Capítulo 3: Secciones y objetivos

- 3.1 Datos masivos
  - Explique el concepto de datos masivos.
    - Describir las fuentes de datos masivos.
    - Explique los desafíos y las soluciones para el almacenamiento de datos masivos.
    - Explique cómo el análisis de datos masivos se utilizan para apoyar las actividades empresariales.

© 2016 Cisco y/o sus filiales. Todos los derechos reservados. Información confidencial de Cisco.

3.1 Datos masivos

© 2016 Cisco y/o sus filiales. Todos los derechos reservados. Información confidencial de Cisco.

¿Qué son los datos masivos o Big Data?

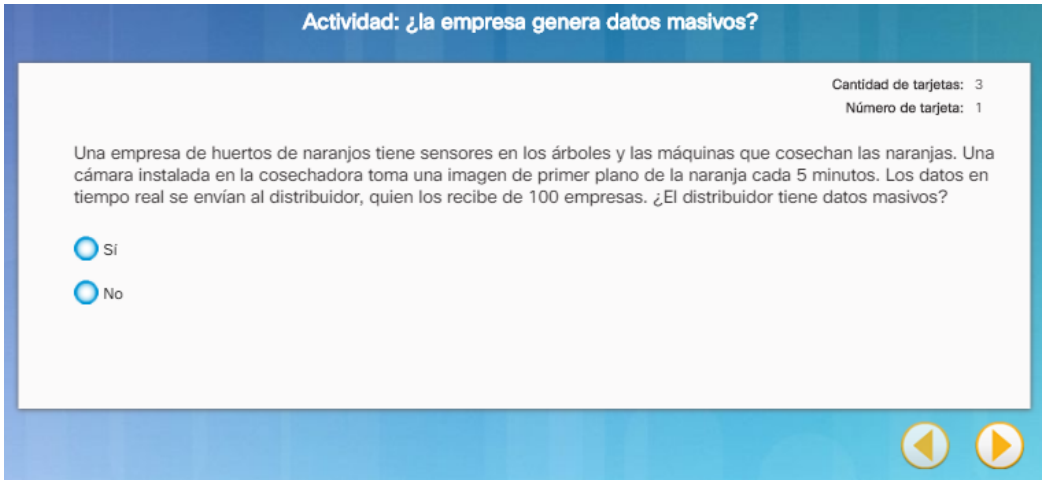
¿Qué son los datos masivos o Big Data?

- Los datos son la información que proviene de una variedad de fuentes, como personas, imágenes, texto, sensores, sitios web y dispositivos de tecnología.
- Hay tres características que indican que una organización puede estar haciendo frente a datos masivos:
  - Una gran cantidad de datos que requiere cada vez más espacio de almacenamiento (volumen).
  - Una cantidad de datos que crece exponencialmente rápido (velocidad).
  - Datos que se generan en diferentes formatos (variedad).
- Ejemplos de volúmenes de datos recopilados por los sensores:
  - Un automóvil autónomo puede generar 4000 gigabits (Gb) de datos por día.
  - Un hogar inteligente conectado puede producir 1 gigabyte (GB) de información de la semana.

© 2016 Cisco y/o sus filiales. Todos los derechos reservados. Información confidencial de Cisco.

¿Qué son los datos masivos o Big Data?

¿La empresa genera datos masivos?



© 2016 Cisco y/o sus filiales. Todos los derechos reservados. Información confidencial de Cisco.

¿Qué son los datos masivos o Big Data?

¿La empresa genera datos masivos?

© 2016 Cisco y/o sus filiales. Todos los derechos reservados. Información confidencial de Cisco.

¿Qué son los datos masivos o Big Data?

¿La empresa genera datos masivos?

© 2016 Cisco y/o sus filiales. Todos los derechos reservados. Información confidencial de Cisco.

¿Qué son los datos masivos o Big Data?

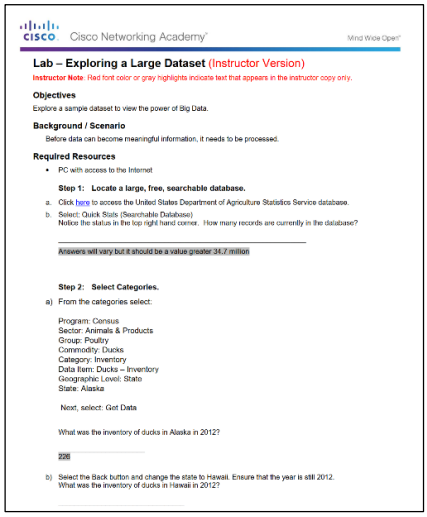
Grandes conjuntos de datos

- Las empresas no necesariamente tienen que generar sus propios datos masivos.
- Hay fuentes de conjuntos de datos gratuitos disponibles y listas para usar y analizar.

