Chapter 14
Big Data Analytics and NoSQL

Learning Objectives

- In this chapter, you will learn:
  - What Big Data is and why it is important in modern business
  - The primary characteristics of Big Data
  - How the core components of the Hadoop framework, HDFs and MapReduce operate
  - The four major approaches of the NoSQL data model and how it differ from the relational model
  - About data analytics, including data mining and predictive analytics

How "*Big Data*" will change your life….

"We swim in a sea of data … and the sea level is rising rapidly."

Data-driven World in the future !

Examples of big data…..

Walmart handles more than 1 million customer transactions ***every hour***, which is imported into databases estimated to contain more than 2.5 petabytes * of data — ***the equivalent of 167 times the information contained in all the books in the US Library of Congress***.

FICO Credit Card Fraud Detection System protects 2.1 billion active accounts world-wide.

The volume of business data worldwide, across all companies, ***doubles every 1.2 years***, according to estimates

(1 Petabyte = 1000000000000000B = 1000^5 B = 10^{15} B = 1 million gigabytes)

*** Think of the hard drive on your computer at home having 500 gigabytes. Now multiply that by 2,000!**

How much data?

- Google processes 20 PB a day (2008)
- Google processes 3.5 billion requests per day, and it stores 10 exabytes of data (10 billion gigabytes!) (2015)
- Facebook has 2.5 PB of user data + 15 TB/day (4/2009)
- eBay has 6.5 PB of user data + 50 TB/day (5/2009)

640K ought to be enough for anybody.

Decimal		
Value	Metric	
1000	kB	kilobyte
1000^2	MB	megabyte
1000^3	GB	**gigabyte**
1000^4	TB	terabyte
1000^5	PB	petabyte
1000^6	EB	exabyte
1000^7	ZB	zettabyte
1000^8	YB	yottabyte

What is collecting all this data?

Smartphones & Apps

Apple's iPhone
(Apple O/S)

Samsung, HTC.
Nokia, Motorola
(Android O/S)

RIM Corp's Blackberry
(BlackBerry O/S)

Search Engines

Google's

Microsoft's

Yahoo's

IAC Search's

What is collecting all this data?

Games Boxes and GPS Systems

Social networks

What is collecting all this data?

Online Shopping Websites

淘宝网
Taobao.com

亚马逊
amazon.cn

JD.COM 京东
多 · 快 · 好 · 省

Banking and phone system

verizon

Can you hear me now?
(Heh heh heh!)

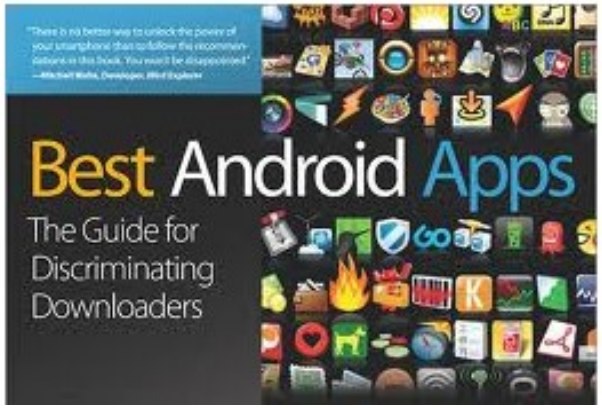
Sprint

T · · Mobile ·

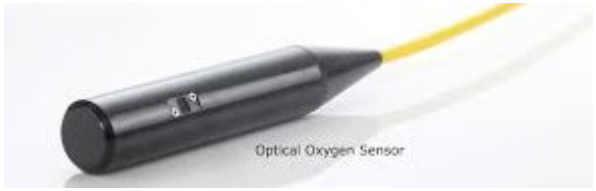
at&t

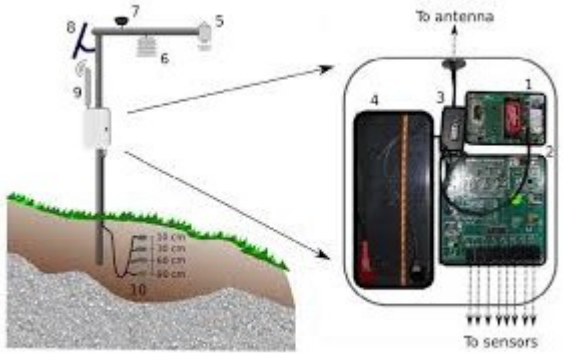
What is collecting all this data?

A real pain in the apps!



