Fundamentos de la Ciencia de Datos
PEC1: ¿Ciencia en los Datos?

Paula Muñoz Lago
paulamlago@uoc.edu
Master en Ciencia de Datos
Universitat Oberta de Catalunya

Octubre 2019

1 Bases de la Ciencia de Datos y diferentes roles

Un dato es un valor obtenido de una observación, que, puesto en contexto, puede darnos una cierta información. Existen varios tipos de datos. Por una parte, los datos simples son datos atómicos, es decir, cadecen de sentido por si mismos y son indivisibles. Un ejemplo puede ser un número, 2. Dicho dato cobra sentido cuando es acompañado de un contexto: 2 manzanas. De esta forma se convierte un dato en información. Los datos estructurados son los que cobran sentido cuando son acompañados de otros datos en una estructura de una o varias dimensiones. Por ejemplo, un archivo en el sistema de archivos de nuestro sistema operativo. En este caso existe una estructura de carpetas y archivos flexible en la que se almacenan nuestros datos. La última fecha de modificación de uno de estos archivos son los llamados metadatos. Finalmente los datos semiestructurados son los generados por humanos.

El valor de los datos para una organización depende de su intencionalidad en cuanto a proyecto futuro. Si su objetivo es convertirse en una organización orientada al dato, tendrán que valorar la limpieza y seguridad de sus datos, además de formar a sus empleados para que desarrollen una metodología analítica y proponer un cambio cultural orientado al dato, entre otras cosas. Si la empresa ya está orientada al dato, es competitiva en el sector, no puede perder de vista la importancia que yace en los datos que dicha organización alberga, ya que conforman el pilar de su posición y continuo desarrollo.

El empleado que explota la información de bases de datos u otras fuentes. Es el encargado de entregar los datos al Ingeniero, quien se ocupa de la plataforma en la que están albergados los mismos. Su tarea de diseñar los cambios que se pueden producir en el sistema, además de implementarlos, en el caso en el que el volumen de datos aumentase. Da orden a los datos desestructurados, y realiza un proceso de limpieza y homogeneización. Finalmente, el Científico de datos, trabaja adaptando los algoritmos utilizados para obtener información a

problemas específicos. Procesa datos "en bruto", es decir, no hace uso de "cajas negras" o programas que procesan los datos para dar un cierto resultado, sin embargo desarrolla sus propios modelos.

El día a día de un Científico de datos se basa en el uso de lenguajes de programación como Python o R. Crean sus propios scripts o programas para realizar el proceso de limpieza y posterior análisis de los datos proporcionados. Si bien es cierto que deben adaptar los algoritmos a los requerimientos del proyecto y del cliente. También elaboran modelos de aprendizaje automático para predecir nuevos datos.

2 Ejemplo de transición hacia una organización orientada al dato

El banco español BBVA ha conseguido una gran mejora en sus ingresos gracias a el proceso de transformación para adaptarse al nuevo entorno de la industria, orientando al dato su modelo de negocio. Siguiendo las tendencias que se están dando en la industria, han implantado dos modelos para convertirse en una organización orientada al dato.

2.1 Implementación del nuevo modelo

Para transformar dicha empresa a una organización orientada al dato se han elaborado dos enfoques.

- *Desde abajo hasta arriba* o *bottom up*. Consiste en formar a los empleados en ciencia de datos, ampliando sus conocimientos técnicos. Concretamente esta medida ha formado a 2000 científicos de datos, además de incorporar a expertos en la materia. También obteniendo los datos de una forma más segura, y estableciendo un mejor modelo de gobernanza de datos, además de actualizando las tecnologías que utilizan en la organización para analizar dichos datos y obtener información de ellos, ya que el auge de estas tecnologías han llevado al aumento de herramientas de análisis de datos, por lo que hay que escoger las que mejor se ajusten a los proyectos de la organización.

