Capítulo 1: Datos e Internet de las cosas

Big Data & Analytics

Cisco | Networking Academy®
Mind Wide Open™

Capítulo 1: Secciones y objetivos

- 1.1 Valor de los datos
  - Demuestre el valor de los datos.
- 1.2 Datos y datos masivos
  - Explique el concepto de datos masivos.
- 1.3 Administración de datos masivos
  - Demuestre el conocimiento de los enfoques de administración de datos en IdT.

1.1 El valor de los datos

Cisco | Networking Academy®
Mind Wide Open™

Los datos en un mundo conectado

- El valor de los datos
  - La cantidad de datos que se guardan y analizan está ampliándose.
  - La variedad de datos alcanzará nuevas áreas.
  - La transformación digital repercutirá en tres elementos principales de nuestras vidas: lo empresarial, lo social y lo ambiental.
- ¿Qué son los datos?
  - Los datos pueden ser muchas cosas.
    - Palabras en un libro, un artículo, o un blog
    - Contenido de una hoja de cálculo o de una base de datos
    - Imágenes o video
    - Un flujo de mediciones de un dispositivo
  - Los datos útiles son información.
  - Determine la cantidad de datos que deben recopilarse.
  - No todos los datos pueden usarse como están.
  - El análisis de datos proporciona información útil o tendencias.

Los datos crecen exponencialmente

- Cálculo de crecimiento exponencial
  - Dos tipos: lineal y exponencial
  - El crecimiento exponencial es más drástico.
- Crecimiento de los datos
  - Los datos actuales están creciendo exponencialmente.
  - Crecimiento de los datos de ejemplo previsto para 2015 a 2020 de Visual Networking Index (VNI) de Cisco
    - El tráfico de datos móviles de los consumidores llegará a 26,1 exabytes por mes en el 2020.
    - El tráfico IP alcanzará 194,4 exabytes por mes en el 2020.
    - El 64 % de todo el tráfico de Internet global cruzará redes de distribución de contenido en el 2020.

El crecimiento de los datos nos cambia la vida

- Impacto del crecimiento de datos
  - Está impulsado por la proliferación de los dispositivos de IdC
  - Esto incluye sensores, terminales inalámbricos y redes móviles
- Ejemplo de negocio: Kaggle
  - Kaggle es una plataforma que conecta las empresas y otras organizaciones que tienen preguntas sobre sus datos con la persona que sabe encontrar las respuestas.
  - Kaggle ejecuta competencias en línea.
- Ejemplo social: DrivenData
  - Ofrece prácticas puntos en ciencia de datos y crowdsourcing a las personas y las organizaciones que estén dirigiendo estos desafíos
- Ejemplo ambiental: Cambio climático
  - Asociación entre NASA y Cisco – Planetary Skin
  - Plataforma global de colaboración en línea de supervisión
  - Captura, recopila, analiza e informa datos en condiciones ambientales

1.2 Datos y datos masivos

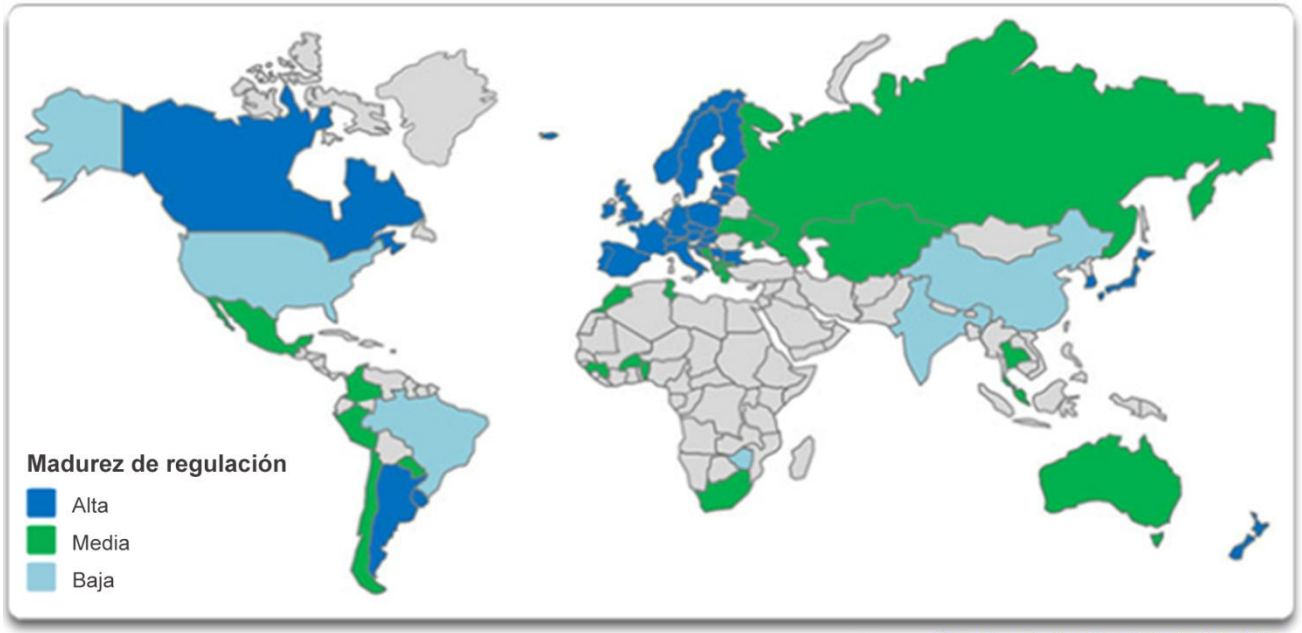
Cisco | Networking Academy®
Mind Wide Open™

