Tipo : Lectura
Capítulo : Introducción al Big Data

I. OBJETIVO

Ampliar sus conocimientos sobre Big Data.

II. LECTURAS COMPLEMENTARIAS

Se sugiere revisar los siguientes enlaces para profundizar en los conceptos tratados en el presente capítulo.

a) Términos y conceptos de Big Data

Big Data: ¿En qué consiste? Su importancia, desafíos y gobernabilidad (Fuente:

PowerData)

Big Data es un término que describe el gran volumen de datos, tanto estructurados como no estructurados, que inundan los negocios cada día. Pero no es la cantidad de datos lo que es importante. Lo que importa con el Big Data es lo que las organizaciones hacen con los datos. Big Data se puede analizar para obtener ideas que conduzcan a mejores decisiones y movimientos de negocios estratégicos.

1. ¿Qué es Big Data?

Cuando hablamos de Big Data nos referimos a conjuntos de datos o combinaciones de conjuntos de datos cuyo tamaño (volumen), complejidad (variabilidad) y velocidad de crecimiento (velocidad) dificultan su captura, gestión, procesamiento o análisis mediante tecnologías y herramientas convencionales, tales como bases de datos relacionales y estadísticas convencionales o paquetes de visualización, dentro del tiempo necesario para que sean útiles.

Aunque el tamaño utilizado para determinar si un conjunto de datos determinado se considera Big Data no está firmemente definido y sigue cambiando con el tiempo, la mayoría de los analistas y profesionales actualmente se refieren a conjuntos de datos que van desde 30-50 Terabytes a varios Petabytes.

La naturaleza compleja del Big Data se debe principalmente a la naturaleza no estructurada de gran parte de los datos generados por las tecnologías modernas, como los web logs, la identificación por radiofrecuencia (RFID), los sensores incorporados en dispositivos, la maquinaria, los vehículos, las búsquedas en Internet, las redes sociales como Facebook, computadoras portátiles, teléfonos inteligentes y otros teléfonos móviles, dispositivos GPS y registros de centros de llamadas.

En la mayoría de los casos, con el fin de utilizar eficazmente el Big Data, debe combinarse con datos estructurados (normalmente de una base de datos relacional) de una aplicación comercial más convencional, como un ERP (Enterprise Resource Planning) o un CRM (Customer Relationship Management).

2. ¿Por qué el Big Data es tan importante?

Lo que hace que Big Data sea tan útil para muchas empresas es el hecho de que proporciona respuestas a muchas preguntas que las empresas ni siquiera sabían que tenían. En otras palabras, proporciona un punto de referencia. Con una cantidad tan grande de información, los datos pueden ser moldeados o probados de cualquier manera que la empresa considere adecuada. Al hacerlo, las organizaciones son capaces de identificar los problemas de una forma más comprensible.

La recopilación de grandes cantidades de datos y la búsqueda de tendencias dentro de los datos permiten que las empresas se muevan mucho más rápidamente, sin problemas y de manera eficiente. También les permite eliminar las áreas problemáticas antes de que los problemas acaben con sus beneficios o su reputación.

El análisis de Big Data ayuda a las organizaciones a aprovechar sus datos y utilizarlos para identificar nuevas oportunidades. Eso, a su vez, conduce a movimientos de negocios más inteligentes, operaciones más eficientes, mayores ganancias y clientes más felices. Las empresas con más éxito con Big Data consiguen valor de las siguientes formas:

- Reducción de coste. Las grandes tecnologías de datos, como Hadoop y el análisis basado en la nube, aportan importantes ventajas en términos de costes cuando se trata de almacenar grandes cantidades de datos, además de identificar maneras más eficientes de hacer negocios.
- Más rápido, mejor toma de decisiones. Con la velocidad de Hadoop y la analítica en memoria, combinada con la capacidad de analizar nuevas fuentes de datos, las empresas pueden analizar la información inmediatamente y tomar decisiones basadas en lo que han aprendido.
- Nuevos productos y servicios. Con la capacidad de medir las necesidades de los clientes y la satisfacción a través de análisis viene el poder de dar a los clientes lo que quieren. Con la analítica de Big Data, más empresas están creando nuevos productos para satisfacer las necesidades de los clientes.

Por ejemplo:

- Turismo: Mantener felices a los clientes es clave para la industria del turismo, pero la satisfacción del cliente puede ser difícil de medir, especialmente en el momento oportuno. Resorts y casinos, por ejemplo, sólo tienen una pequeña oportunidad de dar la vuelta a una mala experiencia de cliente. El análisis de Big data ofrece a estas empresas la capacidad de recopilar datos de los clientes, aplicar análisis e identificar inmediatamente posibles problemas antes de que sea demasiado tarde.
- Cuidado de la salud: El Big Data aparece en grandes cantidades en la industria sanitaria. Los registros de pacientes, planes de salud, información de seguros y otros tipos de información pueden ser difíciles de manejar, pero están llenos de información clave una vez que se aplican las analíticas. Es por eso que la tecnología de análisis de datos es tan importante para el cuidado de la salud. Al analizar grandes cantidades de información - tanto estructurada como no estructurada - rápidamente, se pueden proporcionar diagnósticos u opciones de tratamiento casi de inmediato.
- Administración: La administración se encuentra ante un gran desafío, mantener la calidad y la productividad con unos presupuestos ajustados. Esto es particularmente problemático con lo relacionado con la justicia. La tecnología agiliza las operaciones mientras que da a la administración una visión más holística de la actividad.
- Retail: El servicio al cliente ha evolucionado en los últimos años, ya que los compradores más inteligentes esperan que los minoristas comprendan exactamente lo que necesitan, cuando lo necesitan. El Big Data ayuda a los minoristas a satisfacer esas demandas. Armados con cantidades interminables

