Universidad de Valladolid

ESCUELA DE INGENIERÍA INFORMÁTICA DE SEGOVIA

Grado en Ingeniería Informática de Servicios y Aplicaciones

Un estudio sobre la calidad de datos en entornos empresariales

---

Autora: Miriam Rodríguez Mate

Tutores: Miguel Ángel Martínez Prieto
Jorge Silvestre Vilches

Agradecimientos

En estas primeras líneas de mi Trabajo de Fin de Grado, que supone la finalización de una larga y costosa travesía que con trabajo y esfuerzo, terminará con la obtención del título universitario en el Grado en ingeniería Informática de Servicios y Aplicaciones, unas palabras de agradecimiento a todas aquellas personas que me han apoyado durante todo mi travesía académica y sobre todo en la elaboración de este TFG.

En primer lugar agradecer a mis tutores, **Miguel Ángel Martínez Prieto** y **Jorge Silvestre Vilches**, por todo el apoyo que me han brindado, su fe en mí y su implicación durante el tiempo que he trabajado a su lado.

Agradecer también, al **personal docente de la Escuela de Ingeniería Informática de Segovia** por su compromiso con sus alumnos, por todo lo que aprendí de ellos y que me ha servido para llegar hoy hasta aquí.

Para el desarrollo de este trabajo, ha sido de gran ayuda mi tutor de prácticas de empresa **Juan Rodríguez Jiménez**, que además de aportarme para este proyecto los datos reales a partir de los cuales he podido desarrollar la parte práctica de este trabajo, siempre ha confiado en mí y ha tenido una gran disposición para ofrecerme su ayuda en todo momento.

Me gustaría dar las gracias a mis compañeros y personas cercanas por apoyarme a lo largo de toda la carrera, en especial agradecer a mis amigos **Alicia Martínez** y **Saúl Rodríguez** que han estado siempre en los momentos difíciles y siempre me han dedicado horas de su tiempo.

Por último, y no por eso menos importante, querría dejar constancia del gran papel que ha tenido mi **familia** a lo largo de todo este proceso, que a pesar de la distancia ha estado siempre presente y con una actitud muy positiva. En especial a mi padre Fernando Rodríguez, mi tía Concepción Mate y mi hermano Diego Rodríguez que han supuesto un gran apoyo, sin el cual difícilmente hubiera superado las dificultades que me han ido surgiendo a lo largo de mi trayectoria estudiantil y en especial en la elaboración y aplicación de mi TFG. Pero sobre todo, agradecer de una manera todavía más especial el apoyo incondicional que mi hermana Mabel Rodríguez, mi abuela Concepción González y mi madre Rosa Isabel Mate me han regalado durante este último curso, durante toda mi carrera y durante toda mi vida.

Resumen

Una de las principales y más importantes preocupaciones de las empresas son los datos. Cada día aparecen nuevas tecnologías que facilitan la recolección de datos masivos (Big Data), sin embargo estos grandes volúmenes de datos por sí solos no suponen ningún valor a la empresa, es decir, que los datos deben ser tratados y estudiados para poder realizar cualquier tipo de operación con ellos.

Aquí entra el objeto de este TFG, la calidad de los datos. Las empresas quieren obtener grandes cantidades de información útil, por ello esos datos tienen que estar dotados de la calidad que permita que esa información sea veraz, libre de errores y aporten su máximo valor posible. Contar con datos sin tratar, es decir, sin calidad, perjudica enormemente a las empresas en la toma de decisiones además de repercutir en costos económicos y humanos.

Este Trabajo Fin de Grado trata en profundidad el tema de calidad de los datos, con las respectivas dimensiones que ayudan en la identificación de los datos, la gestión de esta calidad de datos y los problemas y beneficios que conlleva implantar calidad de datos. También trataremos conceptos estrechamente relacionados con la calidad de datos, como son los metadatos, información específica de los datos que ayuda a distinguirlos con mayor grado de individualidad, y el concepto de perfiles de datos, como conjuntos de datos con unas mismas características que nos permiten realizar análisis más específicos a los datos. Se abordan también de forma teórica y con un ejemplo práctico los catálogos de datos, unas herramientas que ayudan a dotar de calidad a lo datos y que recogen todos los conceptos mencionados anteriormente.

Por último, destacamos que este TFG está desarrollado bajo la metodología UVagile, teniendo un enfoque más dinámico que nos permite realizar el trabajo de forma iterativa e incremental, con lo que conseguimos un resultado final más acertado con respecto a los objetivos planteados al inicio del proyecto.

Palabras claves: Datos, calidad de los datos, metadatos, perfiles de datos, catálogos de datos.

Abstract

One of the main and most important concerns of companies is data. Every day new technologies appear that facilitate the collection of massive data (Big Data), however these large volumes of data alone do not represent any value to the company, that is, the data must be treated and studied to be able to perform any type of operation on them.

This is where the object of this Final Degree Project, data quality, comes in. Companies want to obtain large amounts of useful information, so these data must be of a quality that allows this information to be reliable, free of errors and provide the maximum possible value. Having just data, that is, without quality, greatly harms companies in decision making and dealing with wrong information, resulting in economic and human costs.

This Final Degree Project deals in depth with the topic of data quality, with its respective dimensions that help in the identification of data, the management of this data quality and the problems and benefits of implementing data quality. We will also discuss concepts closely related to data quality, such as metadata, specific information of the data that helps to distinguish them with a higher degree of individuality and the concept of data profiles, as sets of data with the same characteristics that allow us to perform more specific analysis of the data. Theoretically and with a practical example, Data Catalogs, tools that help to provide data quality and that include all the concepts mentioned above, are also addressed.

Finally, we emphasize that this TFG is developed under the UVagile methodology, having a more dynamic approach that allows us to perform the work in an iterative and incremental way, so we get a more successful final result with respect to the objectives set at the beginning of the project.

Key words: Data, Data Quality, Metadata, Data Profiles, Data Catalog.

Índice general

Resumen v

Abstract i

Lista de figuras v

Lista de tablas vii

1. Introducción **1**
1.1. Objetivos 2
1.2. Ejemplo práctico 2
1.3. Estructura del documento 5

2. Gestión del proyecto **7**
2.1. Metodología 7
2.1.1. Scrum 9
2.1.2. Relación entre UVagile y Scrum 13
2.2. Planificación 14
2.3. Presupuesto 17
2.4. Balance 19

3. Calidad de datos **23**
3.1. Tipos de Datos que afectan a la calidad de los datos 23
3.2. El concepto de calidad de datos 24
3.2.1. Los Componentes de la calidad de los datos 24
3.3. Dimensiones de la calidad de datos 25
3.3.1. Medidas cuantitativas y cualitativas de la calidad de los datos . . . 27
3.4. La Gestión de la calidad de los datos 30
3.5. Beneficios 34
3.6. Problemas que enfrenta la calidad 34
3.7. Calidad de datos en Big Data 35

4. Los metadatos y los perfiles de datos **39**
4.1. Metadatos . . . 39
4.1.1. Clasificación . . . 40
4.1.2. Funciones principales . . . 42
4.1.3. Ciclo de vida . . . 43
4.1.4. Ventajas y Desventajas . . . 45
4.2. Perfiles de datos . . . 46
4.2.1. Introducción . . . 46
4.2.2. Proceso de los perfiles de datos . . . 48
4.2.3. Tipos de perfiles de datos . . . 50
4.2.4. Tipos de análisis . . . 50
4.2.5. Ámbitos o Escenarios propicios . . . 51
4.2.6. Beneficios . . . 52

5. Catálogos de datos **53**
5.1. Introducción . . . 53
5.1.1. Por qué surgen . . . 55
5.2. Desafíos a los que se enfrentan . . . 56
5.3. Cualidades que deben tener . . . 57
5.4. Características y funciones . . . 57
5.5. Criterios de elección . . . 58
5.6. Beneficios . . . 63
5.7. Estado del arte . . . 64

6. Parte práctica **73**

7. Conclusiones y trabajo futuro **83**
7.1. Conclusiones . . . 83
7.2. Trabajo futuro . . . 84

Índice de figuras

2.1. Proceso Scrum 13
2.2. Calendario de Sprint 15
2.3. Planificación temporal - Sprint 1 15
2.4. Planificación temporal - Sprint 2 16
2.5. Planificación temporal - Sprint 3 16
2.6. Planificación temporal - Sprint 4 17
2.7. Planificación temporal - Sprint 5 17
2.8. Presupuesto Software 18
2.9. Presupuesto Hardware 18
2.10. Nuevo Calendario de Planificación de Sprint 20
2.11. Planificación temporal - Sprint 5 Extra. 20
2.12. Planificación temporal - Sprint 6 Extra. 20
2.13. Presupuesto Hardware Nuevo 21

3.1. Elaboración propia: Métricas cuantitativas 29
3.2. Elaboración propia: Fases del Data Preparation 31

4.1. Elaboración propia: Ciclo de Vida de los metadatos 43

5.1. Elaboración propia: Usuarios para catálogos de datos 54
5.2. Interfaz Collibra Data Catalog 65
5.3. Interfaz Collibra Data Catalog - Búsqueda 65
5.4. Interfaz Watson Knowledge Catalog 66
5.5. Interfaz Oracle Cloud Infrastructure Data Catalog 67
5.6. Interfaz Qlik Catalog 68
5.7. Interfaz SAP Data Intelligence 69
5.8. Interfaz CKAN 70
5.9. Resumen de herramientas de catálogos de datos 71

6.1. Interfaz Talend Studio 73
6.2. Creación de BBDD 75
6.3. Conexión creada en Talend Studio 75
6.4. Análisis de la Vista General del Catálogo 77
6.5. Análisis de Coincidencias 78

6.6. Análisis de Coincidencias 78
6.7. Análisis de Conjunto de Columnas 79
6.8. Análisis de Conjunto de Columnas 2 79
6.9. Análisis Básico de Columnas 80
6.10. Análisis de la Frecuencia del Patrón 81
6.11. Análisis de la Frecuencia del Patrón Resultados 81

Índice de tablas

1.1. Descripción de negocio: Conjunto Clientes. 3
1.2. Descripción de negocio: Conjunto Proveedor. 4
1.3. Descripción de negocio: Conjunto Producto. 4
1.4. Descripción de negocio: Conjunto Cuentas. 5

3.1. Dimensiones de la calidad de datos. 27

5.1. Criterios elección catálogo de datos. 63

Capítulo 1

Introducción

Este Trabajo Final de Grado, desarrollado en el marco del Grado de Ingeniería Informática de Servicios y Aplicaciones, constituye un trabajo teórico-práctico en el ámbito de la gestión de los datos, cuya finalidad es analizar la importancia de la calidad de los datos en el marco empresarial para una gestión óptima en la que el control de la información, la toma de decisiones y el aumento de los beneficios obtenidos (no solo monetarios) contribuya al progreso de las empresas.