- *Desde arriba hasta abajo* o *top down*, focaliza con una visión global los diferentes departamentos e identifican las transformaciones principales que se necesitan para finalmente llegar a ser una organización orientada al dato. Para ello, se ha establecido un modelo de trabajo en la que todas las unidades o departamentos del banco trabajan juntos y así poder abordar proyectos más grandes, aplicando la metodología ágil.

2.2 Características

Para convertirse en lo que denominan "una compañía *data-driven*", en la que todas las decisiones se tomen en base a los datos, previamente han cumplido las

siguientes características.

- Identifican, combinan y gestionan múltiples fuentes de datos. Siempre con el consentimiento de los clientes, han convertido los datos en un activo estratégico.
- Construcción de modelos de análisis avanzado. Además de formar a sus empleados, han contratado a nuevos perfiles expertos en tecnologías Big Data y análisis de datos. Gracias a ello, han podido gestionar la información y datos de sus clientes para aplicar modelos de Inteligencia Artificial, con los que pueden, entre otras cosas, predecir y optimizar resultados.
- Transformación de la organización. Dado el cambio del negocio, han transformado su forma de trabajar, aplicando paulatinamente un cambio cultural orientado al dato. Han hecho uso de la metodología ágil, en la que los equipos construyen su producto desde el *feedback* del cliente, y centran su esfuerzo en entregar en plazo soluciones que funcionen y satisfagan en la actualidad y en el futuro al cliente.

Además de cumplir dichas características, con las cuales podemos afirmar que el banco BBVA es una organización orientada al dato, cuentan con una estrategia clara para el uso y análisis de sus datos.

2.3 Ventajas

La orientación al dato de su modelo de negocio, y en concreto, la fuerza ganada del departamento *BBVA Data&Analytics* han proporcionado al banco una ventaja competitiva a través de la explotación y análisis de sus datos. Han conseguido evaluar al cliente con una mayor precisión para, por ejemplo,predecir el riesgo que supone para la organización darle un crédito.

Gracias a dicha transformación, BBVA trabaja creando productos a partir de los datos, mejorando procesos y tomando decisiones más certeras.

2.4 Perfiles implicados

BBVA ha demandado nuevos perfiles en su proceso de orientar la empresa a un modelo *data-driven*. Para ello, ha formado a parte de sus empleados y ha contratado nuevos. Ha sido el proyecto *Transcendence* el que, internamente, ha sido el motor de la transformación y ha impulsado el personal del departamento de Data&Analitics.

Se han visto involucrados expertos en big data, machine learning e Inteligencia Artificial, que son los denominados *quants*, mientras que el resto de empleados de otros departamentos, que hacen uso de la analítica, son denominados *non-quants*.

3 Integración del departamento financiero a una arquitectura MDM. Evaluación de Madurez e inventario

La implementación de un conjunto de procesos, gobierno, políticas, estándares y herramientas que gestionan los datos maestros de BBVA, en este caso, para el departamento de finanzas, no solo consiste en desarrollar una solución tecnológia.

La implementación de una arquitectura MDM será una pieza vital en la toma de decisiones analíticas, teniendo en cuenta que los datos maestros pueden encontrarse dispersos por cualquier departamento, incluso fuera (clientes). Para que el departamento de IT pueda implementarla con éxito, deberán basarse en la comunicación y conexión entre departamentos, ya que los datos compartidos no pertenecen a un determinado departamento. Deberán determinar a los administradores de los datos, quienes controlaran los valores de los datos de referencia, así como tener en cuenta los *golden records*, es decir, la información más exacta, actual y relevante de las entidades de negocio.

Para implementar dicha arquitectura MDM en BBVA deberán pasar, por lo general, por un total de 11 fases a la hora de implementar un programa de gestión de datos maestros, que resumiremos a continuación.

