



POLITÉCNICA
"Ingeniamos el futuro"

CAMPUS
DE EXCELENCIA
INTERNACIONAL



Graduado en Ingeniería Informática

Universidad Politécnica de Madrid

Escuela Técnica Superior de
Ingenieros Informáticos

TRABAJO FIN DE GRADO

Gestión de la calidad de los datos en la empresa

Autor: Alejandro Jiménez Martín

Director: Ernestina Menasalvas Ruiz

MADRID, JULIO 2017

Agradecimientos

En primer lugar, me gustaría dar las gracias a mi tutora, Ernestina, por el interés mostrado en ayudarme con el desarrollo de este proyecto y por esas conversaciones vía “Skype” desde otros puntos del planeta para resolver todas las dudas que tenía. Recalcar que sin tu ayuda esto no hubiese sido posible. De verdad, GRACIAS.

A mis amigos, gracias por todos los momentos vividos a lo largo de estos años (fiestas, vacaciones, risas, más risas...). Espero que sigamos añadiendo momentos durante mucho tiempo.

Millones de gracias a mi familia, en especial a mis padres y mi hermana. Ellos son la verdadera razón del por qué estoy aquí. Son los responsables de este éxito. Han soportado mis alegrías, mis penas, mis enfados, mis risas, etc. Siempre han sabido motivarme cuando las cosas no salían bien y disfrutar de mis aprobados como si fuesen suyos. Dicen que la familia no se elige y qué orgulloso estoy de que me haya tocado esta. ¡Ya veo la luz al final del túnel!

Por supuesto no puedo olvidarme de mi cuñado Pablo y de lo más bonito de este mundo, mi sobrino. Ellos también han estado siempre ahí en todo momento, uno más tiempo que otro, y me han ayudado a sacar esto adelante. ¡Gracias por todo!

GRACIAS creo que es una palabra que se queda corta para agradecer a la persona que ha estado SIEMPRE apoyándome, dándome consejos, sacándome una sonrisa cuando más lo necesitaba, alegrándose de mis éxitos, sufriendo con mis fallos, etc. En definitiva, gracias por cruzarte en mi camino aquel 7 de noviembre... ¡TE ADORO AMANDA!

Por último, agradecer a aquellas personas que forman parte de mi vida y que, aunque no estén aquí conmigo para disfrutarlo, sé que lo celebrarán allí donde estén... Gracias abuela Lola por enseñarme que la vida se basa en disfrutarla y ser feliz y, a ti, Josefa, mi todo, GRACIAS por haberme dado la oportunidad de conocer a la persona más increíble que he conocido nunca. Cuidaste de mí como sólo tú podías hacerlo. Tu esfuerzo, paciencia, alegría y ganas de vivir me han ayudado a seguir avanzando y sonreír cada mañana. Os estaré muy agradecido toda la vida. ¡Nos vemos!

Resumen

El crecimiento exponencial de los datos en los últimos años es una realidad que hay que afrontar. Según el informe IDC de este último año (2016), se estima que el mercado de datos en 2020 tendrá un valor de 111 mil millones de euros por lo que es necesario implantar nuevas tecnologías que aseguren una rápida adaptación y proporcionen valor a la sociedad.

Hoy en día, disponer de datos de calidad resulta imprescindible en cualquier organización para tomar decisiones adecuadas por lo que es necesario gestionar los datos de forma correcta. Por ello, el uso de herramientas que ayuden a mantener la consistencia, integridad, completitud y precisión de los datos permiten que la calidad y el gobierno de los datos formen parte de la estructura empresarial lo que supone un crecimiento sustancial económico y competitivo.

A pesar de la importancia que presenta la gestión de los datos en una organización, numerosas empresas no lo han afrontado todavía lo que está dando lugar a grandes problemas no sólo a nivel operativo sino también a nivel institucional.

El presente proyecto recoge el análisis acerca de la gestión de la calidad de los datos de las empresas en la actualidad. Una vez analizada la problemática existente provocada por la gestión de datos y el soporte en los procesos decisionales de las organizaciones, este trabajo presenta el desarrollo de una posible solución para afrontar alguno de los retos analizados.

Abstract

The exponential growth of data in recent years is a reality that must be realized. According to the IDC report for the last year (2016), it is estimated that data market in 2020 will have a value of 111 billion euros, so it is necessary to implement new technologies that ensure quick adaptation and provide value to society.

Currently, having quality data is essential in any organization to make adequate decisions so it is necessary to manage data correctly. For this reason, the use of tools that maintain the consistency, completeness, and accuracy of the data allow the quality and governance of the data to be part of the business structure, which means substantial economic and competitive growth.

Despite the importance of data management in an organization, many companies do not pay attention to this area which gives rise to major problems not only at the operational level but also at the institutional level.

This project includes the analysis about data quality management of the companies at present. After analyzing the existing problems caused by data management and support in decision-making processes organizations, this document presents the development of a possible solution to address some of the challenges analyzed.

Índice

Capítulo 1: Introducción.....	1
1.1 La importancia de la calidad de datos en la empresa.....	1
1.2 Motivación y objetivos	4
1.3 Estructura del trabajo	5
Capítulo 2: Estado de la cuestión	7
2.1 Introducción	7
2.2 Big Data y las Vs del Big Data	8
2.3 Cadena de valor del dato.....	9
2.4 ETL	11
2.5 Herramientas ETL.....	14
2.5.1 Informática PowerCenter.....	15
2.5.2 Pentaho Kettle	16
Capítulo 3: Planteamiento del problema	17
3.1 Requisitos planteados	19
Capítulo 4: Solución.....	21
4.1 Situación inicial	21
4.2 Diseño de la solución.....	23
4.3 Desarrollo.....	26
4.3.1 Calidad de Datos.....	26
4.3.2 Data Governance	34
4.3.3 Desarrollo de la ETL	37
4.4 Evaluación y validación.....	44
4.4.1 Lanzamiento de pruebas	44
4.4.2 Establecimiento de requisitos	45
Capítulo 5: Conclusiones y líneas futuras	46
5.1 Conclusiones	46
5.2 Líneas futuras.....	47
Bibliografía.....	48

Capítulo 1: Introducción

1.1 La importancia de la calidad de datos en la empresa

En la actualidad, en cualquier organización, institución u organismo público, trabajar con datos de calidad resulta imprescindible para realizar una buena toma de decisiones. Sin embargo, seguimos haciéndonos la siguiente pregunta ¿las empresas le dan la importancia suficiente al análisis de la calidad del dato? ¿Se invierte para conseguir datos de calidad?

El continuo crecimiento de los datos junto a la facilidad de acceso a los mismos y el uso de potentes sistemas TIC han dado lugar a la realización de actividades complejas a partir de poderosas herramientas que permiten recoger, almacenar, analizar, procesar y visualizar grandes cantidades de datos. El mercado de datos en la actualidad es un negocio de miles de millones de euros que está en continuo crecimiento. Según el informe IDC de este último año (2016)[10], se estima que el mercado de datos en 2020 tendrá un valor de aproximadamente 111 mil millones de euros por lo que la explotación de grandes volúmenes de datos en diversos sectores presentará un especial potencial socioeconómico. Por ello, es necesario adoptar nuevas tecnologías, aplicaciones, casos de uso y modelos de negocio entre varios sectores y dominios que aseguren una rápida adopción por parte de los individuos y proporcionen un importante crecimiento y competitividad.

Esta evolución es una realidad que las empresas deben afrontar y explotar de una manera estructurada, agresiva y ambiciosa para crear valor. Las actividades económicas y sociales han dependido por mucho tiempo de los datos, pero el aumento del volumen, velocidad, variedad y valor social y económico de los datos señalan un cambio de paradigma hacia un modelo de sistema socioeconómico.

Por ello, la información se ha transformado en un recurso clave para las organizaciones lo que supone que el análisis y gestión de los datos se hayan convertido en un factor imprescindible. En un mercado cada vez más competitivo, toda empresa debe tener claro cuál es su principal objetivo (inversiones, reducción de problemas, etc.). Todo ello se puede conseguir con una buena toma de decisiones teniendo en cuenta la información recibida.

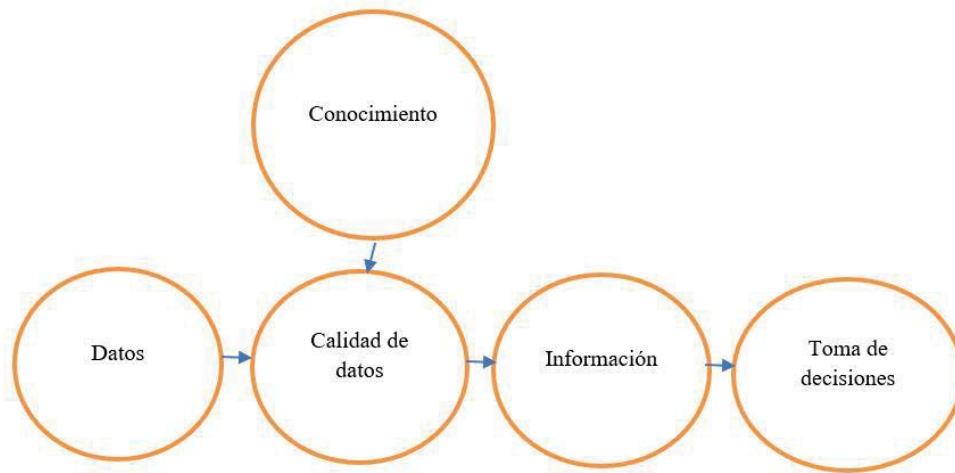


Figura 1: Calidad del dato

Cabe destacar que la calidad del dato depende del objetivo que busque cada organización. Mientras que para una empresa un dato puede resultar de especial interés, para otra resulta insignificante, por lo que es fundamental que los datos a tratar sean útiles y que el objetivo que persigan sea el mismo ya que de este modo, conseguiremos que los datos puedan ser un valor de la organización al poder ser usados como soporte de sus procesos decisionales. Si somos capaces de lograrlo, aumentaremos la escalabilidad de los mismos y obtendremos un ecosistema único que permita que el dato sea visible para todos. Por el contrario, el hecho de no alcanzar el objetivo puede conllevar a los siguientes problemas:

- Incremento de los costes en los procesos.
- Impacto fuerte en el grado de competitividad de la organización.
- Errores de información, análisis e investigación.

Uno de los objetivos en un proceso de calidad de datos es determinar su consistencia y homogeneidad. Para que este objetivo se cumpla, los analistas deben encontrar la inconsistencia del dato, aislarla, realizar un análisis y proporcionar una solución al mismo mediante la utilización de una herramienta o empleando un proceso para el uso de datos de forma adecuada.

A pesar de todo, la gestión de la información es uno de los primeros proyectos que se quedan apartados.

