Big Data

Técnicas, herramientas y aplicaciones

María Pérez Marqués

Big Data. Técnicas, herramientas y aplicaciones
María Pérez Marqués

ISBN: 978-84-943055-5-9
EAN: 9788494305559
IBIC: UNC

RC Libros
Calle Mar Mediterráneo, 2. Nave 6
28830 SAN FERNANDO DE HENARES, Madrid
Teléfono: +34 91 677 57 22
Fax: +34 91 677 57 22
Correo electrónico: info@rclibros.es
Internet: www.rclibros.es
Diseño de colección, y pre-impresión: Grupo RC
Diseño de cubierta: Cuadratín
Impresión y encuadernación: Arvato
Depósito Legal: M-8381-2015
Impreso en España

18 17 16 15 (1 2 3 4 5 6 7 8 9 10 11 12)

INTRODUCCIÓN

Ante el boom actual de la información, las organizaciones han tratado de abordar el problema de analizar grandes volúmenes de datos desde muchos ángulos diferentes. Las herramientas de BIG DATA utilizan tecnologías multinúcleo para ofrecer mayor capacidad de procesamiento a través de altas prestaciones, en base de datos y de análisis en memoria que ofrecen un mayor conocimiento más rápidamente de grandes volúmenes de datos y flujo de datos. Y todo ello independientemente de los formatos y las fuentes de los orígenes de datos. Con las herramientas de BIG DATA se puede procesar información online proveniente de múltiples orígenes como pueden ser las redes sociales o grandes bases de datos no estructuradas. También se pueden tratar los datos de múltiples fuentes y formatos, ya sean texto, datos, imágenes o mezcla de todo ello. Actualmente es posible implementar herramientas de BIG DATA en la forma que mejor se adapte a las necesidades de los usuarios.

El término Big Data suele aplicarse a la información que no puede ser procesada o analizada usando procesos o herramientas tradicionales. Las organizaciones de hoy en día se enfrentan cada vez más a menudo a retos Big Data. Las empresas tienen acceso a una gran cantidad de información, pero no saben cómo obtener valor añadido de la misma, ya que la información aparece en su forma más cruda o en un formato semi-estructurado o no estructurado. Una encuesta de IBM demostró que más de la mitad de los líderes empresariales de hoy en día se dan cuenta de que no tienen acceso a los conocimientos que necesitan para analizar sus datos. Las empresas se enfrentan a estos retos en un clima en el que tienen la capacidad de almacenar cualquier cosa, que están generando datos como nunca antes en la historia y, sin embargo, tienen un verdadero desafío con el análisis de la información.

Las técnicas de Big Data persiguen complementar el manejo de grandes volúmenes de datos con las técnicas de análisis de la información más avanzadas y efectivas para extraer de modo óptimo el conocimiento contenido en los datos.

La base que actualmente caracteriza a las herramientas de BIG DATA es el paquete de código abierto llamado Hadoop para el análisis masivo de datos. Hadoop también se incluye como parte de las herramientas de todo el software de BIG DATA, como SAS, IBM, MICROSOFT y ORACLE. Por ejemplo, SAS incorpora Hadoop en sus aplicaciones (SAS Base SAS Data Integration, Sas Enterprise Guide, SAS Enterprise Miner, ...). También SAS permite trabajar en memoria a través de Hadoop (SAS Visual Analytics y SAS Visual Statistics). IBM trabaja con Hadoop en su plataforma IBM InfoSphere BigInsights (BigInsights). Microsoft incluye Hadoop en SQL Server 2014, Windows Server 2012, HDInsight and Polybase. Oracle incluye Hadoop en Oracle Big Data Appliance, Oracle Big Data Connectors y Oracle Loader for Hadoop.

Este libro presenta las posibilidades de trabajo que ofrecen las herramientas de BIG DATA para procesar y analizar grandes volúmenes de datos de una manera ordenada. A su vez, estas herramientas también permiten extraer el conocimiento contenido en los datos.

CAPÍTULO 1

CONCEPTOS DE BIG DATA

DEFINICIÓN, NECESIDAD Y CARACTERÍSTICAS DE BIG DATA

El término "Big data" suele aplicarse a conjuntos de datos que superan la capacidad del software habitual para ser capturados, gestionados y procesados en un tiempo razonable y por los medios habituales de procesamiento de la información. Este término suele referirse a los siguientes tipos de datos:

Datos de la empresa tradicional: incluye información de los clientes en sistemas de CRM, datos transaccionales ERP, las transacciones de tienda web, los datos contables, etcétera.