En la actualidad, uno de los activos más importantes para las empresas son los datos. Los datos son *"hechos que describen sucesos o entidades, que por sí mismos no proporcionan significado alguno, pero tienen la capacidad de asociarse dentro de un contexto para transmitir información"* [19].

Los datos son sumamente importantes tanto para las personas como para las empresas, ya que, como afirmaba Francis Bacon, padre del método científico y del empirismo filosófico, *"el conocimiento es poder"* [21]. Las personas generan conocimiento gracias a la adquisición y procesamiento de la información, lo que las permite, entre otras cosas, poder elegir entre más opciones y utilizar más y mejores estrategias para enfrentar distintas situaciones que puedan surgir en sus vidas. En el caso de las empresas, generar más conocimiento es sinónimo de progreso, eficacia y eficiencia.

En este sentido, el procesamiento de datos es fundamental ya que se basa en *"la acumulación y manipulación de elementos de datos para producir información significativa"* [69], que después será utilizada por personas y empresas para generar diferentes conocimientos, seleccionar los más adecuados y aplicarlos a diferentes ámbitos.

Cierto es que los datos son muy valiosos en la toma de decisiones, pero no cualquier dato es igual de válido. Para que los datos sean útiles y poder usarlos de manera adecuada, estos deben ser de calidad, es decir, la información que transmiten estos datos debe representar el mundo real de manera correcta, y esta optimización de los datos se consigue gracias a la implementación y gestión de calidad de los datos. La calidad de datos es un conjunto definiciones, reglas y métricas que permiten evaluar y verificar que los datos estudiados se ajusten a la realidad, además consigue que las empresas puedan utilizar los datos obteniendo el máximo valor para la empresa.

En este trabajo nos centraremos en este aspecto, profundizando en su análisis y ahon-

dando en la implementación y beneficios que trae consigo este tipo de gestión de los datos en las empresas, por la gran importancia que tiene la calidad para el buen funcionamiento y progreso de este tipo de organizaciones.

1.1. Objetivos

El **objetivo general** o principal de este trabajo es demostrar los beneficios que trae consigo la gestión de la calidad de los datos en las empresas.

Para conseguir el objetivo general nos hemos propuesto alcanzar los siguientes objetivos específicos:

OG-01 Comprender qué es la calidad de los datos y sus conceptos asociados.
OG-02 Analizar herramientas de catálogos de datos para la gestión de la calidad de los datos.
OG-03 Analizar la calidad de los datos y su repercusión en los entornos empresariales.

1.2. Ejemplo práctico

Para ilustrar de manera concreta la implementación de la calidad de datos en las empresas y entender con mayor facilidad el proceso, en este trabajo vamos a utilizar un ejemplo práctico basado en un proyecto que se enmarca dentro de un contexto real y actual, que ayudará al lector a comprender mejor qué significado y beneficios tiene la calidad en los datos en una empresa.

Concretamente, este ejemplo práctico que proponemos se trata de un proyecto de migración de datos de un sistema antiguo a un sistema nuevo y más actualizado. El encargo del proyecto viene de parte de una gran empresa de Castilla y León, con más de 30 años de experiencia en el sector de la comercialización de productos para la salud animal. Por razones de privacidad y seguridad, se ha anonimizado cualquier dato que pudiese poner en riesgo a la empresa.

El objetivo que persigue la empresa con la migración de sus datos tiene dos vertientes: utilizar un sistema más amplio y actualizado, donde poder almacenar y gestionar sus datos; Aprovechar la migración para implementar calidad en sus datos, ya que se han experimentado problemas, por no tener en cuenta la calidad de sus datos, en la gestión y en el uso diario de estos datos en el sistema antiguo. Por consiguiente, una vez finalizado el proyecto la empresa contará con datos más completos y certeros a la vez que sean más fácilmente manejables y accesibles.

En este trabajo, nos centraremos en la segunda vertiente: implementar calidad en los datos objeto de la migración, es decir, "limpiar" y aplicar la calidad en todos los datos que la empresa nos ha facilitado para la realización del proyecto.

Los conjuntos de datos que nos han facilitado constituyen los datos maestros,datos considerados clave para las operaciones de la empresa y que permiten un punto de referencia común [18] de la empresa: datos de clientes, datos de proveedores, datos de productos y

datos del plan contable. Cada registro, dentro del conjunto de datos, está identificado por un 'No_' único que permite diferenciar cada dato dentro del conjunto de manera unívoca y un 'Nombre' o 'Descripción' que define como se conocerá al cliente, proveedor, producto o cuenta sobre la que trabajemos. A continuación se observan una serie de tablas para ilustrar estos conjuntos de datos junto a una descripción de negocio de sus atributos (ver Tablas 1.1 a 1.4).

El conjunto de datos de Clientes contiene información sobre aquellas empresas que compran productos proporcionados por nuestra empresa principal y concretamente, además de los mencionados con anterioridad, contamos con una serie de atributos orientados al ámbito financiero: 'Grupo contable de cliente' define la cuenta contable en la que se registran los distintos movimientos que realizan los clientes; 'Términos de pago' indica las condiciones que se deberán cumplir para llevar a cabo las operaciones de venta, como, por ejemplo, el plazo de tiempo en el que se deberá efectuar el pago; 'Forma pago' determina cómo se realiza el pago, es decir, si fue en efectivo, transferencia, etc.; y por último 'Grupo registro IVA negocio' hace referencia al tipo de IVA que le corresponde a cada cliente en cada operación de compra-venta.

ATRIBUTOS CONJUNTO DE DATOS: CLIENTE	
Grupo contable de cliente	Cuenta contable dónde se registran los movimientos contables realizados por el cliente.
Términos de pago	Indica las condiciones que deben cumplirse al realizar una operación de venta.
Forma pago	Refleja como se va a realizar el pago.
Grupo registro IVA negocio	Refleja el tipo de IVA asociado a cada cliente en operaciones de compra-venta.

Tabla 1.1: Descripción de negocio: Conjunto Clientes.

El conjunto de datos Proveedor hace referencia a aquellas entidades que abastecen a nuestra empresa principal con aquellos productos necesarios para el desarrollo de su negocio. Cuenta con atributos muy similares a los clientes como 'Grupo contable proveedor' (a diferencia del atributo en clientes, este hace referencia a la cuenta contable en la cual se registran los movimientos que realizan los proveedores), 'Términos de pago' y 'Forma de pago'. Además, se incluyen dos campos diferentes 'Calle' y 'País', que contienen la localidad donde se encuentran las entidades de los distintos proveedores.

Con respecto al conjunto de datos Productos, podemos decir que la información que contienen es de suma importancia, ya que el negocio principal de nuestra empresa es la comercialización de estos productos para la salud animal. Como atributos este conjunto contiene 'Precio por unidad' y 'Unidad de medida', donde se indica el coste de una unidad del producto y las unidades de medida de ese producto. Aparece el 'Nº Proveedor', que indica qué entidad ha proporcionado un producto concreto, y un campo 'Bloqueado' que muestra si un producto esta listo para vender o no. Equivalente al 'Grupo contable cliente / proveedor' encontramos un atributo 'Grupo contable producto', que nos indica la cuenta

ATRIBUTOS CONJUNTO DE DATOS: PROVEEDOR	
Grupo contable de proveedor	Cuenta contable dónde se registran los movimientos contables realizados por el proveedor.
Términos de pago	Indica las condiciones que deben cumplirse al realizar una operación de venta.
Forma pago	Refleja como se va a realizar el pago.
Calle	Indica la dirección física dónde se encuentra el proveedor.
País	Indica la procedencia del proveedor.

Tabla 1.2: Descripción de negocio: Conjunto Proveedor.

contable a la que corresponde el producto. Al tratarse de una gran variedad de productos encontramos el atributo 'Código Grupo Producto' donde nos indica el tipo del producto, es decir, si se trata de productos alimenticios, inflamables o médicos.

ATRIBUTOS CONJUNTO DE DATOS: PRODUCTO	
Precio por unidad	Indica el coste por una unidad de un producto concreto.
Unidad de medida	Define en qué se mide ese producto (Kg, litros, etc.).
Nº Proveedor	Proporciona el código del proveedor que proporciona el producto.
Bloqueado	Señala si un producto está disponible para venderse en ese momento.
Grupo contable producto	Cuenta contable que tiene asociada un producto.
Código Grupo Producto	Qué tipo de producto puede ser (alimenticio, inflamable, médico).

Tabla 1.3: Descripción de negocio: Conjunto Producto.

Por último, el conjunto de datos Cuentas contiene la información del Plan contable de la empresa, dicho de otro modo, nos ofrece el conjunto de normas establecido por la empresa para llevar un registro de la actividad económica de la misma. Este conjunto contiene únicamente tres atributos: una 'Cuenta' referente al registro donde se identifican todas las operaciones que realiza la empresa, un 'Nombre' por el cual se conoce a la cuenta contable y 'Tipo movimiento' que indica el tipo de cuenta que es, Mayor o Auxiliar. Donde una cuenta de Mayor es una cuenta general que alude a una operación general, por ejemplo, la cuenta 6 corresponde al registro de todas las operaciones referentes a las Compras y Gastos dentro de la empresa. Mientras que una cuenta Auxiliar, se emplea para operaciones específicas dentro del concepto general que se expone en la cuenta de Mayor, y que empiezan por el mismo dígito que la cuenta de Mayor, por ejemplo, la cuenta 600000001 corresponde a Compras de mercaderías.