© 2016 Cisco y/o sus filiales. Todos los derechos reservados. Información confidencial de Cisco.

¿Qué son los datos masivos o Big Data?

Práctica de laboratorio: búsqueda en base de datos



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.

a. Click here to access the United States Department of Agriculture Statistics Service database.

b. 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.

a) 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

b) 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?

© 2016 Cisco y/o sus filiales. Todos los derechos reservados. Información confidencial de Cisco.

¿Dónde se almacenan los datos masivos?

¿Cuáles son los desafíos de los datos masivos?

- Los cálculos de datos masivos de IBM concluyen que “cada día creamos 2,5 trillones de bytes de datos”.
- Hay cinco problemas de magnitud en cuanto al almacenamiento con los datos masivos:
  - Administración
  - Seguridad
  - Redundancia
  - Análisis
  - Acceso

© 2016 Cisco y/o sus filiales. Todos los derechos reservados. Información confidencial de Cisco.

¿Dónde se almacenan los datos masivos?

¿Dónde podemos almacenar los datos masivos?

- Por lo general, los datos masivos se almacenan en varios servidores en centros de datos.
- La computación en la niebla utiliza dispositivos “perimetrales” o de clientes de usuarios finales para ejecutar gran parte del procesamiento previo y almacenamiento.
  - Los datos adquiridos a partir de ese análisis de procesamiento previo pueden introducirse en los sistemas de las empresas para modificar los procesos, de ser necesario.
  - Las comunicaciones hacia y desde los servidores y dispositivos es más rápida y requiere menos ancho de banda que lo que supondría constantemente recurrir a la nube.

© 2016 Cisco y/o sus filiales. Todos los derechos reservados. Información confidencial de Cisco.

¿Dónde se almacenan los datos masivos?

La nube y la computación en la nube

- La nube es una colección de centros de datos o grupos de servidores conectados.
- Los servicios en la nube para las personas incluyen lo siguiente:
  - Almacenamiento de datos, tales como imágenes, música, películas y correos electrónicos.
  - Acceso a muchas aplicaciones en lugar de descargar en el dispositivo local.
  - Acceso a datos y aplicaciones en cualquier lugar, en cualquier momento y en cualquier dispositivo.
- Los servicios en la nube para las empresas incluyen lo siguiente:
  - Acceso a los datos de la organización en cualquier momento y en cualquier lugar.
  - Optimiza las operaciones de TI de una organización.
  - Elimina o reduce la necesidad de equipos, mantenimiento, y administración de TI en el sitio.
  - Reduce el costo de necesidades de equipos, energía, requisitos físicos de la planta y la capacitación del personal.

© 2016 Cisco y/o sus filiales. Todos los derechos reservados. Información confidencial de Cisco.

Soporte de empresas con datos masivos

¿Por qué las empresas analizan datos?

- El análisis de datos permite que las empresas comprendan mejor el impacto de sus productos y servicios, ajusten sus métodos y objetivos, y proporcionen a sus clientes mejores productos más rápido.
- Los valores provienen de los dos tipos de datos procesados principales: transaccionales y analíticos.
- La información transaccional se captura y se procesa a medida que se producen eventos.
  - Se utiliza para analizar informes de ventas y planes de fabricación diarios a fin de determinar cuánto inventario transportar.
- La información analítica permite que se realicen tareas de análisis a nivel gerencial, como determinar si la organización debe instalar una nueva planta de fabricación.

© 2016 Cisco y/o sus filiales. Todos los derechos reservados. Información confidencial de Cisco.

Soporte de empresas con datos masivos

Fuentes de información

- Los datos se originan a partir de sensores y cualquier elemento que se haya explorado, introducido y publicado en Internet.
- Los datos recopilados se pueden clasificar como estructurados o no estructurados.
- Los datos estructurados son creados por aplicaciones que utilizan la entrada de formato "fijo", como las hojas de cálculo. Es posible que se deban manipular en un formato común como CSV.
- Los datos no estructurados se generan en un estilo de “forma libre”, como audio, video, páginas web y tweets.
- Entre los ejemplos de herramientas para preparar datos no estructurados para el procesamiento se encuentran:
  - Las herramientas que «raspan la red» (web scraping) extraen datos de páginas HTML automáticamente.
  - Interfaces del programa de aplicación (API) RESTful.