All the sensors





https://pjreddie.com/darknet/yolo/

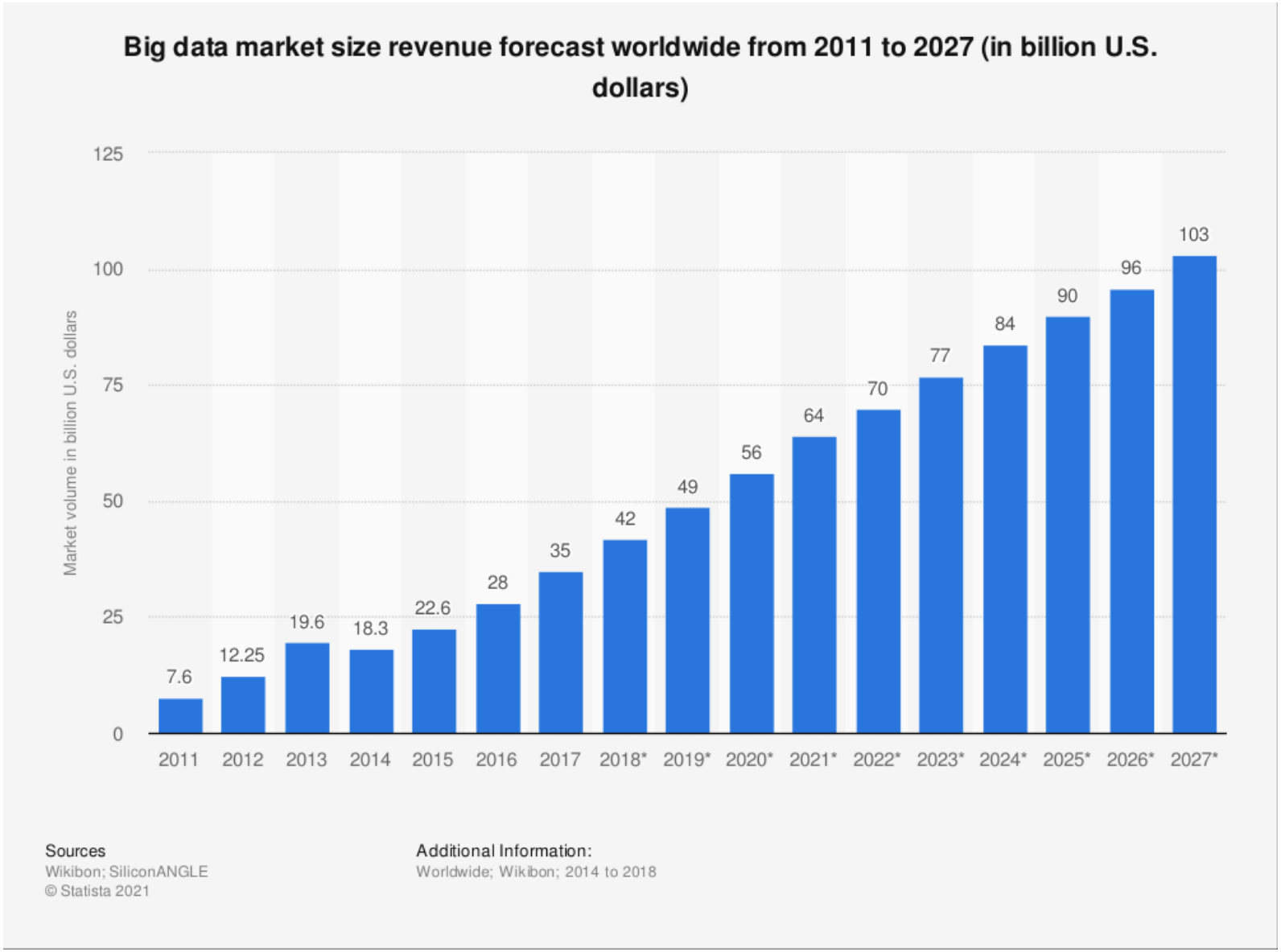
The Future of Big Data

- Today, more than 2.5 exabytes(2.5 billion gigabytes) of data is generated every single day. This is expected to continue growing at a significant rate ……

- Experts now predict that 40 zettabytes of data will be in existence by 2020.

Can you think of running a query on 20,000,000 GB file?

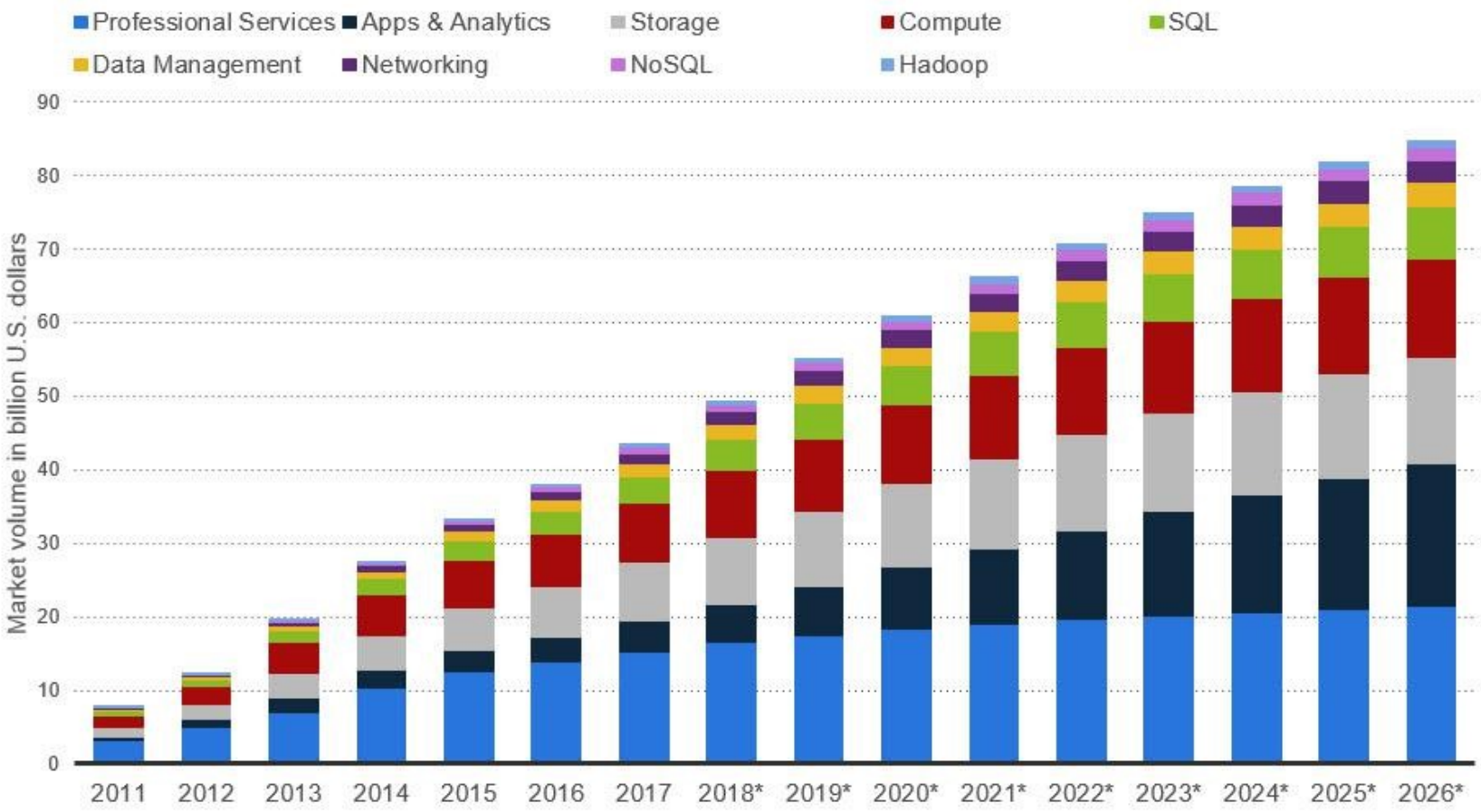
Decimal		
Value	Metric	
1000	kB	kilobyte
1000^2	MB	megabyte
1000^3	GB	**gigabyte**
1000^4	TB	terabyte
1000^5	PB	petabyte
1000^6	EB	exabyte
1000^7	ZB	zettabyte
1000^8	YB	yottabyte



Big Data: changing the way businesses compete and operate

Big Data Market Worldwide Segment Revenue Forecast 2011-2026

Big Data Market Forecast Worldwide from 2011 to 2026, by segment (in billion U.S. dollars)



How Can You Avoid *Big Data*?