Hemos de advertir que, aunque los datos, la empresa y la demanda del proyecto son reales, el proceso que se muestra durante todo el documento es simulado, es decir, es una muestra de cómo debería de aplicarse la calidad de datos según los conceptos, argumentos

ATRIBUTOS CONJUNTO DE DATOS: CUENTAS	
Cuenta	Contiene el número de las diferentes cuentas donde se registran todas las operaciones que realiza la empresa.
Nombre	Nombre identificativo por el cuál se conoce a la cuenta.
Tipo movimiento	Indica si una cuenta es de Mayor (cuenta genérica) o Auxiliar (Cuentas específicas, que pertenecen a subgrupo de una cuenta de Mayor).

Tabla 1.4: Descripción de negocio: Conjunto Cuentas.

y procesos teóricos recopilados y analizados en este trabajo.

Por este motivo, no contamos con todos los datos de la empresa, sino con aquellos que, para el objetivo de nuestro trabajo, puedan ser utilizados sin perjuicio grave para la empresa. Aun así, la ejemplificación de este supuesto práctico aportará concreción a la información teórica expuesta en este documento y ayudará a la comprensión y aplicación de los procesos de implementación de calidad en las empresas si en un futuro se requiriesen estos servicios.

Para reflejar este caso práctico, junto a la teoría habrá apartados donde mostremos el contenido teórico con un ejemplo cogiendo los datos explicados en el caso práctico. Además, en la sección 6 veremos la implementación en su conjunto, la cuál sigue la misma estructura y orden que el desarrollo de este documento.

1.3. Estructura del documento

En esta sección se muestra brevemente un resumen del contenido del proyecto y su organización a lo largo de este documento.

- **Capítulo 1 - Introducción.** Capítulo inicial del proyecto, se expone un ejemplo de un proyecto real en el que nos basamos para explicar los conceptos definidos a lo largo del documento, así como los objetivos a perseguir con este trabajo.
- **Capítulo 2 - Gestión del proyecto.** En este capítulo se explica la metodología usada para este proyecto, la planificación seguida durante el desarrollo del proyecto y los presupuestos y balances asociados a él.
- **Capítulo 3 - Calidad de datos.** Este capítulo comprende aspectos relacionados con la calidad de los datos, que nos permite conocer en profundidad este concepto. Mencionaremos diferentes tipos de datos a tener en cuenta ya que podrían afectar a la calidad de los datos, por qué componentes se forma o las dimensiones en las que se divide. Nos centraremos en la gestión de la calidad de los datos. Listaremos los distintos problemas a los que se enfrenta una empresa a la hora de implantar calidad en los datos y los beneficios que trae consigo hacerlo. Y como último apunte se expondrá brevemente cómo afecta la calidad de los datos en proyectos donde se usen tecnologías Big Data.

- **Capítulo 4 - Los metadatos y los perfiles de datos.** Este capítulo se compone de dos conceptos de suma importancia para la calidad de los datos. Por un lado, se encuentra el concepto de metadato, del cual haremos una clasificación de los distintos tipos que podemos encontrar, funciones principales que aportan, mostraremos el ciclo de vida por el que pasa cualquier metadato y las ventajas y desventajas que se producen por el uso de estos. Por otro lado, abordaremos el concepto de perfil de datos, como herramienta de utilidad para aportar Calidad a los Datos. Mencionaremos el proceso por el cual obtenemos un perfil de datos, distintos tipos de perfiles y distintas técnicas de análisis que usan los perfiles, los ámbitos o escenarios donde el uso de perfiles es especialmente beneficioso y los beneficios que trae el uso de estos perfiles a un proyecto.

- **Capítulo 5 - Catálogos de datos.** En este capítulo explicaremos el concepto de catálogo de datos y el porqué de su aparición como herramienta de calidad de los datos. Veremos algunos de los desafíos a los que se enfrentan estas herramientas al incorporarlas en una empresa. Destacaremos las cualidades que debe tener un buen catálogo de datos para convertirse en una herramienta eficaz, así como las características y funciones más importantes que debería implementar. Señalaremos los criterios por los que se deben guiar las empresas a la hora de elegir el catálogo de datos más apropiado para ellos y sus proyectos. También listaremos los beneficios que trae consigo usar estos catálogos. Por último, definiremos el estado del arte, explicando algunas de las herramientas de catálogos de datos más usadas y conocidas, junto algunas de sus características destacadas.

- **Capítulo 6 - Ejemplo Práctico.** Este es un capítulo que mostrará cómo se gestiona la calidad de los datos de forma práctica, utilizando un catálogo de datos. Además, nos permitirá identificar los distintos conceptos explicados a lo largo del documento de una forma más visual y con un caso de un proyecto real.

- **Capítulo 7 - Conclusiones y visión futura.** Capítulo final del proyecto donde identificaremos distintas conclusiones que surgen de la realización del proyecto y mostraremos una visión de cómo se podría abordar el proyecto a futuro.

Capítulo 2

Gestión del proyecto

En este apartado nos centraremos en definir y comprender la metodología utilizada en el desarrollo del proyecto. Comenzaremos realizando un recorrido por el concepto de esta metodología, para seguir con los detalles de la planificación que se ha llevado a cabo a lo largo del proceso, así como el presupuesto previsto y el balance obtenido al finalizar el proyecto.

2.1. Metodología

En este proyecto hemos decidido optar por una **metodología ágil**. Agile se define como un conjunto de metodologías, valores y principios utilizados para el desarrollo de proyectos cuya esencia radica en seguir un proceso rápido y flexible en el que, a medida que va avanzando este proceso, se puedan ir afianzando y concretando los objetivos de acuerdo a las necesidades del cliente, aclarando las ideas generales y valorando, en el caso de que sea necesario, otras vías de trabajo que se puedan incorporar para que el resultado final obtenga un grado de éxito mayor [5].

Este tipo de metodologías surgieron debido a las necesidades de negocio que han ido ampliándose cada día. Estas necesidades se derivan de problemas que se han ido identificando en el desarrollo de este tipo de proyectos donde, por ejemplo, se ha detectado un elevado grado de incertidumbre por parte de los clientes, a la hora de expresar de forma clara, concreta y concisa sus requisitos, finalidades o expectativas, especialmente al inicio del proyecto. Entre los problemas que podemos encontrar destacamos: (1) el cliente no tiene definidos y/o claros los requisitos al inicio del proyecto, (2) los requisitos definidos al principio del proyecto pueden variar durante el proceso de desarrollo, (3) la existencia de cambios inminentes que no estaban previstos y que provocan consecuencias negativas en el proyecto, como un aumento en la duración de ejecución, sobrepaso del presupuesto establecido inicialmente, etc., (4) no haber sido capaz de visualizar el alcance del proyecto propuesto al inicio del proceso. Estos y otros problemas generan una insatisfacción por parte del cliente [39].

En este contexto, las metodologías ágiles sirven para prever estos problemas tanto al inicio del proyecto, como en su desarrollo. Aportan una cierta flexibilidad y revisión

continua del proyecto, lo que conlleva poder adaptarse a los cambios que puedan surgir y facilitan el poder tomar la mejor decisión en cada momento, atendiendo a los imprevistos, sin comprometer el estado del proyecto.

Su utilización empezó siendo para proyectos software, sin embargo, actualmente podemos encontrar proyectos dirigidos de forma ágil en casi cualquier ámbito de negocio, haciendo que estos proyectos sean más flexibles y adaptables a las nuevas tecnologías y generando un mayor índice de éxito.

En 2001, se estableció un consenso entre diecisiete expertos en ingeniería del software que supuso el boom de las metodologías ágiles como alternativa a las metodologías tradicionales para el desarrollo del software. Este acuerdo se conoce como Manifiesto Ágil[1] y en él se recogen cuatro valores fundamentales:

1. **Individuos e interacciones sobre proceso y herramientas.** Se expone que el elemento fundamental en cualquier tipo de proyecto son las personas, ya que son estas quienes aportan los medios necesarios para llevar a cabo el proyecto con éxito.

2. **Software funcionando sobre documentación extensiva.** Una buena documentación sigue siendo fundamental, sin embargo, las metodologías ágiles abogan por un producto funcional e intuitivo antes que una documentación elaborada ya que esto es lo que realmente valoran los usuarios finales y permite saber con una mayor exactitud sus necesidades.

3. **Colaboración con el cliente sobre negociación contractual.** Uno de los puntos principales es la continua comunicación entre la empresa y sus clientes. Durante el desarrollo del proyecto surgen cambios que ponen en riesgo ciertas partes del proyecto, como pudiera ser que un producto se quede obsoleto en el mercado. Por este motivo es imprescindible que los clientes y la empresa establezcan diálogos frecuentes en las distintas etapas del proyecto, ya que, es beneficioso para ambas partes.

4. **Respuesta ante el cambio sobre seguir un plan.** La gran capacidad de adaptación que tienen este tipo de metodologías hace que las empresas puedan reaccionar de forma rápida ante cualquier imprevisto que pueda surgir. Esta capacidad es esencial en proyectos con un grado de incertidumbre alto.

De estos cuatro valores se derivan doce principios que también se pueden encontrar en el Manifiesto Ágil. En ellos se exponen unas reglas o normas que se han de cumplir para manejar el método Agile y utilizarlo de manera eficaz y eficiente.

Otro de los aspectos que se han de tener en cuenta para trabajar con estas metodologías son los **marcos de trabajo ágiles**, que son conjuntos de herramientas, tareas y procesos utilizados para organizar, planificar y ejecutar los distintos proyectos, desde su inicio hasta su finalización [35]. Estos marcos de trabajo se fundamentan en los cuatro valores y los doce principios de las metodologías Agile. Los más conocidos son *Kanban*, que se centra

[1] `https://agilemanifesto.org/`

en la optimización del flujo de trabajo [74]; *Scrum*, cuyo principal objetivo es maximizar el valor del trabajo realizado [74]; y *XP (Extreme Programming)* que se centra en potenciar las relaciones personales entre la empresa y sus clientes [74].

