IEBS Business School

CURSO 2020-2021

MÁSTER EN BUSINESS INTELLIGENCE Y DATA SCIENCE

MODULO:

Big Data: Captura del dato

Nombre estudiante: Marc Faravelli Rodríguez

No.Matricula: 2047104

E-mail: marcfaravelli@gmail.com

ÍNDICE

1. Diseño y soluciones del caso (Tsara)

En el sector del “Retail”, aun sabiendo que el canal online se utiliza cada vez más y los datos muestran un crecimiento constante en las ventas, es evidente, al menos en los usuarios más habituales que la lealtad y la percepción del valor de la marca son determinadas e influenciadas por las experiencias vividas en las tiendas.

Siguiendo este razonamiento, las tiendas físicas no son solo un punto de partida para la lógica de distribución, sino también para contar la filosofía de la marca y la innovación continua de sus productos, estableciendo una relación directa y exclusiva con los clientes. Por su propia naturaleza, el sector retail debe evolucionar continuamente para adaptarse al crecimiento del tejido conectivo corporativo en el que opera. Eso por tanto cambia y se vuelve eficiente, adquiriendo las características de una lógica omnicanal.

Comprendiendo la necesidad de centrarse no tanto en el producto, sino que, en el consumidor y con el objetivo de mejorar el servicio al cliente, se podría implementar la tecnología de identificación por radiofrecuencia (RFID). De esta manera nos ayudaría a poder responder de forma oportuna y eficiente a las solicitudes de los clientes. Esta tecnología, además de garantizar un mayor control a efectos de seguridad y trazabilidad de la mercancía, gracias a las etiquetas RFID presentes en cada prenda, también nos permitiría extrapolar datos relevantes sobre preferencias de los clientes en las distintas tiendas ubicadas en todo el mundo. Combinando estrategia y tecnología, es posible obtener información relevante en tiempo real sobre múltiples productos comprados o en los que se están agotando en un determinado almacén y deben ser reorganizados. Además, para los clientes que han encontrado artículos en línea que les interesan, pero quieren probarlos en la tienda, se permite saber a qué tienda ir para encontrar el producto de interés, así como las tallas y colores disponibles. La experiencia del cliente adquiere las características de la omnicanalidad, asegurando la continuidad en todo el proceso realizado por el consumidor y optimizando la producción, el tiempo y las inversiones para la compañía.

Con el objetivo de aumentar aún más el compromiso con los clientes durante su estadía en tiendas físicas, se podría desarrollar una aplicación gratuita para dispositivos móviles.

Usando los proprios smartphones, se podría comprar productos sin esperar pagar y escanear los códigos QR, ver su disponibilidad y obtener más información. Gracias a la misma aplicación, nos permitiría analizar los datos que se obtengan de los comportamientos adoptados por los clientes, permitiendo a la empresa Tsara actualizarse y adaptar constantemente las colecciones, elementos indispensables para los departamentos de diseño y producto. Otra herramienta más que iría a complementar la aplicación móvil sería la inserción de funciones de realidad aumentada: al acercar sus smartphones a las ventanas o a unas balizas colocadas en la tienda, sería posible ver modelos vistiendo los artículos en oferta.

Otra implementación diferente en cuanto es una ayuda para el personal de venta seria la tablet o smartphone que les permite interactuar mejor con el cliente, ofreciendo servicios personalizados, accediendo a un sistema CRM que integra y sintetiza todos los datos relacionados con los contactos con ese cliente en particular en los canales de marketing y también facilitando el proceso de compra. Un ejemplo podría ser comprobar la disponibilidad del producto, identificar las tiendas más cercanas, seguir al cliente durante el check-out o entrega a domicilio del producto, evitando las líneas de checkout.

2. Arquitectura para la solución del caso

2.1 Captura del dato

Para definir estrategias de marketing con el objetivo de crear mensajes personalizados y para adquirir nuevos clientes, se deben recopilar datos centrados en el consumidor. Se trata de datos personales, datos de tarjetas de fidelización, datos relacionados con la satisfacción del cliente, datos de servicio al cliente, datos de comportamiento del cliente, datos de compra y preferencias compra, datos de participación del consumidor y datos de marketing digital.

El objetivo de este proyecto es desarrollar una plataforma para Business Intelligence, que permita a Tsara poder monitorear todos los datos de los clientes (compras, preferencias, comentarios, stock…) derivados de la aplicación móvil de la empresa, de

las redes sociales, página web o las tablets de los trabajadores para conocer más las preferencias de los clientes y poder así mejorar su experiencia con la marca.

Además de la variedad de formatos y estructuras, los Big Data que deberemos tener en cuenta para nuestro proyecto, presenta una gran variedad de fuentes. Los datos los podemos clasificar en:

- generado por humanos
- generado por máquina
- generado por el negocio/empresa

Entre las fuentes de datos generados por humanos se encuentran las plataformas de redes sociales (Facebook, LinkedIn, Instagram), blogs de moda, uso compartido de multimedia (YouTube), portales de comercio electrónico (eBay, Amazon), flujos de clics de la página web de Tsara.

Este tipo de fuente también se denomina contenido generado por el usuario, ya que el contenido se produce en gran parte por los usuarios.

Los datos generados por la máquina son producidos por fuentes como RFID y smartphone.

