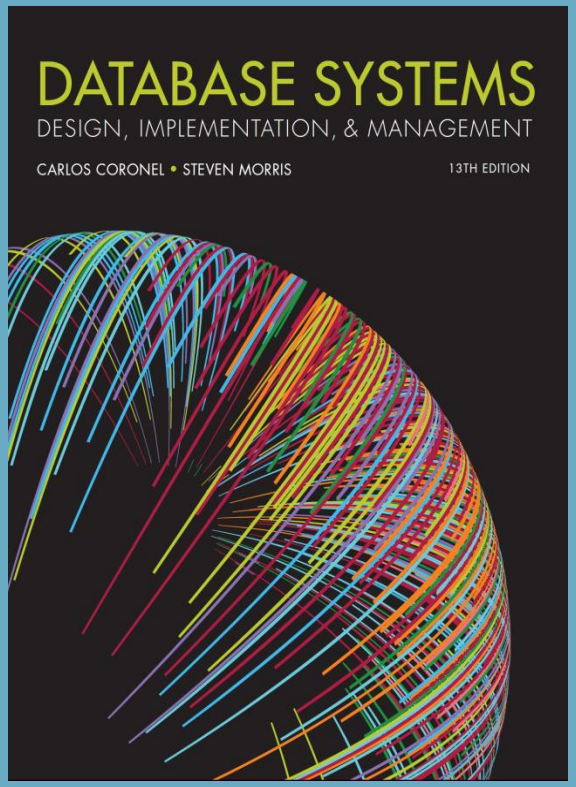


Chapter 14

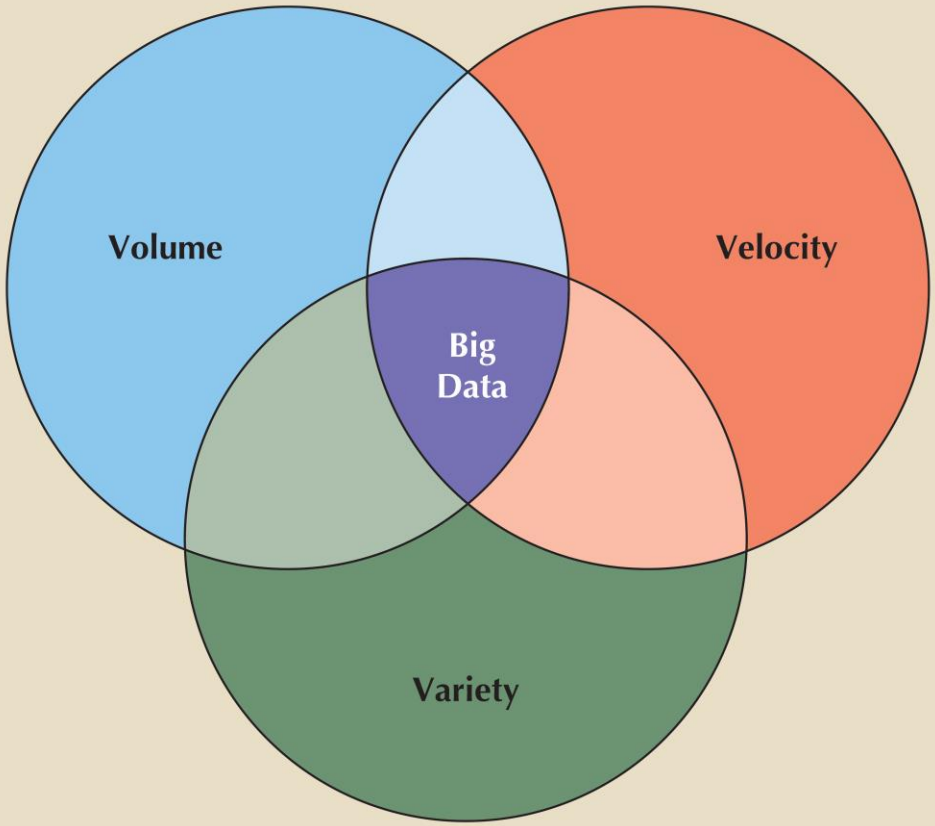
Big Data and NoSQL

Learning Objectives

- After completing this chapter, you will be able to:
  - Explain the role of Big Data in modern business
  - Describe the primary characteristics of Big Data and how these go beyond the traditional “3 Vs”
  - Summarize the four major approaches of the NoSQL data model and how they differ from the relational model
  - Understand how to work with document databases using MongoDB
  - Understand how to work with graph databases using Neo4j

