S.E.P

INSTITUTO TECNOLÓGICO DE ORIZABA

DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN MAESTRIA EN INGENIERÍA ADMINISTRATIVA

“Big Data & Su Poderío”

PRESENTA:

Paul Itai Gómez Palestino

Catedrático

Dr. Fernando Aguirre y Hernández

Orizaba, Ver. Marzo de 2018

Agradecimientos

Agradezco a mi familia por su apoyo en todo momento, a mi madre y mi padre por brindarme todo lo que está a su alcance para que pueda lograr mis aspiraciones profesionales y personales.

De la misma forma al personal docente del Instituto Tecnológico de Orizaba, que me ha brindado las herramientas necesarias para ser un profesional competente.

Al CONACYT por la oportunidad de estudiar en una de las carreras certificadas por su calidad que me permitirán el día de mañana poner en alto el nombre de mi nación y mi ciudad.

Índice

Definición

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 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. (PowerData, 2015).

Big data describe una estrategia holística de gestión de la información que incluye e integra muchos nuevos tipos de datos y de gestión de datos junto con datos tradicionales (ORACLE, 2014).

Dentro de la una definición mucho más amplia, se deben de tomar en cuenta las 4 V para poder entender mejor el alcance del concepto:

Volumen

Se refiere a la cantidad de los datos, cabe resaltar que más volumen no es sinónimo de más datos, por lo que en Big Data es necesario que el procesamiento de volúmenes grandes de información sea de baja densidad. ORACLE dice que los datos de Hadoop[1] deben ser no estructurados (de valores desconocidos), por ejemplo clics realizados en páginas web, mensajes en redes sociales e incluso de aplicaciones móviles, el tráfico existente en la red, entre otros. La labor de Big Data es realizar la conversión de esos datos de en información de utilidad. Refiriéndose

[1] Sistema de código abierto que se utiliza para almacenar, procesar y analizar grandes volúmenes de datos.

al tamaño de almacenamiento puede variar de decenas de terabytes[2] a cientos de petabytes[3], dependiendo de cada organización.

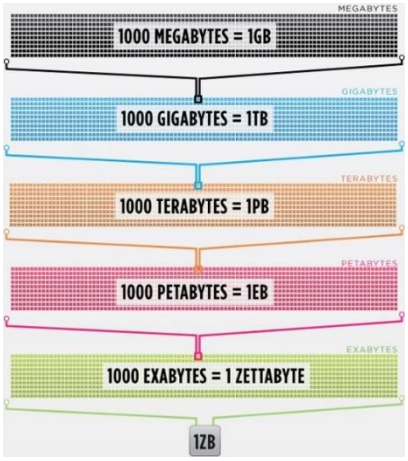


Ilustración 1. Descripción del Tamaño de Datos

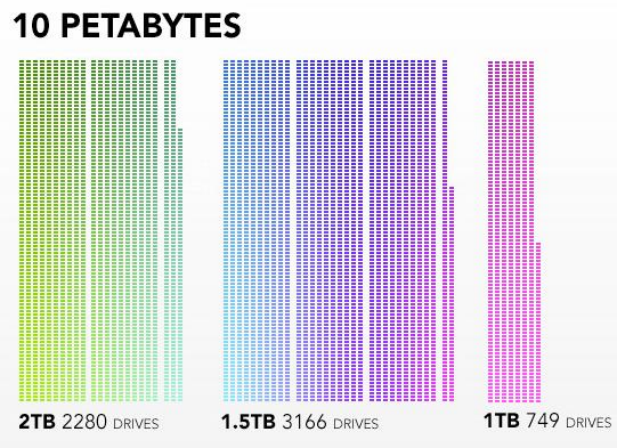


Ilustración 2. Descripción Gráfica de Petabytes en Datos

Hadoop

El sistema Hadoop tiene la función de aligerar el trabajo de los desarrolladores por la dificultad de la programación paralela, brindando un ecosistema que sirve de ayuda al usuario, distribuyendo el fichero[4] en nodos[5], permitiendo ejecutar varios procesos de forma paralela. El sistema Hadoop dispone de módulos de control para la monitorización de los datos, permite la integración de addons[6], que sirven para facilitar el trabajo, manipulación, seguimiento y consulta de la información almacenada.

Valor

No es raro decir que actualmente los datos o información tienen valor, sin embargo el contar con ellos no genera ninguna utilidad, sino que debe descubrirse alguna aplicación para ellos. Existen muchas técnicas de carácter cuantitativo y de investigación que permiten sacarle valor a esos datos, un claro ejemplo de ello es el análisis de las preferencias de los clientes que realizan muchas compañías, que sirve para poder realizar una oferta relevante, en la que se incluyen datos como la ubicación.