de datos de programas de fidelización de clientes, hábitos de compra y otras fuentes, los minoristas no sólo tienen una comprensión profunda de sus clientes, sino que también pueden predecir tendencias, recomendar nuevos productos y aumentar la rentabilidad.

- Empresas manufactureras: Estas despliegan sensores en sus productos para recibir datos de telemetría. A veces esto se utiliza para ofrecer servicios de comunicaciones, seguridad y navegación. Ésta telemetría también revela patrones de uso, tasas de fracaso y otras oportunidades de mejora de productos que pueden reducir los costos de desarrollo y montaje.
- Publicidad: La proliferación de teléfonos inteligentes y otros dispositivos GPS ofrece a los anunciantes la oportunidad de dirigirse a los consumidores cuando están cerca de una tienda, una cafetería o un restaurante. Esto abre nuevos ingresos para los proveedores de servicios y ofrece a muchas empresas la oportunidad de conseguir nuevos prospectos.
- Otros ejemplos del uso efectivo de Big Data existen en las siguientes áreas:
  - Uso de registros de logs de TI para mejorar la resolución de problemas de TI, así como la detección de infracciones de seguridad, velocidad, eficacia y prevención de sucesos futuros.
  - Uso de la voluminosa información histórica de un Call Center de forma rápida, con el fin de mejorar la interacción con el cliente y aumentar su satisfacción.
  - Uso de contenido de medios sociales para mejorar y comprender más rápidamente el sentimiento del cliente y mejorar los productos, los servicios y la interacción con el cliente.
  - Detección y prevención de fraudes en cualquier industria que procese transacciones financieras online, tales como compras, actividades bancarias, inversiones, seguros y atención médica.
  - Uso de información de transacciones de mercados financieros para evaluar más rápidamente el riesgo y tomar medidas correctivas.

3. Desafíos de la calidad de datos en Big Data

Las especiales características del Big Data hacen que su calidad de datos se enfrente a múltiples desafíos. Se trata de las conocidas como 5 Vs: Volumen, Velocidad, Variedad, Veracidad y Valor, que definen la problemática del Big Data.

Estas 5 características del Big Data provocan que las empresas tengan problemas para extraer datos reales y de alta calidad, de conjuntos de datos tan masivos, cambiantes y complicados.

Hasta la llegada del Big Data, mediante ETL podíamos cargar la información estructurada que teníamos almacenada en nuestro sistema ERP y CRM, por ejemplo. Pero ahora, podemos cargar información adicional que ya no se encuentra dentro de los dominios de la empresa: comentarios o likes en redes sociales, resultados de campañas de marketing, datos estadísticos de terceros, etc. Todos estos datos nos ofrecen información que nos ayuda a saber si nuestros productos o servicios están funcionando bien o por el contrario están teniendo problemas.

Algunos desafíos a los que se enfrenta la calidad de datos de Big Data son:

1. Muchas fuentes y tipos de datos

Con tantas fuentes, tipos de datos y estructuras complejas, la dificultad de integración de datos aumenta.

Las fuentes de datos de Big Data son muy amplias:

- Datos de internet y móviles.
- Datos de Internet de las Cosas.
- Datos sectoriales recopilados por empresas especializadas.
- Datos experimentales.

Y los tipos de datos también lo son:

1. Tipos de datos no estructurados: documentos, vídeos, audios, etc.
2. Tipos de datos semi-estructurados: software, hojas de cálculo, informes.
3. Tipos de datos estructurados

Solo el 20% de información es estructurada y eso puede provocar muchos errores si no acometemos un proyecto de calidad de datos.

2. Tremendo volumen de datos

Como ya hemos visto, el volumen de datos es enorme, y eso complica la ejecución de un proceso de calidad de datos dentro de un tiempo razonable.

Es difícil recolectar, limpiar, integrar y obtener datos de alta calidad de forma rápida. Se necesita mucho tiempo para transformar los tipos no estructurados en tipos estructurados y procesar esos datos.

3. Mucha volatilidad

Los datos cambian rápidamente y eso hace que tengan una validez muy corta. Para solucionarlo necesitamos un poder de procesamiento muy alto.

Si no lo hacemos bien, el procesamiento y análisis basado en estos datos puede producir conclusiones erróneas, que pueden llevar a cometer errores en la toma de decisiones.