Big Data (2 of 4)

FIGURE 14.1 ORIGINAL VIEW OF BIG DATA

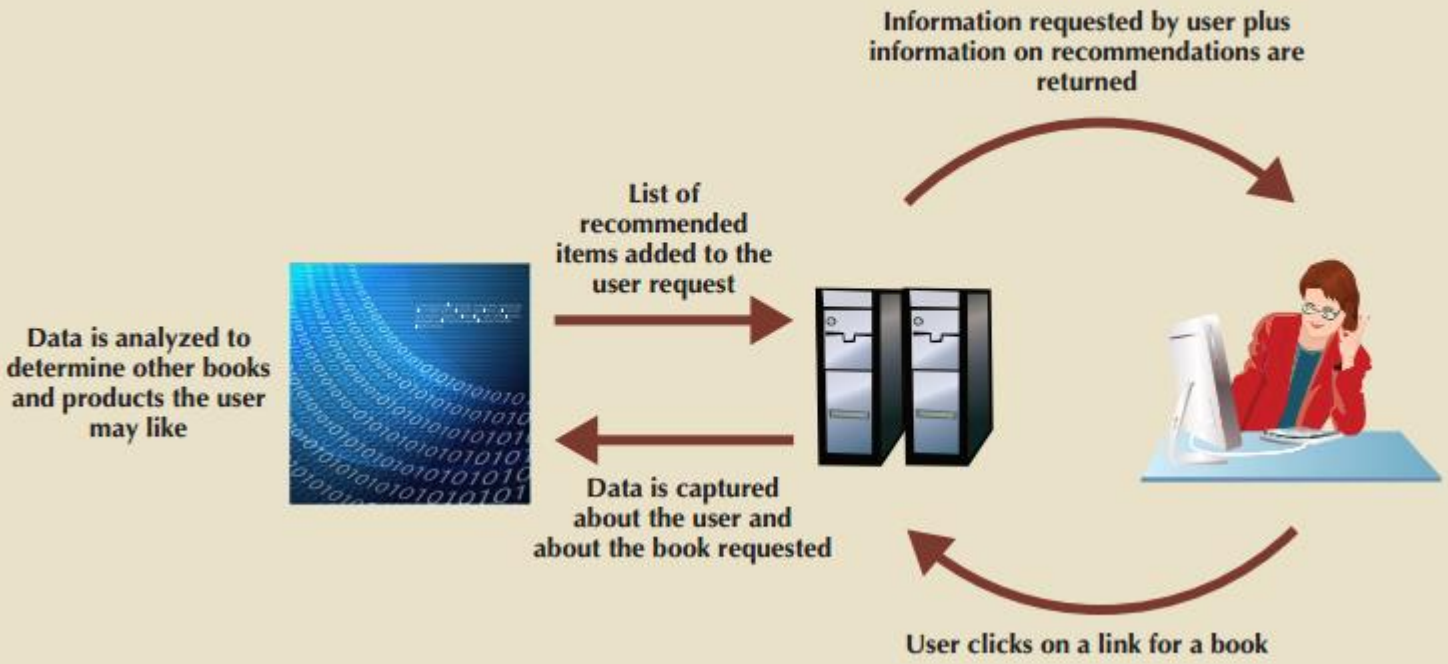


Big Data (1 of 4)

- 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
  - Unstructured data: does not fit into a predefined model

Big Data (3 of 4)