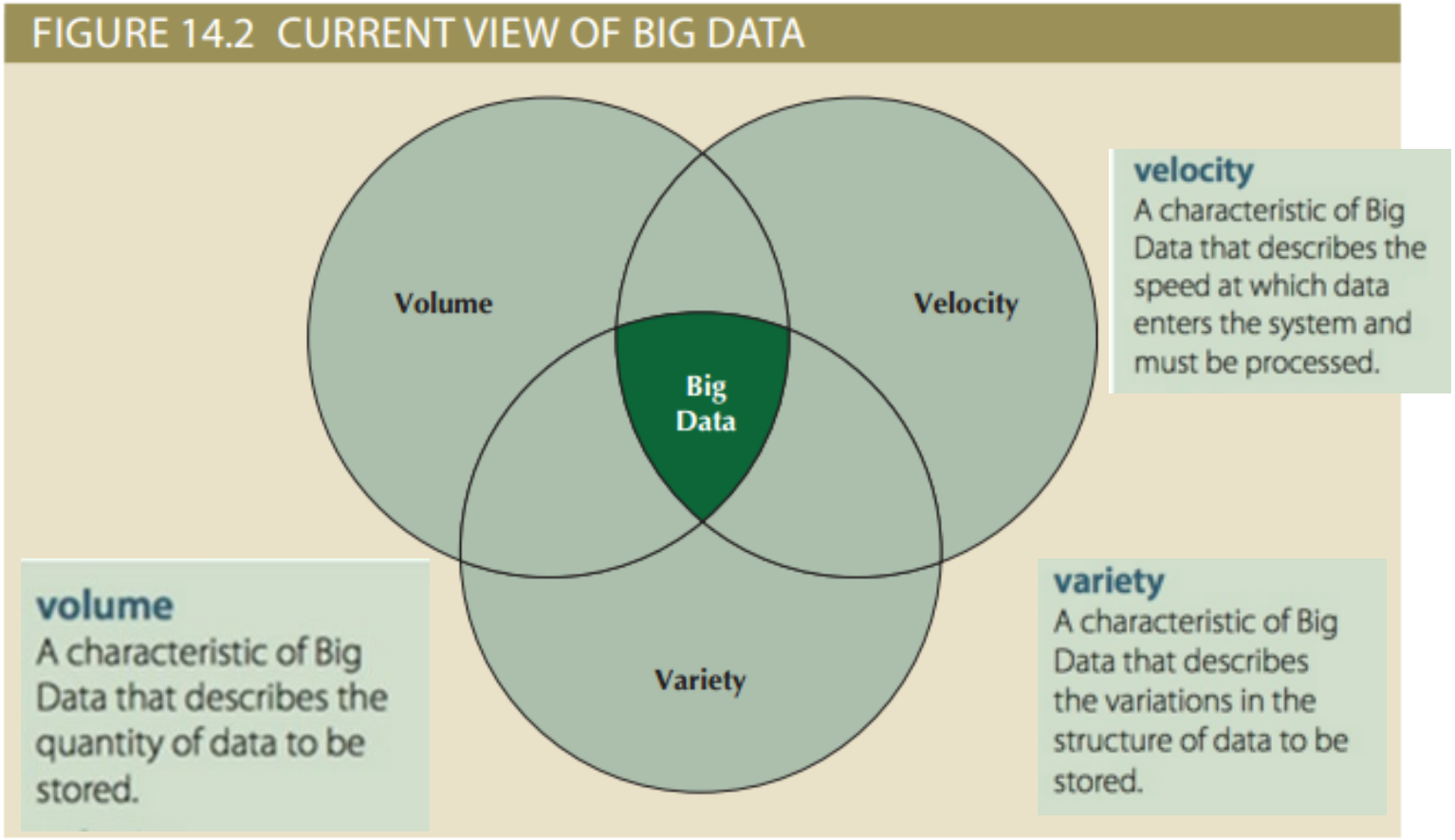
- Pay cash for everything!
- Never go online!
- Don’t use a telephone!
- Never leave your house!

I- What is Big Data

Big Data is used in the singular and refers to a collection of data sets so large and complex, it's impossible to process them with the usual databases and tools.

Because of its size and associated numbers, Big Data is hard to capture, store, search, share, analyze and visualize.

Figure 14.2 – Current View of Big Data



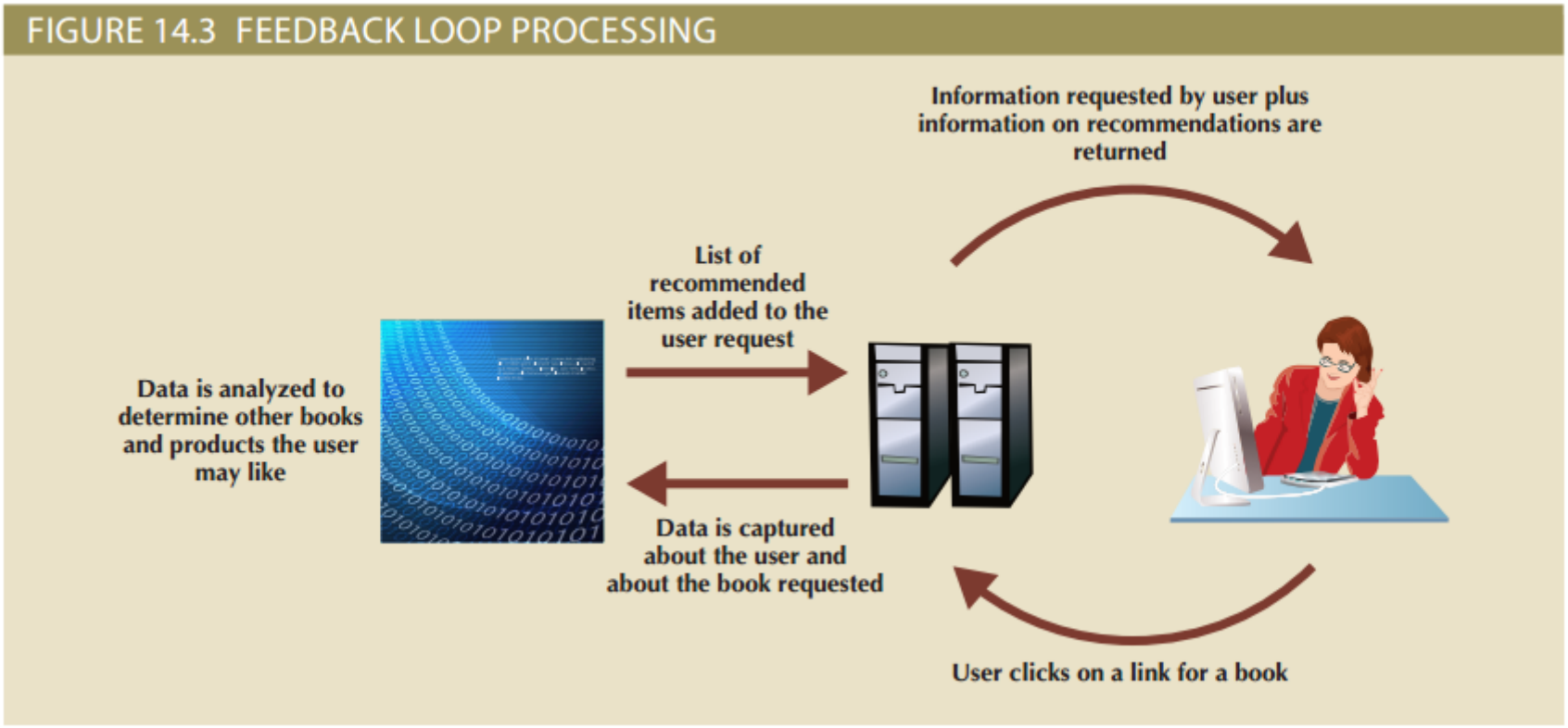
3V-Volumn

- **Volume**: Quantity of data to be stored
  - **Scaling up** is keeping the same number of systems but migrating each one to a larger system. It involves moving to larger and faster system.
    - There are limits to how large and fast a single system can be;
    - The costs increase at a dramatic rate.
    - A variable option as relational databases grows.
  - **Scaling out** means when the workload exceeds server capacity, it is spread out across a number of servers
    - Dominant approach for big data management, and cheaper.
    - For example, eBay enterprise data warehouse, which is over 14PB and spread over hundreds of thousands of nodes.