En nuestro proyecto, nos hemos decantado por un **marco de trabajo Scrum** que ha guiado al equipo de trabajo durante todo el proyecto tanto en la planificación y cronograma de las interacciones y acciones, como en el reparto de tareas, normas que se han de seguir y otras cuestiones que detallaremos en el siguiente subapartado. Debemos destacar que todo el equipo es Scrum Máster cuyo certificado está avalado por la organización *European Scrum*.

Más concretamente hemos seguido una **metodología UVagile** [1], basada en las metodologías ágiles y en el marco de trabajo Scrum descritas en el artículo JENUI 2020: *Hacia la consolidación de las aulas ágiles* [40]. Esta iniciativa surgió durante el curso 2018-2019, en el ámbito universitario, específicamente en el Grado de Ingeniería Informática de Servicios y Aplicaciones en el Campus de Segovia, con el objetivo de implantar metodologías ágiles en el ámbito de la enseñanza universitaria.

2.1.1. Scrum

En esta sección nos centraremos en explicar cómo funciona el marco de trabajo Scrum, con el que hemos estado trabajando. Toda la información contenida en esta sección es aportada por La Guía de Scrum de 2020 [49], escrita por Ken Schwaber y Jeff Sutherland.

Scrum es un framework para el desarrollo y mantenimiento de productos complejos. El equipo de trabajo puede abordar problemas complejos de una forma adaptativa y hacer entregas de forma iterativa y eficiente aportando el máximo valor (incremento). Scrum emplea un proceso iterativo e incremental basándose en procesos empíricos, y se compone de diferentes elementos donde cada uno de ellos tiene un papel específico y es esencial para tener éxito.

Estos elementos son:

1. Equipo Scrum. Roles.
2. Eventos.
3. Artefactos.
4. Reglas asociadas.

1. EQUIPO SCRUM

La esencia de Scrum es pequeños equipos de personas, preferiblemente entre 5 y 11 personas dentro de un mismo grupo. El equipo individualmente es muy flexible y adaptativo. Son auto-organizados y multifuncionales, por lo que tienen las habilidades necesarias para desempeñar las funciones y procesos necesarios para el desarrollo del proyecto sin depender de personas ajenas a él.

Se compone por:

- ***Product Owner* o Propietario del Producto.**

  Este rol lo asume *una única persona* del equipo. Su principal responsabilidad es la de maximizar el valor del producto llevado a cabo por el Equipo de Desarrollo.

  Sus funciones dentro del equipo Scrum son: decide sobre lo que hay que hacer en el proyecto. Debe tener un conocimiento absoluto sobre el producto y las necesidades del cliente y crear una visión general del producto. Hace de intermediario entre el equipo de desarrollo y los *stakeholders*[2]. Crea, mantiene y prioriza el Product Backlog (uno de los artefactos de Scrum) y además debe expresar claramente todos sus elementos al equipo, para facilitar su entendimiento y transparencia. Debe participar en las reuniones organizadas.

- ***Development Team* o Equipo de Desarrollo.**

  Se compone de **entre 3 y 9 profesionales**, los cuales se encargan de entregar los incrementos de producto que potencialmente esté 'Terminado' y puedan aportar valor al producto final. Dentro de sus funciones deben auto-organizarse y gestionar su propio trabajo.

  Entre sus principales características destacan: trabajan de manera colaborativa. Son quien decide cómo cumplir con los objetivos de cada sprint (uno de los eventos de Scrum); administran el *Sprint Backlog* (un artefacto de Scrum); y deben participar en las reuniones que se realizan a diario.

- **Scrum Master.**

  Existe **un único responsable** para adoptar este rol. Su principal papel es ayudar a todo el equipo Scrum y a personas externas a entender qué interacciones pueden ser útiles y cuales no para el éxito del proyecto.

  Sus principales ocupaciones son: está al servicio del equipo, eliminando cualquier impedimento que pueda ocurrir; formar al *Product Owner* y al *Development Team* en las técnicas y pasos a seguir con las metodologías ágiles; liderar sin sobreponerse con autoridad, protege y guía al equipo; encargarse de gestionar que se lleven a cabo todas las ceremonias.

2. EVENTOS

Scrum define una serie de eventos que se llevan a cabo a lo largo del proyecto, cuyo propósito es crear regularidad y minimizar la necesidad de reuniones extra no definidas en Scrum. Todos los eventos se definen en un periodo de tiempo determinado, de acuerdo a su objetivo.

Los objetivos de estos eventos son: ayudar en la inspección y adaptación del proyecto, establecer transparencia en el proyecto, lo que ayuda a comprender mejor en qué estado se encuentra el mismo y las necesidades que se deben abordar.

[2]Todas aquellas personas o entidades afectadas en mayor o menor medida por las decisiones y actividades que se desarrollan en la empresa y que permiten, a su vez, el completo funcionamiento de la misma. Fuente: `https://protecciondatos-lopd.com/empresas/stakeholders/`

Los distintos eventos son:

- **Sprint.**

  Es el evento principal de Scrum. Su duración depende del tipo de proyecto, donde se establece entre dos y cuatro semanas (un mes) de duración, a lo largo del cual se debe crear y entregar al final un incremento de producto 'Terminado'. Su duración, una vez establecida, no puede variar y todos los sprints de los que se compone el proyecto deben tener la misma duración. El proceso es un ciclo, donde una vez acaba un sprint empieza inmediatamente el siguiente. Debe participar todo el Equipo Scrum.

  Los demás eventos están integrados dentro de cada sprint.

- **Planificación del Sprint o *Sprint Planning.***

  Es la reunión realizada al principio de cada sprint, donde se decide qué alcance va a abordar ese sprint, y debe participar todo el Equipo Scrum. Su duración es proporcional a la duración del sprint, de tal forma que si un sprint es de 3 semanas la planificación del sprint tendrá un máximo de 4 horas.

  Al final de la reunión todo el equipo debe tener claro cuál es el objetivo para ese sprint y cómo conseguirán dicho fin. El Equipo de Desarrollo se reparte las tareas de una manera equitativa y proporcional al alcance del sprint.

- **Scrum Diario o *Daily Scrum.***

  Reunión diaria en la que participa el Equipo de Desarrollo y opcionalmente el Scrum Master, donde el equipo de desarrollo se junta para ver en qué punto se encuentran y planear el trabajo de las siguientes 24 horas, optimizando así la colaboración, comunicación y el desempeño del equipo. Se debe realizar siempre en un horario fijo y con una duración máxima de 15 minutos.

  Durante esta reunión cada miembro del equipo de desarrollo deberá contestar a tres preguntas:

  - ¿Qué hice ayer por el equipo?
  - ¿Qué haré hoy por el equipo?
  - ¿Qué bloqueos me impiden conseguirlo?

- **Revisión del Sprint o *Sprint Review.***

  Se lleva a cabo al final de cada sprint con el objetivo de mostrar a los *stakeholders* el progreso conseguido y las tareas finalizadas en ese sprint y estos plantean las dudas que puedan tener. Participa todo el equipo Scrum junto a los *stakeholders*, en un tiempo proporcional a la duración del sprint, en caso de tres semanas de sprint la revisión del sprint durará un máximo de 2 horas.

- **Retrospectiva del Sprint o *Sprint Retrospective*.**

  Es la última reunión del sprint, de carácter interno, ya que participan todos los miembros de equipo Scrum, y cuya finalidad es crear una oportunidad para que el equipo se inspeccione a sí mismo y poder crear un plan de mejoras con vistas al nuevo sprint.

  Su duración máxima sería de una hora y media para un sprint de tres semanas. En este sprint es fundamental que se hable sobre cómo fue el último sprint, identificar qué salió bien y qué se puede mejorar.

3. ARTEFACTOS

Los artefactos de Scrum son herramientas que representan en todo momento el trabajo y valor de las tareas realizadas por el equipo Scrum. Aseguran la transparencia en la información, lo que ayuda al equipo a entender en todo momento en que estado se encuentra cada sprint y por tanto el proyecto.

Se dividen en tres artefactos:

- **Pila del Producto o *Product Backlog*.**

  Lista priorizada que contiene todo lo necesario para completar el producto, y es la única fuente de requisitos para cualquier cambio que pudiera realizarse en el producto. El *Product Owner* es el único responsable de gestionar el *Product Backlog*, incluyendo su contenido, disponibilidad y ordenación.

  Nunca llega a completarse, sino que refleja solo los requisitos conocidos y entendidos al principio y va evolucionando conforme al desarrollo del producto y su entorno. Es un artefacto dinámico, en constante cambio para identificar lo que necesita el producto para ser adecuado, competitivo y útil.

  Se priorizan aquellas tareas que aporten un valor más significativo, acorde a la visión de negocio y las necesidades del cliente.

- **Pila del Sprint o *Sprint Backlog*.**

  Conjunto de elementos del *Product Backlog* escogidos para el sprint actual, más un plan establecido para entregar el incremento de producto y conseguir el objetivo del sprint. Es gestionada por el Equipo de Desarrollo y nos aporta una visión del avance que llevan en el sprint.

- **Incremento o *Increment*.**

  Constituye la suma de todos los elementos 'Terminados' del *Product Backlog* durante el sprint y todos los anteriores completados. Al final del sprint se le pone la marca de 'Terminado' cuando el incremento del producto está en condiciones de ser utilizado y cumple con la definición (*Definition of Done*) que impone el equipo Scrum.