FIGURE 14.3 FEEDBACK LOOP PROCESSING



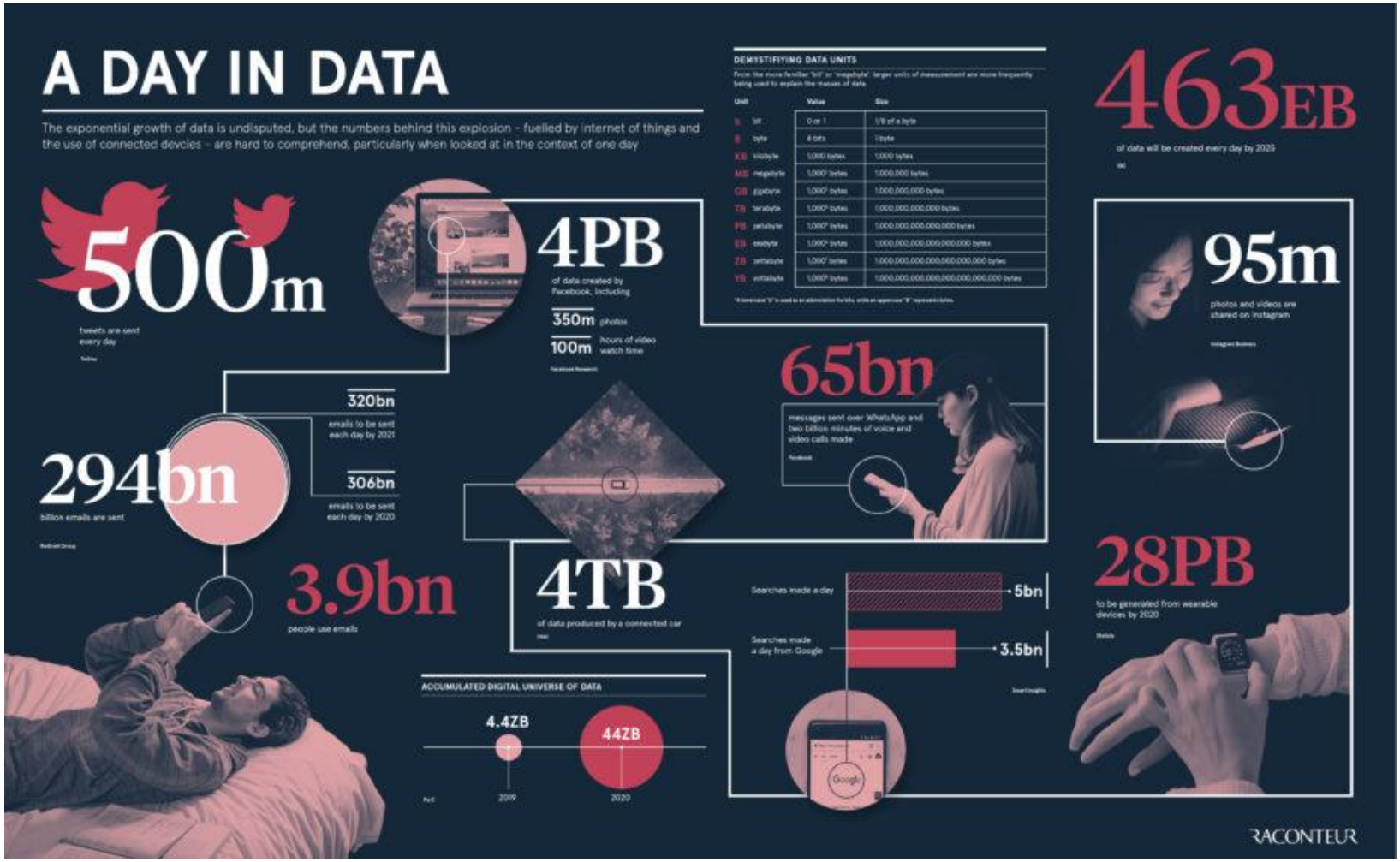
Big Data (4 of 4)