3V-Velocity

- **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
    - The data must be processed and filtered as it enters the system to determine which data to keep and which data to be discard.
    - Focus on input
  - **Feedback loop processing** refers to the analysis of data to produce actionable results
    - Focus on output.

Figure 14.3 – Feedback Loop Processing



The process of capturing the data, processing it into usable information, and then acting on that information is feedback loop.

3V-Variety

- **Variety**: Variations in the structure of data to be stored
  - **Structured data** fits into a predefined data model
    - Relational Database
  - **Unstructured data** dies not fit into a predefined model
    - Maps, Satellite Images, Emails, Texts, Tweets, Videos…
- Other characteristics:
  - **Variability:** Changes in meaning of data based on context
  - **Veracity:** Trustworthiness of data
  - **Value:** Degree data can be analyzed for meaningful insight
  - **Visualization:** Ability to graphically present data to make it understandable to users

RDBMS and Big Data

- Big data represents a new wave in data management challenges, but it does not mean that relational database technology is going away.
- Structured data that depends on ACID transactions, will always be critical to business operations. Relational databases are still the best way for storing and managing this type of data.
- Relational databases not necessarily best for storing and managing all organizational data

II-Hadoop

- De facto standard for most Big Data storage and processing
- Java-based framework for distributing and processing very large data sets across vast clusters of computers
- Most important components:
  - **Hadoop Distributed File System (HDFS)**: Low-level distributed file processing system that can be used directly for data storage
  - **MapReduce:** Programming model that supports processing large data sets

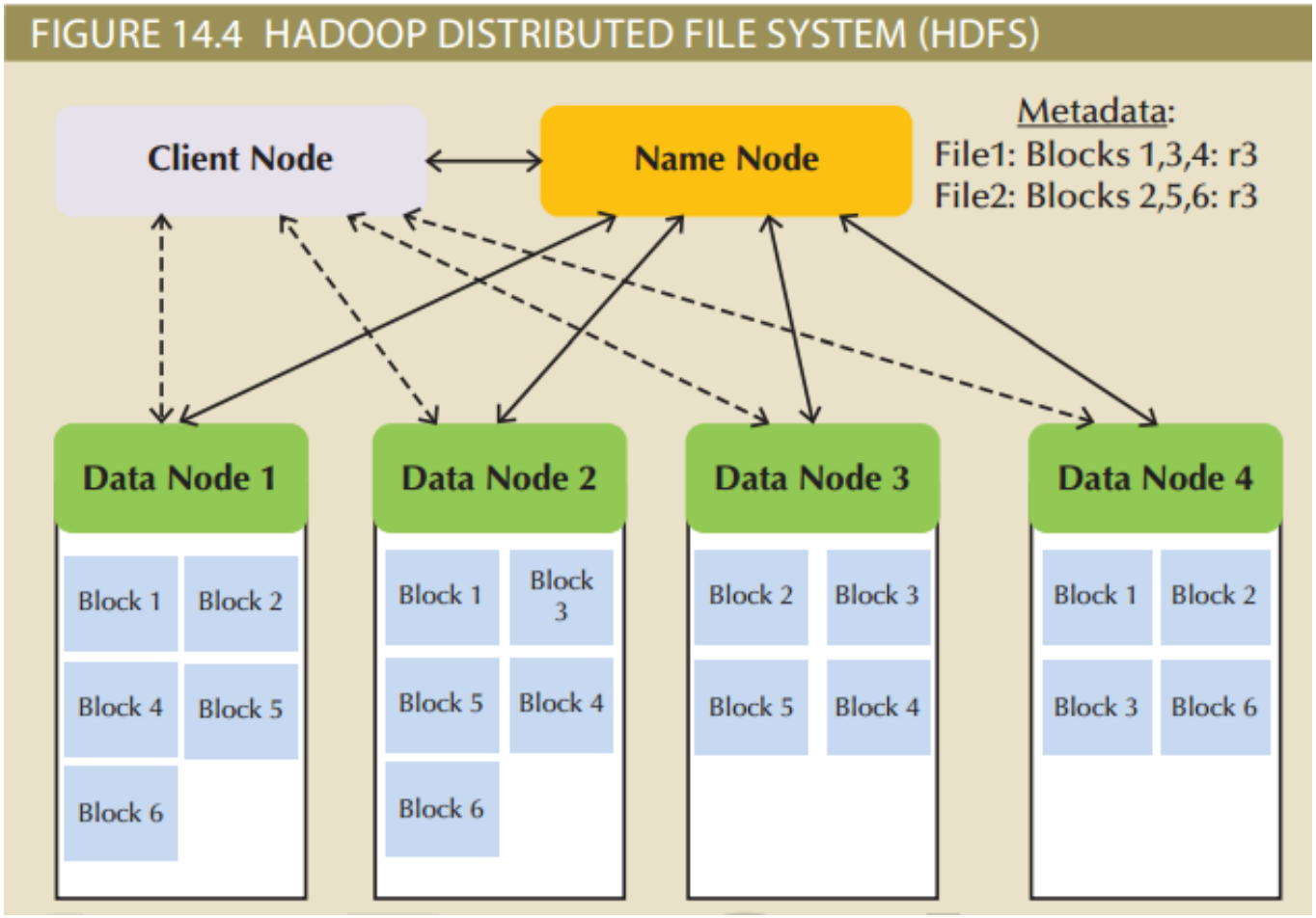
Hadoop Distributed File System (HDFS)

- Approach based on several key assumptions:
  - *High volume* - Default block sizes is 64 MB and can be configured to even larger values
  - *Write-once, read-many* - Model simplifies concurrent issues and improves data throughput
  - *Streaming access* - Hadoop is optimized for batch processing of entire files as a continuous stream of data
  - *Fault tolerance* – HDFS is designed to replicate data across thousands of low lost computers so that when one fails, data is still available from another device

Hadoop Distributed File System (HDFS)

- Uses several types of nodes (computers):
  - **Data node: one or more data node** store the actual file data
  - **Name node: one name node** contains file system metadata
  - **Client node: one client node** makes requests to the file system, makes requests to the file system, either to read files or to write new files
  - Data node communicates with name node by regularly sending **block reports (**data info**)** and **heartbeats (**every 3 seconds, availability**)**

