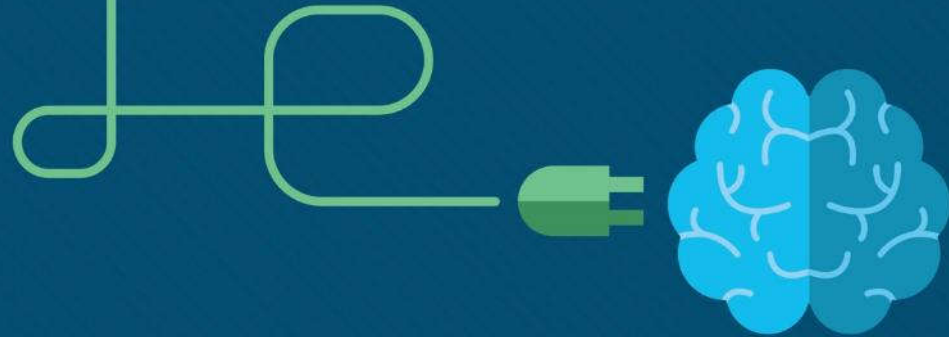




Chapter 2. Big Data Analytics

Anthony Maina



Sections & Objectives

- 2.1 Big Data
 - Explain the concept of Big Data.
 - Describe the sources of Big Data.
 - Explain the challenges and solutions to Big Data storage.
 - Explain how Big Data analytics are used to support Business.

2.1 Big Data

What is Big Data?

What is Big Data?



- **Data** is information that comes from a variety of sources, such as people, pictures, text, sensors, web sites and technology devices.
- Three characteristics that indicate an organization may be dealing with Big Data:
 - A **large amount of data** that increasingly requires more storage space (volume).
 - An amount of data that is **growing exponentially** fast (velocity).
 - Data that is generated in **different formats** (variety).
- Examples of data amounts collected by sensors:
 - One **autonomous car** can generate 4,000 gigabits (Gb) of data per day.
 - One **smart connected home** can produce as much as 1 gigabyte (GB) of information a week.⁴

What is Big Data?

What is Big Data? Attributes of Big Data

- **Volume**
 - Vast quantities of data being generated.
- **Velocity**
 - The rapid rate at which data is being data is being generated
- **Variety**
 - The heterogeneous nature of bid data; it comes from multiple sources e.g. smart phones, social media etc
- **Veracity**
 - Quality and accuracy of the data
- **Value**
 - Ability to generate trends, associations and patterns previously not possible using traditional analytical techniques.

What is Big Data?

Does the Business Generate Big Data?

Activity: Does the business have big data?

Number of Cards: 3
Card Number: 1

An orange grove company has sensors in the trees and on the machines that harvest the oranges. A camera mounted on the harvester takes a close-up picture of the orange every 5 minutes. Live data is sent to the distributor who gets this data from 100 companies. Does the distributor have big data?

☐ Yes

☐ No

Navigation buttons: Previous (left arrow) and Next (right arrow)

What is Big Data?


Large Datasets

- Companies do not necessarily have to generate their own **Big Data**.
- There are sources of free data sets available, ready to be used and analyzed.



What is Big Data?

Lab – Database Search

 Cisco Networking Academy®Mind Wide Open®

Lab – Exploring a Large Dataset (Instructor Version)

Instructor Note: Red font color or gray highlights indicate text that appears in the instructor copy only.

Objectives

Explore a sample dataset to view the power of Big Data.

Background / Scenario

Before data can become meaningful information, it needs to be processed.

Required Resources

- PC with access to the Internet

Step 1: Locate a large, free, searchable database.

- Click [here](#) to access the United States Department of Agriculture Statistics Service database.
- Select: Quick Stats (Searchable Database)
Notice the status in the top right hand corner. How many records are currently in the database?

Answers will vary but it should be a value greater 34.7 million

Step 2: Select Categories.

- From the categories select:

Program: Census
Sector: Animals & Products
Group: Poultry
Commodity: Ducks
Category: Inventory
Data Item: Ducks – Inventory
Geographic Level: State
State: Alaska

Next, select: Get Data

What was the inventory of ducks in Alaska in 2012?

226

- Select the Back button and change the state to Hawaii. Ensure that the year is still 2012.
What was the inventory of ducks in Hawaii in 2012?



Where is Big Data Stored?

What are the Challenges of Big Data?



- IBM's Big Data estimates conclude that "each day we create 2.5 quintillion bytes of data".
- Five major storage problems with Big Data:
 - Management
 - Security
 - Redundancy
 - Analytics
 - Access

Where is Big Data Stored?