© 2016 Cisco y/o sus filiales. Todos los derechos reservados. Información confidencial de Cisco.

Soporte para empresas con datos masivos

Visualización de datos

- La minería de datos es el proceso por el cual los datos sin procesar se transforman en información significativa.
- Los datos sometidos a minería de datos se deben analizar y presentar a los administradores y las personas responsables de tomar decisiones.
- La determinación de las mejores herramientas de visualización que se deben usar variará en función de lo siguiente:
  - Cantidad de variables
  - Cantidad de puntos de datos en cada variable
  - Representan los datos una línea de tiempo
  - Los elementos requieren comparaciones
- Entre los gráficos populares se incluyen gráficos circulares, de líneas, de columnas, de barras y de dispersión.

© 2016 Cisco y/o sus filiales. Todos los derechos reservados. Información confidencial de Cisco. 16

Soporte para empresas con datos masivos

Tipos de gráficos

sco. 17

Soporte de empresas con datos masivos

Análisis de datos masivos para el uso eficaz en la empresa

- El análisis de datos es el proceso de inspección, limpieza, transformación y creación de modelos de datos para descubrir información útil.
- Tener una estrategia permite que una empresa determine el tipo de análisis requerido y la mejor herramienta para realizar el análisis.
- Las herramientas y aplicaciones varían desde el uso de una hoja de cálculo de Excel o Google Analytics para muestras de datos de pequeñas a medianas, hasta las aplicaciones dedicadas a la manipulación y al análisis de conjuntos de datos realmente masivos.
- Entre los ejemplos se incluyen a Knime, OpenRefine, Orange y RapidMiner.

© 2016 Cisco y/o sus filiales. Todos los derechos reservados. Información confidencial de Cisco.

3.2 Resumen del capítulo

© 2016 Cisco y/o sus filiales. Todos los derechos reservados. Información confidencial de Cisco.

Resumen del capítulo

Resumen

- Las tres características de los datos masivos son las siguientes:
  - gran cantidad de datos que requiere cada vez más espacio de almacenamiento (volumen)
  - rápido crecimiento exponencial (velocidad)
  - generados en diferentes formatos (variedad)
- La computación en la niebla utiliza dispositivos “perimetrales” o de clientes de usuarios finales para ejecutar el procesamiento previo y almacenamiento.
  - Se diseñó con el fin de mantener los datos más cerca del origen para su procesamiento previo.
- La nube es un conjunto de centros de datos o grupos de servidores conectados que ofrecen acceso a software, almacenamiento y servicios, en cualquier lugar y en cualquier momento, mediante una interfaz de navegador.
  - Proporciona un aumento del almacenamiento de datos y reduce la necesidad de equipos de TI en el sitio, mantenimiento y administración.
- El procesamiento de datos distribuidos toma grandes volúmenes de datos de una fuente y los divide en partes más pequeñas, y los distribuye en muchas ubicaciones para que se procesen.
  - Cada computadora de la arquitectura distribuida analiza su parte del total de datos masivos.

© 2016 Cisco y/o sus filiales. Todos los derechos reservados. Información confidencial de Cisco.

Resumen del capítulo

Resumen (continuación)

- Las empresas obtienen valor mediante la recopilación y el análisis de datos para comprender el impacto de los productos y servicios, ajustar los métodos y objetivos, y proporcionar a sus clientes mejores productos con mayor rapidez.
- Los datos estructurados se crean mediante aplicaciones que utilizan entradas de formato “fijo”, como hojas de cálculo o formularios médicos.
- Los datos no estructurados se generan en un estilo de “forma libre”, como audio, video, páginas web y tweets.
- Ambas formas de datos deben manipularse en un formato común para su análisis.
- La minería de datos es el proceso que se utiliza para convertir los datos sin procesar en información significativa al detectar patrones y relaciones en los grandes conjuntos de datos.
- La visualización de datos es el proceso que se utiliza para captar los datos analizados y usar gráficos como línea, columna, barra, diagrama o dispersión para presentar la información importante.

© 2016 Cisco y/o sus filiales. Todos los derechos reservados. Información confidencial de Cisco.


CISCO