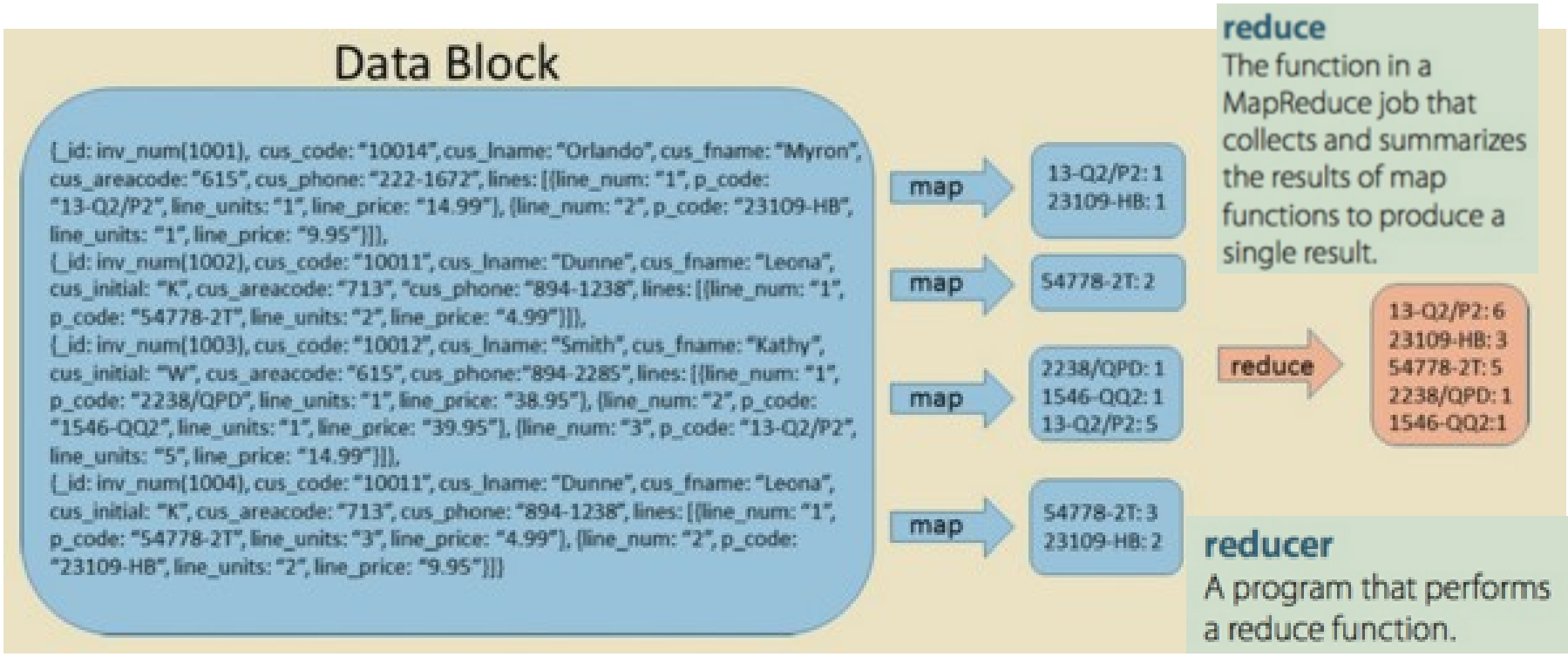
Figure 14.4 – Hadoop Distributed File System (HDFS)



MapReduce

- Framework used to process large data sets across clusters
  - Breaks down complex tasks into smaller subtasks, performing the subtasks and producing a final result
  - **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** summaries results of map function to produce a single result
    - **Reducer** program performs the reduce function
  - **Mapper and Reducer are written as procedure-oriented Java program.**