Dé donde provienen los datos masivos

- Definición de datos masivos
  - Datos que son tan grandes, rápidos o complejos que se vuelve imposible guardarlos, procesarlos y analizarlos con aplicaciones tradicionales de almacenamiento y análisis de datos.
- Características de los datos masivos
  - Las cuatro grandes V de los datos masivos: volumen, velocidad, variedad y veracidad.
  - Volumen: cantidad de datos
  - Velocidad: tasa a la que se generan los datos
  - Variedad: tipo de datos
  - Veracidad: evitar que los datos imprecisos estropeen un conjunto de datos
- Qué cantidad de datos hace que se consideren datos masivos
  - Paul Zikopaulos de IBM estableció que se necesitan entre 200 y 600 terabytes para que los datos se consideren masivos

Datos abiertos y datos privados

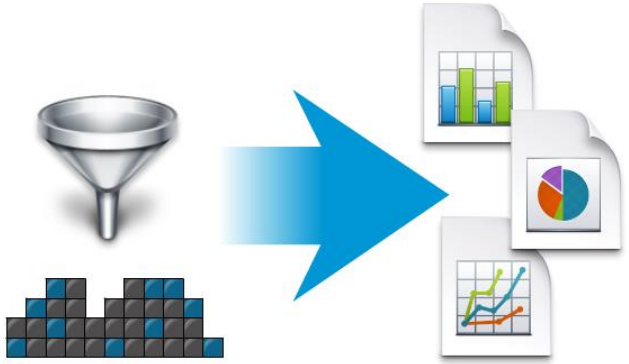
- Datos abiertos
  - La Open Knowledge Foundation describe los datos abiertos como "todo contenido, información o dato que el público puede utilizar, reutilizar y redistribuir de forma gratuita y sin restricción legal, tecnológica ni social".
- Datos privados
  - Datos relacionados con una expectativa de privacidad y regulados por un país o un gobierno determinados



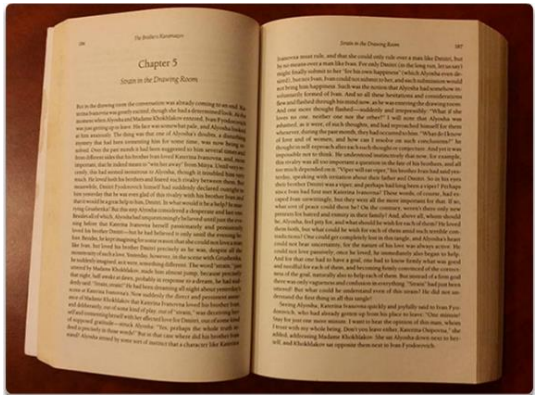
Foro Económico Mundial, abril de 2014

Datos estructurados y no estructurados

- Datos estructurados
  - Son los que se introducen y se mantienen en campos fijos dentro de un archivo o un registro.
  - Se introducen, clasifican, consultan y analizan fácilmente
  - Bases de datos u hojas de cálculo relacionales