- Other characteristics
  - 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
  - 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 are not necessarily 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

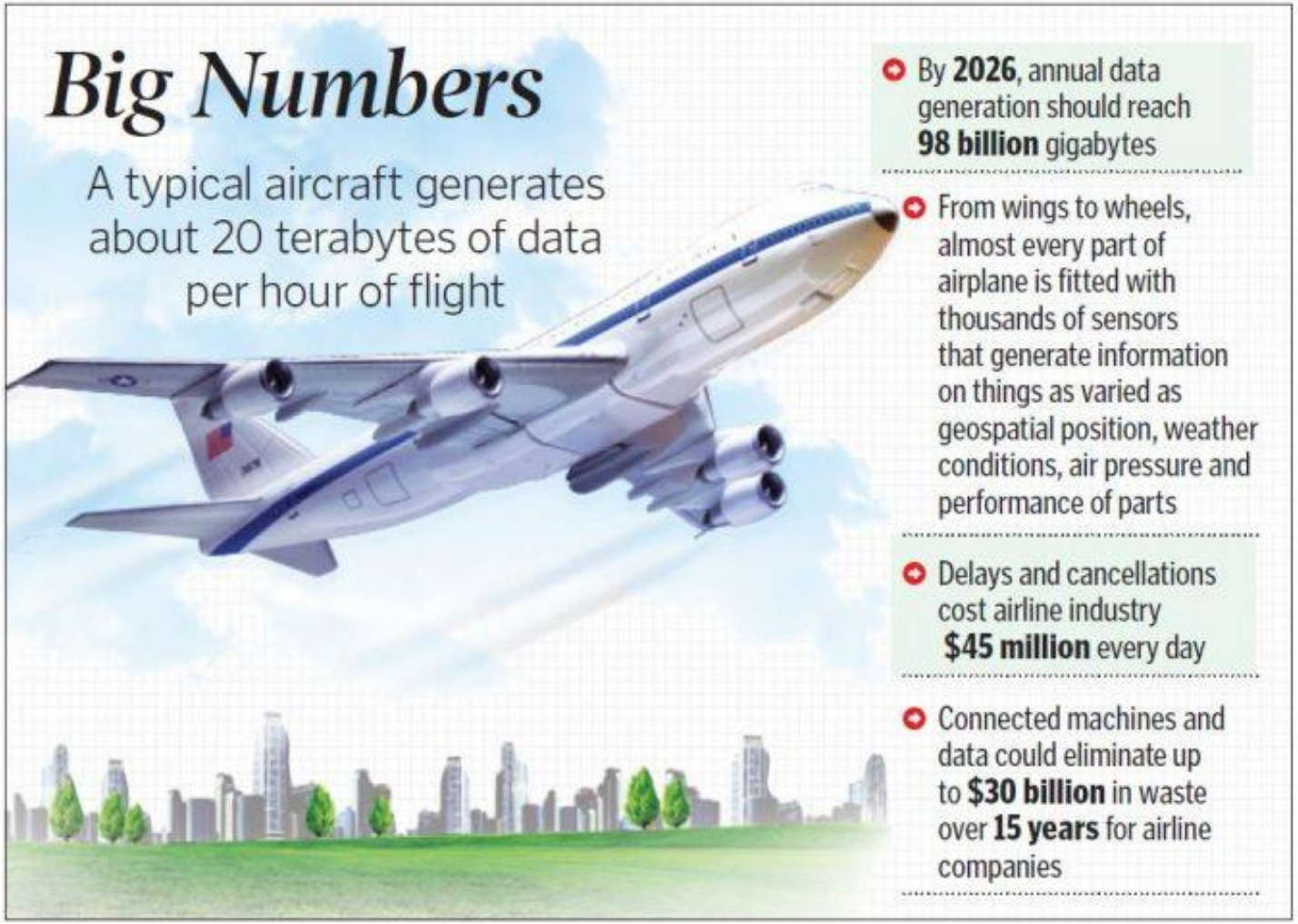
Examples of Big Data

- General Stats: Per Minute Ratings
  - Here are some of the per minute ratings for various social networks:
  - Snapchat: Over 527,760 photos shared by users
  - LinkedIn: Over 120 professionals join the network
  - YouTube: 4,146,600 videos watched
  - Twitter: 456,000 tweets sent or created
  - Instagram: 46,740 photos uploaded
  - Netflix: 69,444 hours of video watched
  - Giphy: 694,444 GIFs served
  - Tumblr: 74,220 posts published
  - Skype: 154,200 calls made by users

How Much Data



Airline Example IOT



NoSQL Applications You Might be Using

- Facebook
- LinkedIn
- Gmail

NoSQL (1 of 7)

- Nosql: non-relational database technologies developed to address Big Data challenges
  - Name does not describe what the NoSQL technologies are, but rather what they are not (poor job of that as well)