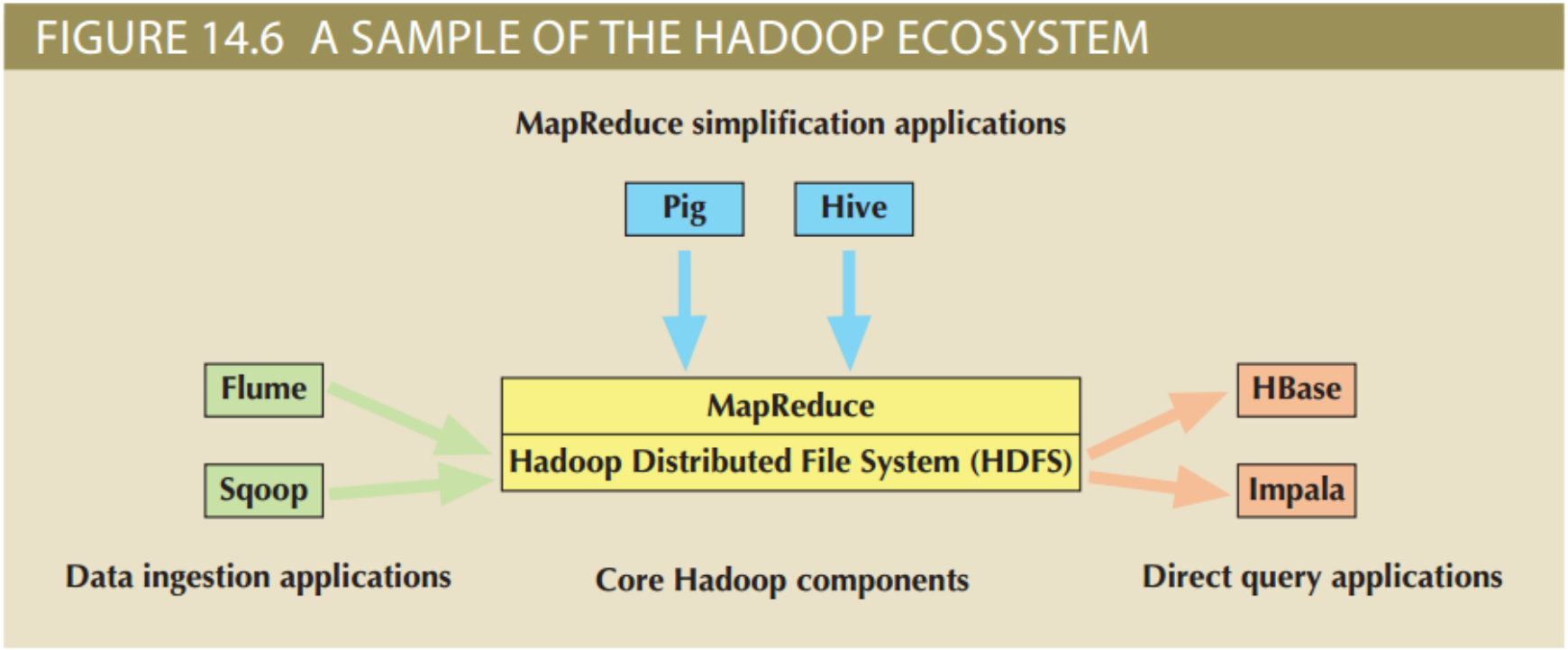
An Example


Data Block
{_id: inv_num(1001), cus_code: "10014", cus_lname: "Orlando", cus_fname: "Myron", cus_areacode: "615", cus_phone: "222-1672", lines: [{line_num: "1", p_code: "13-Q2/P2", line_units: "1", line_price: "14.99"}, {line_num: "2", p_code: "23109-HB", line_units: "1", line_price: "9.95"}]},
{_id: inv_num(1002), cus_code: "10011", cus_lname: "Dunne", cus_fname: "Leona", cus_initial: "K", cus_areacode: "713", "cus_phone: "894-1238", lines: [{line_num: "1", p_code: "54778-2T", line_units: "2", line_price: "4.99"}]},
{_id: inv_num(1003), cus_code: "10012", cus_lname: "Smith", cus_fname: "Kathy", cus_initial: "W", cus_areacode: "615", cus_phone: "894-2285", lines: [{line_num: "1", p_code: "2238/QPD", line_units: "1", line_price: "38.95"}, {line_num: "2", p_code: "1546-QQ2", line_units: "1", line_price: "39.95"}, {line_num: "3", p_code: "13-Q2/P2", line_units: "5", line_price: "14.99"}]},
{_id: inv_num(1004), cus_code: "10011", cus_lname: "Dunne", cus_fname: "Leona", cus_initial: "K", cus_areacode: "713", cus_phone: "894-1238", lines: [{line_num: "1", p_code: "54778-2T", line_units: "3", line_price: "4.99"}, {line_num: "2", p_code: "23109-HB", line_units: "2", line_price: "9.95"}]}
map
13-Q2/P2: 1
23109-HB: 1
map
54778-2T: 2
map
2238/QPD: 1
1546-QQ2: 1
13-Q2/P2: 5
map
54778-2T: 3
23109-HB: 2
reduce
13-Q2/P2: 6
23109-HB: 3
54778-2T: 5
2238/QPD: 1
1546-QQ2:1
reduce
The function in a MapReduce job that collects and summarizes the results of map functions to produce a single result.
reducer
A program that performs a reduce function.

Hadoop Ecosystem

- Map Reduce Simplification Applications:
  - *Hive* is a data warehousing system that sites on top of HDFS and supports its own SQL-like language
  - *Pig* compiles a high-level scripting language (Pig Latin) into MapReduce jobs for executing in Hadoop
- Data Ingestion Applications:
  - *Flume* is a component for ingesting data in Hadoop
  - *Sqoop* is a tool for converting data back and forth between a relational database and the HDFS
- Direct Query Applications:
  - *HBase* is a column-oriented NoSQL database designed to sit on top of the HDFS that quickly processes sparse datasets
  - *Impala* was the first SQL-on-Hadoop application

Figure 14.6 – A Sample of the Hadoop Ecosystem



III-NoSQL

- Name given to non-relational database technologies developed to address Big Data challenges.
  - The term ”NoSQL” was never meant to imply that products in this category should never include support for SQL. In fact, many such products support query languages that mimic SQL in important ways.
  - More recently, some industry observers have tried to interject that “NoSQL” could stand for “not only SQL”.

Characteristics of volume, velocity, and variety to an extent that Makes the big data unsuitable for management by a relational database management system. Google : BigTable; Amazon: Dynamo

Four Categories

- There are hundreds of products that can be considered as being under the broadly defined term NoSQL. Most of these fit roughly into one of four categories:
  - key value data stores,
  - document databases,
  - column-oriented databases,
  - graph databases.

Key-value (KV) databases

- **Key-value (KV) databases** store data as a collection of key-value pairs organized as **buckets** which are the equivalent of tables.
  - ❖No foreign keys, relationship can not be tracked.
  - ❖Extremely fast and scalable for basic processing.

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"

Key-value (KV) databases

- Key values must be unique within a bucket, but they can be duplicated across buckets.
- All operations are based on the bucket plus the key.
- Operation: get, store and delete.
  - Get: to retrieve the value component of the pair.
  - Store: to place a value in a key.
  - Delete: to remove a key-value pair.

Document databases

- **Document databases** store data in key-value pairs in which the value components are tag-encoded documents grouped into logical groups called **collections.** XML, JSON etc.
  - Tags inside the document are accessible to the DBMS, which makes sophisticated querying possible.

