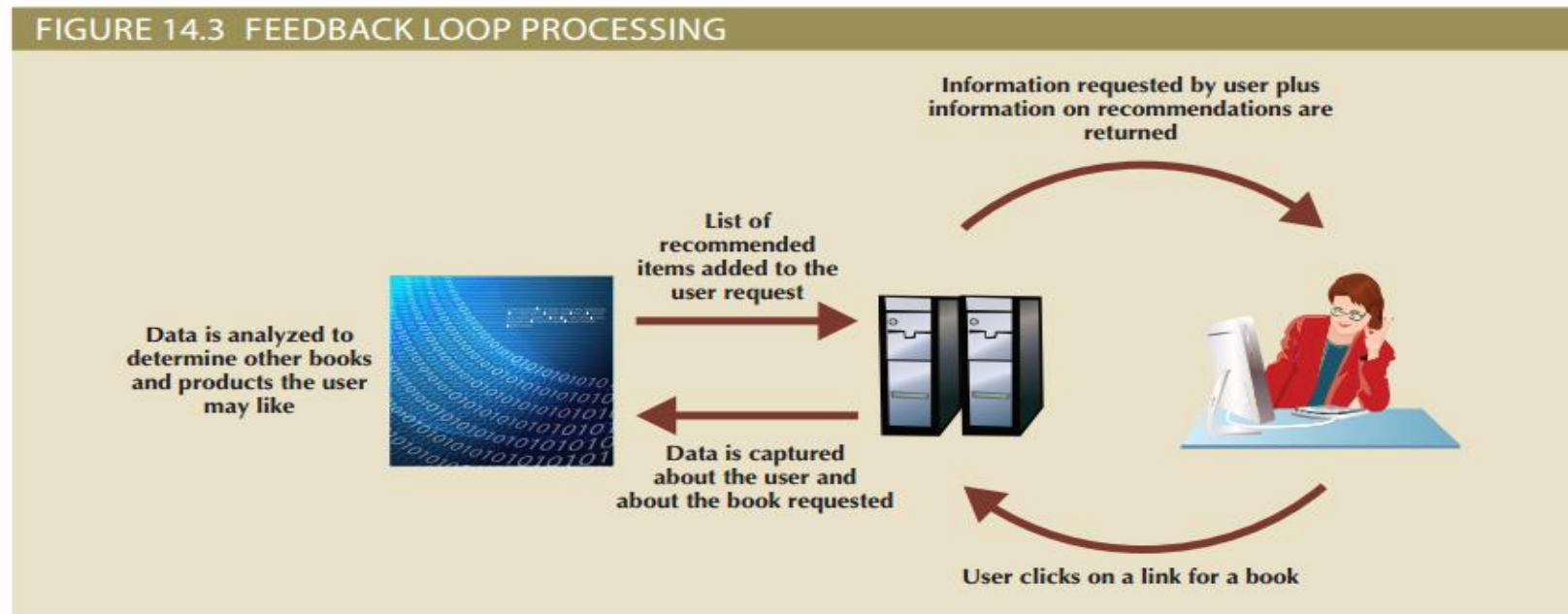
Hadoop

- De facto standard for most Big Data storage and processing

- Hadoop is not a database.

- Hadoop is a Java-based framework for distributing and processing very large data sets across clusters of computers.

- Two most important components:

  – **Hadoop Distributed File System (HDFS)**: low-level distributed file processing system that can be used directly for data storage

  – **MapReduce**: is a programming model that supports processing large data sets in a highly parallel, distributed manner.

Based on several key assumptions

- **High volume:** Hadoop assumes that files will be extremely large.
  - Data in the HDFS is organized into physical blocks, just as in other types of file storage systems.

- **Write-once, read-many**:
  - Using a write-once, read-many model simplifies concurrency issues and improves overall data throughput.
  - Using this model, a file is created, written to the file system, and then closed. Once the file is closed, changes cannot be made to its contents.