4. No existen estándares de calidad de datos unificados

En 1987 la Organización Internacional de Normalización (ISO) publicó las normas ISO 9000 para garantizar la calidad de productos y servicios. Sin embargo, el estudio de los estándares de calidad de los datos no comenzó hasta los años noventa, y no fue hasta 2011 cuando ISO publicó las normas de calidad de datos ISO 8000.

Estas normas necesitan madurar y perfeccionarse. Además, la investigación sobre la calidad de datos de Big Data ha comenzado hace poco y no hay apenas resultados.

La calidad de datos de Big Data es clave, no solo para poder obtener ventajas competitivas sino también impedir que incurramos en graves errores estratégicos y operacionales basándonos en datos erróneos con consecuencias que pueden llegar a ser muy graves.

4. Cómo construir un plan de Data Governance en Big data

Gobernabilidad significa asegurarse de que los datos estén autorizados, organizados y con los permisos de usuario necesarios en una base de datos, con el menor número posible de errores, manteniendo al mismo tiempo la privacidad y la seguridad.

Esto no parece un equilibrio fácil de conseguir, sobre todo cuando la realidad de dónde y cómo los datos se alojan y procesan está en constante movimiento.

A continuación, veremos algunos pasos recomendados al crear un plan de Data Governance en Big Data.

1. Acceso y Autorización Granular a Datos

No se puede tener un gobierno de datos efectivo sin controles granulares.

Se pueden lograr estos controles granulares a través de las expresiones de control de acceso. Estas expresiones usan agrupación y lógica booleana para controlar el acceso y autorización de datos flexibles, con permisos basados en roles y configuraciones de visibilidad.

En el nivel más bajo, se protegen los datos confidenciales, ocultándolos, y en la parte superior, se tienen contratos confidenciales para científicos de datos y analistas de BI. Esto se puede hacer con capacidades de enmascaramiento de datos y diferentes vistas donde se bloquean los datos en bruto tanto como sea posible y gradualmente se proporciona más acceso hasta que, en la parte superior, se da a los administradores una mayor visibilidad.

Se pueden tener diferentes niveles de acceso, lo que da una seguridad más integrada.

2. Seguridad perimetral, protección de datos y autenticación integrada

La gobernabilidad no ocurre sin una seguridad en el punto final de la cadena. Es importante construir un buen perímetro y colocar un cortafuego alrededor de los datos, integrados con los sistemas y estándares de autenticación existentes. Cuando se trata de autenticación, es importante que las empresas se sincronicen con sistemas probados.

Con la autenticación, se trata de ver cómo integrarse con LDAP [Lightweight Directory Access Protocol], Active Directory y otros servicios de directorio. También se puede dar soporte a herramientas como Kerberos para soporte de autenticación. Pero lo importante es no crear una infraestructura separada, sino integrarla en la estructura existente.

3. Encriptación y Tokenización de Datos

El siguiente paso después de proteger el perímetro y autenticar todo el acceso granular de datos que se está otorgando es asegúrese de que los archivos y la información personalmente identificable (PII) estén encriptados y tokenizados de extremo a extremo del pipeline de datos.

Una vez superado el perímetro y con acceso al sistema, proteger los datos de PII es extremadamente importante. Es necesario encriptar esos datos de forma que, independientemente de quién tenga acceso a él, puedan ejecutar los análisis que necesiten sin exponer ninguno de esos datos.

4. Constante Auditoría y Análisis

La estrategia no funciona sin una auditoría. Ese nivel de visibilidad y responsabilidad en cada paso del proceso es lo que permite a la TI "gobernar" los datos en lugar de simplemente establecer políticas y controles de acceso y esperar lo mejor. También es cómo las empresas pueden mantener sus estrategias actualizadas en un entorno en el que la forma en que vemos los datos y las tecnologías que utilizamos para administrarlos y analizarlos están cambiando cada día.

Estamos en la infancia de Big Data e IoT (Internet de Cosas), y es fundamental poder rastrear el acceso y reconocer patrones en los datos.

La auditoría y el análisis pueden ser tan simples como el seguimiento de los archivos de JavaScript Object Notation (JSON).

5. Una arquitectura de datos unificada

En última instancia, el responsable de TI que supervisar la estrategia de administración de datos empresariales, debe pensar en los detalles del acceso granular, la autenticación, la seguridad, el cifrado y la auditoría. Pero no debe detenerse ahí. Más bien debe pensar en cómo cada uno de estos componentes se integra en su arquitectura de datos global. También debe pensar en cómo esa infraestructura va a necesitar ser escalable y segura, desde la recolección de datos y almacenamiento hasta BI, analítica y otros servicios de terceros. La gobernanza de los datos es tanto acerca de repensar la estrategia y la ejecución como sobre la propia tecnología.

Va más allá de un conjunto de reglas de seguridad. Es una arquitectura única en la que se crean estos roles y se sincronizan a través de toda la plataforma y todas las herramientas que se aportan a ella.