Existen varios motivos por los cuales las empresas no invierten en la calidad del dato:

- **Compleitud de datos:** Las organizaciones consideran que los datos que presentan son suficientes, sin embargo, muchos estudios defienden lo contrario.
- **Miedo a lo desconocido:** Las empresas no saben cómo afrontar un proyecto de calidad de datos. En este punto, es fundamental que deleguen este proyecto en un servicio de consultoría para poder integrar los datos de la manera más adecuada posible en los sistemas de los clientes.
- **Conocimiento insuficiente del cliente:** A pesar de tener un gran volumen de datos, las empresas no cuentan con toda la información relevante de sus clientes lo que da lugar a numerosos problemas y sus posteriores conflictos internos.
- **Necesidad de datos:** Es otro de los motivos más comunes. Se piensa que los datos no se necesitan. Esto es un pensamiento totalmente erróneo. Comprobar la completitud, consistencia y detección de errores de los datos puede ayudar en la toma de decisiones y, lo que es más importante, ahorrar mucho dinero a la entidad.
- **Problemas económicos:** Muchas empresas tienen la idea de que invertir en calidad del dato supone una inversión muy costosa. Sin embargo, esto no es así. Actualmente existen herramientas asequibles que implementan soluciones en poco tiempo y con una efectividad muy alta.

En definitiva, el análisis de la calidad de los datos es una tarea fundamental en el desarrollo de una organización para poder alcanzar todo el potencial del mismo.

1.2 Motivación y objetivos

Todo lo anteriormente expuesto ha motivado el presente trabajo fin de grado. De esta manera abordaremos la problemática de la gestión de los datos en las organizaciones. El origen del problema viene causado en parte por la naturaleza de los datos que en ocasiones no es del todo clara lo que da lugar a la necesidad de realizar un análisis completo sobre la procedencia, métodos de adquisición, usos, etc. lo que provoca una pérdida de tiempo considerable a la hora de buscar una solución concreta y lo que es peor, grandes problemas internos de comunicación y entendimiento ya no sólo a nivel departamental sino a nivel organizacional. En la actualidad, los datos crecen exponencialmente por lo que, si no controlamos esto y lo gestionamos correctamente, surgirán muchos problemas en un futuro no muy lejano que pondrán en peligro a la organización que genera, gestiona y usa estos datos y en última instancia a la sociedad por las consecuencias que estos pueden tener.

En cuanto a los objetivos que se buscan con la realización de este proyecto son los siguientes:

- Análisis de los problemas de gestión de calidad de datos en una empresa.
- Establecimiento de una posible solución a la problemática planteada.
- Análisis y evaluación de los resultados obtenidos.
- Establecimiento de conclusiones a partir del desarrollo del proyecto y explicación de líneas futuras que ayuden a mejorar la calidad del dato y, por consiguiente, el beneficio a una empresa.

1.3 Estructura del trabajo

En este apartado se mostrarán los capítulos de los que se compone este proyecto:

- **Introducción:** En este primer capítulo se ha establecido la importancia que presenta la calidad del dato, así como los objetivos a cubrir en la realización de este proyecto.
- **Estado de la cuestión:** En el segundo capítulo se muestran cuáles son los puntos destacados del problema a resolver además de las diferentes herramientas utilizadas para su tratamiento.
- **Establecimiento del problema:** En el tercer capítulo se detalla el planteamiento del problema, realizando un análisis para su posterior solución.
- **Solución:** En el cuarto capítulo se plantea la solución llevada a cabo y se establecerá un análisis y evaluación de resultados obtenidos.
- **Conclusiones y líneas futuras:** En este último capítulo se determinarán las conclusiones alcanzadas tras la realización del proyecto, así como las posibles líneas futuras.

Capítulo 2: Estado de la cuestión

2.1 Introducción

Hoy en día, la mayor parte de las empresas necesitan obtener información acerca de sus datos, los cuales deben ser fiables y precisos de tal forma que permitan un análisis exhaustivo y, por tanto, válido. Sin embargo, esto no sucede así. La mayoría de las empresas consideran que los datos que obtienen de sus clientes no son del todo precisos lo que afecta a sus bases de datos. Por ello, la calidad de datos es fundamental para establecer valores de negocio. Las empresas que se encargan de analizar millones de datos a través de un proceso de calidad presentan una ventaja competitiva frente a aquellas que no lo hacen.

Existen muchos problemas que dificultan el empleo de calidad de datos:

- Empleo de diferentes bases de datos para una misma operación.
- Falta de entendimiento sobre la implementación de reglas que deben utilizarse para incorporar los datos a los sistemas.
- Falta de procedimientos internos de procesamiento y estandarización de datos.
- Inexistencia de un único responsable de los datos.

A pesar de los riesgos que supone no disponer de un proceso de calidad de datos, muchas empresas reconocen que no desarrollan un plan de calidad hasta que no se producen problemas específicos. Todo ello provoca conflictos tanto internos como externos ya que si la empresa espera a que sus empleados encuentren un problema, la efectividad del análisis y de la toma de decisiones ya se han visto afectados mientras que, si son los clientes los que detectan el problema antes que los empleados, esto provoca la insatisfacción de los mismos lo que determina la falta de confianza en la propia imagen de la empresa.

El principal objetivo para establecer una buena calidad de datos pasa por crear equipos de personas capaces de desarrollar una estrategia de gestión de datos, implementarla y asegurar su cumplimiento por parte de la empresa; crear procesos que muestren que la gestión de datos funciona correctamente y emplear la tecnología para evitar el error humano y de esta forma, reducir la inconsistencia.

2.2 Big Data y las Vs del Big Data

Existen numerosas definiciones acerca del concepto Big Data, sin embargo, muchas de ellas presentan una gran confusión con respecto a su significado. Dan Ariely en su obra [1] comentó: *“Big Data es como el sexo en la adolescencia: todo el mundo habla de él, nadie sabe cómo hacerlo, todos creen que los demás lo están haciendo y, claro, todos dicen que lo hacen.”*

Por ello, es conveniente tener claro cuál es su verdadero significado. Una de las definiciones más claras acerca del término Big Data viene dada por la consultora Gartner [6] *“Big Data es un recurso de información de gran volumen, alta velocidad y / o alta variedad que exige nuevos métodos de procesamiento de información rentables e innovadores que permiten una mejor comprensión, toma de decisiones y automatización de procesos.”*

Este concepto define las características principales que presenta o debe tener una arquitectura de Big Data, las denominadas “Vs del Big Data”. Son las siguientes:

- **Volumen:** Su principal objetivo es tratar grandes cantidades de datos. Recoger y organizar absolutamente toda la información que nos llega es esencial para tener registros completos, y que las conclusiones que obtengamos sirvan eficientemente a la hora de tomar decisiones.
- **Velocidad:** Es una característica fundamental ya que el tiempo de reacción ante un problema debe ser inmediato. Cuando tratamos con grandes cantidades de datos, reaccionar rápidamente a un problema resulta imprescindible para evitar problemas mayores.

- **Variedad:** Muestra la diversidad que presentan los datos y sus diferentes fuentes de obtención de los mismos. Esto provoca un aumento de complejidad tanto en su almacenamiento como en su procesamiento y análisis. Por tanto, el objetivo reside en conjugar y combinar cada tipo de información y su tratamiento específico para alcanzar un conjunto homogéneo.
- **Veracidad:** Es el grado de confianza que se establece a los datos a emplear. Esta característica determina la calidad de los resultados por lo que su estudio resulta imprescindible para cualquier compañía.
- **Valor:** Se debe dar valor a la sociedad, a las empresas y a los gobiernos, en definitiva, a las personas; todo el proceso tiene que ayudar a impulsar el desarrollo, la innovación y la competitividad, pero también mejorar la calidad de vida de las personas.

2.3 Cadena de valor del dato

Una vez explicado qué es el Big Data y cuáles son sus características, vamos a definir el concepto de cadena de valor del dato.

Porter en su obra [21] define la cadena de valor como la *“herramienta de ayuda a la decisión para modelar la cadena de actividades que una organización realiza con el fin de entregar un producto o servicio valioso para el mercado”*

La cadena de valor categoriza las actividades principales de una organización con el fin de comprender el funcionamiento de las mismas y optimizarlas en caso necesario. Una cadena de valor se compone de una serie de subsistemas cada uno con entradas, procesos de transformación y salidas. Como herramienta analítica, la cadena de valor puede ser aplicada a flujos de información para comprender la creación de valor en tecnología de datos. Se usa para modelar actividades de alto nivel en sistemas de información.

Existen 5 fases en la cadena de valor del dato. Son las siguientes:

2. Estado de la cuestión

- **Data Acquisition:** La adquisición de datos es el proceso de recolección, filtrado y limpieza de datos antes de que se almacene en una base de datos o cualquier otra solución de almacenamiento en la que pueda llevarse a cabo. La adquisición de datos es uno de los grandes desafíos en términos de requisitos de infraestructura. La infraestructura debe ser predecible tanto en la captura de datos como en la ejecución de consultas, ser capaz de manejar volúmenes de transacción muy altos y apoyar las estructuras de datos flexibles y dinámicas.

- **Data Analysis:** El análisis de datos se ocupa de transformar los datos en bruto adquiridos en útiles para la toma de decisiones. El análisis de los datos implica la transformación y el modelado de datos con el objetivo de resaltar los datos pertinentes, sintetizarlos y extraer información oculta útil con alto potencial desde un punto de vista empresarial.

- **Data Curation:** Es la gestión activa de datos a través de su ciclo de vida para asegurarse de que cumple con los requisitos necesarios de calidad de datos para su uso efectivo. Este proceso se puede clasificar en diferentes actividades tales como la creación de contenidos, selección, clasificación, transformación y validación. Los responsables de llevar a cabo dicho proceso son los anotadores de datos. Son los encargados de mejorar la accesibilidad y calidad del dato y asegurar que los datos son confiables, visibles, accesibles y reutilizables.

- **Data Storage:** Es el proceso de gestión de datos de forma escalable cuyo objetivo es satisfacer las necesidades de las aplicaciones que requieran un acceso rápido a los datos.

- **Data Usage:** Abarca las actividades basadas en el acceso a datos, su análisis y las herramientas necesarias de integración para introducir los datos en la empresa. El uso de datos en la toma de decisiones de negocio puede mejorar la competitividad mediante la reducción de costes, el aumento de valor añadido, o cualquier otro parámetro que pueda ser medido con criterios de rendimiento existentes.