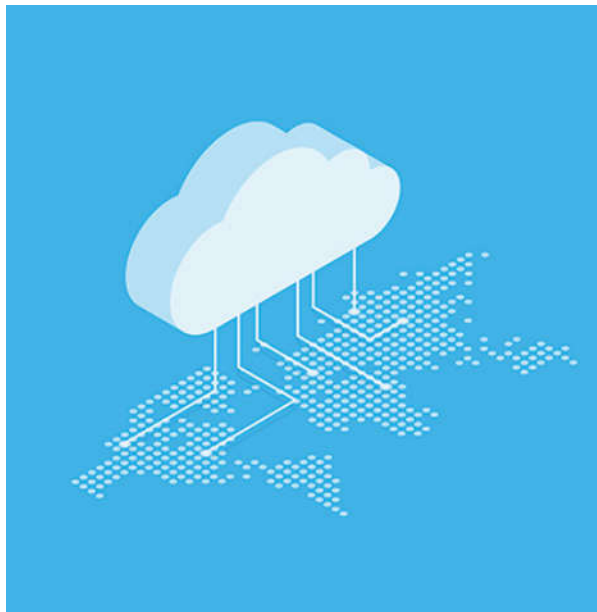
Where Can We Store Big Data?

- Big data is typically stored on multiple servers, in **data centers**.
- **Fog computing** utilizes end-user clients or “edge” devices to do a substantial amount of the pre-processing and storage.
 - Data from that pre-processed analysis can be fed back into the companies’ systems to modify processes if required.
 - Communications to and from the servers and devices is quicker and requires less bandwidth than constantly going out to the cloud.



Where is Big Data Stored?

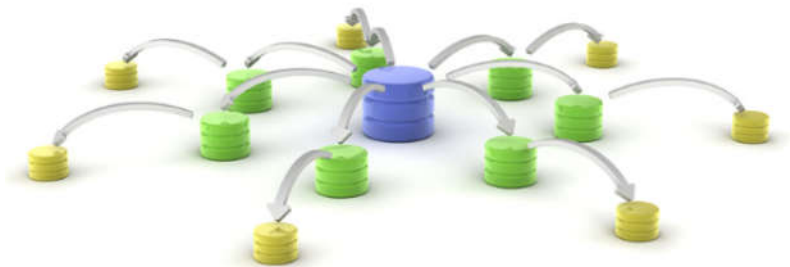
The Cloud and Cloud Computing



- The **cloud** is a collection of data centers or groups of connected servers.
- Cloud services for individuals include:
 - **Storage of data**, such as pictures, music, movies, and emails.
 - **Access many applications** instead of downloading onto local device.
 - **Access data and applications** *anywhere, anytime*, and on *any device*.
- Cloud Services for an Enterprise include:
 - Access to organizational data anywhere and at any time.
 - Streamlines the IT operations of an organization.
 - Eliminates or reduces the need for onsite IT equipment, maintenance, and management.
 - Reduces cost for equipment, energy, physical plant requirements, and personnel training needs.

Where is Big Data Stored?

Distributed Processing



- **Distributed data processing** takes the large volume of data and breaks it into smaller pieces.
- These smaller pieces are distributed in many locations to be processed by many computers.
- Each computer in the distributed architecture analyzes its part of the Big Data picture (horizontal scaling).
- **Hadoop** (open source software utilities that facilitate using a network of many computers to solve problems involving massive amounts of data and computation) was created to deal with these Big Data volumes. It has two main features that has made it the industry standard:
 - **Scalability** - Larger cluster sizes improve performance and provide higher data processing capabilities.
 - **Fault tolerance** – Hadoop automatically replicates data across clusters.

Supporting Business with Big Data

Why Do Businesses Analyze Data?

- **Data analytics** allows businesses to better understand the impact of their products and services, adjust their methods and goals, and provide their customers with better products faster.
- **Value** comes from two primary types of processed data, *transactional* and *analytical*.
- Transactional information is captured and processed as events happen.
 - Used to analyze daily sales reports and production schedules to determine how much inventory to carry.
- Analytical information supports managerial analysis tasks like determining whether the organization should build a new manufacturing plant.



Supporting Business with Big Data

Sources of Information



- **Data** originates from sensors and anything that has been scanned, entered, and released to the Internet.
- Collected data can be categorized as structured or unstructured.
- Structured data is created by applications that use “fixed” format input such as spreadsheets. May need to be manipulated into a common format such as CSV.
- Unstructured data is generated in a “freeform” style such as audio, video, web pages, and tweets.
- Examples of tools to prepare unstructured data for processing are:
 - “Web scraping” tools automatically extract data from HTML pages.
 - RESTful application program interfaces (APIs).