- Key-value (KV) databases: 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
- Document databases: similar to key-value databases and can almost be considered a subtype of KV databases
  - Store data in key-value pairs in which the value components are encoded documents grouped into large groups called collections

NoSQL (2 of 7)

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"

NoSQL (3 of 7)

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"}

NoSQL (4 of 7)

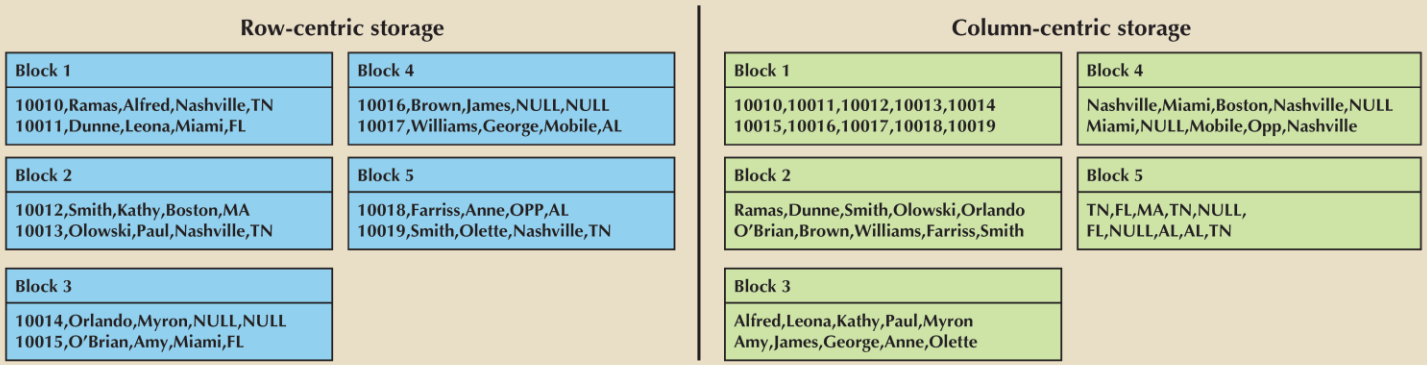
- Column-oriented databases refers to two technologies
  - Column-centric storage: data stored in blocks which hold data from a single column across many rows
  - Row-centric storage: data stored in block which hold data from all columns of a given set of rows
- Graph databases store data on relationship-rich data as a collection of nodes and edges
  - Properties: like attributes; they are the data that we need to store about the node
  - Traversal: query in a graph database

NoSQL (5 of 7)