Deberán **identificar las fuentes de datos maestros** creando un catálogo. Obtendrán fuentes de todos los departamentos, además de fuera de la empresa (clientes). **Identificar productores y consumidores**, identificando quienes generarán los datos maestros y quienes los utilizarán. A continuación deberán **recopilar y analizar los metadatos de los datos maestros**, es decir, la información que acompaña a los datos, como el tipo de datos, los atributos que lo componen etc. Los empleados con conocimiento de los datos fuente y capacees de transformar la fuente de datos en el formato de los datos maestros deben ser **nombrados administradores de los datos** o *Data stewards*. Dado que necesitarán tomar decisiones sobre los datos maestros, deben **implementar un programa y gobierno del dato**.

Llegados a este punto, deberán **desarrollar un modelo de datos maestros**, definiendo los datos maestros (su tipo, rango...), deben ser prudentes en esta fase e incluir todos los atributos de los datos maestros. Por ejemplo, en el caso en el que los datos personales de los clientes se considerasen datos maestros, los atributos serían: nombre, apellidos, edad, fecha de nacimiento, género, estatura y estado civil, entre otros. En este punto en el que se debe definir el modelo, se podrían evitar los atributos "estatura" y "edad" ya que o no son relevantes para la finalidad de la empresa o pueden calcularse a partir de otro dato.

Partiendo de la definición del modelo de dato maestro, deberán **elegir un conjunto de herramientas** que se ajuste a sus necesidades, además de **diseñar una infraestructura** que albergue todos los datos y a la que puedan acceder todos los departamentos de la empresa. A continuación deberán seguir un proceso de **curación de datos**, utilizando las herramientas seleccionadas y

comprobar si se dan excepciones.

Finalmente, para terminar de implantar la arquitectura MDM, tendrán que **modificar los sistemas que producen, mantienen o consumen** datos maestros e **implementar los procesos de mantenimiento**. Una de las claves de MDM es la seguridad de la limpieza de los datos, por lo que han de establecer métodos para encontrar posibles problemas.

3.1 Evaluación de Madurez

Puesto que la empresa BBVA quizás previamente tuviese alguna fase implantada, la evaluación de madurez permite saber en qué nivel de cada fase se encuentra la organización. Existen cinco niveles de madurez en los que puede encontrarse la empresa, respecto a su capacidad de arquitectura, gobierno, gestión, identificación, integración y la gestión de los procesos de negocio.

A continuación analizaremos las capacidades de BBVA para evaluar en qué nivel de madurez se encuentran.

- Arquitectura: Nivel estratégico (5). Dada la información presente en los artículos informativos de la situación actual de BBVA, habiendose convertido en una organización orientada al dato, supondremos que todos sus departamentos tienen total comunicación entre ellos, y que la arquitectura implementada en la empresa distribuye los cambios a todas las fuentes.

- Gobierno: Nivel estratégico (5). La gobernanza de los datos maestros está integrada con otras iniciativas de gobierno de datos como una función más y asegura una comunicación segura entre departamentos.

- Gestión: Nivel proactivoo (4). Todas las instancias de datos maestros tienen una identificación única cuya gestión está integrada en todos los departamentos de la empresa.

- Identificación: Nivel proactivo (4). Todas las aplicaciones disponen de una herramienta de búsqueda, coincidencia y resolución por identificador. La fusión de datos duplicados y la consolidación de los mismos se produce de manera automática a excepción de casos puntuales, en los que los expertos de negocio deberán tomar acción.

- Integración: Nivel estratégico (5). Cuando en BBVA surge un nuevo proyecto y se desarrolla una aplicación nueva, ésta se vincula con los datos maestros y se integra con el ya implementado sistema de gestion de dichos datos.

- Gestión de procesos de negocio: Nivel estratégico (5). La gestión de datos maestros está presente en la gestión de negocio en todos los ámbitos. Además, se usa para el perfilado de productos, proveedores y clientes.

Para concluir, en la Tabla 1 se muestran los niveles de madurez asignados a cada capacidad que debe tener una empresa a la hora de enfrentarse a la gestión de datos maestros.