---

[2] Terabyte (TB), equivalente a 10^{12} bytes, es decir, 1 000 000 000 000 (un mil millones) de bytes.
[3] Petabyte (PB) equivale a 10^{15} bytes, es decir, 1 000 000 000 000 000 bytes.
[4] Conjunto lógico de información o de datos que se designa con un nombre y se configura como una unidad autónoma completa para el sistema o el usuario.
[5] Punto de intersección o unión de varios elementos que confluyen en el mismo lugar.
[6] Extensión o añadidura puede referirse a una mejora instalable para proyectos en informática.

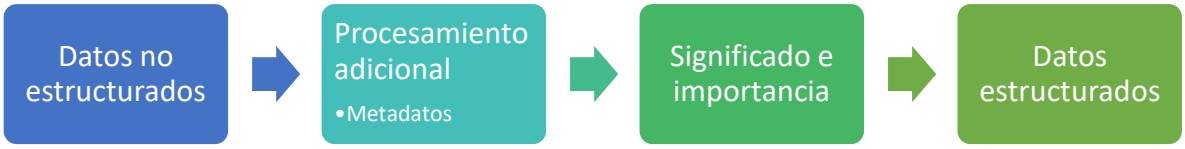
Poder almacenar y procesar toda la información tiene un costo, sin embargo, debido a la creciente demanda del análisis de comportamiento online, los precios en materia de computación y almacenamiento de datos ha disminuido, de modo que se puede realizar el análisis estadístico de una gran cantidad de información sin necesidad de segmentarla o de usar una muestra únicamente.

El hecho de poder procesar toda la información en conjunto supone una innovación para la toma de decisiones, permitiendo que sean más exactas. Dentro del proceso de descubrimiento de información de valor se requiere la participación de analistas o especialistas en la materia, usuarios y ejecutivos. De esta manera Big Data dbe aprender a predecir el comportamiento humano, reconociendo los patrones, para poder ofrecer una predicción de los comportamientos.

Variedad

Este aspecto es referente a los datos no estructurados y a aquellos que pueden catalogarse como semiestructurados, entre los que se encuentran los textos, audios y vídeos. Todos estos datos demandan un procesamiento adicional para poder generar algún significado, así como la utilización de metadatos[7] de apoyo. Es decir, este aspecto intenta cuantificar la complejidad de la información y reducirla.

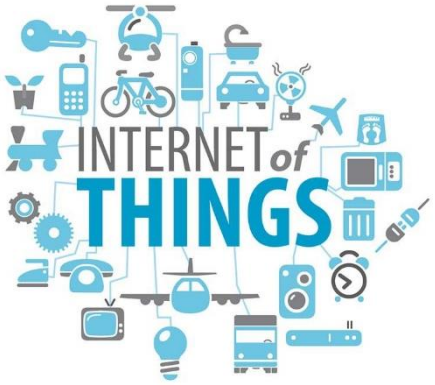
Cuando se ha logrado comprenderlos, los datos no estructurados pueden ser procesados como datos estructurados, es decir, se pueden resumir, alinear y trazar para auditorías. Sin embargo, existe una mayor complejidad cuando los datos que son obtenidos desde un origen conocido cambian sin previo aviso, esto produce un lastre para el análisis.



Velocidad

Es el ritmo al que son recibidos los datos y en el cual se les aplica alguna acción como analizarlos o procesarlos. Para obtener una mayor velocidad se requiere una capacidad de memoria alta, no solo en bytes sino en el poder de lectura, por lo que la importancia de tecnologías como el almacenamiento en la nube y la velocidad de internet son fundamentales.

[7] Grupo de datos que describen el contenido informativo de un objeto al que se denomina recurso.



Por ejemplo, algunas aplicaciones del Internet de las Cosas[8] (Internet of Things), tienen agregados de estado y seguridad, estas requieren que se realicen acciones en tiempo real, así como evaluaciones.

Ilustración 3. Internet de las Cosas

Otro ejemplo son los productos inteligentes que están preparados para utilizar Internet, estos funcionan en tiempo real brindando información relevante como estadísticas de uso, seguridad, localización, entre otros. Así es como las aplicaciones de comercio electrónico intentan utilizar estas variables, mezclando la ubicación de un Smartphone con las preferencias personales para realizar ofertas por medio de publicidad. Desde un punto de vista operativo, las aplicaciones diseñadas para celulares tienen una base de usuarios enorme y un tráfico de red más amplio, por lo que la experiencia y expectativa de respuesta debe ser inmediata.