En la siguiente imagen recogida de [2] podemos ver la estructura que presenta la cadena de valor del dato:

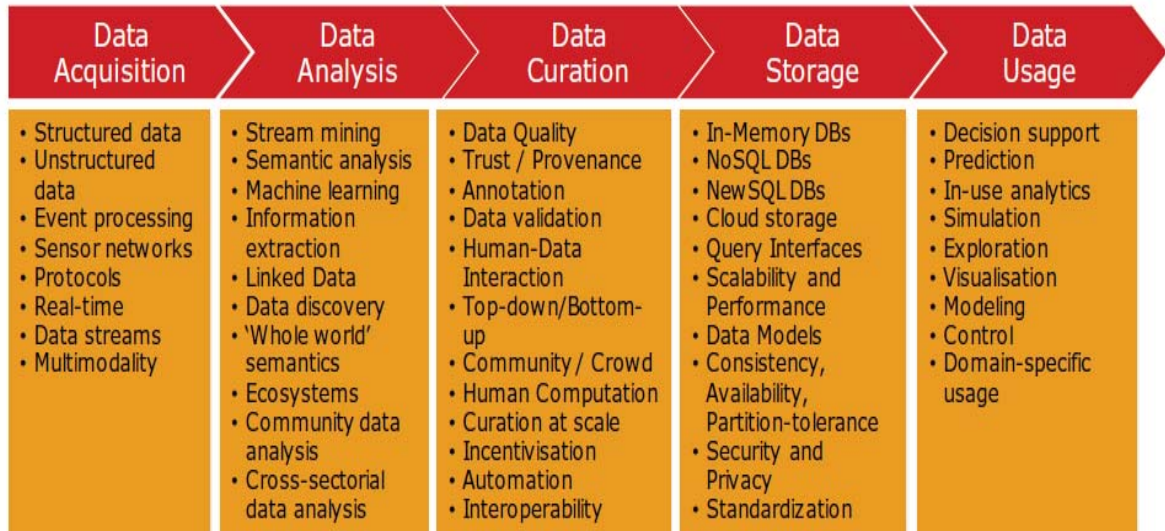


Figura 2: Cadena de Valor del Dato

2.4 ETL

El concepto de ETL proviene de las siglas en inglés “Extract, Transform and Load”. Se trata de un proceso del “Data Warehouse” responsable de extraer datos de sistemas fuente y colocarlos en un almacén de datos. Este proceso permite a las empresas pasar datos provenientes de otras fuentes, transformarlos y cargarlos en otras bases de datos, data warehouse o data mart para analizarlos o enviarlos a otro sistema operacional para ayudar a su proceso de negocio.

Extracción: Se trata del primer paso de una ETL. Consiste en extraer datos de los sistemas origen y hacerlos accesibles para su posterior procesamiento. El objetivo principal es recuperar todos los datos requeridos del sistema fuente con el menor número de recursos posibles. Para ello, debe diseñarse de manera que no afecte negativamente al sistema fuente en términos de rendimiento, tiempo de respuesta o cualquier tipo de bloqueo.

Existen varias formas de realizar la extracción:

- **Notificación de actualización:** si el sistema fuente es capaz de proporcionar una notificación de que se ha cambiado un registro y describir el cambio, esta es la forma más fácil de obtener los datos.
- **Extracción incremental:** algunos sistemas no son capaces de notificar que se ha producido una actualización, pero son capaces de identificar qué registros han sido modificados y proporcionar una extracción de dichos registros.
- **Extracción completa:** algunos sistemas no son capaces de identificar qué datos se han cambiado, por lo que este tipo de extracción es la única manera de obtener los datos del sistema. Requiere guardar una copia de su última extracción en el mismo formato para poder identificar los cambios.

Transformación: Es el segundo paso de un proceso ETL. Este proceso establece un conjunto de reglas de negocio sobre los datos extraídos en la fase anterior para posteriormente, convertirlos en datos con una estructura común para su procesamiento y análisis. Para ello, es necesario que los datos empleen el mismo formato por lo que es en este paso donde resulta imprescindible realizar este proceso para evitar duplicidades o impedir la conexión entre las diferentes fuentes.

En algunos casos, es necesario realizar una manipulación de datos o establecer alguna transformación como:

- Selección de determinadas columnas.
- Codificación de valores.
- Generación de campos clave en el destino.
- Unión de datos de múltiples fuentes.
- Realización de operaciones.
- Traducción de códigos.

Existen muchas más y su aplicación dependerá de la finalidad de cada caso.

2. Estado de la cuestión

Carga: Este es el último paso de un proceso ETL. Se basa en recoger los datos procedentes del proceso de transformación y cargarlos en el sistema destino. Este proceso varía en función de la finalidad que busque la empresa. Además, es importante asegurarse de que la carga se realiza correctamente y lo más rápido posible. Para ello, se debe desactivar cualquier restricción antes de la carga y habilitarlos de nuevo tras completar la carga.

Existen dos formas de cargar datos en un proceso ETL:

- **Carga simple:** Se basa en establecer un resumen de todas las transacciones realizadas durante un periodo determinado de tiempo y enviar dicha información como una única transacción hacia la fuente destino almacenando su valor calculado. Es el método más sencillo de emplear en un proceso de carga.
- **Rolling:** A diferencia del anterior, en este caso se almacena la información resumida en distintos niveles asociados a un instante de tiempo determinado o distintos niveles jerárquicos de alguna o varias dimensiones de magnitud almacenada.

En la siguiente imagen recogida de [19] podemos ver la estructura general de una ETL:

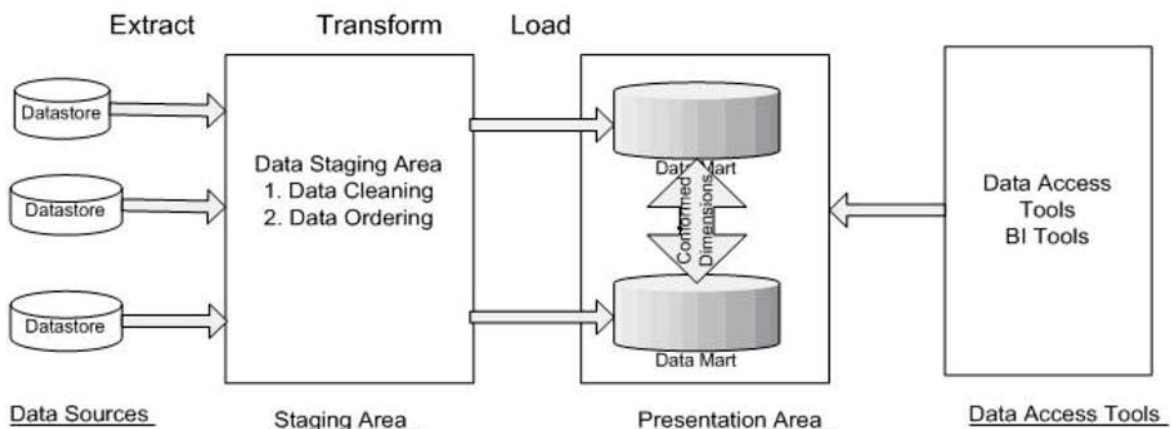


Figura 3: Esquema ETL

2.5 Herramientas ETL

Las herramientas ETL ayudan a las organizaciones a ahorrar dinero y tiempo a la hora de desarrollar un Data Warehouse. Existen numerosos motivos por los cuales las empresas requieren del uso de estas herramientas, sin embargo, pueden resumirse en tres puntos:

- **Rendimiento:** A pesar de que la codificación manual requiere horas de desarrollo para la creación de un Data Warehouse eficiente, este sigue siendo el método de integración de datos más utilizado.
- **Problemas de conexión:** Resulta muy complicado conectar diferentes sistemas de bases de datos sin usar una herramienta externa.
- **Manejo de actualizaciones:** Si se produce una modificación en la base de datos o una integración de datos, gran parte del código manual deberá rehacerse.

Además de ser utilizadas para el desarrollo de sistemas Data Warehouse, son útiles para otros objetivos como:

- Migración de datos de diferentes aplicaciones.
- Actualización de usuarios a sistemas paralelos.
- Sincronización de diferentes sistemas operacionales.
- Consolidación de grandes volúmenes de datos para mantener históricos o para crear procesos de borrado en los sistemas origen.

Existen numerosas herramientas ETL, pero nos centraremos básicamente en dos:

2.5.1 Informática PowerCenter

Informática PowerCenter [20] es la plataforma ETL líder del mercado. Permite acelerar y desarrollar procesos de integración de datos en proyectos de Business Intelligence, Data Governance, etc. Las características principales que convierten a esta herramienta en una referencia a nivel mundial son las siguientes:

- **Confianza operacional:** Establece pruebas de validación de datos automatizadas, repetibles y auditables que advierten cuando se producen problemas en desarrollo y en procesos operativos de tal forma que se puedan identificar rápidamente y solucionarlos antes de que empeoren y puedan provocar problemas más graves.
- **Agilidad:** Se emplean funciones que comparten metadatos comunes que ayudan a la colaboración entre usuarios de negocio e IT.
- **Reutilización:** Permite aprovechar las opciones de transformación preintegradas lo que ayuda a mejorar el rendimiento y efectividad de los usuarios de negocio.
- **Autonomía:** Permite mejorar la agilidad del proceso.
- **Escalabilidad:** Esta es una de las características más importantes que diferencia a esta herramienta del resto. Es capaz de soportar grandes volúmenes de datos gracias a su interfaz gráfica, particionado dinámico, balanceo de carga adaptable, alta disponibilidad y optimización “push down”.
- **Datos en tiempo real:** Esta característica resulta de vital importancia en cuanto a la eficiencia operativa de negocio.
- **Metadatos:** Esta herramienta proporciona gráficas completas que ayudan a mejorar el gobierno, la auditabilidad y la gestión de cambios.
- **Conectividad universal:** Ofrece acceso a los datos desde cualquier tipo de fuente mediante conectores de alto rendimiento.

2.5.2 Pentaho Kettle

Es una herramienta de Pentaho [4] responsable de los procesos de extracción, transformación y carga (ETL). Para poder trabajar en ello, se emplea un diseñador gráfico de transformaciones específico denominado “Spoon”.

Spoon es una interfaz gráfica de usuario que permite el diseño de transformaciones, las cuales pueden ser ejecutadas a través de herramientas de Kettle [4] (Pan y Kitchen):

- **Pan:** motor de transformación que se encarga de la lectura, manipulación y escritura de datos desde la fuente origen hacia la fuente destino. En definitiva, es el encargado de ejecutar las transformaciones diseñadas por Spoon [4].
- **Kitchen [4]:** programa encargado de ejecutar los trabajos diseñados por Spoon en XML o en un catálogo de base de datos.