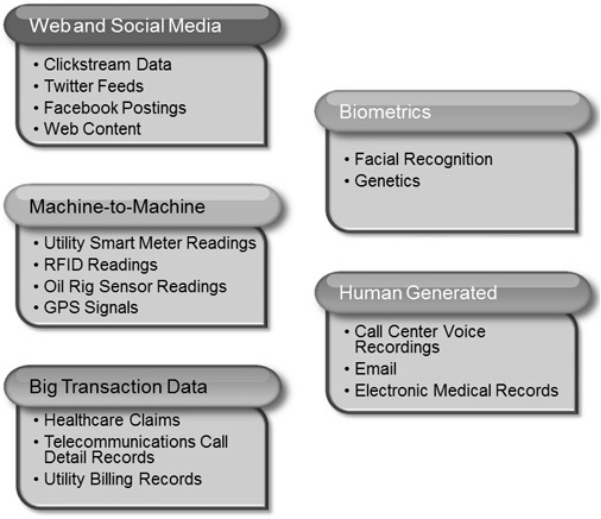
Machine-generated /sensor data: incluye registros de detalles de llamadas ("Call Detail Records, CDR"), los weblogs, los medidores inteligentes, los sensores de fabricación, registros de equipos, datos de sistemas comerciales, etc.

Datos de medios sociales: Incluye datos sobre blogs, Twiter, plataformas de Social Media como Facebook, etc.

Grandes bases de datos: con información multidimensional, relacional y no relacional.

Grandes conjuntos de datos no estructurados con mezcla de fuentes de origen y tipos de datos: numéricos, textuales, gráficos, etc.

El esquema siguiente amplía un poco más los tipos de datos a tener en cuenta en el tratamiento con técnicas de Big Data.



1.- *Web and Social Media*: incluye contenido web e información que es obtenida de las redes sociales como Facebook, Twitter, LinkedIn, etc., blogs.

2.- *Machine-to-Machine (M2M)*: M2M se refiere a las tecnologías que permiten conectarse a otros dispositivos. M2M utiliza dispositivos como sensores o medidores que capturan algún evento en particular (velocidad, temperatura, presión, variables meteorológicas, variables químicas como la salinidad, etc.), los cuales transmiten a través de redes alámbricas, inalámbricas o híbridas a otras aplicaciones que traducen estos eventos en información significativa.

3.- *Big Transaction Data*: incluye registros de facturación, en telecomunicaciones registros detallados de las llamadas (CDR), etc. Estos datos transaccionales están disponibles en formatos tanto semiestructurados como no estructurados.

4.- *Biometrics*: información biométrica en la que se incluye huellas digitales, escaneo de la retina, reconocimiento facial, genética, etc. En el área de seguridad e inteligencia, los datos biométricos han sido información importante para las agencias de investigación.

5.- *Human Generated*: las personas generamos diversas cantidades de datos como la información que guarda un call center al establecer una llamada telefónica, notas de voz, correos electrónicos, documentos electrónicos, estudios médicos, etc.

Dentro del sector de tecnologías de la información y la comunicación, Big Data es una referencia a los sistemas que manipulan grandes conjuntos de datos. Las dificultades más habituales en estos casos se centran en la captura, almacenamiento, búsqueda, compartición, análisis y visualización.

Además del gran volumen de información, existe en una gran variedad de datos que pueden ser representados de diversas maneras en todo el mundo, por ejemplo de dispositivos móviles, audio, video, sistemas GPS, incontables sensores digitales en equipos industriales, automóviles, medidores eléctricos, veletas, anemómetros, etc., los cuales pueden medir y comunicar el posicionamiento, movimiento, vibración, temperatura, humedad y hasta los cambios químicos que sufre el aire, de tal forma que las aplicaciones que analizan estos datos requieren que la velocidad de respuesta sea lo demasiado rápida para lograr obtener la información correcta en el momento preciso. Estas son las características principales de las aplicaciones típicas de Big Data.

Dado el gran avance que existe día a día en las tecnologías de información, las organizaciones se han tenido que enfrentar a nuevos desafíos que les permitan analizar, descubrir y entender más allá de lo que sus herramientas tradicionales reportan sobre su información. La necesidad del Big Data surge al mismo tiempo que el gran crecimiento durante los últimos años de las aplicaciones disponibles en internet (geo-referenciamiento, redes sociales, etc.) que han sido parte importante en las decisiones de negocio de las empresas.