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

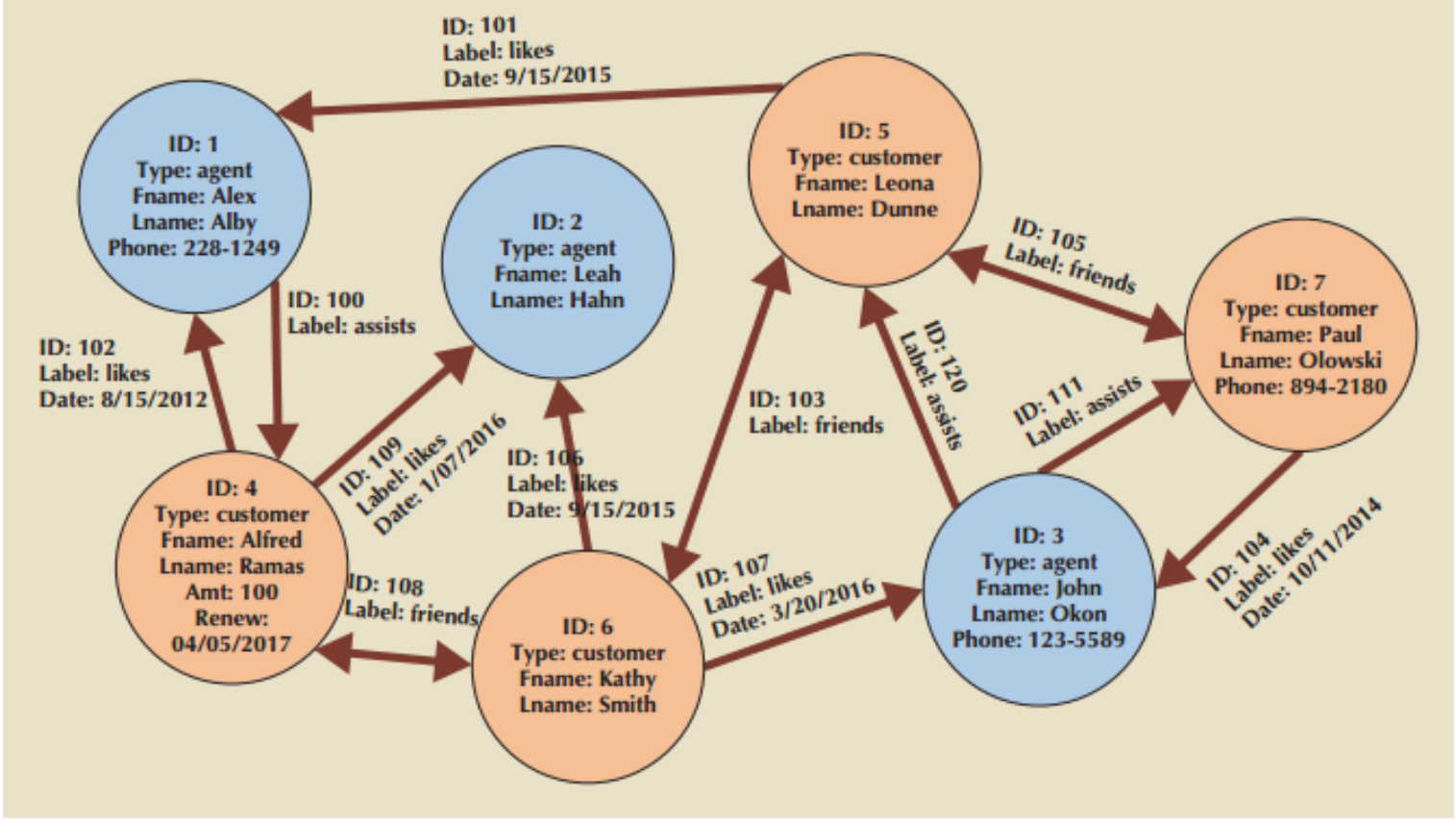
CUSTOMER relational table

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



NoSQL (6 of 7)

FIGURE 14.11 GRAPH DATABASE REPRESENTATION



NoSQL (7 of 7)

- Aggregate awareness: data is collected or aggregated around a central topic or entity
  - Aggregate aware database models achieve clustering efficiency by making each piece of data relatively independent
- Graph databases, like relational databases, are aggregate ignorant
  - Do not organize the data into collections based on a central entity

Working with Document Databases Using MongoDB (1 of 2)