Las características principales que presenta Pentaho Kettle son las siguientes:

- Empleo de tecnologías estándar como Java, XML y JavaScript.
- Facilidad de instalación y configuración.
- Servicio multiplataforma (Windows, Macintosh, Linux).
- Presenta un entorno gráfico de desarrollo.
- Basado en dos tipos de objetos: Transformaciones y trabajos (colección de transformaciones).
- Facilidad de uso: Cada proceso se crea con una herramienta gráfica donde se especifica qué hacer sin escribir código para indicar cómo hacerlo.
- Orientado a metadatos.
- Aplicación independiente.
- Soporte de diferentes formatos de entrada y salida (archivos de texto, hojas de datos, motores de bases de datos, etc.).
- Manipulación de datos sin limitaciones.

Capítulo 3: Planteamiento del problema

En el departamento de bases de datos de una empresa dedicada al sector financiero se ha determinado que es necesario reforzar la orientación al cliente a partir de sus datos de tal forma que la organización pueda mantener una visión global e integral del mismo y mejorar su marco de negocio. Hasta ahora, esta empresa ha ido almacenando en su base de datos los registros de sus clientes de forma gradual sin tener en cuenta ningún tipo de condición o regla que ayude a establecer un análisis fiable acerca del mismo lo que ha provocado las siguientes debilidades y riesgos para la misma:

- **Rendimiento:** Rendimiento ineficiente de sus procesos (excesivo tiempo de proceso y elevado consumo de recursos).
- **Negocio:** Problemas existentes en cuanto a tratamiento de clientes para ser aprovechados por las unidades de negocio.
- **Eficiencia:** Información obsoleta y falta de conocimiento. Esto provoca un mayor grado de complejidad y un coste excesivo de recursos.
- **Competitividad:** El hecho de no presentar información útil sobre sus clientes da lugar a que no pueda competir frente a otras empresas que sí disponen de datos de calidad lo que supone una pérdida económica importante para la empresa e incluso la disolución de la misma.
- **Confianza:** No disponer de datos consistentes implica la insatisfacción por parte del cliente y, consecuentemente, una posible marcha del mismo manchando la imagen de la organización.
- **Entendimiento:** La falta de información precisa da lugar a problemas de comunicación no sólo a nivel externo sino también a nivel interno.

3. Planteamiento del problema

En conclusión, esta empresa dispone de datos inconsistentes sobre sus clientes lo que provoca que los ejecutivos de la misma no puedan realizar un análisis exhaustivo en busca de soluciones y, por consiguiente, tomar decisiones de negocio correctas que ayuden a la economía de la organización. Este hecho da lugar a problemas potencialmente peligrosos que ponen en riesgo a la entidad.

Por ello, es fundamental que la organización disponga de un sistema de análisis de datos capaz de mostrar datos útiles que ayuden a su análisis y posteriormente ayuden a establecer los siguientes puntos:

- Conocer mejor a sus clientes.
- Disponer de información más fiable sobre el cliente (controlar mejor los campos calculados, unicidad de criterios y racionalización).
- Aumentar la capacidad analítica, disponer de información histórica y facilitar la explotación de información para que sea ágil su respuesta.
- Obtener el valor del cliente.
- Disponer de información normalizada de los datos de contacto de clientes.

Si somos capaces de cumplir con los objetivos marcados, los beneficios para la empresa serán considerables ya que se ganará en competitividad, fidelidad del cliente y productividad dando lugar a un mayor control en calidad de datos, identificación rápida de información errónea, ahorro de costes, explotación de datos para gestión del conocimiento y mejor toma de decisiones.

3.1 Requisitos planteados

Los requisitos previos planteados para hacer frente a la problemática establecida son los siguientes:

- **Perfilado:** El objetivo que se busca es obtener datos fiables que ayuden a identificar los orígenes de información de datos incorrectos e incompletos como punto de partida para establecer una toma de decisiones correcta acerca de los datos de los clientes.

- Conocer la calidad de los datos en la empresa a través de un análisis previo, así como identificar el origen de los datos.
- Analizar los campos “nombre”, “apellidos”, “DNI”, “teléfono”, “email”, “dirección” y “sexo”.

- **Estandarización y normalización:** Disponer de información normalizada de los datos de contacto de los clientes.

- Estandarización de los campos “nombre”, “apellidos”, “DNI”, “teléfono”, “email” y “dirección”.
- Verificación del campo “sexo”.
- Normalización y actualización de teléfono de contacto, email y dirección.

- **Enriquecimiento:** Permitir obtener una mejor visión del cliente.

- Obtener información procedente de fuentes externas que ayuden a comprobar su situación actual (situación laboral, actividad profesional).
- Ampliación del modelo de información para incorporación de nuevas bases de datos con sus correspondientes procesos de carga (ETL).

3. Planteamiento del problema

- **Deduplicación:** Tener un identificador único de cliente.

- Definición de estrategia de deduplicación óptima de acuerdo a la tipología de datos obtenida como resultado del perfilado.
- Cuando un cliente se dé de baja, ese identificador quedará bloqueado para que no pueda ser usado de forma que el alta de clientes sea autoincremental.

- **Agrupación de clientes:** Permitir identificar otras relaciones de clientes.

- Creación de nivel de agrupación en los que se identificarán clientes con la misma dirección.
- Creación de nivel de agrupación en relaciones múltiples (ejemplo: Si A y B están relacionados, entonces B y C también, de forma que creemos un grupo A-B-C).

Capítulo 4: Solución

4.1 Situación inicial

En la actualidad, la estructura organizativa a nivel operacional que presenta esta empresa es la siguiente:

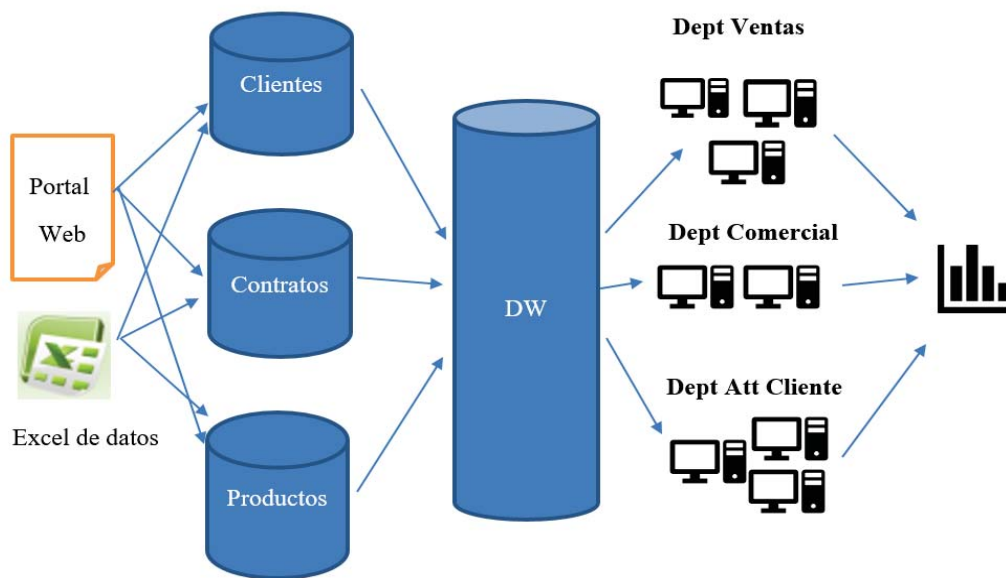


Figura 4: Esquema de la situación inicial de la empresa

Un usuario a través del portal web o en una de las múltiples oficinas que ofrece esta empresa puede registrarse y firmar un contrato determinado en función de sus necesidades. Para ello, debe ofrecer los siguientes datos (nombre, apellidos, DNI, email, teléfono, dirección, cuenta bancaria y producto). Una vez completado este paso, el servicio web o uno de los empleados se encargan de rellenar los datos correspondientes y de registrarlos en las bases de datos de la entidad. Las bases de datos que se tienen son las siguientes:

- **Cientes:** base de datos encargada de almacenar los datos de los clientes (nombre, apellidos, DNI, email, teléfono, dirección).
- **Contratos:** Base de datos encargada de almacenar los datos de los contratos asociados a cada cliente (contrato, tipo de contrato, DNI cliente, duración contrato, fecha alta, fecha baja).
- **Productos:** Base de datos encargada de almacenar los datos de los productos que han sido adquiridos asociados a un contrato y un cliente determinado (producto, DNI cliente, contrato, precio, fecha alta, fecha baja).

Estas bases de datos se almacenan a su vez en una base de datos masiva (Data warehouse) en donde se insertan todos los datos procedentes de clientes, contratos asociados y productos adquiridos por ese cliente. Los datos quedan almacenados en bruto, sin ningún tipo de filtro lo que hace que los datos recogidos desde los ficheros origen no sean modificados en ningún momento dando lugar a datos inconsistentes en el sistema.

Desde la base de datos masiva, los diferentes departamentos de la entidad (comercial, ventas y atención al cliente) irán recogiendo los datos que requieran en función de sus necesidades. Posteriormente, cada semestre se realiza un informe de mercado entre los diferentes departamentos para obtener una visión global acerca de la situación de la entidad y establecer una toma de decisiones con los resultados obtenidos.

En los últimos dos años, los beneficios de esta empresa han caído sustancialmente debido a la inconsistencia e integridad de los datos procedentes de su base de datos masiva lo que ha dado lugar a toma de decisiones erróneas que han supuesto pérdidas importantes para la entidad. Por ello, han llegado a la conclusión que la solución más próxima para poder obtener los beneficios anteriores es la implantación de una ETL que ayude a limpiar los datos que no sean relevantes y actualizar aquellos que hayan sufrido modificaciones. Este proceso permite la sincronización de los datos de forma que el resultado sea visible por cualquier departamento desde el mismo instante en el que se cargan. Así, nos aseguramos de que los datos almacenados sean persistentes. Otra de las medidas a implantar es la creación de data

martas en sustitución a la base de datos masiva que ayuden a agrupar los datos en función de las necesidades de cada departamento. Así conseguimos que la información sea independiente para cada departamento, pero accesible para cualquiera de ellos pudiendo consultar datos sin ningún problema. Con este planteamiento se busca mantener una estructura adecuada que identifique el origen de los datos, que mejore la calidad de los mismos y que ayude a la toma de decisiones.