En la figura 2.1 [3] mostramos un gráfico ilustrando cómo sería el proceso Scrum general:

[3] Fuente imagen: `https://xn--zoraidaceballosdemario-4ec.info/scrum/zoraida-ceballos-de-marino-scrum-que-es-y-para-que-sirve-esta-metodologia/`

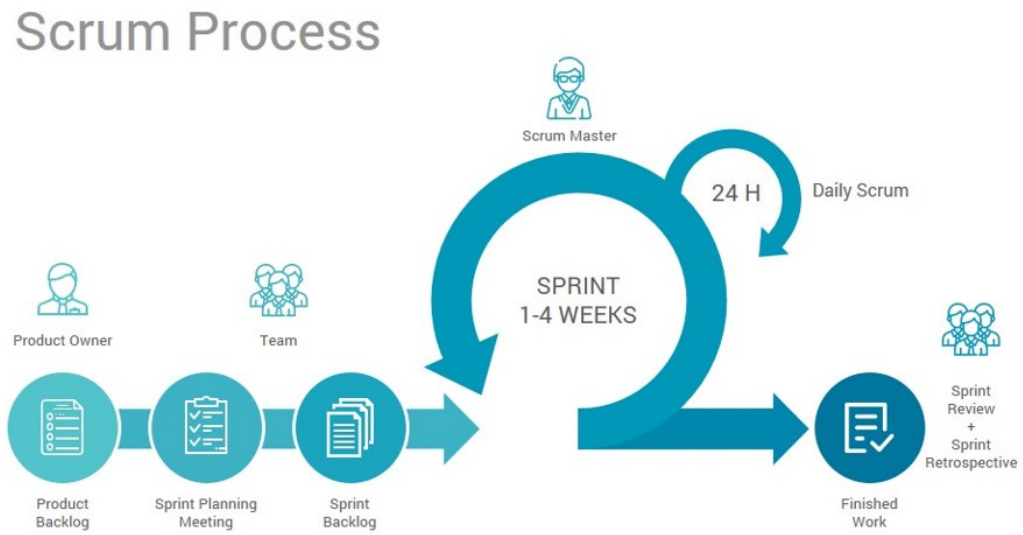

Figura 2.1: Proceso Scrum

2.1.2. Relación entre UVagile y Scrum

UVagile se basa en el marco de trabajo Scrum por lo que toma parte de sus conceptos. Sin embargo, UVagile trabaja aspectos que difieren con el marco de trabajo estándar.

Los cambios que presenta UVagile frente Scrum son [40]:

- UVagile adapta el marco de trabajo Scrum para su utilización en **metodologías docentes**. UVagile se lleva a cabo en un **aula ágil**, un entorno de enseñanza aprendizaje colaborativo.
- El **equipo de trabajo** esta formado por el Profesor realiza el rol de *Product Owner*, quien se encarga de facilitar el proceso de aprendizaje y mantener un entorno de confianza mutua. Y el Alumno, que cumple el papel de Scrum Master y Equipo de desarrollo cuya función es abordar el proceso de aprendizaje de forma activa, desde un punto de vista colaborativo y comunicación constante y fluida.
- Con respecto a los artefactos se organizan en el **tablero de aprendizaje**, accesible por todos los miembros del equipo de trabajo, muestra una visión actualizada del estado del aula ágil. Se divide en dos componentes principales: el ***Backlog de Aprendizaje***, similar al *Product Backlog* de Scrum, donde se organizan las historias de aprendizaje, y el ***Sprint Backlog*** donde se planifican las tareas de un sprint. Además, en este tablero puede aparecer un número variable de columnas que muestre en más detalle el estado de cada una de las historias de aprendizaje, por ejemplo, *Pendiente*, *En curso* y *Terminado*.
- La duración de estos proyectos equivale al tiempo que dura el TFG, y éstas se organizan, al igual que en Scrum, por sprints. Al inicio se empieza con una **reunión**

inicial para conocer el aprendizaje que se va a abordar y dividir los futuros sprint y su contenido, cada sprint corresponde con un objetivo de aprendizaje, hisotrias de aprendizaje. Se repite esta reunión al inicio de cada sprint para planificar ese sprint concreto. Se realizan **reuniones semanales** de seguimiento (similares a las *Daily Scrum*), *Weekly*, fijando un mismo día y hora, donde todo el Equipo de trabajo participa. Por último, se fijan las **retroplaning**, una reunión al final de cada sprint donde obtenemos *feedback* en tiempo real, que nos permite identificar rápidamente posibles problemas y carencias en el aprendizaje, para así solucionarlo y, además, se fijan los objetivos para el siguiente sprint.

2.2. Planificación

La planificación de este trabajo se ha realizado bajo la metodología UVagile, definida anteriormente. Al inicio del trabajo, realizamos una reunión de Inicio del proyecto, con una visión global de todo el trabajo, donde acordamos los objetivos en los que nos íbamos a centrar. Definimos, que para conseguir el objetivo final distribuiríamos el trabajo en 5 sprint, de 3 semanas de duración y 60 horas de trabajo cada uno. Las reuniones de seguimiento (Weekly) y final de sprint (Retroplaning) se fijaron a los martes por la mañana. La organización y gestión de las tareas, se hizo a través de la herramienta Trello, pues nos daba una visión global de los artefactos Product Backlog y Sprint Backlog. Al inicio de cada sprint, en la reunión de planificación, concluimos los objetivos de ese sprint y cómo se iba a abordar, identificando y dividiendo distintas tareas que sean asumibles para entregar un incremento de valor al final de ese sprint, y se describen como Historias de usuario. A estas historias de usuario se les asigna un fecha de vencimiento y unos criterios de aceptación, los cuales se deberán cumplir para considerar que este completamente acabada. Además, se les asigna un nivel de dificultad medida en puntos de historia, que se rigen por la escala de Fibonacci (1-2-3-5-8-13), donde una historia de usuario con un gran número de puntos de historia (13) deberá ser replanteada y dividida en historias más simples, mostrando la magnitud de dicha historia de usuario. Estos puntos de historia, representan el grado de dificultad que pueda tener una historia de usuario y advierte al equipo de este hecho, pero no equivale a las horas que puedan pasar para completar la historia de usuario.

El desarrollo de un proyecto es complejo de estimar y planificar, aunque nos guiaremos por metodologías ágiles, sin embargo, usar este tipo de metodologías nos da la capacidad de ir abordando todo el trabajo a medida que avanzamos en él. A continuación, mostraremos un calendario con la distribución que se planteó inicialmente, Figura 2.2 y detallaremos brevemente cómo estaba organizado cada sprint.

Una vez establecido el calendario con los sprints y las tareas dentro de cada sprint, junto a sus fechas de duración y puntos de historia estimados, debemos estimar las horas que nos podría llevar cada historia de usuario. Sin embargo, la estimación de las horas en Scrum se basa en experiencia obtenida de otros proyectos parecidos y en nuestro caso no existe tal experiencia, por lo que la estimación puede presentar diferencias con la realidad.

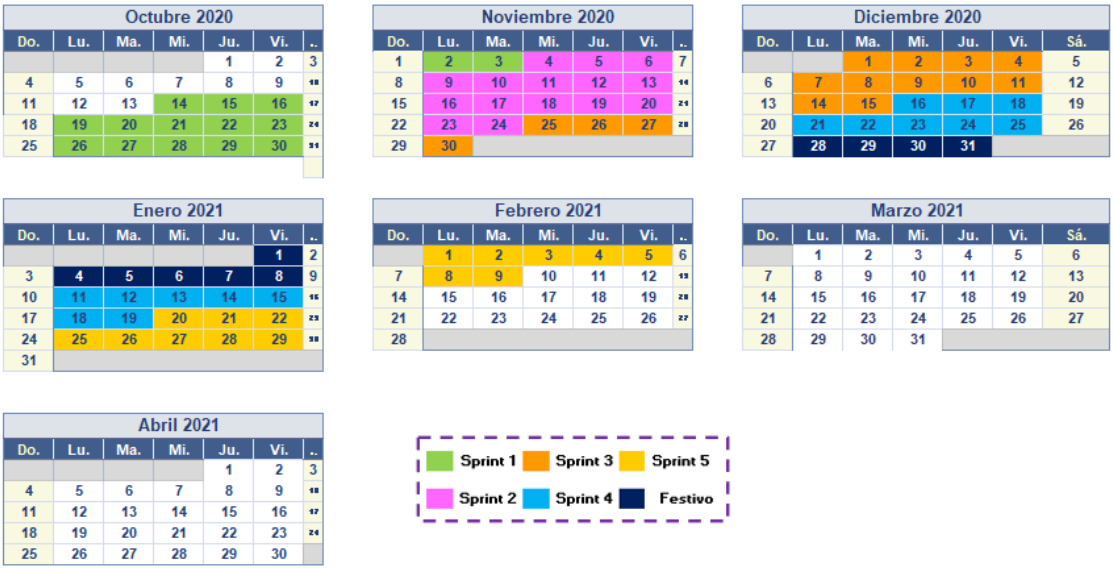

Figura 2.2: Calendario de Sprint

Para nuestro proyecto repartimos 60 horas para cada sprint, las cuales se distribuyeron proporcionalmente entre las tareas basándonos en su complejidad. Este fue uno de los puntos por los que se escogió esta metodología y no una tradicional, que se basa en el tiempo y el alcance para estimar los costes de un proyecto. Gracias a ello podemos ir avanzando en el trabajo de forma continua, marcando los objetivos más cercanos y fácilmente planificables. En las siguientes tablas (Figuras 2.3 a 2.7) mostraremos el desglose de cada uno de los sprint, con las distintas tareas, sus puntos de historia y las horas estimadas. Además para facilitar la visualización y entendimiento del presupuesto, calcularemos el coste de cada tarea y sprint en función de las horas estimadas.

- **Sprint 1**

Sprint 1

Historias de Usuario	Puntos de historia	Fecha Inicio	Fecha Fin	Horas estimadas
Lectura documentación sobre Calidad de datos	13	14/10/2020	24/10/2020	29
Punto Calidad de Datos	13	25/10/2020	02/11/2020	31
Total Ptos.	**26**		Total horas.	**60**

Figura 2.3: Planificación temporal - Sprint 1

El primer sprint del proyecto se compone de dos historias de usuario complejas que han sido dividas en varias subtareas. El objetivo de este sprint es tomar contacto con el tema que se va a tratar a lo largo del documentom y concretamente abordar el tema de calidad de datos. El incremento de este sprint es el entendimiento del tema principal del trabajo y la documentación del conocimiento adquirido en la memoria.