- Popular document database
  - Among the NoSQL databases currently available, MongoDB has been one of the most successful in penetrating the database market
- MongoDB, comes from the word humongous as its developers intended their new product to support extremely large data sets
  - High availability
  - High scalability
  - High performance

Working with Document Databases Using MongoDB (2 of 2)

- Importing Documents in MongoDB
  - Refer to the text for an importation example and considerations
- Example of a MongoDB Query Using find()
  - Methods are programed functions to manipulate objects
    - Find() method retrieves objects from a collection that match the restrictions provided
    - Pretty() method is used to improve readability of the documents by placing key:value pairs on separate lines
  - Refer to the text for a query example

Working with Graph Databases Using Neo4j (1 of 3)

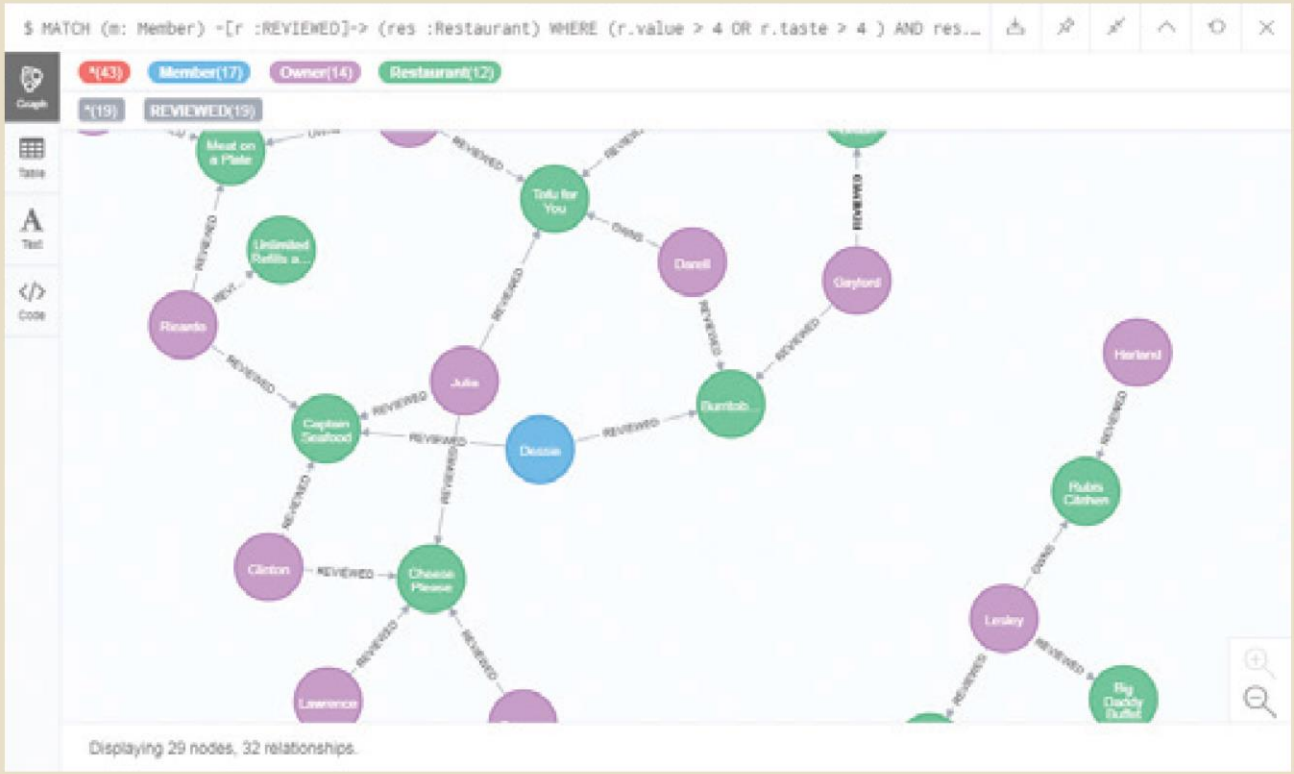
- Even though Neo4j is not yet as widely adopted as MongoDB, it has been one of the fastest growing NoSQL databases
  - Graph databases still work with concepts similar to entities and relationships
    - Focus is on the relationships
  - Graph databases are used in environments with complex relationships among entities
    - Heavily reliant on interdependence among their data
  - Neo4j provides several interface options
    - Designed with Java programming in mind
- Creating nodes in Neo4j
  - Nodes in a graph database correspond to entity instances in a relational database
  - Cypher is the interactive, declarative query language in Neo4j
  - Nodes and relationships are created using a CREATE command