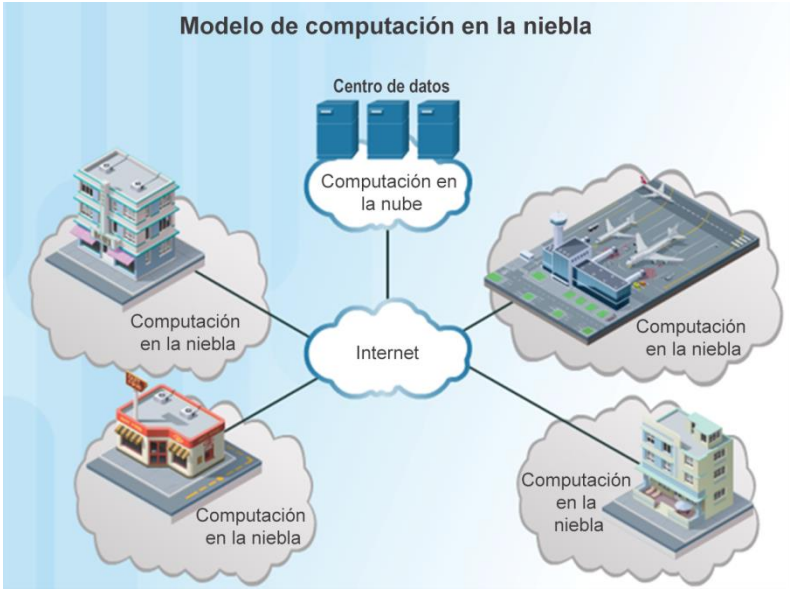


- Datos no estructurados
  - Carecen de organización
  - Datos sin procesar
  - Contenido formado por fotos, audio, video, sitios web, blogs, libros, diarios, informes técnicos, presentaciones de PowerPoint, artículos, correo electrónico, wikis, documentos de procesador de textos y texto en general.



Datos almacenados y datos en movimiento

- Datos almacenados
  - Datos almacenados en una ubicación física, como una unidad de disco duro en un servidor o dentro de un centro de datos
  - Siguen el flujo de análisis tradicional: **Almacenar > Analizar > Notificar > Actuar**
- Datos en movimiento
  - Son datos dinámicos que requieren procesamiento en tiempo real antes de que se transformen en irrelevantes u obsoletos
  - El análisis y la acción se producen más temprano que tarde
  - El flujo de análisis de los datos es: **Analizar > Actuar > Notificar > Almacenar**



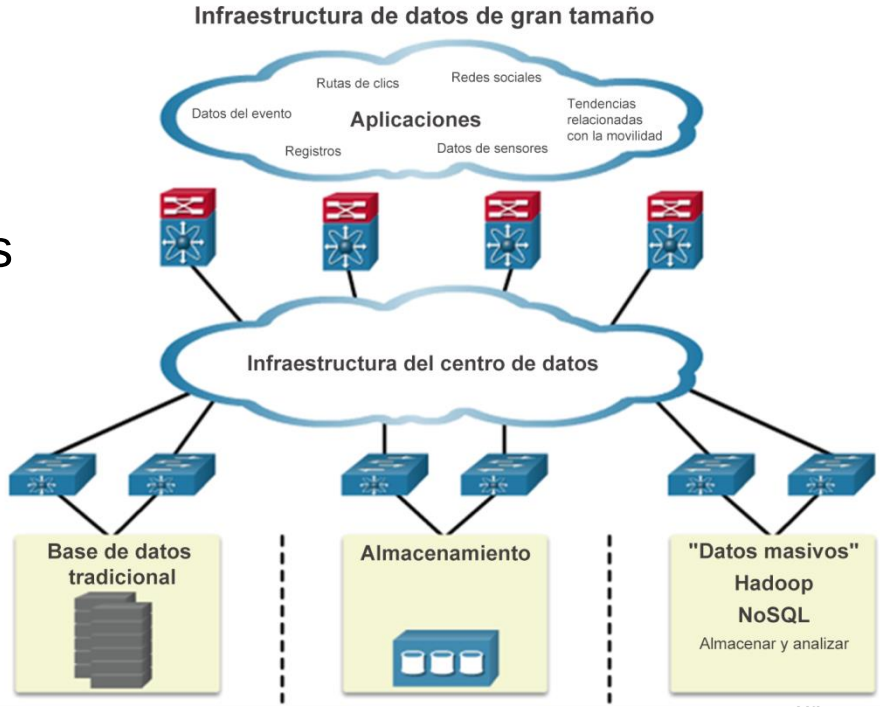
1.3 La evolución hacia los datos masivos

Cisco | Networking Academy®
Mind Wide Open™

La evolución hacia los datos masivos

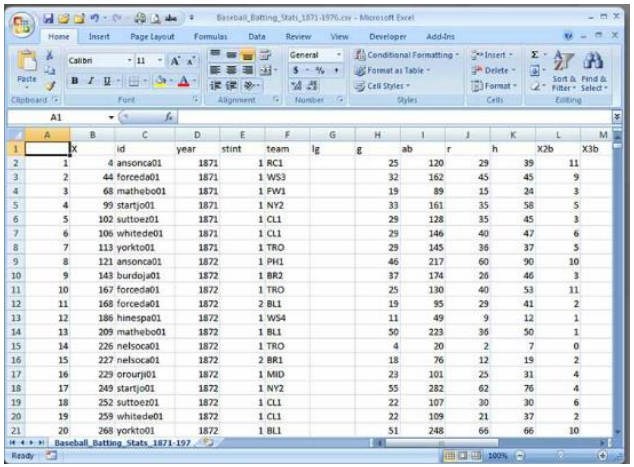
Infraestructura tradicional a la de datos masivos