Descripción

Una vez teniendo en claro todos estos principios, se debe puntualizar que entonces Big Data es un conjunto de datos, que a su vez son combinaciones de dichos conjuntos de datos cuyo ***volumen, valor, variedad y velocidad*** dificultan la captura, registro, administración, procesamiento y análisis por medio de tecnologías y herramientas convencionales, tales como bases de datos de relación, estadísticas y paquetes de visualización, dentro de un tiempo necesario para que puedan resultar de utilidad.

No está definido cuál es el tamaño que debe poseer un conjunto de datos específico para ser considerado como Big Data, ya que continua cambiando conforme avanza el tiempo, actualmente la mayoría de los analistas y profesionales en la materia

[8] El Internet de las cosas potencia objetos que antiguamente se conectaban mediante circuito cerrado, como comunicadores, cámaras, sensores, y demás, y les permite comunicarse globalmente mediante el uso de la red de redes.

dicen que son conjuntos de datos que parten desde 30 Terabytes. Por tanto, posee una naturaleza sumamente compleja, debido a la naturaleza no estructurada de gran parte de los datos generados por las tecnologías que se usan actualmente, como las búsquedas por Internet de información, las redes sociales e interacciones que en ellas ocurren (Facebook, Twitter, Google, entre otras), los registros en páginas, los sensores de los dispositivos (mediciones, ubicación GPS), las computadoras portátiles, los Smartphones (teléfonos inteligentes) y registros de centros de llamadas, incluso la maquinaria y los vehículos.

Ilustración 4. Interacción de datos

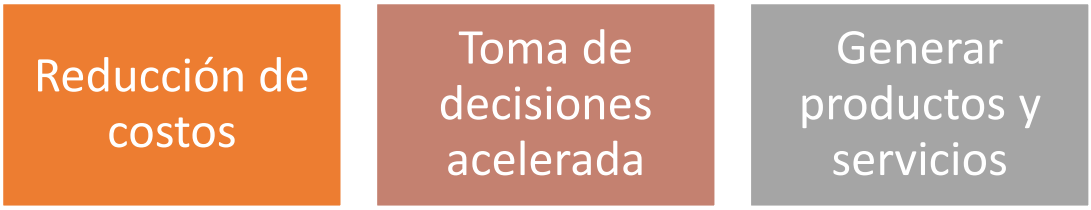
Para poder utilizar Big Data de manera eficaz debe combinarse con los datos estructurados (base de datos relacional) de una aplicación comercial convencional, tales como un ERP o un CRM.

Importancia

El hecho de que Big Data proporciona respuestas a muchas preguntas que en ocasiones las compañías no sabían que deben contestar, es lo que hace esta herramienta sumamente útil a nivel empresarial, porque brinda un punto de referencia. El volumen de información que se requiere permite que los datos pueden ser moldeados de cualquier manera que las compañías requieran. Al hacerlo son capaces de identificar los problemas de una forma más comprensible.

El hecho de poder recopilar grandes cantidades de datos y permitir encontrar tendencias especificas dentro de estos, permiten que las empresas puedan tomar decisiones de manera ágil, eficiente y sin problemas. Algo sumamente importante de resaltar es que permite eliminar áreas problemáticas mucho antes de que los problemas afecten la reputación de la compañía o lastren sus beneficios.

Big Data ayuda a las organizaciones a aprovechar su información por medio del análisis, utilizándola para identificar oportunidades de crecimiento o mejora. Esto permite realizar movimientos de negocios inteligentes, tener operaciones más eficientes, generar mayores ganancias y lograr la satisfacción de los clientes. Se deben considerar beneficios por medio de esta herramienta como los siguientes:



Reducción de costos

Se deben echar en mano las tecnologías de datos más potentes y que presentan potencial, como el caso del sistema Hadoop y el análisis basado en la nube. Estos generan una ventaja en los costos, puesto que cuando se trata de almacenar grandes cantidades de datos, existe una gran cantidad de oferta que muestra un crecimiento exponencial dentro de los siguientes años, permitiendo además identificar maneras más eficientes de comercialización.

Toma de decisiones más rápida

Haciendo referencia al sistema Hadoop, su velocidad y analítica de información, fusionada con la capacidad de analizar nuevas fuentes de datos, sirve a las empresas para disponer de la información de manera inmediata (ya sea a manera de resumen o como datos específicos que se requieran) y de esta manera tomar decisiones basadas en lo que han aprendido (inteligencia artificial).