4.2 Diseño de la solución

El diseño de la solución propuesta es el siguiente:

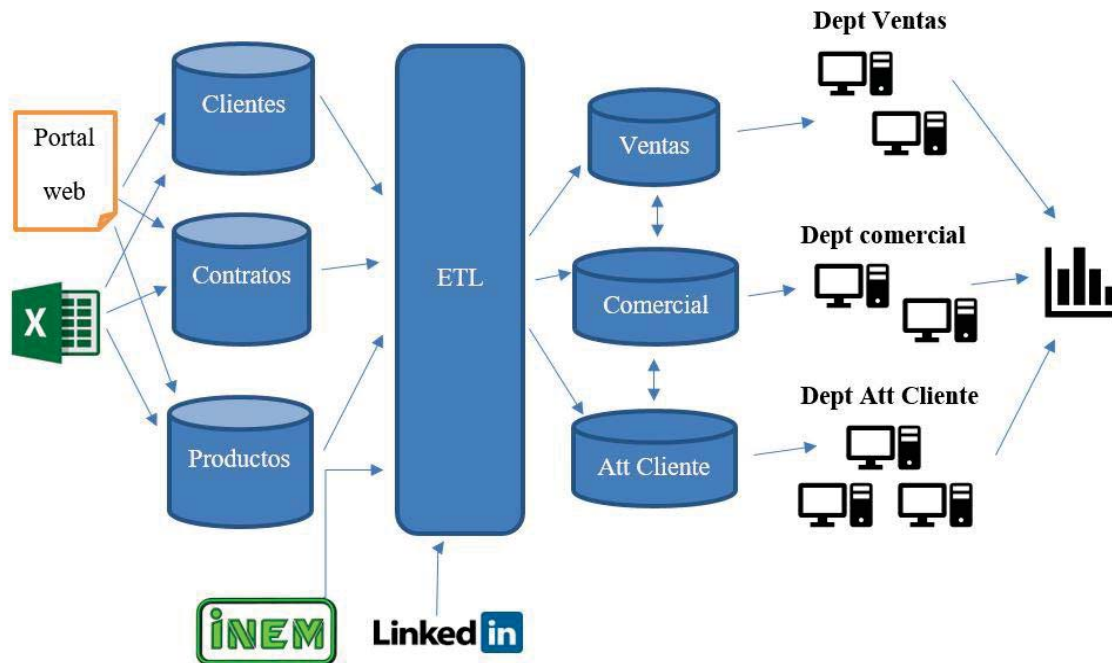


Figura 5: Esquema propuesto

A diferencia de la estructura implantada en la actualidad y en función de los requisitos previos planteados, se ha introducido un proceso ETL para mejorar en gran medida el funcionamiento operacional de la entidad y conseguir una toma de decisiones adecuada a los datos filtrados. Además de este proceso, se han establecido fuentes externas de información asociadas al proceso ETL (INEM, LinkedIn) que ayuden a comprender y establecer un filtrado en función de la situación laboral del cliente. El funcionamiento de este sistema es el siguiente:

Un usuario determinado desea firmar un contrato con esta compañía. Para ello debe registrarse a través de la página web de la entidad o en una de las múltiples oficinas que ofrece esta empresa. El requisito necesario por parte del cliente es ofrecer los siguientes datos (nombre, apellidos, DNI, email, teléfono, dirección, cuenta bancaria y producto). Una vez completado este paso, el servicio web o uno de los empleados rellenarán los datos correspondientes y los almacenará en las bases de datos de la entidad mencionadas anteriormente (Clientes, Contratos y Productos).

Una vez almacenados los datos del cliente, el contrato asociado al mismo y los productos adquiridos, se establece un proceso de extracción, transformación y carga de datos. Los datos extraídos provienen de las bases de datos de cliente, contratos y productos. Una vez identificados las tablas origen, se realiza el proceso de transformación en el que se filtrarán los datos en función de las condiciones establecidas por la entidad y, posteriormente se realizará la carga de los mismos en cada uno de los “data marts” correspondientes. Son los siguientes:

- **Ventas:** Base de datos departamental en la que se almacenan los datos correspondientes a las ventas de los diferentes productos comprados por los clientes. Los campos que presenta esta base de datos son: producto, cliente, fecha, precio.
- **Comercial:** Base de datos departamental en la que se almacenan los datos provenientes de los productos adquiridos por los clientes y el contrato asociado a dicha compra. Los campos que presenta esta base de datos son: cliente, producto, contrato, fecha.

- **Att Cliente:** Base de datos departamental en la que se almacenan los datos correspondientes a los clientes y contratos asociados a los mismos. Los campos que presenta esta base de datos son: Nombre, Apellidos, DNI, email, sexo, teléfono, dirección, contrato, fecha alta y fecha_baja.

Estos tres data marts están asociados entre sí de forma que la información que presenta cada uno de ellos es compartida, pudiendo consultar los datos que requieran cada uno de los departamentos para establecer las conclusiones oportunas. Además, los datos que sean modificados en un departamento quedarán reflejados en los departamentos restantes a tiempo real por lo que la sincronización será una nueva característica en la implantación de este nuevo sistema.

Por último, se realizará un informe de mercado semestral entre los diferentes departamentos para obtener una visión global acerca de la situación de la entidad y establecer una toma de decisiones con los resultados obtenidos. Con este nuevo sistema, se estima que la toma de decisiones y las conclusiones alcanzadas ayuden a mejorar la situación económica de esta organización.

4.3 Desarrollo

Para desarrollar la solución propuesta, se ha realizado un análisis previo sobre la calidad del dato en la empresa para comprobar las inconsistencias que ofrecen los datos almacenados en las bases de datos de la entidad y encontrar el origen de los mismos.

4.3.1 Calidad de Datos

Se entiende por calidad de datos a las técnicas y procesos destinados a mejorar la eficacia de los datos almacenados en nuestras bases de datos. Si queremos que un proceso de calidad de datos sea eficaz, debe ser repetible para poder establecer ciclos de mejora que ayuden a obtener datos de mayor calidad. Hasta ahora, los datos tratados en la organización no han pasado por ningún proceso de calidad, lo que ha ocasionado problemas tanto operacionales como de comunicación. Por ello, los datos deben adaptarse al uso que se les pretende dar y satisfacer las necesidades del usuario.

Los beneficios que presenta el uso de la calidad de datos son los siguientes:

- Ahorro de costes.
- Optimización de recursos en la gestión de datos gracias a la automatización de la información.
- Mejora de resultados con datos fiables.
- Incremento económico.
- Compartición de responsabilidades entre la calidad y el gobierno de los datos.
- Limpieza de datos.

En resumen, se puede determinar que el término “calidad de datos” es un concepto confuso ya que, en función de los objetivos, las características que los datos deben tener para ser adecuados son diferentes.

4.3.1.1 Dimensiones de la calidad del dato

Para poder comenzar con el análisis de calidad de los datos, es necesario tener un punto de partida que permita identificar su estado actual. Esta identificación se consigue a través de la realización de un proceso de perfilado de datos que ayude a detectar el estado en el que se encuentran los mismos y a partir de ahí, corregir y determinar parámetros de control que ayuden a medir los procesos de calidad. Dichos parámetros de control son conocidos como dimensiones y son considerados puntos clave para asegurar la calidad del dato.

Las dimensiones que cubren la calidad del dato son las siguientes:

- **Completitud:** Los datos deben contener información completa que ayude al seguimiento de los mismos.

En el caso de esta organización, en la base de datos “CLIENTES”, el hecho de no disponer de un campo “SEXO” supone un problema considerable si tenemos en cuenta las personas que compran un determinado producto. Es importante conocer el sexo del cliente que firma un contrato ya que, dependiendo de ello, se pueden realizar informes eficientes que ayuden a captar nuevos clientes en función de los productos ofertados. Por ello, es necesario añadir dicho campo.

Por otro lado, destacar que ninguno de los campos pertenecientes a cada una de las tablas “CLIENTES”, “CONTRATOS” y “PRODUCTOS” pueden estar vacíos a excepción del campo “FECHA_BAJA”. El hecho de no disponer de datos completos puede suponer problemas graves a la hora de establecer consultas por lo que los resultados obtenidos no son fiables. Por tanto, todos los registros a insertar en cada una de las bases de datos deben estar rellenos.

- **Conformidad:** Los datos que están en los campos de la tabla deben ser legibles y con un formato adecuado. En este punto se establecen los siguientes casos:
 - **Clientes:** Actualmente, el campo DNI presenta un formato de 8 dígitos sin letra y 8 dígitos con letra. Esto puede ser un problema debido a la inconsistencia de formato que provoca que un cliente pueda estar dos veces con formatos distintos. Por ello, el formato a aplicar será de 8 dígitos seguido de una letra. Además, el campo sexo se identificará con una “M” o una “F”, las fechas vendrán dadas por el formato “DD/MM/YYYY” y las direcciones empezarán por “C/” o “Av”. Todos los campos de la tabla serán de tipo “string” a excepción del campo fecha que será de tipo “date”.
 - **Contratos:** En este momento, los contratos no presentan un formato concreto. Por ello, se establece que los contratos tendrán una longitud de 20 dígitos. Además del formato del campo DNI mencionado anteriormente, los campos “DURACION”, “FECHA_ALTA” y “FECHA_BAJA” serán de tipo date y presentarán el formato “DD/MM/YYYY”.
 - **Productos:** La base de datos de productos presenta registros duplicados debido al formato que presentan los campos “DNI” y “CONTRATO” por lo que se establecerán los formatos mencionados anteriormente para los campos comentados. Los campos de la tabla serán de tipo “string” a excepción del campo fecha.
- **Consistencia:** Se debe evitar información contradictoria. Los datos tienen que ser fiables. Por tanto, es fundamental que los datos recogidos sean eficientes y que provengan de fuentes seguras. En este caso, la entidad cumple con dicha dimensión. Sin embargo, deberemos tener especial cuidado al recoger datos provenientes de la red social “Linkedin” ya que puede provocar inconsistencias en los datos de los que disponemos.

- **Precisión/exactitud:** Los datos deben ser precisos para que puedan ser utilizados. Debido a los problemas de formato y completitud mencionados anteriormente, la información obtenida no es persistente. Si logramos desarrollar los cambios establecidos, los datos a utilizar serán válidos.
- **Duplicación:** Los datos no deben estar duplicados. Para ello, se introducirá un nuevo campo en cada una de las bases de datos de la entidad (clientes, contratos, productos) denominado “ID_CLIENTE”, “ID_CONTRATO” e “ID_PRODUCTO” respectivamente que ayude a evitar este problema de duplicación. Además, se tendrá en cuenta los siguientes aspectos:
 - El campo “ID_CLIENTE” y el campo “DNI” de la tabla “CLIENTES” serán clave primaria (PK) de forma que no podrá existir ningún registro con el mismo id de cliente y el mismo DNI.
 - El campo “ID_CONTRATO” y los campos “ID_CONTRATO” e “ID_CLIENTE” de la tabla “CONTRATOS” serán claves primarias (PK) de forma que no podrá existir dos contratos iguales asociados a un mismo cliente y con el mismo id.
 - El campo “ID_PRODUCTO” y los campos “ID_CLIENTE” e “ID_CONTRATO” de la tabla “PRODUCTOS” serán claves primarias (PK) de forma que no podrá haber dos productos iguales asociados a un mismo contrato y cliente y con el mismo id.
- **Integridad:** Toda la información relevante de un registro debe estar presente para que pueda utilizarse. Para asegurar esta dimensión, los datos deberán estar actualizados de forma que cualquier departamento que quiera acceder a un dato en concreto tenga la certeza que dicho dato es fiable.