El concepto de Big Data se aplica para toda aquella información que no puede ser procesada o analizada utilizando procesos o herramientas tradicionales. Hay cuatro características clave que definen la información relativa al Big Data:

- *Volumen*. Los datos relativos al Big data se producen en cantidades mucho más grandes que los datos tradicionales. Por ejemplo, un solo motor a reacción puede generar 10 TB de datos en 30 minutos. Con más de 25000 vuelos de aerolíneas por día, el volumen diario de solo esta única fuente de datos se ejecuta en petabytes. Los medidores inteligentes y equipos industriales pesados como las refinerías de petróleo y plataformas de perforación generan volúmenes de datos similares, lo que agrava el problema.

- *Velocidad*. Los flujos de datos de medios sociales, aunque no es tan masivo como los datos generados por máquinas, producen una gran afluencia de opiniones y valiosas relaciones para la gestión de clientes. Incluso a 140 caracteres por tweet, la alta velocidad (o frecuencia) de los datos de Twitter proporciona grandes volúmenes de información (más de 8 TB por día).

- *Variedad.* Los formatos de datos tradicionales tienden a ser relativamente bien definidos por un esquema de datos. En contraste, los formatos de datos no tradicionales exhiben un ritmo vertiginoso del cambio. A medida que se añaden nuevos servicios, nuevos sensores desplegados, o nuevas campañas de marketing, se necesitan nuevos tipos de datos para capturar la información resultante.

- *Valor*. El valor económico de los diferentes datos varía significativamente. Por lo general hay buena información embebida en un gran conjunto más amplio de datos no tradicionales: El desafío esencial es identificar la información valiosa, transformarla y extraer los datos para su análisis. A partir de los datos convenientemente extraídos y transformados se analiza el conocimiento contenido en los mismos.

APLICACIONES TÍPICAS DE BIG DATA

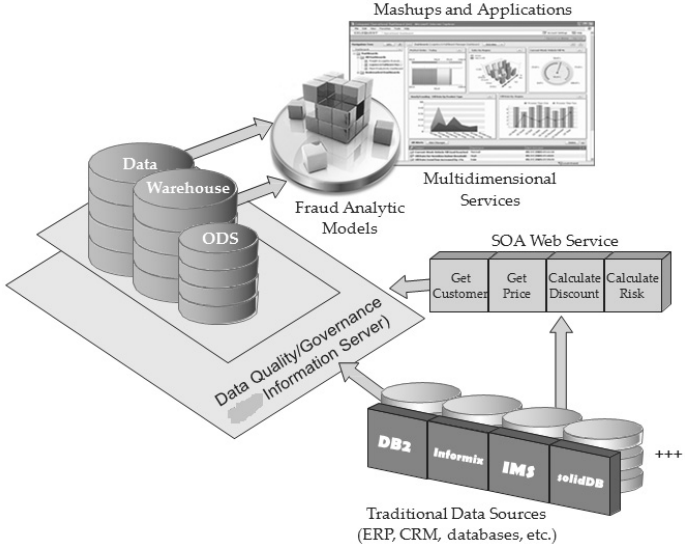
Existe una gran variedad de aplicaciones de las técnicas de Big Data. Siempre que sea necesario extraer el conocimiento inmerso en grandes volúmenes de datos estructurados, semiestructurados o no estructurados, tienen cabida las aplicaciones de Big Data. Pero estas técnicas no solo se aplican en la fase de análisis de la información, sino también en su propia recogida, transformación y puesta a disposición para los analistas. En los párrafos siguientes se citan algunos de los campos donde las técnicas de Big Data tienen más aplicación.

Patrones de detección del fraude

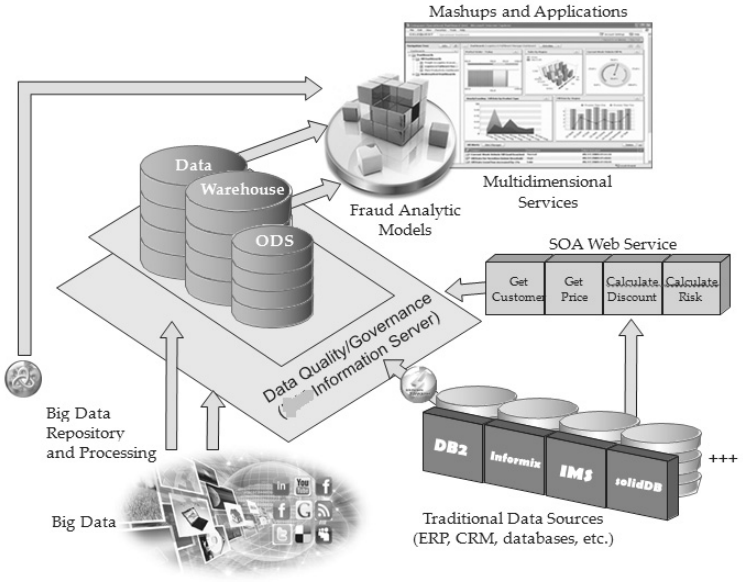
La detección de fraude es un problema típico en los servicios financieros verticales, pero se encuentra en cualquier tipo de transacciones (subastas en línea, juego online, reclamaciones de seguros, fraude fiscal, etc.). Prácticamente en cualquier lugar donde haya transacciones financieras está involucrado el fraude. Este tipo de transacciones presenta un potencial para el abuso y está omnipresente el fantasma del fraude. Una plataforma Big Data puede aportar la oportunidad de hacer más de lo que se ha hecho hasta ahora para identificar y paliar el fraude.