Generar nuevos productos/servicios

Big Data ofrece la capacidad de analizar y medir las necesidades de los clientes, por lo tanto, la satisfacción de los mismos se da a través del análisis su información, con la que se puede saber con certeza que es lo que quieren o necesitan. Por medio de la analítica, las empresas crean nuevos productos y servicios para satisfacer las exigencias de sus clientes. Pueden llegar incluso a generarse nuevas necesidades que los mismos no sabían que tenían.

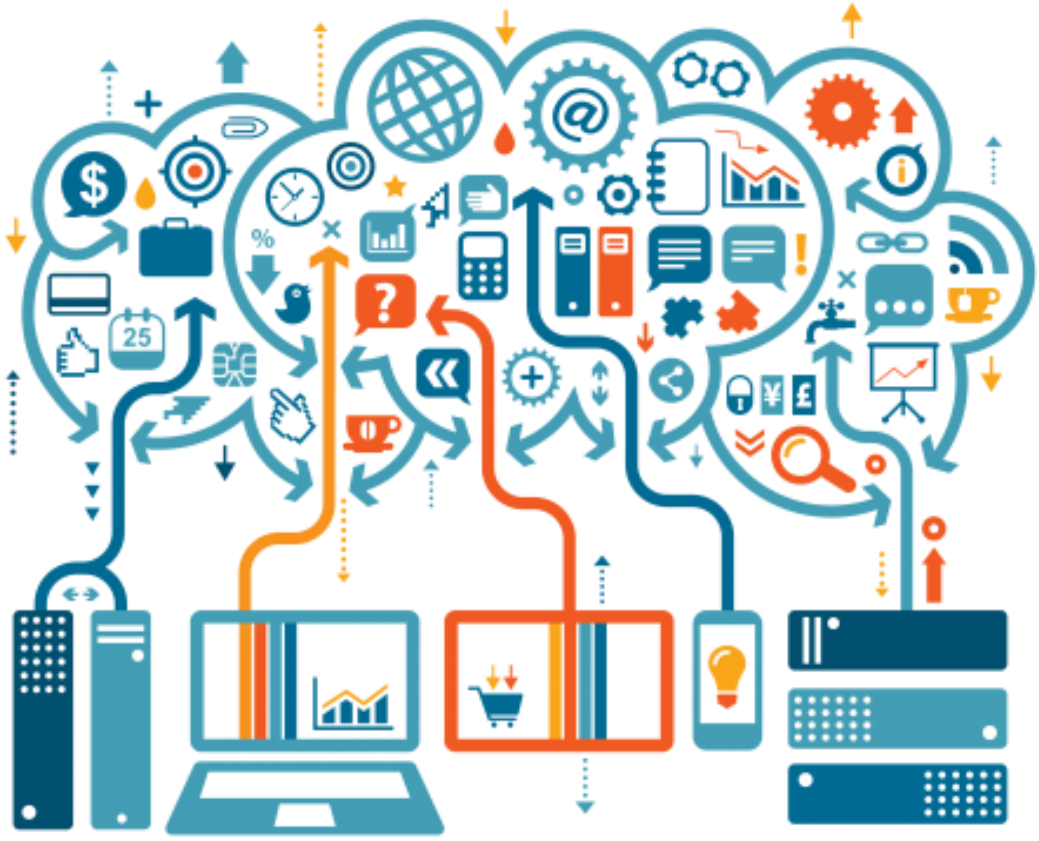


Ilustración 5. Big Data y su importancia

Aplicación

Como se pudo observar anteriormente, el poder de alcance de Big Data es inimaginable, realmente los límites son puestos por las mismas empresas, ya que depende de ellas que hacer con la información. A continuación se presentan formas en que puede ser utilizada esta herramienta en diversos sectores:

Salud

Big Data contiene grandes cantidades de información en la industria de la sanidad. Parte primordialmente de los registros de pacientes, los planes de salud general y especializada, información de los seguros y alcances y también información difícil de manejar. Todos estos datos brindan información que resulta clave cuando se aplica un análisis. Es por eso que la tecnología analítica de datos es de vital importancia para el cuidado de la salud. Al analizar estas grandes cantidades de información, se pueden proporcionar diagnósticos a los pacientes y opciones de tratamiento casi de inmediato, generando con esto posibilidades de atacar enfermedades antes de que sean irreparables.