Si somos capaces de identificar y separar los defectos de los datos en estas dimensiones, podremos emplear técnicas adecuadas para mejorar la información y los procesos que crean y utilizan dicha información.

4.3.1.2 Consecuencias de los errores en los datos

El hecho de disponer de una mala calidad de datos implica amenazas destacables en la toma de decisiones y, en consecuencia, en la operación y gestión de la organización. Una calidad de datos pobre, además de ser uno de los principales problemas en errores del proceso, es el causante de decisiones incorrectas en una organización. Por ello, las pérdidas económicas provocadas en la entidad por este motivo son cuantiosas. Además de esto, el impacto que provoca sobre los clientes es muy alto y, por consiguiente, el grado de insatisfacción de los mismos aumenta ya que al caer en errores como datos personales incorrectos, contratos erróneos, direcciones incorrectas, etc. genera una pérdida de tiempo considerable para solucionar el problema.

Una posible solución a la detección de errores de forma sencilla es mediante el uso de dos alternativas que son capaces de identificar de forma eficaz los datos erróneos:

- **Verificación manual:** Esta alternativa encuentra los errores más comunes verificando en base a la fuente original todos sus valores, permitiendo determinar qué valores son correctos y cuáles no.
- **Análisis automático:** Emplea tanto el software como el conocimiento del analista de calidad para detectar los errores. Este análisis se puede aplicar tanto a transacciones, bases de datos que están cambiando y bases de datos en producción.

Las técnicas analíticas bien aplicadas, detectan suficientes errores como para tener una idea clara del estado del dato.

4.3.1.3 Estrategia de calidad

Cada vez existen más datos provenientes de fuentes externas (web, DataWareHouse, etc.) lo que provoca que la cantidad de información aumente. Los datos que anteriormente eran introducidos con un propósito concreto, ahora se han aplicado a otras finalidades. Mientras que la calidad del dato puede ser adecuada para sistemas transaccionales, no ocurre lo mismo para sistemas BI (un contrato incorrecto en un sistema transaccional afecta a un solo cliente mientras que en un sistema BI puede suponer un problema muy grave).

Por ello, se debe implementar niveles de calidad de datos para procesos automatizados ya que esto evitaría problemas. Todos los procesos se automatizan y la intervención humana queda en un segundo plano. Esto puede traer defectos en el servicio debido a que un sistema automatizado no es capaz de cancelar un proceso erróneo, en cambio, una persona sí. Los datos defectuosos conllevan una pobre gestión del cliente.

4.3.1.4 Conclusiones

A través de la cadena de valor del dato podremos detectar cualquier anomalía que se pudiera mostrar a lo largo del proceso, de ahí, que el seguimiento sea un punto clave. Si nos preocupamos verdaderamente de la importancia que tiene la calidad del dato, obtendremos beneficios claves que obtengan valor (minimizar riesgos, ahorro de tiempo y recursos, toma de decisiones oportunas, adaptación a estándares, mejora de confianza y relaciones ante sus clientes frente a la competencia).

4.3.1.5 Retos de la calidad de datos

En este apartado se analizarán los diferentes retos existentes en la calidad de los datos.

- La diversidad de fuentes de datos aporta abundantes tipos de datos y estructuras de datos complejas y aumenta la dificultad de integración de datos.

En el pasado, las empresas sólo utilizaban los datos generados a partir de sus propios sistemas de negocio, tales como datos de ventas e inventario. Pero ahora, los datos recogidos y analizados por las empresas han superado este ámbito. Las fuentes de datos grandes son muy amplias, incluyendo:

- Conjuntos de datos de internet y móvil a Internet.
- Datos recogidos por diversas industrias.
- Datos experimentales y de observación científica tales como datos de alta energía experimental, datos biológicos y datos de observación espacial.
- Datos de IoT (Internet of things).

Estas fuentes producen los siguientes tipos de datos:

- **Datos no estructurados:** Poseen un formato tal y como fueron recolectados, los cuales carecen de un formato específico (documentos, vídeo, audio, etc.)
- **Datos semi-estructurados:** No se limitan a campos determinados, mantienen marcadores para separar elementos. Pueden contener información poco regular como para ser gestionada de una forma estándar (paquetes / módulos de software, hojas de cálculo e informes financieros.)
- **Datos estructurados:** Los datos ingresados tienen bien definido su longitud y formato (Fechas, números, cadenas de caracteres, etc.). Se almacenan en tablas.

La cantidad de datos no estructurados ocupa más del 80% de la cantidad total de datos existentes. Para las empresas, la obtención de grandes volúmenes de datos con estructura compleja de diferentes fuentes resulta una tarea complicada ya que existen datos inconsistentes o contradictorios entre los datos de diferentes fuentes.

- Los datos cambian muy rápido y el ciclo de vida de los mismos es muy corto, lo que requiere mayores necesidades de tecnología de procesamiento.

Si las empresas no son capaces de recoger los datos requeridos en tiempo real o hacer frente a las necesidades de datos durante un tiempo muy largo, entonces se puede obtener información obsoleta y no válida. El procesamiento y análisis basados en estos datos producirá conclusiones inútiles o engañosas, que eventualmente conducirán a errores de toma de decisiones por parte de gobiernos o empresas. En la actualidad, el procesamiento en tiempo real y el software de análisis para grandes datos todavía está en fase de desarrollo o mejora.

- El volumen de datos es muy grande y es difícil juzgar la calidad de los datos dentro de un tiempo razonable.

Actualmente, la cantidad global de información se duplica cada dos años. Por ello, es difícil recolectar, limpiar, integrar y, finalmente, obtener los datos de alta calidad necesarios dentro de un plazo razonable. Debido a que la proporción de datos no estructurados en datos grandes es muy alta, se necesitará mucho tiempo para transformar tipos no estructurados en tipos estructurados y procesar los datos. Este es un gran reto para las técnicas existentes de calidad de procesamiento de datos.

4.3.2 Data Governance

Una vez establecido el análisis de la calidad del dato en la empresa, es necesario conocer el origen de los mismos con el objetivo de ofrecer una solución rápida y eficaz en caso de detectar errores en los datos.

Llamamos Data Governance al sistema que ejerce un control sobre cómo se usan los datos en una empresa; en otras palabras, se trata de una disciplina empresarial cuyo objetivo se basa en ofrecer un control sobre el mantenimiento, almacenamiento, creación, uso e intercambio de información vital para el negocio. En la actualidad, cualquier tipo de compañía, organismo público o institución tiene que trabajar con un gran volumen de información recogida en datos los cuales tienen que ser tratados, recopilados y extraídos para poder realizar un análisis preciso del negocio. Aunque el uso de Data Governance es importante para organizaciones de todos los tipos y tamaños; se convierte en un elemento crucial a medida que crece la cantidad de datos generados y el número de fuentes en los que se originan y residen. Por ello, la calidad de los datos debe ser analizada ya que poseer datos erróneos puede desembocar en decisiones equivocadas que reduzcan el rendimiento y la rentabilidad de la empresa e incluso, su disolución.

De ahí que el uso del Data Governance se haya convertido en una herramienta fundamental ya que gracias a ello conseguimos una gestión productiva y un eficiente proceso de generación de informes pudiendo reducir los costes y el tiempo empleado en las tareas de análisis de información.

4.3.2.1 Implantación del Data Governance

La implantación de un programa de Data Governance en una organización no es sencilla. La mayoría de los problemas vienen dados por la coordinación y la comunicación de los distintos departamentos de la empresa (ventas, comercial y atención al cliente), así como la ausencia de roles de responsabilidad en los activos de la información y falta de planificación. Por ello, se establecen unos pasos básicos para llevarla a cabo:

En primer lugar, se debe priorizar las áreas en las que se va a aplicar. Para ello, es necesario determinar los departamentos que más pueden beneficiarse de su implantación y establecer objetivos unidos a su estrategia corporativa. Todo ello facilita la coordinación entre las distintas partes implicadas. En este caso, los tres departamentos requieren de la implantación de un sistema de Data Governance puesto que son los pilares básicos en los que se desarrolla la estrategia operacional de la entidad. En el caso del departamento de ventas es fundamental conocer el origen de los productos y contratos asociados, así como el importe de los mismos. De esta forma nos aseguramos de que ante cualquier error en alguno de los datos, podamos ver rápidamente de dónde viene el problema. En el caso del departamento comercial, es necesario conocer el origen de los clientes, productos y contratos asociados ya que son los encargados de vender estos productos a clientes potenciales que ayuden a mantener la economía de la entidad. Por último, en el caso del departamento de atención al cliente, al igual que el departamento comercial, necesitan conocer el origen de los clientes, productos y contratos asociados de forma que ante cualquier problema o incidencia de un cliente, puedan ver el motivo del mismo e implantar una solución rápida.

Una vez completado el ciclo anterior, se debe determinar qué datos están disponibles de forma que los empleados puedan acceder a los datos de manera rápida y eficaz. Cada departamento dispone de sus propios datos:

- En el departamento de ventas se disponen de los datos de productos y contratos.
- En el departamento comercial se disponen de los datos de clientes, productos y contratos.
- En el departamento de atención al cliente se disponen de los datos de clientes, productos y contratos.

Cabe destacar que los datos son compartidos de forma que cualquier cambio producido en un departamento quedará reflejado en los departamentos restantes.

Cuando la información sea accesible, se crearán los roles, responsabilidades y reglas para determinar el trabajo de cada uno. Este punto es imprescindible ya que cualquier iniciativa en la que no exista una cooperación entre áreas estará condenada al fracaso. Como se ha comentado anteriormente, cada departamento tendrá su funcionalidad. El departamento de ventas se encargará de comprobar los ingresos que lleva la empresa en un periodo determinado de tiempo, el departamento comercial llevará a cabo la venta de productos y el departamento de atención al cliente dará soporte a los clientes.

El siguiente paso es asegurar la integridad de los datos ya que la información evoluciona y se actualiza constantemente por lo que debe ser monitorizada para controlarla y garantizar la integridad de los datos recogidos mediante procesos de análisis y normalización. En este proceso de control es necesario determinar quiénes son los responsables del mantenimiento y control de dichos datos y dotar a los mismos de la tecnología necesaria para que, en caso de fallo, la integridad y la calidad de los datos se mantengan a salvo. Cada departamento tendrá un responsable encargado de mantener la integridad de los datos. A continuación, se transforma las transacciones de datos en datos maestros, es decir, datos esenciales que definen el negocio: clientes, proveedores, productos, etc. Si dichos datos se mantienen sincronizados y vinculados, todos los usuarios del sistema tendrán acceso a la misma información lo que supondrá una base sólida y coherente al funcionamiento de la organización.