Varios desafíos en el patrón de detección de fraude son directamente atribuibles exclusivamente utilizando las tecnologías convencionales. El tema más común y recurrente que se observa en todos los patrones de Big Data es el relativo a los límites de almacenamiento de datos. Asimismo, son de gran importancia los recursos computacionales disponibles para procesar información relativa al fraude. Sin las tecnologías Big Data, estos factores limitan la información que puede ser analizada. Es más, entornos altamente dinámicos tienen patrones de fraude cíclico que van y vienen en horas, días o semanas. Si los datos utilizados para identificar o impulsar nuevos modelos de detección de fraude no está disponibles con inmediatez, el descubrimiento de los patrones de fraude puede llegar tarde cuando ya se haya ejecutado el daño.

Tradicionalmente, en casos de fraude, se utilizan muestras y modelos para identificar a los clientes que caracterizan a un determinado tipo de perfil. El problema con esta aproximación es que aunque funciona, está perfilando un segmento y no el microtratamiento a nivel transacción o persona individual. Sencillamente, hacer una previsión basada en un segmento es bueno, pero tomar una decisión basándose en los datos reales de una transacción individual es obviamente mejor. Para hacer esto, necesitamos trabajar con un conjunto mayor de datos que en el caso de la aproximación convencional tradicional. Se estima que a través de las herramientas tradicionales solo se está utilizando un 20 por ciento de la información disponible que podría ser útil para el modelado del fraude. El enfoque tradicional se muestra en la figura siguiente:



Es posible utilizar herramientas de Big Data para proveer un repositorio elástico y rentable para utilizar el 80 por ciento restante de la información y transformarla en útil para modelar el fraude. Posteriormente esta información alimentará la elaboración del modelo de fraude. En la figura siguiente se presenta el esquema. Se trata de un moderno sistema de detección de fraude típico de una plataforma de Big Data de bajo costo para modelado de exploración y descubrimiento. Se observa cómo pueden aprovecharse los datos mediante sistemas tradicionales directamente o a través de la integración en protocolos de calidad y gestión de datos existentes.



Patrones de Social Media

Tal vez el patrón de uso de Big Data más comentado son los medios de comunicación social y el sentimiento del cliente. Puede utilizar Big Data para averiguar lo que los clientes opinan sobre uno mismo (y tal vez lo que están diciendo acerca de la competencia). Además, se puede utilizar este resultado recién encontrado para averiguar cómo esta información repercute en las decisiones y la forma en que su empresa se comporta. Más específicamente, puede determinar qué factores están impactando a las ventas, la efectividad o la receptividad de sus campañas de marketing, la exactitud de su marketing (producto, precio, promoción y colocación) y así sucesivamente.

Aunque los accesos básicos a las redes sociales pueden aportar la tendencia de las opiniones, no pueden responder lo que en definitiva es una cuestión más importante: "¿por qué dice la gente lo que están diciendo y comportándose de la manera que se están comportando?". La necesidad de este tipo de respuesta obliga a enriquecer el acceso a los medios de comunicación social con información adicional y en forma diferente que es probable que residen en múltiples sistemas empresariales. En pocas palabras, es necesaria la analítica de los medios de comunicación social utilizando también los repositorios de datos tradicionales (SAP, DB2, Teradata, Oracle, SAS, etc.). No obstante, es necesario mirar más allá de solo los datos. Hay que observar la interacción de las personas con sus comportamientos, tendencias financieras, transacciones reales y así sucesivamente. Ventas, promociones, programas de fidelización, acciones de mercado e incluso variables tales como el clima pueden ser conductores por los que podemos detectar el comportamiento de los consumidores para poder modelizarlo. Llegar a la base de por qué sus clientes están comportando de una determinada manera requiere tipos de información en forma dinámica y rentable, especialmente durante las fases de exploración inicial del proyecto.