Arquitectura	Gobierno	Gestión	Identificación	Integración	Gestión de negocio
5	5	4	5	4	5

Table 1: "Niveles asignados a las diferentes capacidades"

3.2 Gestión de Inventario

Para realizar la gestión de inventario, es necesario que previamente se haya creado una lista de datos maestros, combinando datos de varias fuentes, y se hayan limpiado, estén actualizados y ajustados al modelo de datos maestros. Para ello se utilizan herramientas ETL (*Extraction, Transform, Load*) para normalizar y estandarizar los datos, reemplazar los que faltan y generar un mapa de atributos. Por ejemplo, del dato de acceso a la aplicación web por parte de un usuario los atributos podrían ser: nombre, apellidos, cuenta del usuario, día y hora de acceso.

En BBVA gestionan la lista de datos maestros con una combinación continua, es decir, cada departamento tendrá su lista de datos cuyas aplicaciones van modificando y finalmente fusionarán los cambios con la copia maestra. Los datos financieros provendrán del ERP (*Enterprise resource planning*), sistema dedicado a la administración de recursos de toda la empresa. Sin embargo, el departamento de finanzas no solo hará uso de estos datos, también necesitará los datos de clientes, obtenidos del CRM (*Customer relationship management*), sistema que administra la comunicación con los clientes. Así como los datos de los empleados, obtenidos del departamento de recursos humanos y los datos de productos o localización, como los datos de las direcciones de las oficinas de BBVA, extendidas por todo el mundo.

Para que la organización trabaje en un entorno VUCA (*Volatility, uncertainty, complexity, ambiguity*), debe gestionar los datos maestros, incluyendo catálogos de productos en el proyecto MDM.

4 Ciclo de vida de la entidad Cliente añadida en la arquitectura MDM

El ciclo de vida de la entidad cliente está comprendida entre su nacimiento, es decir, cuando el cliente se abre una cuenta en el banco, hasta su fin, cuando deja de estar vinculado a BBVA. En BBVA deben gestionar el ciclo que sigue el dato desde su incorporación hasta su descarte, venta o reciclaje, basándose en políticas para gestionar el camino que sigue el dato a través de los sistemas de la empresa.

El ciclo de vida de los datos de los clientes pasan por diferentes fases:

1. **Captura del dato**. Adquiridos mediante la introducción de datos, cuando un cliente se abre una nueva cuenta a través de Internet o acudiendo a

una sucursal bancaria, donde un trabajador con permisos para crear o modificar dichos datos pueda hacerlo a través de un software seguro.

2. **Mantenimiento del dato**. Los datos de los clientes antes de esta fase aún no tienen valor para el BBVA, deberán pasar por procesos de integración, para que tengan el mismo formato que otros anteriormente introducidos, limpieza (por ejemplo, si el nombre tiene acentos poco comunes o caracteres no incluidos en el formato del dato) o enriquecimiento (por ejemplo, rellenar campos que falten, como confirmando el correo electrónico del cliente, mandándole un email o completar el campo edad con la fecha de nacimiento). También deberán pasar por procesos de extracción, transformación y carga.

3. **Síntesis del dato**. Aplicando un proceso de lógica inductiva, la empresa dará valor a los datos de Clientes previamente adquiridos, haciendo uso de otros datos. Por ejemplo, de podría deducir la probabilidad de que el nuevo cliente recientemente introducido cometa un fraude, es decir, la seguridad de dar un préstamo a ese cliente induciendolo de datos de préstamos bancarios anteriores.

4. **Uso del dato**. En esta fase el dato se convierte en beneficio para la entidad, y puede ser utilizado con diversos fines.

5. **Publicación del dato**. Se puede publicar de forma interna o externa, teniendo en cuenta de que si se publica fuera de la organización, no podrá recuperarse para ser modificado. Por ejemplo, BBVA publica el crecimiento económico del año pasado en cifras. Puede hacerlo tanto para sus trabajadores, como para publicitarse.

6. **Archivado de datos**. Cuando ya se les ha dado uso a los datos, deben ser almacenados en un entorno seguro en caso de que se vuelvan a necesitar.

7. **Eliminación del dato**. Cuando el cliente deje de tener una cuenta en BBVA, sus datos serán destruidos del sistema.