Por último, se desarrollará un mecanismo de retroalimentación que asegure la mejora de los procesos; corrigiendo fallos y permitiendo el desarrollo.

El uso de Data Governance ayuda a todos los empleados de una organización, desde directivos hasta el personal de IT, gestión de datos y otros departamentos a cooperar para lograr un objetivo común, lo que ocasiona la toma de buenas decisiones. De este modo, se pueden crear reglas de forma más eficiente, garantizando el cumplimiento de las normas y haciendo frente a los problemas.

La implantación del Data Governance busca:

- Mejorar en la toma de decisiones.
- Asegurar la transparencia de los procesos.
- Construir procesos repetibles.
- Reducir costes y aumentar la eficacia a partir de la coordinación.
- Asegurar que los datos cumplen con la demanda.

4.3.3 Desarrollo de la ETL

En este apartado se explicará detalladamente el funcionamiento del proceso ETL implementado para este sistema:

El objetivo que se busca con la implantación de esta ETL es el filtrado, limpieza e inserción de datos útiles en cada uno de los data marts de la entidad. Para ello se obtendrán los datos pertinentes de las tablas origen, se establecerá un cruce entre las mismas para agrupar los datos requeridos, se aplicará el formato pertinente a cada uno de los campos y se insertarán en las tablas T_VENTAS, T_COMERCIAL y T_ATT_CLIENTE.

Orígenes

Las entradas de datos empleadas para el proceso son 3:

- T_CLIENTES
- T_CONTRATOS
- T_PRODUCTOS

La estructura y definición de cada campo empleados para la tabla T_CLIENTES es la siguiente:

VARIABLE	FORMATO	DESCRIPCIÓN
ID_CLIENTE	Autonumérico	Identificador del cliente
NOMBRE	Texto (20)	Nombre del cliente
APELLIDOS	Texto (50)	Apellidos del cliente
DNI	Texto (10)	Clave única que muestra el documento nacional de identidad del cliente
EMAIL	Texto (40)	Correo asociado al cliente
TELEFONO	Texto (20)	Teléfono fijo o móvil
DIRECCION	Texto (50)	Dirección del cliente
SEXO	Texto (1)	Sexo del cliente

La estructura y definición de cada campo empleados para la tabla T_CONTRATOS es la siguiente:

VARIABLE	FORMATO	DESCRIPCIÓN
ID_CONTRATO	Texto (30)	Identificador del contrato
TIPO	Texto (20)	Tipo de contrato
ID_CLIENTE	Numérico	Identificador del cliente asociado al contrato
DURACIÓN	Fecha DD/MM/YYYY	Duración del contrato
FECHA_ALTA	Fecha DD/MM/YYYY	Fecha de alta del contrato
FECHA_BAJA	Fecha DD/MM/YYYY	Fecha de baja del contrato

La estructura y definición de cada campo empleados para la tabla T_PRODUCTOS es la siguiente:

VARIABLE	FORMATO	DESCRIPCIÓN
ID_PRODUCTO	Autonumérico	Identificador del producto
TIPO_PRODUCTO	Texto (20)	Tipo de producto
ID_CLIENTE	Numérico	Identificador del cliente asociado al producto
ID_CONTRATO	Texto (30)	Identificador del contrato asociado al producto
PRECIO	Numérico	Precio del producto adquirido por el cliente
FECHA_ALTA	Fecha DD/MM/YYYY	Fecha de alta del contrato
FECHA_BAJA	Fecha DD/MM/YYYY	Fecha de baja del contrato

Transformaciones

Las transformaciones utilizadas para la realización de esta ETL son las siguientes:

- **Join T_CLIENTES con T_CONTRATOS:** Se realiza un cruce de las tablas T_CLIENTES con la tabla T_CONTRATOS por la clave ID_CLIENTE para determinar los contratos que presenta cada cliente.
- **Join T_CLIENTES y T_CONTRATOS con T_PRODUCTOS:** Una vez que conocemos los contratos que presenta cada cliente, se realiza un cruce por ID_ENTIDAD e ID_CONTRATO para determinar cuáles son los productos que presentan los contratos asociados a cada cliente.
- **Expresión de cambio de formato:** En esta transformación, se establecen los formatos requeridos para alcanzar la conformidad de los mismos:
 - El campo DNI presentará el formato de 8 dígitos seguidos de una letra.
 - El campo SEXO vendrá dado por una “F” o una “M”.
 - Los campos DURACIÓN, FECHA_ALTA y FECHA_BAJA serán de tipo fecha con el formato “DD/MM/YYYY”.

- El campo DIRECCIÓN tendrá que empezar por “C/” o por “Av”
- El campo CONTRATO tendrá longitud 20.
- **LookUp situación laboral:** Esta transformación se encargará de obtener información procedente de bases de datos externas provenientes de empresas como “INEM” y “Linkedin” y obtener como resultado la situación laboral de cada cliente, de forma que podamos sacar conclusiones a partir de este dato.

Salidas

Las tablas en las que se insertarán los datos tratados son las siguientes:

- T_VENTAS
- T_ATT_CLIENTE
- T_COMERCIAL

La tabla T_VENTAS presenta la siguiente estructura:

VARIABLE	FORMATO	DESCRIPCIÓN
ID_CLIENTE	Autonumérico	Identificador del cliente
ID_CONTRATO	Texto (20)	Identificador del contrato
TIPO_PRODUCTO	Texto (20)	Tipo de producto
PRECIO	Numérico	Precio del producto adquirido por el cliente

La tabla T_COMERCIAL presenta la siguiente estructura:

VARIABLE	FORMATO	DESCRIPCIÓN
ID_CLIENTE	Autonumérico	Identificador del cliente
NOMBRE	Texto (20)	Nombre del cliente
APELLIDOS	Texto (50)	Apellidos del cliente
SEXO	Texto (1)	Sexo del cliente

4. Solución

DNI	Texto (9)	Clave única que muestra el documento nacional de identidad del cliente
TELÉFONO	Texto (20)	Teléfono del cliente
DIRECCIÓN	Texto (40)	Dirección del cliente
ID_CONTRATO	Texto (20)	identificador del contrato
TIPO_PRODUCTO	Texto (20)	Tipo de producto
SITUACIÓN_LABORAL	Texto (20)	Situación Actual del cliente
PRECIO	Numérico	Precio del producto adquirido por el cliente
FECHA_ALTA	Fecha DD/MM/YYYY	Fecha de alta del contrato
FECHA_BAJA	Fecha DD/MM/YYYY	Fecha de baja del contrato

La tabla T_ATT_CLIENTE presenta la siguiente estructura:

VARIABLE	FORMATO	DESCRIPCIÓN
ID_CLIENTE	Autonumérico	Identificador del cliente
NOMBRE	Texto (20)	Nombre del cliente
APELLIDOS	Texto (50)	Apellidos del cliente
DNI	Texto (9)	Clave única que muestra el documento nacional de identidad del cliente
TELÉFONO	Texto (20)	Teléfono del cliente
DIRECCIÓN	Texto (40)	Dirección del cliente
EMAIL	Texto (40)	Correo del cliente
ID_CONTRATO	Texto (20)	identificador del contrato
TIPO_PRODUCTO	Texto (20)	Tipo de producto
PRECIO	Numérico	Precio del producto adquirido por el cliente
FECHA_ALTA	Fecha DD/MM/YYYY	Fecha de alta del contrato
FECHA_BAJA	Fecha DD/MM/YYYY	Fecha de baja del contrato

Proceso

- 1) Cruzamos las tablas origen T_CLIENTES, T_CONTRATOS y T_PRODUCTOS para obtener los contratos y productos asociados al cliente.
- 2) Una vez realizado el paso anterior, buscaremos en las bases de datos externas de las empresas “INEM” y “Linkedin” la situación laboral de cada cliente a través de un “lookup” filtrando por su nombre, apellidos y DNI.
- 3) Una vez obtenido la situación laboral del cliente, estableceremos el formato necesario para cada uno de los campos.
- 4) Por último, se insertarán en las tablas T_VENTAS, T_COMERCIAL y T_ATT_CLIENTE los campos asociados a cada tabla.

Como resultado de este proceso ETL conseguiremos que los datos que habían sido introducidos en bruto a partir de las tablas origen se inserten en las tablas destino con un formato específico, evitando problemas de duplicidad y cumpliendo con las dimensiones que cubren la calidad del dato.

4.3.3.1 Contenido de los data marts

Los data marts de los que dispondrá la empresa con la implantación de esta nueva estructura serán los siguientes:

- **Ventas:** formará parte del departamento de ventas. Los campos que presenta esta base de datos son ID_CLIENTE, ID_CONTRATO, ID_PRODUCTO y PRECIO. Este departamento sólo requiere los datos de las ventas de productos asociados a clientes a través de un contrato. Además, es fundamental conocer el precio para establecer conclusiones y obtener resultados favorables que ayuden a tomar decisiones fiables. Los datos almacenados provienen de la carga del proceso ETL creado anteriormente lo que implica que la información que se almacene en este data mart será información tratada y filtrada.

- **Comercial:** formará parte del departamento comercial de la entidad. Los campos que presenta esta base de datos son ID_CLIENTE, NOMBRE, APELLIDOS, SEXO, DNI, TELÉFONO, DIRECCIÓN, ID_CONTRATO, TIPO_PRODUCTO, SITUACIÓN_LABORAL, PRECIO, FECHA_ALTA y FECHA_BAJA. Este departamento se encargará de recopilar toda la información posible acerca de los productos de la entidad, los datos del cliente y los contratos asociados al mismo para, posteriormente, ofrecer nuevos productos o mejoras de contrato al cliente además de intentar captar nuevos clientes. Al igual que el departamento anterior, los datos almacenados serán tratados y filtrados en el proceso ETL y almacenados tras su carga.
- **Atención al cliente:** formará parte del departamento de atención al cliente de la entidad. Los campos que presenta esta base de datos son ID_CLIENTE, NOMBRE, APELLIDOS, DNI, TELÉFONO, DIRECCIÓN, EMAIL, ID_CONTRATO, TIPO_PRODUCTO, PRECIO, FECHA_ALTA, FECHA_BAJA. Este departamento se encargará de disponer de toda la información del cliente con el objetivo de resolver cualquier incidencia o duda que presente el cliente. Al igual que el departamento anterior, los datos almacenados serán tratados y filtrados en el proceso ETL y almacenados tras su carga.

Los tres data marts, a pesar de ser bases de datos independientes, compartirán sus datos con el resto de departamentos ofreciendo transparencia y estarán sincronizados a tiempo real de forma que, ante cualquier cambio o modificación por parte de alguno de los departamentos, dicho cambio quedará reflejado, consiguiendo así la integridad de los mismos y evitando inconsistencias y problemas de comunicación interna.