- **Sprint 2**

Sprint 2

Historias de Usuario	Puntos de historia	Fecha Inicio	Fecha Fin	Horas estimadas
Lectura documentación Catálogos de datos	13	05/11/2020	10/11/2020	17
Punto Catálogo de Datos (Parte 1)	8	11/11/2020	23/11/2020	30
Herramientas de Catálogo de Datos	8	17/11/2020	23/11/2020	13
Total Ptos.	29		Total horas.	60

Figura 2.4: Planificación temporal - Sprint 2

Para este segundo sprint nos centramos en un punto principal del proyecto como son los catálogos de datos, abordando conceptos relacionados con estos y un estudio breve sobre su estado del arte, mostrando algunas de las herramientas más conocidas. El incremento es obtener el conocimiento necesario sobre catálogos de datos y dejarlo plasmado en la memoria, junto a un breve estudio sobre distintas herramientas de catálogos de datos.

- **Sprint 3**

Sprint 3

Historias de Usuario	Puntos de historia	Fecha Inicio	Fecha Fin	Horas estimadas
Punto Catálogo de Datos (Parte 2)	8	25/11/2020	08/12/2020	20
Lectura Documentación Metadatos	8	09/12/2020	12/12/2020	10
Punto de Metadatos	8	09/12/2020	14/12/2020	30
Total Ptos.	24		Total horas.	60

Figura 2.5: Planificación temporal - Sprint 3

El sprint 3, se centra en acabar el capítulo dedicado a catálogos de datos y recolectar e incluir información sobre metadatos, quedando redactado en la memoria y consiguiendo así el incremento de este sprint.

- **Sprint 4**

En el sprint 4 nos centraremos en el último punto de teoría sobre el que tratará el trabajo, los perfiles de datos, y en realizar un ejemplo práctico a través de una de las herramientas de catálogo de datos estudiadas con anterioridad. El incremento para este sprint consistirá en dejar constancia en la memoria de todos los conceptos relacionados sobre los perfiles de datos, además del entendimiento y demostración del funcionamiento de una herramienta de catálogo de datos sobre un conjunto de datos concreto.

Sprint 4

Historias de Usuario	Puntos de historia	Fecha Inicio	Fecha Fin	Horas estimadas
Lectura Documentación Perfiles de Datos	8	15/12/2020	21/12/2020	10
Punto Perfiles de Datos	8	16/12/2020	25/12/2020	20
Parte práctica de la memoria	13	11/12/2020	19/01/2021	30
Total Ptos.	29		Total horas.	60

Figura 2.6: Planificación temporal - Sprint 4

Sprint 5

Historias de Usuario	Puntos de historia	Fecha Inicio	Fecha Fin	Horas estimadas
Completar datos faltantes en la memoria	8	20/01/2021	01/02/2021	30
Revisión de la memoria	8	01/02/2021	08/02/2021	20
Crear Presentación de la memoria	5	04/02/2021	09/02/2021	10
Total Ptos.	21		Total horas.	60

Figura 2.7: Planificación temporal - Sprint 5

- **Sprint 5**

  Para este último sprint se prevé que trate con los puntos comunes del trabajo, como son la documentación de un presupuesto y balance, la metodología usada y las conclusiones sacadas. Además se realizará una revisión exhaustiva del documento entero y por último una breve presentación para ilustrar el contenido de la memoria durante la defensa del proyecto.

Una vez mostrado el desglose de cada sprint con sus respectivas tareas y horas estimadas, podemos decir que este proyecto deberá suponer un **esfuerzo de 300 horas y 129 puntos de historia totales** de trabajo. Las horas estimadas coinciden con lo pactado en la Planificación del sprint y se corresponden con las horas máximas que debe llevar un Trabajo Fin de Grado.

2.3. Presupuesto

La realización de un presupuesto en nuestro caso es una tarea compleja, ya que no tenemos la suficiente información sobre que herramientas vamos a necesitar y no tenemos la experiencia suficiente en proyectos similares para prever situaciones inesperadas que nos lleven a tener que dedicar más horas de trabajo.

Por ello para construir este presupuesto nos basamos en la información de la cual tenemos certeza al inicio del proyecto, que nos ayudará a forjarnos una idea inicial del coste que nos podría suponer este trabajo.

Obtener el presupuesto supone tener en cuenta **tres componentes principales**, Costes de Software, Costes de Hardware y Costes Humanos. Los costes se calculan en función del porcentaje de uso y el tiempo de uso.

1. **Costes Software:** En la tabla siguiente podemos observar los componentes software que se van a utilizar y el coste que nos supondrá su utilización. Al contar con software libre los costes serán nulos.

Programas	Coste/mes	%Uso	Meses de Uso	Coste Real
Talend Open Studio for Data Quality	0	100	2	0
phpMyAdmin	0	100	2	0
Overleaf	0	100	5	0
Microsoft Teams	0	100	5	0
Trello	0	100	5	0
			Coste total:	**0**

Figura 2.8: Presupuesto Software

2. **Costes Hardware:** En cuanto al coste que proviene del hardware, primero debemos considerar el precio por el que nos saldría la utilización del ordenador por meses. Para este proyecto se considerará una frecuencia de reemplazo de 6 años. El precio medio de un ordenador, que pueda rendir en condiciones óptimas, sería de unos 700 €, por lo que el coste al mes por ese PC sería de 9,72 €/mes.

Componentes	Coste/mes	%Uso	Meses de Uso	Coste Real
Internet	36	100	5	180
Ordenador	9'72	100	5	48'6
			Coste total:	**228'6**

Figura 2.9: Presupuesto Hardware

3. **Coste Humano:** Para obtener los costes humanos, hemos tomado como referencia el sueldo medio de un analista de datos junior, ya que consideramos que es el perfil más adecuado en este proyecto. El sueldo corresponde con 26.800 €/año [4],partiendo de ese sueldo incluyendo costes sociales y salario del trabajador [5] el coste real sería 34.840 €/año. Con ello calculamos el sueldo promedio por hora, basándonos en las horas laborables máximas que un trabajador puede asumir según lo dicta el Estatuto de Trabajadores y que son unas 1.826 horas, lo que supone un sueldo bruto de

[4] `https://www.linkedin.com/salary/`
[5] `https://factorialhr.es/blog/coste-empresa-trabajador/`

19,08 €/h. Con el sueldo por horas y las horas obtenidas en la planificación de los diferentes sprints podemos obtener el coste total humano.

Coste Humano Total = Nº horas estimadas * Sueldo / hora

C.Humano Total = 300 h * 19,08 €/h = 5.723 €

Una vez obtenidos los presupuestos individuales, debemos calcular el precio total final que nos saldría para este proyecto:

Total Presupuesto = C.Software + C.Hardware + C.Humanos

Total Presupuesto = 0 + 228,6 €+ 5.723 € = **5.952,59 €**

2.4. Balance

Una vez avanzado el proyecto, siguiendo la planificación especificada, surgieron una serie de complicaciones que tuvieron un impacto directo en el proyecto. Durante el desarrollo del trabajo, tuvimos que coordinarlo junto a otra actividad exigente como son las Prácticas de Empresa, por lo que no se pudo cumplir las 60 horas por cada sprint y en consecuencia la carga de trabajo se volvió menor. Este imprevisto no planteó un cambio en el alcance ni una planificación de horas, sino una reordenación del trabajo previsto a lo largo de más sprints, añadiendo dos sprints extra.

Antes de determinar la decisión de incluirlos, tuvimos que tener presente que esto no afectará a los tiempos y objetivos finales del proyecto. Teniendo la certeza que no afectaría gravemente al proyecto pudimos añadir los nuevos sprints y reorganizar la planificación inicial, trasladando el alcance previsto entre los sprints 5, 6 y 7.

A continuación, podemos observar cómo quedaría el nuevo calendario de la planificación de sprint:

Este cambio en el calendario afecta al proyecto en un incremento de las horas de trabajo, por lo que consecuentemente aumentan los gastos. Para anotar este aumento mostraremos el desglose de estos nuevos sprints y las consecuencias que traen consigo. Estos sprints se corresponden con el sprint 5 extra y el sprint 6 extra, ya que decidimos incluir los nuevos sprints antes de el último sprint previsto en la planificación. Para el **nuevo sprint 5** trabajamos refinando los conceptos planteados anteriormente de calidad de datos, metadatos y catálogos de datos, mientras que en el **sprint nuevo 6** refinamos el concepto de perfil de datos y realizamos el ejemplo práctico basado en una herramienta de catálogo de datos.

Con estos nuevos sprints, como hemos mencionado, se aumenta el tiempo para la finalización del proyecto. Concretamente se han incrementado las horas de trabajo en 55 horas, así como los gastos de software y hardware.

Con estos nuevos datos, debemos calcular la diferencia entre el presupuesto inicial y el presupuesto definitivo. Para mostrar el impacto que ha tenido este cambio en el proyecto a nivel económico calculamos la desviación resultante de la diferencia entre el coste antiguo

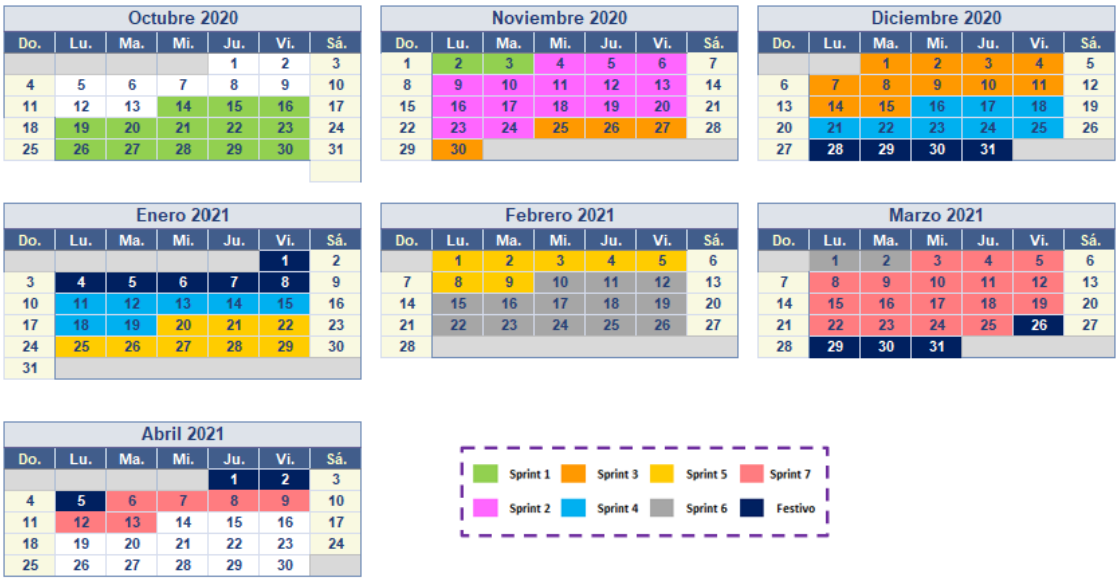

Figura 2.10: Nuevo Calendario de Planificación de Sprint

Sprint 5 (extra)

Historias de Usuario	Puntos de historia	Fecha Inicio	Fecha Fin	Horas estimadas
Refinar punto Calidad de Datos	8	20/01/2021	28/01/2021	20
Refinar punto Metadatos	8	29/01/2021	03/02/2021	15
Refinar punto Catálogo de Datos	8	03/02/2021	09/02/2021	30
Total Ptos.	24		Total horas.	65

Figura 2.11: Planificación temporal - Sprint 5 Extra.