Working with Graph Databases Using Neo4j (2 of 3)

- Refer to the text for examples
  - Using the CREATE command to create a member node
  - Retrieving node data with MATCH and WHERE
  - Retrieving relationship data with MATCH and WHERE

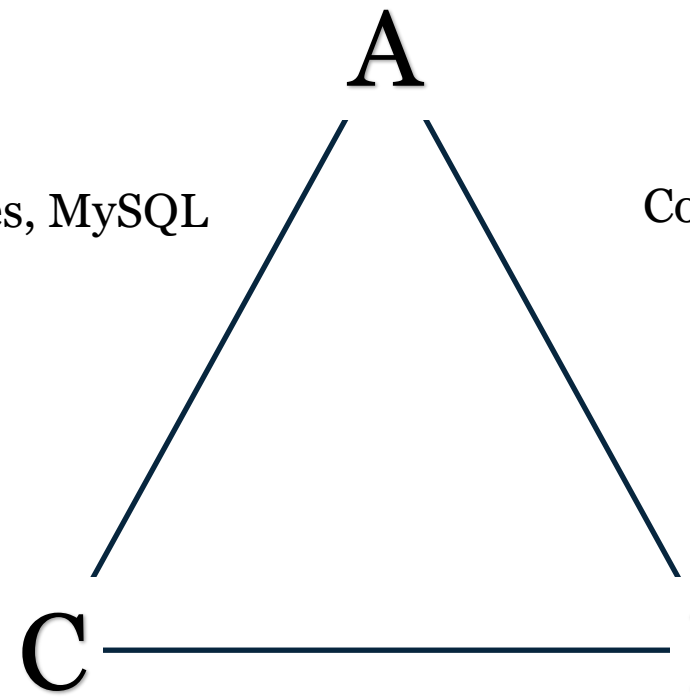
Working with Graph Databases Using Neo4j (3 of 3)

FIGURE 14.13 NEO4J QUERY USING MATCH/WHERE/RETURN



Where does it fit in CAP Theorem

Availability: Each client can always read and write



RDBMS, Postgres, MySQL

Couch DB, Cassandra, Dynamo

Consistency: All clients always have the same view of data

Big Table, Hbase, MongoDB, Redis

Partition Tolerance: Works well despite physical network partitions

Summary (1 of 2)

- Big Data is characterized by data of such volume, velocity, and/or variety that the relational model struggles to adapt to it
- Volume, velocity, and variety are collectively referred to as the 3 Vs of Big Data
- The Hadoop framework has quickly emerged as a standard for the physical storage of Big Data
- NoSQL is a broad term to refer to any of several nonrelational database approaches to data management
- Key-value databases store data in key-value pairs
- Document databases also store data in key-value pairs, but the data in the value component is an encoded document

Summary (2 of 2)

- Column-oriented databases, also called column family databases, organize data into key-value pairs in which the value component is composed of a series of columns, which are themselves key-value pairs
- Graph databases are based on graph theory and represent data through nodes, edges, and properties
- NewSQL databases attempt to integrate features of both RDBMS (providing ACID-compliant transactions) and NoSQL databases (using a highly distributed infrastructure)
- MongoDB is a document database that stores documents in JSON format
- Neo4j is a graph database that stores data as nodes and relationships, both of which can contain properties to describe them