FIGURE 14.8 DOCUMENT DATABASE TAGGED FORMAT

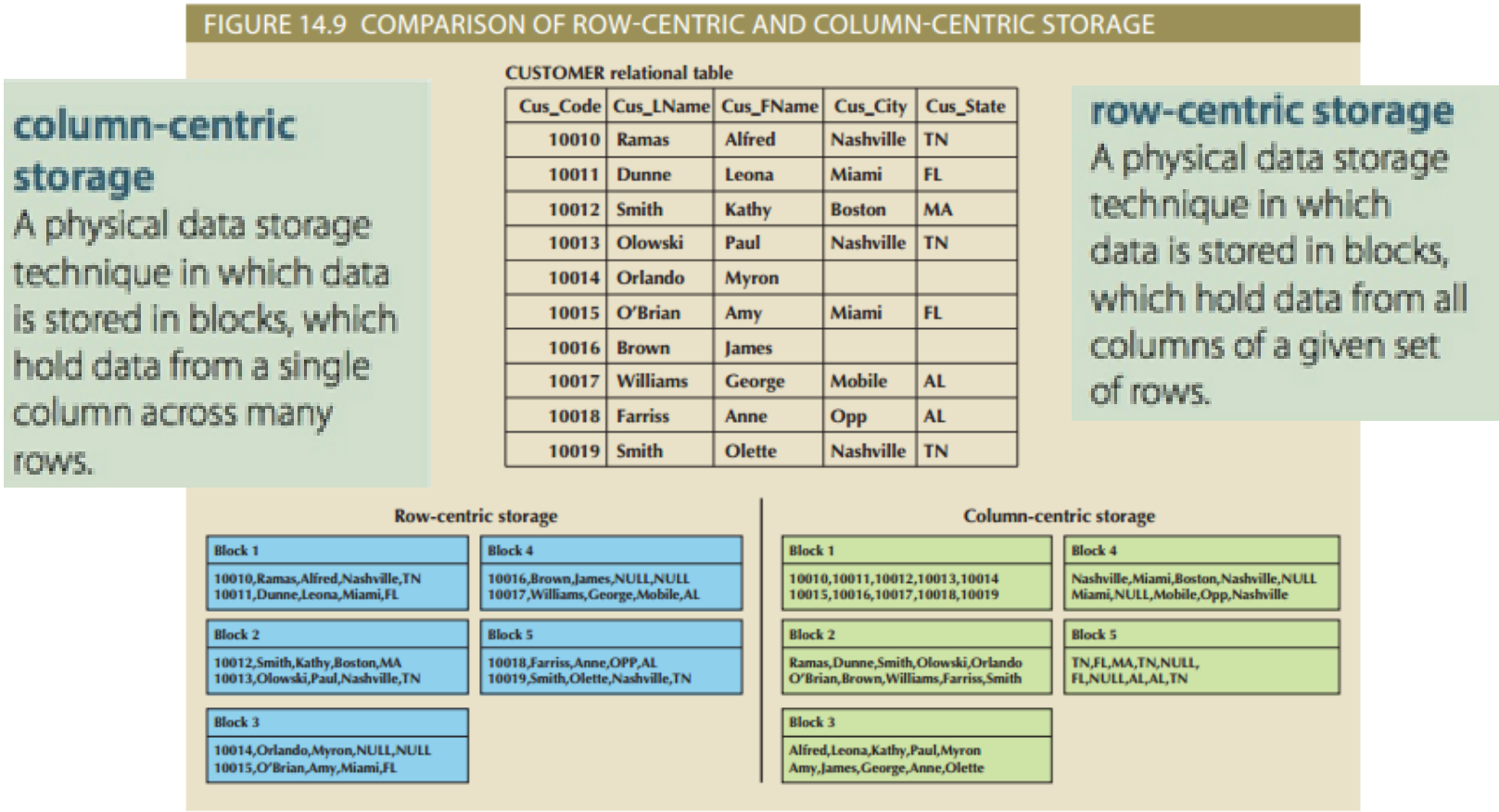
Collection = Customer

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"}

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

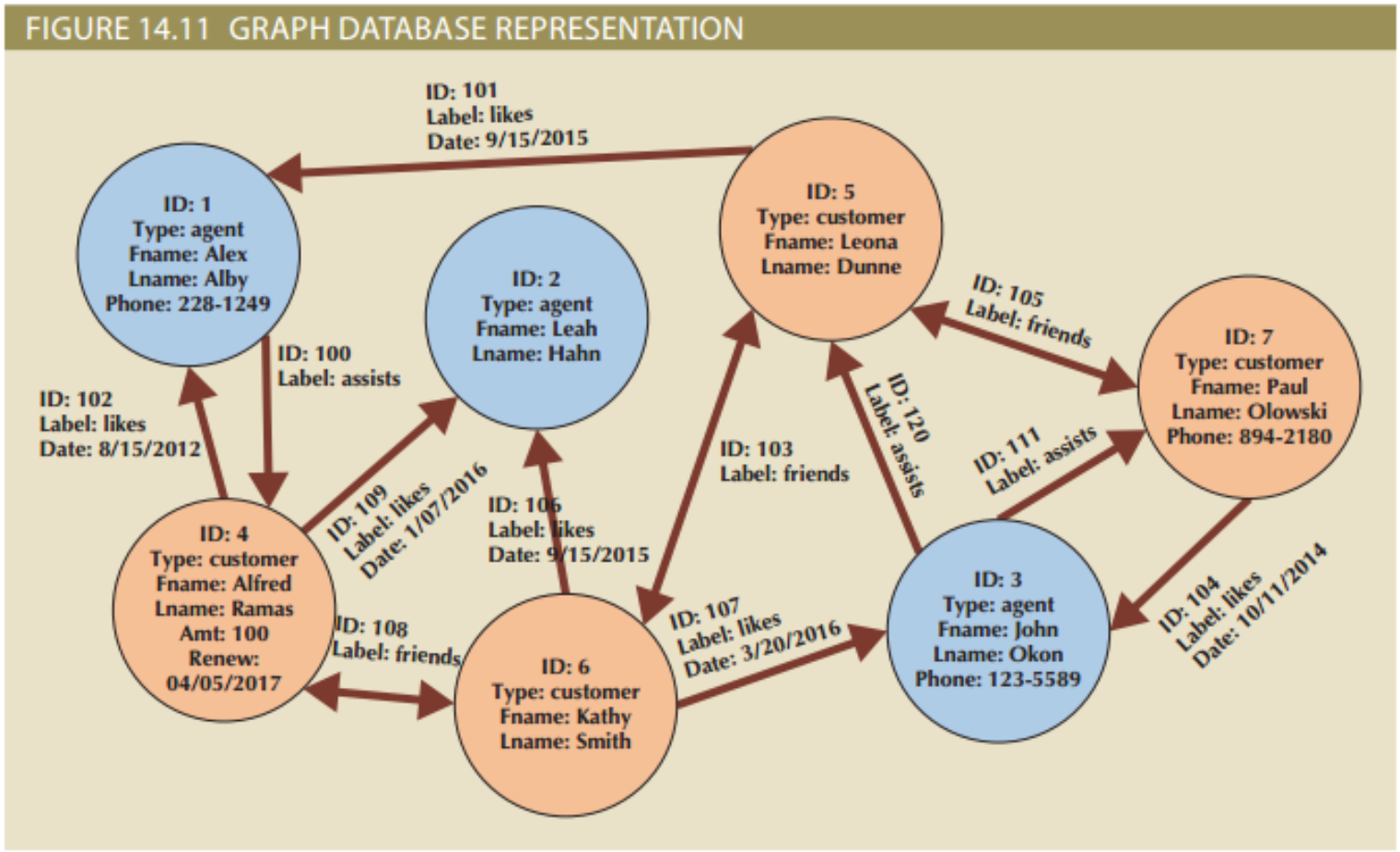
Figure 14.9- Comparison of Row-Centric and Column-Centric Storage



Graph databases

- **Graph databases** store data on relationship-rich data as a collection of nodes and edges
  - Properties are the attributes of a node or edge of interest to a user
  - Traversal is a query in a graph database

Figure 14.10- Graph Database Representation



NewSQL Databases

- Database model that attempts to provide ACID-compliant transactions across a highly distributed infrastructure
  - SQL as the primary interface
  - Latest technologies to appear in the data management area to address Big Data problems
  - No proven track record
  - Have been adopted by relatively few organizations

IV-Data Analytics

- Subset of business intelligence (BI) functionality that encompasses mathematical, statistical, and modeling techniques used to extract knowledge from data
  - Continuous spectrum of knowledge acquisition that goes from discovery to explanation to prediction
- **Explanatory analytics** focuses on discovering and explaining data characteristics based on existing data
- **Predictive analytics** focuses on predicting future data outcomes with a high degree of accuracy

Data Mining

- Focuses on the discovery and explanation stages of knowledge acquisition by:
  - Analyzing massive amounts of data to uncover hidden trends, patterns, and relationships
  - Forming computer models to simulate and explain findings and using them to support decision making
- Can be run in two modes:
  - *Guided* – End-user decides techniques to apply to data
  - *Automated* – End-user sets up the tool to run automatically and the data-mining tool applies multiple techniques to find significant relationships