Sprint 6 (extra)

Historias de Usuario	Puntos de historia	Fecha Inicio	Fecha Fin	Horas estimadas
Refinar punto Perfiles de Datos	5	10/02/2021	16/02/2021	20
Parte práctica de la memoria	13	17/02/2021	02/03/2021	45
Total Ptos.	18		Total horas.	65

Figura 2.12: Planificación temporal - Sprint 6 Extra.

de Hardware y Software y el nuevo mostrado anteriormente. Además, los Costes Humanos debemos calcularlos añadiendo las nuevas horas de trabajo que supuso la ampliación de dos nuevos sprints.

1. **Nuevo Coste de Software**

   Los gastos de Software no supondrían ningún cambio, ya que inicialmente eran cero.

2. **Nuevo Coste de Hardware**

Los gastos en el Hardware, supondrían un aumento en el coste por dos meses más de uso, quedando de la siguiente manera:

Componentes	Coste/mes	%Uso	Meses de Uso	Coste Real
Internet	36	100	7	252
Ordenador	9'72	100	7	68'04
			Coste total:	**320'04**

Figura 2.13: Presupuesto Hardware Nuevo

3. **Nuevo Coste Humano**

Coste Humano Total = (Nº horas estimadas + Horas extras) * Sueldo / hora

C.Humano Total = (300h + 55h) * 19,08 €/h = 6.773,38 €

- **Impacto económico:**

Desviación Coste HW + SF = Total Nuevo - Total Antiguo

D.C. HW + SF = (320,04 + 0) - (228,6 + 0) = 91,44 €

Desviación Total = (C. Humano + D.C. HW + SF) Nuevo - (C. Humano + D.C. HW + SF) Antiguo

D.Total = (6.773,38 € + 320,04 €) - (5.952,59 € + 228,6 €) = 912,26 €

Esto nos dice que el impacto que tuvo el cambio impuesto en la planificación inicial supone de un **coste de 912,26 €**, con una **desviación de 55 horas** de trabajo, del plan original.

Con los nuevos costes calculados podemos saber el **coste total** que nos ha supuesto el desarrollo de nuestro proyecto:

Coste Total Proyecto = C.Software + C.Hardware + C.Humano

Coste Total = 0 + 320,07 + 6.773,38 = **7.093,45 €**

Capítulo 3

Calidad de datos

A lo largo de este capítulo abarcaremos aspectos relevantes sobre la calidad de datos, como la definición del concepto, sus componentes y dimensiones, así como un apartado en el que analizaremos la Gestión de la calidad de datos. También se incluye un apartado de "calidad de datos en Big Data" ya que es un tema muy recurrente que también se ve especialmente afectado por la calidad.

3.1. Tipos de Datos que afectan a la calidad de los datos

Antes de comenzar a analizar los conceptos citados anteriormente, debemos conocer cuáles son los tipos de datos que existen y pueden afectar a la calidad provocando mayores efectos adversos [7].

Datos oscuros o *Dark data*

Datos que se recopilan, procesan y almacenan como parte de las actividades del negocio cotidianas, pero que no se utilizan con ningún otro fin. Es decir, es información sin estructurar, ni etiquetar, lo que provoca que sean difíciles de localizar entre toda la información disponible y son desaprovechados dentro de la organización.

Según un estudio de IBM el 80 % de los datos generados en Big Data son *Dark data* [1].

Datos sucios o *Dirty data*

Datos imprecisos o incompletos que causan un daño real a la empresa, ya que pueden llegar a repercutir en un costo económico causado por las decisiones que se toman a partir de estos datos inválidos.

Datos no estructurados

Datos que no tienen un formato específico, que no se gestionan en un sistema transaccional.

Como ejemplo para mostrar estos conceptos, cogeremos algunos de los datos proporcionados en el caso práctico. Datos oscuros son aquellos que se encuentran desactualizados o desestructurados: en los datos maestros de Clientes, existen clientes a los que se les prestó

[1] `https://www.eleconomista.es/opinion-blogs/noticias/8833843/12/17/DArK-DATA-LA-OPORTUNIDAD-DE-LAS-COMPANIAS.html`

un servicio una vez y luego no se volvió a hacer operaciones sobre ese cliente, y sin embargo quedó guardado en la base de datos. Como ejemplo de Datos sucios, tanto en Proveedores, como en Clientes existe un campo "Grupo Contable"donde existen inconsistencias para un mismo valor: para definir el grupo contable Nacional definen el valor 'NACIONAL' y 'SP', lo que crea una incongruencia a la hora de realizar operaciones basadas en ese valor, ya que al filtrar por uno de los dos valores el otro no quedaría incluido, aunque los dos definan lo mismo.

3.2. El concepto de calidad de datos

Según Elternativa, consultoría experta en Data Management: *"la calidad de datos hace referencia a una percepción o una evaluación de la idoneidad de los datos para cumplir su propósito en un contexto dado"* [68]

Por otra parte, la Organización Internacional de Normalización (ISO) establece la norma ISO 9000:2000, que define la calidad como *"el grado en el que un conjunto de características inherentes cumple con los requisitos, esto es, con la necesidad o expectativa establecida, generalmente implícita u obligatoria"* [29].

En otras palabras, la calidad de datos es un análisis previo a la utilización de los datos, que debe garantizar que son verídicos, relevantes y útiles, evitando la toma de decisiones erróneas dentro de las empresas y que puedan traer consigo consecuencias fatídicas.

En toda ciencia se desea llegar a un estándar, es decir, partir de una base común desde la cual se pueda comenzar a aplicar los principios de dicha ciencia en el trabajo práctico. En el caso de la calidad de datos no existe un estándar de uso fijo, sin embargo, para llegar a crearlo y poder determinar una "calidad de datos estándar" se deben conocer previamente los siguientes requisitos [7]:

- Quién es el encargado de crear los requisitos de calidad.
- Cuál es el proceso por el que se define la calidad.
- Cuáles son los límites establecidos en los que esa calidad ha de moverse para que el cumplimiento de los requisitos se considere aceptable.

3.2.1. Los Componentes de la calidad de los datos

La calidad de datos se forma a raíz de tres componentes que son indispensables en el proceso de gestión de calidad de datos [14].

Data monitoring o Monitoreo de Datos

Acto continuo de establecer estándares de calidad de datos mediante un conjunto de métricas significativas para la empresa, revisar los resultados de manera recurrente y tomar medidas correctivas que puedan superar los umbrales aceptables de calidad.

Data correction o Corrección de Datos

Acto de corregir los datos que se encuentran fuera del estándar establecido.

Data profiling o Perfiles de Datos

Proporciona a las empresas la capacidad de analizar grandes cantidades de datos rápidamente en un proceso sistemático y repetible. Trataremos este componente más adelante, en la Sección 4.2, dedicado a los perfiles de datos.

3.3. Dimensiones de la calidad de datos

Las dimensiones de la calidad de datos pueden variar según el criterio del experto que las plantee. En este apartado las definiremos según el criterio de la presidenta y directora de Granite Falls Consulting, Danette McGilvray, publicadas en su libro *Executing Data Quality Projects. Ten Steps to Quality Data and Trusted Information*, en 2008 [36].

Las dimensiones son las características de la calidad que nos aportan diferentes métricas a través de los cuales podremos medir y gestionar la calidad de los datos y de la información. Cada proyecto es único, por lo que no es necesario escoger todas y cada una de las dimensiones explicadas a continuación, sino que se deben elegir aquellas que aporten un mejor resultado a las necesidades de negocio de la empresa.

Un estudio de estas características hará mejorar el alcance de un proyecto, ya que se tendrá un mejor conocimiento del esfuerzo real de aplicar calidad de datos y se podrá definir una línea base. Además, con este estudio se podrá definir una secuencia de las actividades empresariales, así como las limitaciones en recursos y tiempo.

Para su clasificación se han dividido en 12 dimensiones:

1. **Especificación de datos**

   Estudio exhaustivo del Linaje de los datos o *Data Lineage*, así como la documentación, normas de los datos, reglas de negocio, metadatos, etc.

   Esta dimensión proporcionará una norma que permita comparar los resultados de la evaluación de la calidad de los datos, así como una serie de instrucciones sobre el manejo de los datos y definición de manuales.

2. **Fundamentos de la integridad de los datos**

   Se definen las características básicas de los datos, como su estructura o contenido.

   Es una dimensión imprescindible, ya que las demás se basan en ella. Aporta una serie de medidas de la calidad, la validez de los datos, patrones[2] que siguen, rangos en los que se mueven o integridad referencial.

3. **Duplicación**

   Define la capacidad de saber si existen datos duplicados. La existencia de estos datos podría traer consigo grandes costos, por lo que es importante tenerlos "vigilados".