Es un hecho que el análisis de los tweets es un indicador revelador sobre el impacto potencial del sentimiento del cliente sobre los productos. Este tipo de registros es muy elocuente, no solo por el volumen y la velocidad de su crecimiento, sino también porque el sentimiento está siendo expresado para cualquier producto o servicio. Además, todo el mundo es capaz de expresar la reacción y sentimiento en segundos y sin filtros ni trabas geográficas.

Patrones de modelado y gestión de riesgo

El modelado para la gestión de riesgos es otro patrón de aplicación y uso común del Big Data. La crisis financiera de 2008, la crisis de las hipotecas "subprime" asociadas y sus secuelas han hecho del modelado de riesgos y su gestión, un área clave de interés para las instituciones financieras. Como se sabe por los mercados financieros de hoy, una carencia de entender el riesgo puede tener efectos devastadores de creación de riqueza. Además, conocidas las normas reguladoras que afectan a las instituciones financieras en todo el mundo, es necesario asegurarse rápidamente de que los niveles de riesgo caen dentro de límites aceptables.

Como fue el caso en el patrón de detección de fraude, las empresas utilizan entre el 15 y 20 por ciento de los datos estructurados disponibles en sus modelos de riesgo. No es que no se reconozca que hay un montón de datos que están

potencialmente subutilizados, sino que no saben dónde puede encontrarse la información relevante en el resto de los datos. Además, puede ser demasiado caro en la infraestructura actual de muchas empresas analizar a muchos clientes para investigar.

También es típico analizar lo que pasa al final de una jornada bursátil en una firma financiera. Es esencial conseguir una instantánea de sus posiciones a la clausura de la jornada. Instantáneamente, las empresas pueden derivar e identificar su posición financiera usando sus modelos en poco tiempo e informar a los reguladores para el control de riesgos internos.

Dos problemas iniciales se asocian a este patrón de uso de modelado y gestión de riesgo: "¿cuántos datos se van a utilizar para el modelo?" y "¿cuál es la velocidad de los datos?". Desafortunadamente, la respuesta a la segunda pregunta es a menudo difícil. Finalmente, se persigue considerar la tendencia de servicios financieros para mover el modelo de riesgo y ajustar las posiciones del día a día. Este desafío no puede ser resuelto con los sistemas tradicionales. Otra característica de los mercados financieros de hoy es que hay enormes volúmenes de comercio. Si mezclamos los picos de volumen con los requisitos para construir el mejor modelo y gestionar el riesgo adecuadamente con ejecución diaria, tenemos un problema de Big Data delante de nosotros.

Big Data y el sector de la energía

El sector de la energía ofrece muchos retos de casos de uso de Big Data. El problema principal consiste en cómo hacer frente a los grandes volúmenes de datos de los sensores de las instalaciones remotas. Muchas empresas están utilizando solo una fracción de los datos, ya que carecen de la infraestructura necesaria para almacenar o analizar la escala de los datos disponibles.

Tomemos por ejemplo una plataforma de perforación de petróleo típico que puede tener de 20000 a 40000 sensores a bordo. Todos estos sensores están fluyendo los datos sobre la calidad de la plataforma petrolera y otras variables. No todos los sensores están en acción en todo momento, pero algunos están reportando muchas veces por segundo. Se necesita tener una pista sobre qué porcentaje de esos sensores se utilizan activamente, aunque conocer todo el problema sea imposible.

De manera similar los clientes no están utilizando toda la información de datos que están disponibles para ellos en su proceso de toma de decisiones. Por supuesto, cuando se trata de datos de energía, tasas de recaudación o variables

similares, lo que realmente nos preguntamos es si hemos hecho todo lo posible para la captura y el aprovechamiento de la información que se está recopilando.

Con la idea de la ganancia, la seguridad y la eficiencia en mente, las empresas deben estar constantemente en busca de señales y ser capaces de relacionar esas señales con sus resultados potenciales o probables. Si se descarta el 90 por ciento de los datos de los sensores, no es posible que se puedan comprender o modelar las correlaciones existentes.

Big Data en el Call Center

El reto de la eficiencia del centro de llamadas es similar al caso de la detección del patrón de fraude. Al igual que la dinámica apropiada en información de fraude es crítica para los modelos de fraude robustos, en un centro de llamadas si no se gestiona bien la relación entre el tiempo de la resolución de la llamada y la gestión posterior de los patrones de descontento, la información recogida va a perder su valor. Es vital poder aplicar un patrón de respuesta óptima dinámicamente de modo que los desfases de tiempo de respuesta no resulten nocivos. Esta gestión exige el uso de herramientas de Big Data.