Ing. Carlos Ormella Meyer y Asoc.
Gestión y Auditoría de Riesgos y Seguridad de la Información
Tel: +54-11-3979-7220 – Cel: (+54-911) 6513-2751
E-mail: ingcomyasoc@gmail.com

Breve introducción a Big Data

© Ing. Carlos Ormella Meyer (*)

En la empresa moderna, los datos son la moneda que guía todas las decisiones y acciones. Esos datos, hoy en día, constituyen conjuntos muy grandes, que se mueven rápidamente y carecen de una estructura común.

Justamente el concepto básico de Big Data responde a datos que son lo bastante variados, grandes y generados con gran rapidez y de diferentes fuentes, como para ser procesados por métodos convencionales.

Por sus características más destacadas se suele definir Big Data con las **3V: Volumen, Variedad y Velocidad**.

Especialmente en el tema de la **Variedad** se destacan:

a) **Datos estructurados**. Los conocidos datos alojados en bases de datos relacionales y fácilmente interrogados con SQL
b) **Datos no estructurados**. Los producidos en redes sociales como Facebook y YouTube, smartphones, textos de mensajes de email, multimedia, e incluso datos de GPS y de sensores...

La analítica de Big Data hace una suerte de destilado de los datos en bruto de data warehouses, sensores de dispositivos, transacciones, video, audio, y redes sociales en patrones que pueden interpretarse como tendencias predictivas.

Efectivamente, en este nuevo escenario la analítica predictiva juega un papel importante constituyendo el proceso de usar un conjunto de sofisticadas herramientas analíticas para desarrollar modelos y estimaciones de cómo será el futuro del rubro analizado en base a la información pasada.

Y todo se complementa porque hoy se puede con herramientas adecuadas hacer query con una sintaxis extendida del SQL en bases de datos no estructuradas como la de Hadoop.

Podemos mencionar algunas aplicaciones de Big Data como las que siguen:

- Retail. Recolección y análisis de patrones de compras y transacciones para campañas de marketing más efectivas.
  Reducción del overhead, Mejores estrategias de stock por medio de la gestión en línea del inventario y mejora/restructuración de la cadena de aprovisionamiento.

- Servicios financieros. Predicción más efectiva del mercado, mejoras en las operaciones y rapidez en las transacciones, reducción del riesgo con tarjetas de crédito, análisis del riesgo crediticio por la historia de las transacciones, mejor seguridad especialmente por medio del registro de transacciones anteriores en cuanto a la identificación de actividades anormales, comportamiento fraudulento, etc.

- Seguros. La tendencia es a nuevas técnicas centradas en las personas en lugar de estar centradas en los reclamos.
  Una mayor efectividad y beneficios se logran por medio de información de patrones climáticos, redes sociales como Facebook y YouTube, logrando una mejor comprensión del comportamiento de los clientes y el perfil de riesgo de cada uno.
  Esto incluye la posibilidad de generar primas de seguro personalizadas por medio del análisis de riesgos incluyendo hábitos de frenado y acelerado, distancias recorridas, calles, caminos y rutas transitadas (mediante GPS).

Además, un mejor conocimiento de las necesidades y comportamiento de los clientes permiten aumentar la retención y satisfacción de los mismos.

- Telecomunicaciones. Mayor efectividad en campañas en línea de marketing en base a una plataforma para recolectar, almacenar y analizar datos de clientes y sus transacciones creando patrones de uso celular, flujo de ingresos de promociones en tiempo basado en la ubicación, etc. De hecho, gracias a la analítica se puede realizar una reingeniería de la medición del mercado por medio de procesos de integración de datos, desarrollo de nuevas interfaces para el usuario, y la construcción de informes para este escenario.
  Todo esto incluso puede ayudar a controlar el mayor fraude actual en las comunicaciones debido a la mayor cantidad de dispositivos inteligentes.

- Salud. Registro Electrónico de Salud (EHR) en línea con la historia de las dolencias, operaciones y medicaciones de un paciente, y los resultados de los chequeos correspondientes. De hecho el carácter no estructurado propio de una base de datos de EHR es el que precisamente puede manejar Big Data.
  Además de la gestión de los EHR se tiene la posibilidad de realizar análisis predictivo, que incluso permite a los médicos tomar decisiones en minutos y mejorar el tratamiento de los pacientes. Adicionalmente se puede implementar la característica de alertas en tiempo real y proveer consejo para tomar decisiones prescriptivas.
  Una extensión importante es la disponibilidad de pruebas clínicas en proceso o realizadas universalmente, efectos secundarios de los medicamentos, datos de enfermedades comunes prevalecientes en ciertas partes del mundo. Todo esto permite la identificación de fuentes de infección, cuidados preventivos más eficaces, seguimiento de epidemias, etc.

- Gobierno. En el área de educación: desempeño estudiantil, análisis de gastos, planificación del crecimiento escolar. En servicios sociales, eficiencia y gestión eficaz de beneficios. En el área legal: identificación del fraude en servicios médicos, análisis predictivo para evaluar la delincuencia.

La seguridad de la información plantea requisitos propios en Big Data además de los tradicionales.

En general se propone el uso de cifrado basado en atributos, así como la correlación de eventos para detectar intrusiones.

Las principales herramientas que complementan la seguridad son SIEM (Gestión de Eventos e Información de Seguridad) especialmente con el soporte de la Inteligencia de Amenazas (TI) que recoge evidencias, analizándolas y filtrándolas, para establecer sus características.

Especialistas en Big Data

En este nuevo escenario han surgido diferentes tipos de especialistas que, de hecho, señalan para los estudiosos las tendencias que implican mejores perspectivas de trabajo más retributivo.

Al tradicional Analista de Datos ahora se han agregado el Ingeniero de Datos y el Científico de Datos.

Ingeniero de Datos (Data Engineer). Profesional que diseña, construye, realiza ETL (extracción, transformación y carga) de grandes conjuntos de datos y de diferentes fuentes, integra datos de distintos recursos, gestiona big data y crea almacenes de big data que se pueden usar para informes o análisis por parte de los Científicos de Datos.

Algunas herramientas que se usan para este nivel son: Hadoop, Pig, MySQL, bases de datos NoSQL como MongoDB o Cassandra, data streaming, y SQL.

Científico de Datos (Data Scientist). Su función principal es ayudar a las organizaciones a convertir sus volúmenes de big data en información valiosa y procesable.

Para lograr esos objetivos, las habilidades para la resolución de problemas de un Científico de Datos implican una comprensión de los métodos tradicionales y los nuevos análisis de datos para construir modelos estadísticos o descubrir patrones en los datos.

Para todo ello, los Científicos de Datos aplican estadísticas, modelado estadístico y predictivo, machine learning, (aprendizaje automático), y enfoques analíticos para resolver problemas críticos de negocio.

Algunas herramientas que se manejan son: lenguajes de programación como Python, R, o Scala, marco de trabajo bajo Spark Apache, Hadoop, herramientas de minería de datos y algoritmos, machine learning, estadísticas y análisis multivariante.

* Ing. Carlos Ormella Meyer. Cursos y Soporte Digital - Asesoramiento - @meyerormella