Figure 14.12- Extracting Knowledge From Data

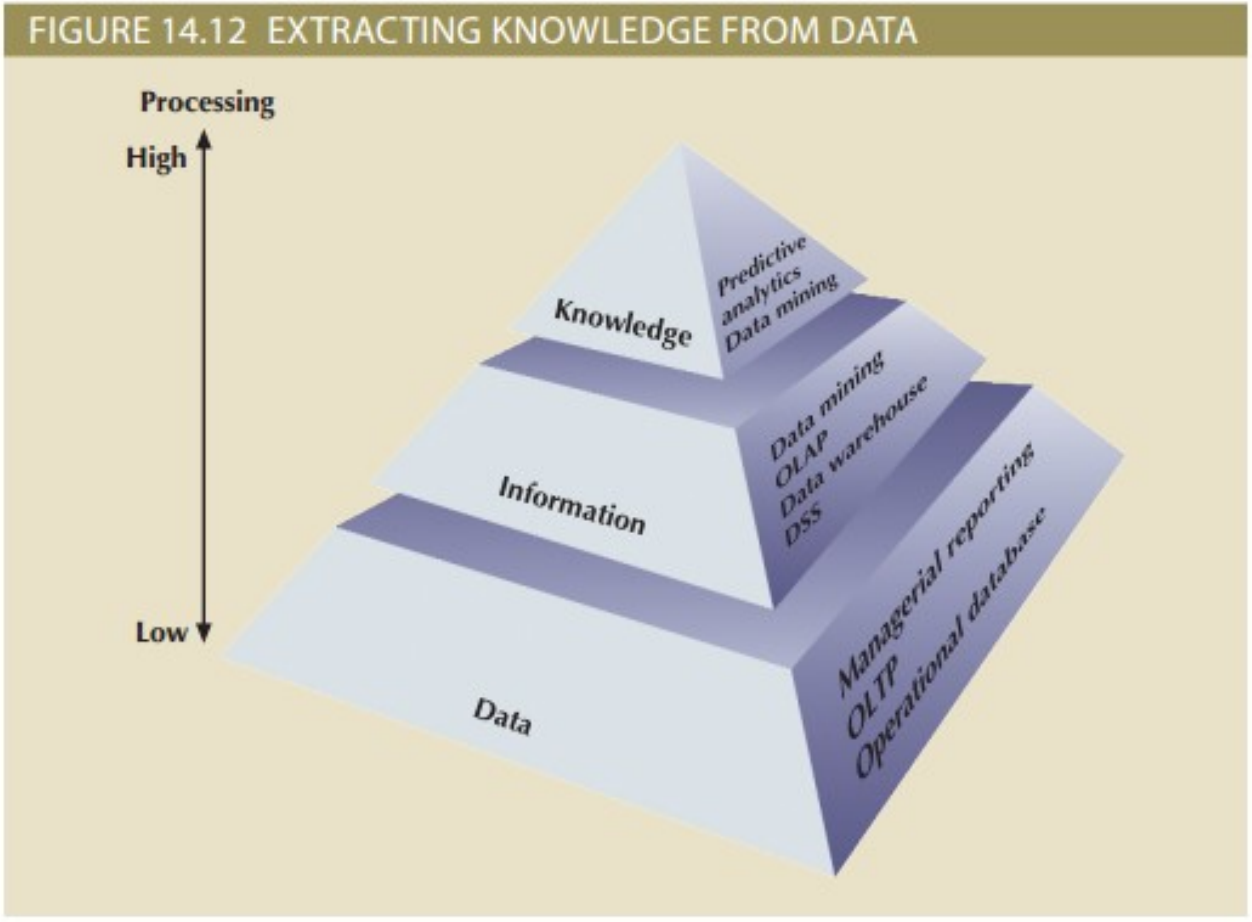
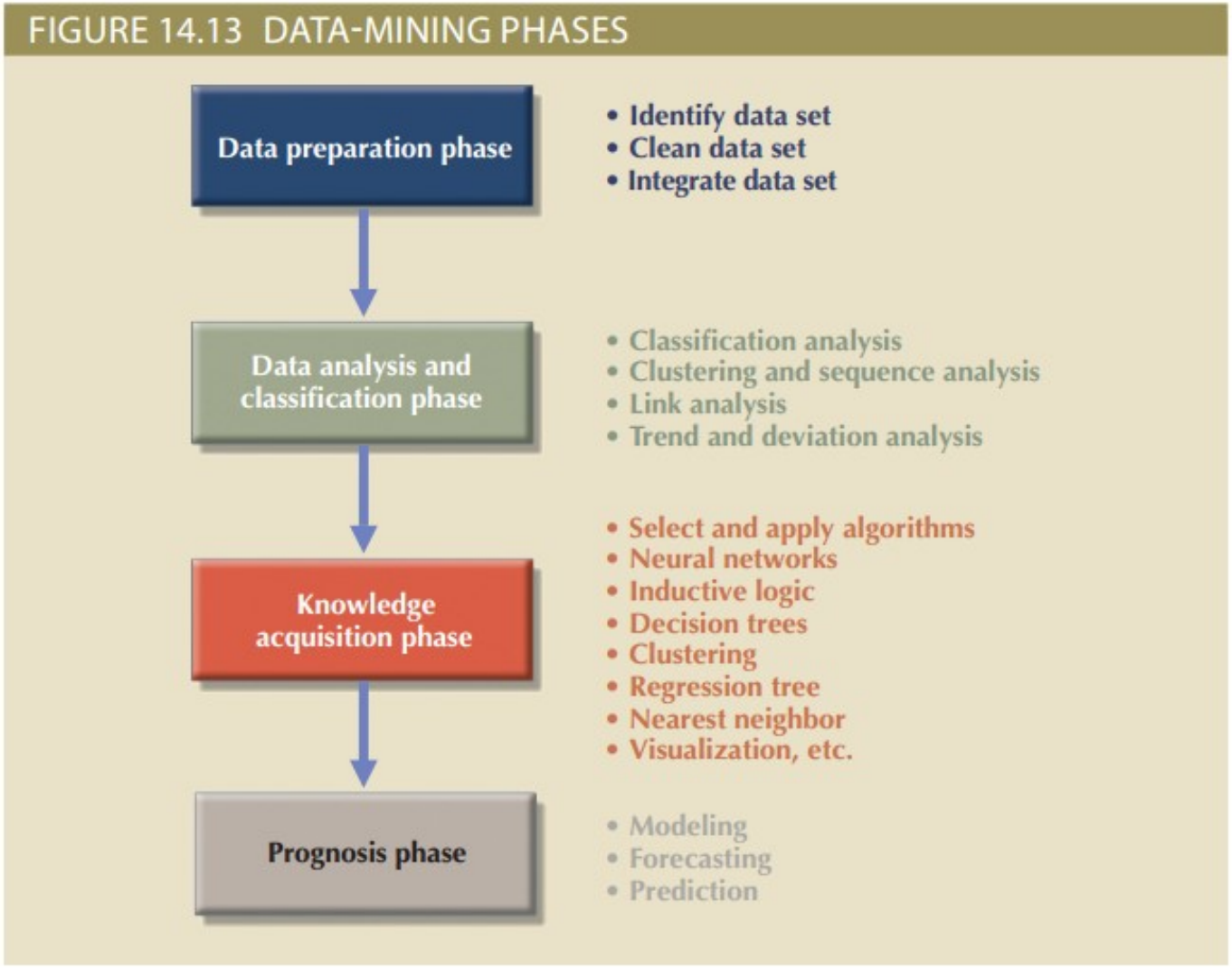


Figure 14.13- Data-Mining Phases



Predictive Analytics

- Refers to the use of advanced mathematical, statistical, and modeling tools to predict future business outcomes with a high degree of accuracy
  - Focuses on creating actionable models to predict future behaviors and events
  - Most BI vendors are dropping the term *data mining* and replacing it with *predictive analytics*
- Models used in customer service, fraud detection, targeted marketing and optimized pricing
  - Can add value in many different ways but needs to be monitored and evaluated to determine return on investment

How is Big Data Used in Practice?

1. Understanding and targeting customers.
2. Understanding and Optimizing Business Processes
3. Personal Quantification and Performance Optimization
4. Improving Healthcare and Public Health
5. Improving Sports Performance
6. Improving Science and Research
7. Optimizing Machine and Device Performance
8. Improving Security and Law Enforcement.
9. Improving and Optimizing Cities and Countries
10. Financial Trading

- http://www.ap-institute.com/big-data-articles/how-is-big-data-used-in-practice-10-use-cases-everyone-should-read.aspx

Summary

- What is the big data? And 3V characteristics: Volume, Velocity, Variety.
- Hadoop: two most important components
- NoSQL: four categories
- Data Analytics