- **Streaming access:** optimized for batch processing of entire files as a continuous stream of data

- **Fault tolerance:** designed to replicate data across many different devices so that when one fails, data is still available from another device

MapReduce

- Framework used to process large data sets across clusters
- An open-source API that provides fast data analytics services
- One of the main big-data technologies to process massive data stores.
- Breaks down complex tasks into smaller subtasks, performing the subtasks, and producing a result
  - divide and conquer
- Combination of two functions
  - **Map function** takes a collection of data and sorts and filters it into a set of key-value pairs
    - Mapper program performs the map function
  - **Reduce function** takes a collection of key-value pairs, all with the same key value, and summaries them to a single result
    - Reducer program performs the reduce function

NoSQL

- A new generation of database management systems that is not based on traditional relational database model developed to address Big Data challenges

Popular NoSQL Database

TABLE 14.3

NoSQL DATABASES

NoSQL CATEGORY	EXAMPLE DATABASES	DEVELOPER
Key-value database	Dynamo Riak Redis Voldemort	Amazon Basho Redis Labs LinkedIn
Document databases	MongoDB CouchDB OrientDB RavenDB	MongoDB, Inc. Apache OrientDB Ltd. Hibernating Rhinos
Column-oriented databases	HBase Cassandra Hypertable	Apache Apache (originally Facebook) Hypertable, Inc.
Graph databases	Neo4J ArangoDB GraphBase	Neo4j ArangoDB, LLC FactNexus

Key-Value (KV) database

- conceptually the simplest of the NoSQL data models
- Store data as a collection of key-value pairs organized as **buckets** which are the equivalent of tables
- Buckets are logical grouping of keys
- Operations: *get/fetch*, *store*, and *delete*

FIGURE 14.7 KEY-VALUE DATABASE STORAGE

Bucket = Customer

Key	Value
10010	"LName Ramas FName Alfred Initial A Areacode 615 Phone 844-2573 Balance 0"
10011	"LName Dunne FName Leona Initial K Areacode 713 Phone 894-1238 Balance 0"
10014	"LName Orlando FName Myron Areacode 615 Phone 222-1672 Balance 0"

Document Databases

- Store data in key-value pairs
- Value components are tag-encoded documents
- Encoded documents are grouped into large groups called **Collections**
- The document can be in any encoded format, such as XML, JSON

FIGURE 14.8 DOCUMENT DATABASE TAGGED FORMAT

Collection = Customer

Key	Document
10010	{LName: "Ramas", FName: "Alfred", Initial: "A", Areacode: "615", Phone: "844-2573", Balance: "0"}
10011	{LName: "Dunne", FName: "Leona", Initial: "K", Areacode: "713", Phone: "894-1238", Balance: "0"}
10014	{LName: "Orlando", FName: "Myron", Areacode: "615", Phone: "222-1672", Balance: "0"}

JavaScript Object Notation (JSON)

- open-standard file format

- human-readable text

- data objects consisting of:

  - attribute–value pairs

  - array data types
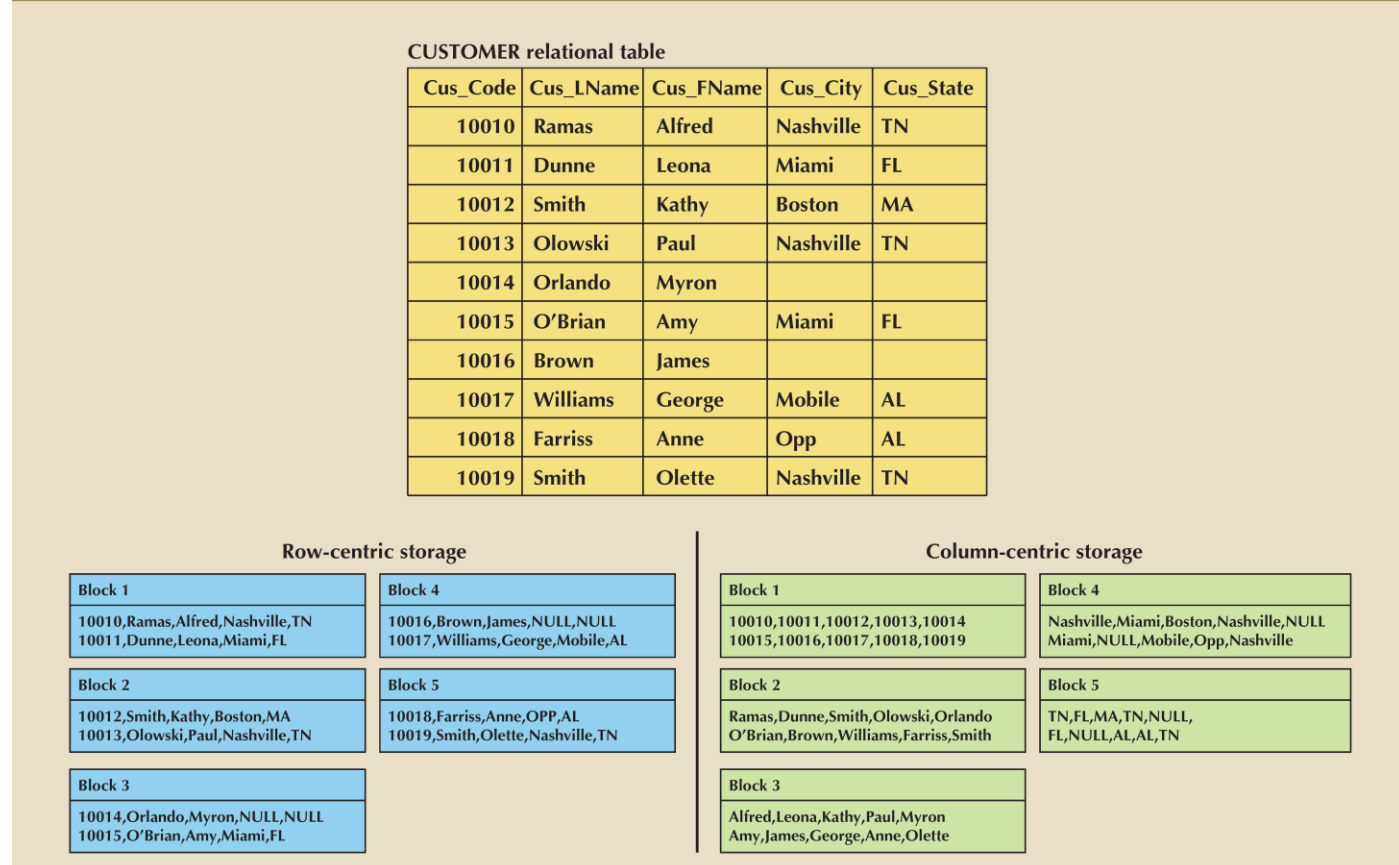
```
{
  "firstName": "John",
  "lastName": "Smith",
  "age": 25,
  "address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021"
  },
  "phoneNumber": [
    {
      "type": "home",
      "number": "212 555-1234"
    },
    {
      "type": "fax",
      "number": "646 555-4567"
    }
  ],
  "gender": {
    "type": "male"
  }
}
```

Column-Oriented Databases

- Column-oriented databases refers to **two technologies**
  - **Column-centric storage**: data stored in blocks which hold data from a single column across many rows
  - **Row-centric storage**: data stored in block which hold data from all columns of a given set of rows

FIGURE 14.9 COMPARISON OF ROW-CENTRIC AND COLUMN-CENTRIC STORAGE

CUSTOMER relational table

Cus_Code	Cus_LName	Cus_FName	Cus_City	Cus_State
10010	Ramas	Alfred	Nashville	TN
10011	Dunne	Leona	Miami	FL
10012	Smith	Kathy	Boston	MA
10013	Olowski	Paul	Nashville	TN
10014	Orlando	Myron		
10015	O'Brian	Amy	Miami	FL
10016	Brown	James		
10017	Williams	George	Mobile	AL
10018	Farriss	Anne	Opp	AL
10019	Smith	Olette	Nashville	TN