4.4 Evaluación y validación

4.4.1 Lanzamiento de pruebas

Para comprobar el funcionamiento de la ETL se han establecido una serie de pruebas a cumplir para asegurar la calidad del dato:

- 1) Todos los campos almacenados en la tabla T_CLIENTES deben estar rellenos.
- 2) Todos los campos almacenados en la tabla T_CONTRATOS deben estar rellenos.
- 3) Todos los campos almacenados en la tabla T_PRODUCTOS deben estar rellenos.
- 4) El campo “DIRECCIÓN” cargado en las tablas T_COMERCIAL y T_ATT_CLIENTE debe empezar por “C/” o “Av”.
- 5) El campo “SEXO” almacenado en la tabla T_CLIENTES debe ser de tipo texto y tener valor “M” o “F”.
- 6) El campo “FECHA_ALTA” cargado en las tablas T_COMERCIAL y T_ATT_CLIENTE debe ser de tipo fecha y con el formato “dd/mm/yyyy”.
- 7) El campo “DNI” cargado en las tablas T_COMERCIAL y T_ATT_CLIENTE debe presentar un formato de 8 dígitos seguido de una letra.
- 8) El campo “ID_CONTRATO” debe presentar una longitud de 20 dígitos.
- 9) Todos los campos de la tabla T_VENTAS deben estar rellenos.
- 10) Todos los campos de la tabla T_COMERCIAL deben estar rellenos.
- 11) Todos los campos de la tabla T_ATT_CLIENTE deben estar rellenos.
- 12) No puede haber registros duplicados en ninguna de las tablas destino (T_VENTAS, T_COMERCIAL y T_ATT_CLIENTE)
- 13) Las modificaciones realizadas en una tabla deben ser visibles en las tablas restantes.

Cada una de las pruebas enunciadas han sido probadas secuencialmente y todas ellas han cumplido con su objetivo. Por motivos de confidencialidad, los resultados obtenidos no pueden mostrarse, pero si se puede comentar que la implantación de la ETL ha ayudado en gran medida a la limpieza de datos inconsistentes y, consecuentemente, ha conseguido la integridad de los mismos.

4.4.2 Establecimiento de requisitos

Los requisitos alcanzados tras la realización de este proyecto han sido los siguientes:

- **Perfilado de datos:** Se han cumplido con todos los requisitos planteados en este punto, tanto el análisis previo de la calidad de los datos en el sistema como la obtención de datos fiables que ayuden a identificar los orígenes de información de datos incompletos.
- **Estandarización y normalización:** Se ha cumplido con la estandarización del campo “DNI” y “DIRECCIÓN” así como la verificación del campo “SEXO”. Los campos “NOMBRE”, “APELLIDOS”, “TELÉFONO” e “EMAIL” no han sido necesarios actualizarlos.
- **Enriquecimiento:** Se ha cumplido con la obtención de información procedente de fuentes externas (INEM, LinkedIn).
- **Deduplicación:** Se ha cumplido con los objetivos planteados en este punto (identificador único para cada registro).
- **Agrupación de clientes:** Por motivos de tiempo no ha sido posible cumplir con este requisito planteado.
- Se ha establecido un planteamiento real del problema de análisis de datos en una empresa.
- Se ha explicado la problemática que presenta en la actualidad el uso de datos erróneos.
- Se ha desarrollado una posible solución a la problemática planteada.

Capítulo 5: Conclusiones y líneas futuras

5.1 Conclusiones

En este trabajo fin de grado se ha analizado la problemática de la gestión de los datos y se ha propuesto una solución para un caso particular de una organización. En primer lugar, a pesar del esfuerzo que supone analizar la gestión de los datos en las empresas, he sido capaz de realizar dicho estudio y entender mejor la problemática que presenta disponer de datos no fiables.

Una vez alcanzado este objetivo, se ha analizado la problemática de una empresa en particular y se han identificado los principales riesgos y debilidades que suponen para esta empresa la falta de datos de calidad y se han establecido una serie de requisitos a cumplir durante el desarrollo de la solución (perfilado, estandarización y normalización, enriquecimiento, deduplicación). Estos requisitos intentan abarcar todos los problemas existentes en una organización con una pobre calidad de datos.

Posteriormente, se ha desarrollado una posible solución al problema planteado alcanzando todos los requisitos. Esto ha dado lugar a una mejora considerable en la obtención de datos persistentes y completos. Para ello, se ha lanzado una serie de pruebas para comprobar que los resultados obtenidos se asemejan a los resultados esperados.

Consiguientemente podemos decir que los objetivos planteados al principio de este proyecto se han cumplido satisfactoriamente.

Las conclusiones alcanzadas tras el análisis, el estudio de la problemática de la gestión de datos en las empresas y el desarrollo de la solución planteada me han ayudado en gran medida a comprender que la implantación de la calidad de datos en una organización no es necesaria sino fundamental. Las ventajas que supone tener datos de calidad vienen ligadas a los beneficios económicos, de rendimiento, eficiencia, eficacia y toma de decisiones que aporta a la empresa. Destacar también que los procesos de extracción, transformación y carga son

una parte imprescindible en cualquier proyecto de calidad de datos ya que son el motor encargado de filtrar, limpiar, actualizar, codificar y extraer datos que ofrezcan valor al empleado. Otro de los puntos importantes ha sido el establecimiento de un control de datos (Data Governance) ya que permite conocer el origen de los mismos de forma que ante cualquier incidencia o búsqueda, el tiempo empleado en encontrar dicho dato sea mínimo.

En conclusión, cualquier empresa que presente un volumen de datos elevado, necesita un proceso de calidad de datos para poder seguir creciendo y ser competitiva frente a otras empresas del sector ya que, en caso de no hacerlo, posiblemente quede en un segundo plano, las pérdidas económicas sean sustanciales y llegue incluso al cierre de la misma.

5.2 Líneas futuras

Se ha desarrollado una posible solución a un problema real pero lo recomendable sería su implantación en la organización. Considero que este es un punto de partida que poco a poco va tomando importancia y que, dentro de muy poco tiempo, la calidad del dato será obligatorio en cualquier empresa. La realización de este proyecto también invita a la creación de nuevas herramientas ETL que mejoren en gran medida a las actuales del mercado lo que ayudará a mejorar el rendimiento actual y hacer frente al crecimiento exponencial de la información de los últimos años.


Bibliografía

- [1] Ariely, Dan. *Predictably irrational*. New York: HarperCollins, 2008.
- [2] Cavanillas, José M., Edward Curry, and Wolfgang Wahlster. *New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe*. Springer, 2016.
- [3] Data Governance [en línea] < <http://www.informationbuilders.es/data-governance> > [Consulta: 10 Marzo 2017]
- [4] Bouman, Roland, and Jos van Dongen. "Pentaho solutions." *Business Intelligence and Data Warehousing with Pentaho and MYSQL* (2009).
- [5] Amanpartap Singh, P. A. L. L., and Jaiteg Singh Khaira. "A comparative review of extraction, transformation and loading tools." *Database Systems Journal BOARD* (2013): 42.
- [6] Gartner, Big Data [en línea] < <https://research.gartner.com/definition-what-is-big-data?resId=3002918&srcId=1-8163325102> > [Consulta: 12 Abril 2017]
- [7] Berthold, Michael R., et al. *Guide to intelligent data analysis: how to intelligently make sense of real data*. Springer Science & Business Media, 2010.
- [8] Alfonso del Gallo, Estudio sobre calidad y gestión de datos 2016 [en línea] < <http://www.experian.es/assets/servicios-de-marketing/white-papers/ExperianCalidadDatos2016.pdf> > [Consulta: 15 Marzo]
- [9] Cai, L. & Zhu, Y., (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*. 14, p.2. DOI: <http://doi.org/10.5334/dsj-2015-002>
- [10] Big Data Value Association, Big Data Value Strategic Research and Innovation Agenda, Version 3.0 (Enero 2017) [En línea] < <http://www.bdva.eu/> > [Consulta: 20 Abril 2017]
- [11] Procesos ETL: Definición, características, beneficios y retos [en línea] < <http://blog.powerdata.es/el-valor-de-la-gestion-de-datos/bid/312584/procesos-etl-definicion-caracteristicas-beneficios-y-retos> > [Consulta: 30 Abril 2017]

-
- [12] Abinitio, La importancia de la calidad de los datos (Noviembre, 2015) [en línea] < <http://www.abinitio.es/blog/la-importancia-de-la-calidad-de-los-datos/> > [Consulta: 22 Marzo 2017]
- [13] Wikipedia, Calidad de datos [en línea] < https://es.wikipedia.org/wiki/Calidad_de_datos > [Consulta: 7 Marzo 2017]
- [14] Powerdata, La importancia del Data Governance [en línea] < <http://blog.powerdata.es/el-valor-de-la-gestion-de-datos/bid/406824/la-importancia-del-data-governance-case-study> > [Consulta: 15 Marzo 2017]
- [15] James Orr, El valor de los datos: La importancia del Data Governance (2012) [en línea] < <http://www.silicon.es/el-valor-de-los-datos-la-importancia-del-data-governance-19589> > [Consulta: 20 Marzo 2017]
- [16] Xaime Méndez Baudot, Volumen de datos Almacenados en 2020 [en línea] < <http://www.elmundo.es/elmundo/2013/07/30/navegante/1375199227.html> > [Consulta: 25 marzo 2017]
- [17] Las 5 Vs que caracterizan el concepto de Big Data (2014) < <https://bigdata400.wordpress.com/2014/11/11/las-5-vs-que-caracterizan-el-concepto-de-big-data/> > [Consulta: 14 Abril]
- [18] Ron Bisio, The power of a data value chain for your business [en línea] < <http://dataconomy.com/2017/02/power-of-data-value-chain/> > [Consulta: 20 abril 2017]
- [19] Santiago Eibe, Arquitectura del Data Warehouse (2017) [en línea] < <https://moodle.upm.es/titulaciones/oficiales/course/view.php?id=6951> > [Consulta: 24 abril 2017]
- [20] Informática Powercenter [en línea] < <https://www.informatica.com/es/products/data-integration/powercenter.html> > [Consulta: 30 Abril 2017]
- [21] Porter, Michael E. *Competitive advantage: Creating and sustaining superior performance*. Simon and Schuster, 2008.



Este documento esta firmado por

	Firmante	CN=tfgm.fi.upm.es, OU=CCFI, O=Facultad de Informatica - UPM, C=ES
	Fecha/Hora	Tue Jul 04 17:15:49 CEST 2017
	Emisor del Certificado	EMAILADDRESS=camanager@fi.upm.es, CN=CA Facultad de Informatica, O=Facultad de Informatica - UPM, C=ES
	Numero de Serie	630
	Metodo	urn:adobe.com:Adobe.PPKLite:adbe.pkcs7.sha1 (Adobe Signature)