Ilustración 6. Big Data en el sector de la salud

Administración

Uno de los principales desafíos que afronta la administración es asegurar la calidad y aumentar la productividad de las operaciones con presupuestos que generalmente se encuentran ajustados. Big Data puede permitir la agilización de las operaciones por medio de la tecnología, brindándole a la administración una visión mucho más amplia de las actividades.

Publicidad

La creciente utilización de smartphones (teléfonos inteligentes), así como de dispositivos con integración de GPS, permite a los anunciantes dirigirse a los consumidores cuando están cerca de una tienda específica, por ejemplo un restaurante, una librería o una cafetería. Esto genera oportunidades para los proveedores de servicios como obtener mayor ingresos, conseguir nuevos prospectos, posicionarse y lograr el éxito.

Ilustración 7. Big Data en la Publicidad

Ventas

El servicio al cliente se ha vuelto sumamente importante para todas las empresas, de igual manera los clientes se han vuelto exigentes hasta en el más mínimo de los detalles, por lo que las ventas han evolucionado, ya que los compradores más inteligentes esperan que los minoristas comprendan exactamente lo que necesitan y cuando lo necesitan.

Big Data puede permitir a los minoristas a satisfacer dichas 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.

Turismo

Debe permitir obtener la satisfacción de los clientes, puesto que es clave para la industria del turismo, pero esta característica es 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.

Ilustración 8. Big Data en el Turismo

Desafíos

Las características particulares de Big Data hacen que su calidad de datos se enfrente a muchísimos desafíos:

Diversidad en las 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.

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.

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.

No existen estándares de calidad de datos unificados

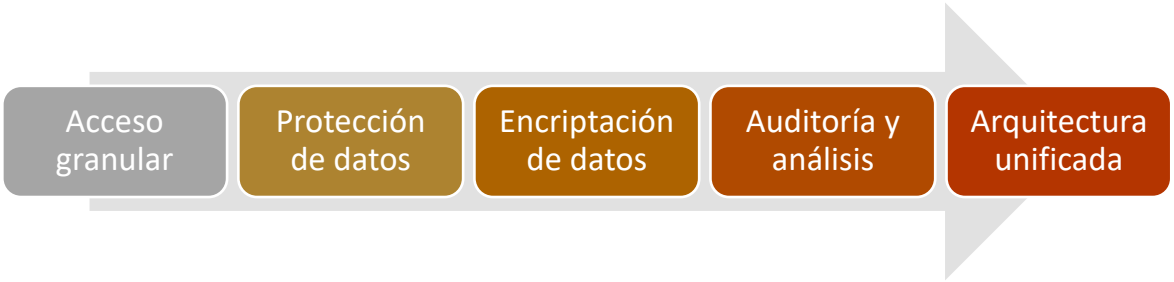
En 1987 la Organización Internacional de Normalización (ISO) publicó las normas ISO 9000 para garantizar la calidad de productos y servicios. Sin embargo, el estudio de los estándares de calidad de los datos no comenzó hasta los años noventa, y no fue hasta 2011 cuando ISO publicó las normas de calidad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

Plan de Gobernabilidad de Datos

La gobernabilidad es referente a asegurarse de que los datos se encuentren 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. Conseguir equilibrio fácil dentro de estas características es difícil, sobre todo cuando la realidad de dónde y cómo los datos se alojan y procesan está en constante movimiento.



Acceso 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.

Protección de Datos

La gobernabilidad no ocurre sin una seguridad en el punto final de la cadena. Es importante construir un buen perímetro y colocar un cortafuegos 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.

Encriptación

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.

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).

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.

Propuesta de tesis

Propuesta 1

Utilizar Big Data para analizar la información de la sociedad de Veracruz y poder prevenir delitos, por medio del monitoreo de actividad en redes que facilite el encaminamiento y corrección de los individuos.

Propuesta 2

Generar propuestas para el mejoramiento del tejido social, escalando desde los sectores más rezagados para lograr una integración más rápida.

Fuentes de Consulta

Especialistas en Gestión de Datos. (Octubre, 2012). Big Data: ¿En qué consiste? Su importancia, desafíos y gobernabilidad. Marzo, 2018, de PowerData Sitio web: https://www.powerdata.es/big-data

ORACLE. (Agosto, 2014). Big data empresarial. Marzo, 2018, de ORACLE Latinoamérica Sitio web: https://www.oracle.com/lad/big-data/index.html

Quer, A. (Septiembre 05, 2013). ¿Cómo se relacionan Big Data y Hadoop?. Marzo, 2018, de PowerData Sitio web: https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/bid/328879/c-mo-se-relacionan-big-data-y-hadoop