Row-centric storage

Block 1
10010,Ramas,Alfred,Nashville,TN
10011,Dunne,Leona,Miami,FL

Block 2
10012,Smith,Kathy,Boston,MA
10013,Olowski,Paul,Nashville,TN

Block 3
10014,Orlando,Myron,NULL,NULL
10015,O'Brian,Amy,Miami,FL

Block 4
10016,Brown,James,NULL,NULL
10017,Williams,George,Mobile,AL

Block 5
10018,Farriss,Anne,OPP,AL
10019,Smith,Olette,Nashville,TN

Column-centric storage

Block 1
10010,10011,10012,10013,10014
10015,10016,10017,10018,10019

Block 2
Ramas,Dunne,Smith,Olowski,Orlando
O'Brian,Brown,Williams,Farriss,Smith

Block 3
Alfred,Leona,Kathy,Paul,Myron
Amy,James,George,Anne,Olette

Block 4
Nashville,Miami,Boston,Nashville,NULL
Miami,NULL,Mobile,Opp,Nashville

Block 5
TN,FL,MA,TN,NULL,
FL,NULL,AL,AL,TN

Graph Databases

- Store relationship-rich data as a collection of nodes and edges
  - **Propertie**s: like attributes; they are the data that we need to store about the node
  - **Traversal**: query in a graph database
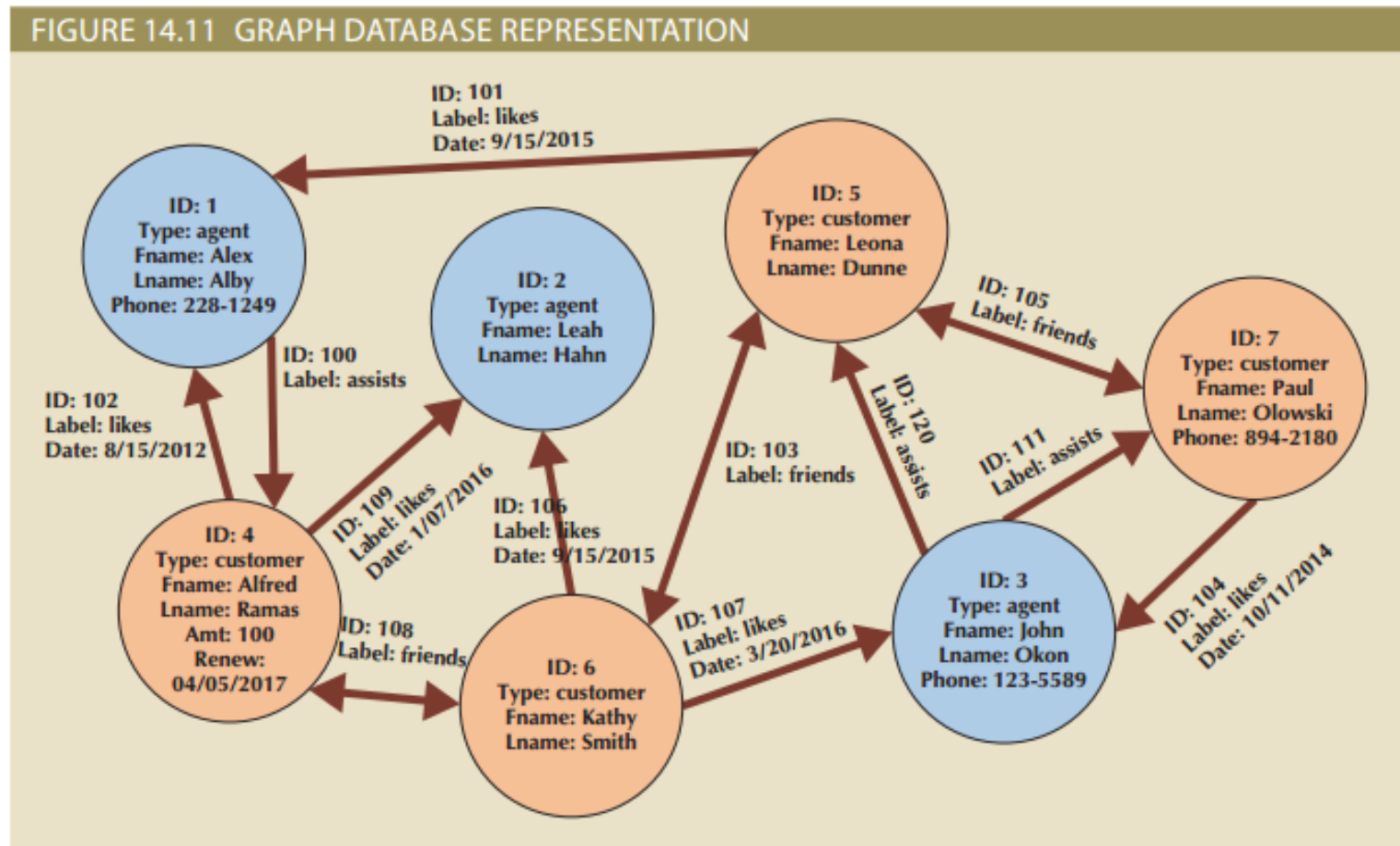
FIGURE 14.11 GRAPH DATABASE REPRESENTATION

Summary

- Operational and Decision Support Data

- Business Intelligence

- Data Warehouse and Data Marts

- Data Analytics and Data Mining

- Data Visualisation

- Big Data

- NoSQL Databases

This Week's OnTrack Task

- 9.1P Simple Business Intelligence using Excel

- 9.2D Interaction with a database via an user interface

Next Week

- Data Security and Unit Review

Thank you

See you next week

Any questions/comments?

Readings and References:

- Chapter 13 and 14

Database Systems : Design, Implementation, & Management 13TH EDITION, by Carlos Coronel, Steven Morris