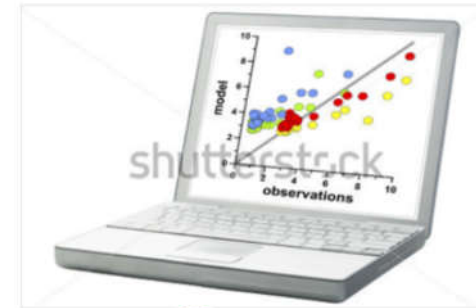
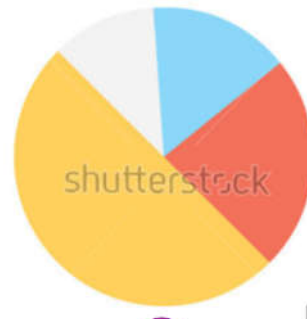
Supporting Business with Big Data

Data Visualization

- **Data mining** is the process of turning raw data into meaningful information.
- The mined data must be analyzed and presented to managers and decision makers.
- Determining the best visualization tools to use will vary based on the following:
 - Number of variables
 - Number of data points in each variable
 - Is the data representing a timeline
 - Items require comparisons
- Popular charts include line, column, bar, pie, and scatter.



Supporting Business with Big Data Chart Types



Supporting Business with Big Data

Analyzing Big Data for Effective Use in Business



- **Data analysis** is the process of inspecting, cleaning, transforming, and modeling data to uncover useful information.
- Having a strategy helps a business determine the type of analysis required and the best tool to do the analysis.
- Tools and applications range from using an Excel spreadsheet or Google Analytics for small to medium data samples, to the applications dedicated to manipulating and analyzing really big datasets.
- Examples include Knime, OpenRefine, Orange, and RapidMiner.

Supporting Business with Big Data

Excel lab: Forecasting



Cisco Networking Academy®

Mind Wide Open™

Lab – Using Excel to Forecast (Instructor Version – Optional Lab)

Instructor Note: Red font color or gray highlights indicate text that appears in the instructor copy only. Optional activities are designed to enhance understanding and/or to provide additional practice.

Objectives

Part 1: Input the Data

Part 2: Execute a Data Forecast

Background / Scenario

Forecasting is a way of predicting values in the future based on data. Managers want data instantly in order to make decisions and they rely on techniques such as forecasting to make those decisions. With big data, volumes of data are produced instantaneously. This presents a challenge to collect and process this data in real time.

This lab is very basic and is designed to just show you how forecasting is performed in Microsoft Excel. You will be inputting a set of weekly grades and using the forecast feature to see what grades are predicted for the next few weeks.

Note: The Forecast menu option is available in the 2016 version of Excel. If you do not have this version, the formula is provided. You might do better copying the formula from the lab than inputting it.

Note: If you do not have the Forecast icon available in the Data menu option, but have the 2016 version of Excel, select the **File** menu option > **Options** > **Add-Ins** > **Go** > enable the checkbox beside **Analysis ToolPak** > **OK**. If you return to the **File** > **Options** > **Add-Ins** window, you should see the Analysis ToolPak in the top section where the active add-ins list.

2.2 Chapter Summary

Chapter Summary

Summary

- Three characteristics of Big Data:
 - large amount of data that increasingly requires more storage space (volume)
 - growing exponentially fast (velocity)
 - generated in different formats (variety)
- Fog computing utilizes end-user clients or “edge” devices to do pre-processing and storage.
 - Designed to keep the data closer to the source for pre-processing.
- The cloud is a collection of data centers or groups of connected servers giving anywhere, anytime access to software, storage, and services using a browser interface.
 - Provide increased data storage and reduce the need for onsite IT equipment, maintenance, and management.
- Distributed data processing takes large volumes of data from a source and breaks it into smaller pieces and distributes to many locations to be processed.
 - Each computer in the distributed architecture analyzes its part of the Big Data picture.

Chapter Summary

Summary (Cont.)

- Businesses gain **value** by *collecting and analyzing data to understand the impact of their products and services, adjust their methods and goals, and provide their customers with better products faster.*
- **Structured data** is created by applications that use “fixed” format input such as spreadsheets or medical forms.
- **Unstructured data** is generated in a “freeform” style such as audio, video, web pages, and tweets.
- Both forms of data need to be manipulated into a common format to be analyzed.
- **Data mining** is the process of turning raw data into meaningful information by discovering patterns and relationships in large data sets.
- **Data visualization** is the process of taking the analyzed data and using charts such as line, column, bar, pie, or scatter to present meaningful information.