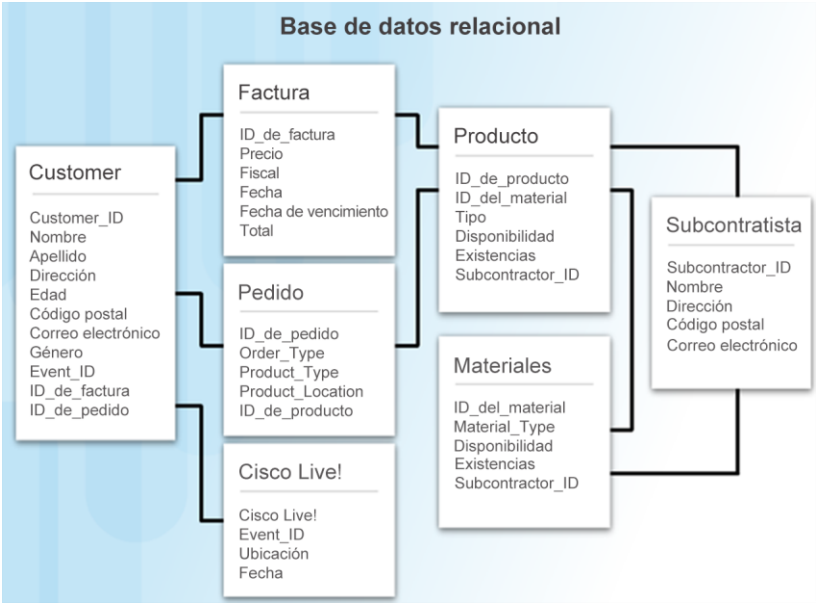
- Servidores de bases de datos y herramientas de procesamiento de datos tradicionales
- Sistemas de datos distribuidos en recursos independientes conectados horizontalmente, a fin de lograr la escalabilidad necesaria para el procesamiento eficiente de conjuntos de datos extensos
- Soluciones de computación in situ y en la nube



Tecnologías de administración de datos básicas

- Base de datos de archivo plano: almacena los registros en un solo archivo sin estructura jerárquica, como una hoja de cálculo
- Base de datos relacional: captura las relaciones entre diferentes conjuntos de datos, lo que crea información más útil





Tecnologías de administración de datos básicas

- El sistema de administración de bases de datos relacional es la tecnología de bases de datos predominante, imbatible durante más de 30 años
- El análisis de datos masivos se vuelve cada vez más difícil de administrar con un sistema de administración de bases de datos relacionales (RDBMS)
- Hadoop Distributed File System (HDFS) es un sistema de archivos distribuido con tolerancia a fallas que se creó para manejar grandes volúmenes de datos
- Estructura de base de datos NoSQL creada para acelerar y simplificar el diseño de bases de datos Satisface las demandas de las aplicaciones web
- SQLite: motor de bases de datos SQL simple y fácil de usar; es la base de datos más usada del mundo

1.4 Resumen

Cisco | Networking Academy®
Mind Wide Open™

Resumen

- Los datos pueden ser palabras en un libro, el contenido de una hoja de cálculo, fotos, archivos o flujos de mediciones enviadas desde un dispositivo.
- El crecimiento de datos puede ser lineal y exponencial. El exponencial implica un incremento más drástico.
- Las cuatro V de los datos masivos son: volumen, velocidad, variedad y veracidad.
- Los datos estructurados son los que se introducen en campos fijos dentro de un archivo o un registro. Los datos no estructurados no tienen un esquema fijo que identifique el tipo de datos.
- Los datos almacenados son datos estáticos almacenados en una ubicación física.
- Los datos en movimiento analizan y extraen valor de los datos antes de almacenarse.
- Una base de datos de archivo plano es como una hoja de cálculo que almacena los registros en un solo archivo sin estructura jerárquica.
- Una base de datos relacional captura las relaciones entre distintos conjuntos de datos y puede proporcionar información más útil.

Resumen

- Hadoop se creó para manejar volúmenes de datos masivos.
- Una base de datos NoSQL almacena y accede a los datos de manera diferente que las bases de datos relacionales.
- SQLite es un motor de bases de datos SQL simple y fácil de usar; es la base de datos más usada del mundo.