Finalmente, los datos generados por el negocio son todos los datos generados por humanos o máquinas, producidos internamente por la empresa Tsara y relacionados con las actividades basadas en datos de los procesos comerciales corporativos, es decir, aquellas actividades cuyas decisiones se determinan a partir de consideraciones sobre los datos disponibles. Muchos de ellos son datos históricos, almacenados por separado en las bases de datos relacionales del área funcional relevante.

Se podrían adquirir datos útiles de numerosas fuentes internas, como del departamento de marketing, que proporciona información sobre los aspectos demográficos y psicográficos de los clientes, el comportamiento de compra y las visitas al sitio web; servicio al cliente (tablets de los trabajadores en tienda,…) que verifica el nivel de satisfacción del cliente y registra cualquier problema de asistencia; contabilidad, que elabora estados financieros y registra sistemáticamente datos de ventas, costos y flujos

de efectivo; producción, que proporciona datos sobre volúmenes de producción, envíos y existencias.

El hecho de recibir una gran cantidad de datos de diferentes fuentes conlleva también una gran tipología de datos:

- Estructurados: bases de datos de la empresa Tsara
- Semiestructurados: HTML, …
- No estructurados: videos, imágenes, …

Para la ingesta de datos usaremos el Apache Nifi. Este si bien es excelente para transportar información entre varios sistemas, para trabajar en BigData y para importar datos, también es muy interesante porque gracias a la gestión de carga, evita que colapsen los distintos recursos implicados. Por lo tanto, también es útil en casos de grandes cantidades de datos y sobre todo de diferentes tipologías, como es nuestro caso. Por último, tiene una interfaz web que brinda una vista completa de los sistemas y su estado, carga y errores.

2.2 Repositorio de datos

En mi perspectiva, la velocidad del big data requiere tener un data lake en la empresa, que, sin embargo, debe ir acompañado de un data warehouse para analizar la situación general y producir informes de análisis sencillos.

Todo esto para decir que los dos sistemas de almacenamiento tienen objetivos muy diferentes entre sí. Los data lakes nos proporcionan una base de datos que se pueden analizar en tiempo real con diferentes objetivos de análisis cada vez (ej. CRM, redes sociales, webs,…). El data warehouse nos permitiría producir posteriormente informes estandarizados que involucran procesos ETL y de transformación muy complejos y sofisticados.

Dado nuestro objetivo de incrementar las ventas en las tiendas físicas de Tsara, después de un análisis ELT, los componentes y niveles que caracterizan la arquitectura de la

plataforma Business Intelligence, implementada para la monitorización de los datos, se gestionan mediante un proceso ETL.

En general, ETL no es más que un proceso de extracción, transformación y carga de datos en una base de datos integrada, en un Data Warehouse. Los datos que analizaremos se extraen de fuentes OLAP y se basan en el lenguaje de consulta estructurado (SQL), un lenguaje estandarizado que nos permite comunicar, gestionar y administrar las distintas bases de datos, así como leer y modificar los datos. Luego se someten a un proceso de transformación, que consiste, en seleccionar solo los datos de interés para el sistema, eliminar los duplicados, derivar nuevos datos calculados, etc. Esta transformación tiene como finalidad consolidar los datos, es decir, homogeneizar los datos procedentes de distintas fuentes, y asegurarse de que estén lo más cerca posible de nuestra lógica de negocio del sistema de análisis para el que se desarrolla.

La gestión de todos los flujos, que rigen la plataforma de Business Intelligence implementada, se han desarrollado con Knime. Esta es una de las herramientas más poderosas y permite la integración de innumerables procesos de aprendizaje automático y minería de datos. También es particularmente eficaz en el preprocesamiento de datos, por lo tanto, en la extracción, transformación y carga de datos. Gracias a una tubería de datos modular, el software se configura sobre todo como una herramienta de minería de datos orientada al flujo de datos.

2.3 Arquitectura para el procesado

La necesidad de gestionar una gran cantidad de datos, evitando que los tiempos de procesamiento crezcan exponencialmente, lleva al diseño de una solución capaz de maximizar el procesamiento de datos, pero, sobre todo capaz de eliminar los puntos dentro del flujo de procesamiento de datos que, a medida que los datos crecen, podrían ser críticos.

La solución fue diseñada e implementada sobre la base de la plataforma Hadoop. Este programa informático es:

- una de las tecnologías más utilizadas en Big Data;

- utilizado como núcleo en las plataformas de los principales proveedores, como Cloudera, Amazon, Hortonwork, etc;
- utilizado tanto en la nube como en contextos internos

Estas y muchas otras se encuentran entre las principales razones por las que optamos por diseñar y desarrollar una solución utilizando la tecnología Hadoop, analizando sus características y lógica.

Una cuestión para tener en cuenta es que a medida que aumenta la cantidad de datos, hay algunas de estas actividades que se ven afectadas, en términos de rendimiento, lo que impacta negativamente en toda la solución producida. Las actividades en cuestión se refieren al muestreo y la organización de datos por un lado y a la ejecución de algoritmos, por otro. La arquitectura, por lo tanto, ha sido diseñada sobre la lógica MapReduce de Hadoop, de manera que se aproveche el cálculo ofrecido por el clúster de Hadoop.

Con respecto a la entrada y la salida, se podría pensar en utilizar una base de datos, desde la que se alimenta HDFS y almacenar Big Data. Una posible elección podría ser la de Amazon S3, un sistema de archivos distribuido en la nube de Amazon, capaz de escalar a medida que aumentan los datos. Esto podría permitir adoptar una solución capaz de mantener la totalidad de los datos brutos recopilados.