[2]Patrones: conjuntos de cadenas con las que se puede definir el contenido, la estructura y la calidad de datos de alta complejidad.

4. **Exactitud / Precisión**

   Define el grado de exactitud que tiene un dato con respecto a lo que representa en la realidad.

   Es una medida que permite identificar el dato y verificar su fuente de origen. Este proceso de evaluación requiere que se haga de forma automática a través de una herramienta que normalmente son los perfiles de datos.

5. **Coherencia y sincronización**

   Los datos se obtienen de diversas fuentes por lo que es necesario que existan procesos de equivalencia entre los datos que verifiquen que son conceptualmente iguales, indistintamente del sistema que se esté utilizando para extraer dichos datos.

6. **Puntualidad y disponibilidad**

   Mide el grado en el que los datos están actualizados y disponibles en el momento de su uso.

   Se debe prestar atención al periodo de tiempo que está comprendido entre que el objeto se modifica en el mundo real y el momento en el que se actualiza en la base de datos y es trasladada a los usuarios, ya que es una pequeña brecha que puede provocar anomalías en los datos.

7. **Facilidad de uso y mantenimiento**

   Para satisfacer las necesidades de los usuarios, se debe estudiar el grado en el que se puede acceder a los datos, utilizarlos y mantenerlos. Además, se debe tener en cuenta la facilidad de uso de estos datos para cualquier tipo de usuario.

8. **Cobertura de datos**

   Se estudia la forma en la que la base de datos interactúa con los datos y cómo lo refleja a los distintos usuarios.

   Mide la disponibilidad y exhaustividad de los datos en comparación con la totalidad de los datos que se puedan llegar a capturar.

9. **Calidad de presentación**

   Estudia la forma más efectiva y fácil de presentar los datos e información para aquellos usuarios que la utilizan. Se presta atención al formato y apariencia de estos datos.

10. **Percepción, relevancia y confianza.**

    Establece qué datos son más prioritarios que otros en base a las necesidades de la empresa.

    Mide el valor de los datos según cómo los perciben los usuarios comparándolos con lo que expresan en realidad.

11. **Decaimiento de datos.**

    Mide la tasa de cambio negativo que sufren los datos, es decir, mide el grado de desactualización de un dato.

    Conocer estas tasas permite establecer medidas y mecanismos para mantener los datos completos y precisos. Cuanto mayor es esta tasa de deterioro mayor tiempo y actualizaciones se deben emplear en estos mecanismos.

12. **Transaccionalidad**

    Mide el grado en que los datos producirán una transacción comercial o el resultado deseado. Dicho de otro modo, los datos deben producir el resultado esperado. Incluso si las personas adecuadas han definido los requisitos de negocio y han preparado los datos para cumplirlos, es importante que los datos produzcan el resultado esperado: ¿Se puede generar la factura correctamente? ¿Se puede completar una orden de venta?

Dimensiones de la calidad de los datos	
1. Especificación de datos	7. Facilidad de uso y mantenimiento.
2. Fundamentos de la integridad de los datos.	8. Cobertura de datos.
3. Duplicación.	9. Calidad de presentación.
4. Exactitud / Precisión.	10. Percepción, relevancia y confianza.
5. Coherencia y sincronización.	11. Decaimiento de datos.
6. Puntualidad y disponibilidad	12.Transaccionalidad.

Tabla 3.1: Dimensiones de la calidad de datos.

Evaluar e implantar estas características de forma adecuada indica que los datos pueden considerarse de calidad, por lo que se asegura que éstos sean auténticos y válidos. En este contexto, el Fondo Monetario Internacional ha creado un método común de evaluación de los datos para todas las empresas, denominado Marco de Evaluación de la calidad de los datos (DQAF, *Data Quality Assessment Framework*) para estimar la calidad de los datos, cuya última versión data de mayo de 2012 [16].

3.3.1. Medidas cuantitativas y cualitativas de la calidad de los datos

Apoyando el análisis de las dimensiones de la calidad, existen una serie de **medidas cuantitativas** que ayudan a determinar con exactitud y medir numéricamente la calidad de los datos [43]. Las métricas nos dan una forma de evaluar una dimensión determinada, y pueden asociarse varias de ellas para una misma dimensión [4].

Estas medidas (Ver Figura 3.1) son:

- **Completitud.** Establece la medida en que los datos pueden representar todos los estados significativos de una realidad.

  La completitud de una dimensión se mide comparando: (nº de datos no nulos / total datos)

- **Validez.** Representa el grado en que el dato aporta valor al conjunto que pertenece.

  La validez de una dimensión se mide comparando: (nº de datos con valor (no nulos) en el conjunto / total datos conjunto)

- **Unicidad.** Detecta la presencia de datos duplicados.

  La unicidad de una dimensión se mide comparando: (nº datos no duplicados / total datos)

- **Integridad.** Especifica el grado de conformidad de los datos y si se ajustan a las reglas establecidas de la relación entre los datos.

  La integridad de una dimensión se mide comparando: (nº datos no ajustados / total datos)

- **Precisión.** Indica la medida en que los datos representan la realidad a la que se refiere.

  La precisión de una dimensión se mide comparando: (nº de datos sin errores / total datos)

- **Coherencia.** Representa el grado en que un mismo dato contiene la misma información indistintamente de la fuente de la que se obtenga.

  La coherencia de una dimensión se mide comparando: (dato original / dato real)

- **Oportunidad.** Indica si un dato está disponible en el momento preciso en el que un usuario requiere de él.

  La oportunidad de una dimensión se mide comparando: (tiempo de entrega / tiempo en que lo ve el usuario)

- **Representación.** Está vinculado a la presentación de los datos frente a los usuarios.

Aparte de estas medidas cuantitativas si se quiere adquirir una perspectiva real y más completa de la situación de la empresa, es necesario considerar las **medidas cualitativas**. Estas medidas hacen referencia a elementos como la satisfacción del cliente y usuarios de negocio, índices de cumplimiento de metas, la aparición de redundancias en los procesos o la identificación de oportunidades de negocio.

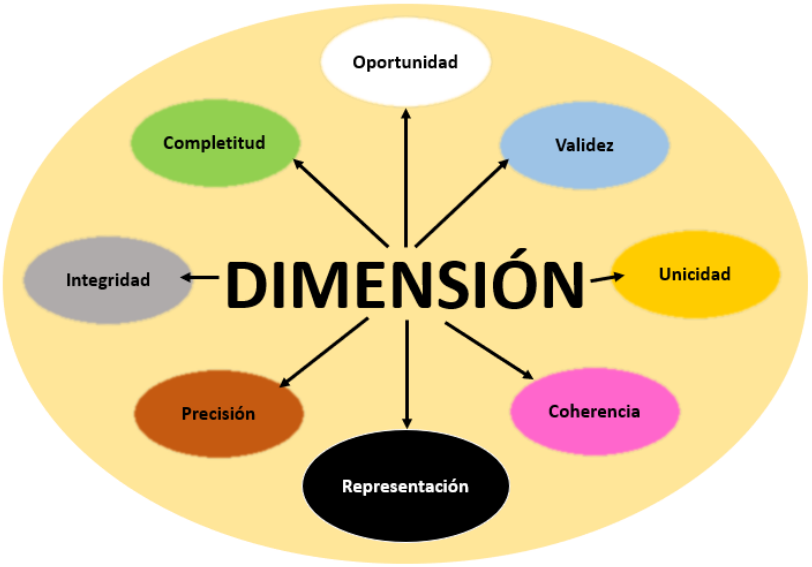

Figura 3.1: Elaboración propia: Métricas cuantitativas

Ejemplo práctico Con respecto al supuesto práctico, hemos escogido las dimensiones que aportarán un beneficio directo al objetivo del proyecto y por tanto su estudio es imprescindible. El análisis de las demás dimensiones también añadirían un valor importante a los datos, sin embargo, debemos tener en cuenta la operatividad y tiempo de análisis que supondría realizar este estudio, mientras que el valor que aportan es complementario y se puede realizar más adelante.

Las dimensiones escogidas son:

- 2. Fundamentos de la integridad de los datos.

  El análisis de esta dimensión es imprescindible para cualquier proyecto porque nos da un visión de cómo es la estructura y el contenido de todos los datos que vamos a manejar. Con el estudio de esta dimensión observamos, por ejemplo, que los datos maestros de Productos tienen siempre una misma estructura, es decir, contemplan siempre los mismos campos: *identificador, descripción, unidad de medida, precio por unidad, proveedor que proporciona ese producto, si está bloqueado o no, Código de Grupo de Producto, Grupo contable de producto, Grupo Registro IVA Producto*. Y su contenido es variable, haciendo cada producto único. De igual forma, ocurre con los demás conjuntos de datos.

  Este análisis de la dimensión "Fundamentos de la integridad de los datos" ya ha sido realizada por la empresa previamente, los resultados se pueden ver en las Tablas 1.1 a 1.4, en la Sección 1.2. El proceso que se debe realizar sería el siguiente: partiendo de un gran número de datos necesarios y recogidos por la empresa deben identificar, para cada dato, aquellos atributos que aporten más valor de negocio y en consecuencia diferenciar entre los distintos grupos de datos que tengan una misma estructura. De esta forma conseguimos diferenciar los cuatro conjuntos de datos

descritos, junto a su estructura y metadatos que contienen.

- 3. Duplicación y 4. Exactitud / Precisión.

  Con respecto al análisis de estas dimensiones nos aporta información sobre aquellos datos que pueden afectar a la empresa de manera negativa. Esto lo conseguimos con procesos automáticos, gracias a los perfiles de datos, donde podemos identificar de manera efectiva si existe redundancia en los datos y si éstos son verídicos. Nos permite conocer aquellos datos que restarán calidad cuando tratemos con datos anómalos.

  Estas dos dimensiones son analizadas a través del catálogo de datos Talend Studio descrito en la Sección 5.7. Gracias a esta herramienta podemos realizar análisis sobre nuestros conjuntos de datos, de forma que localicemos aquellos datos anómalos ya sean porque están duplicados o no son lo suficientemente precisos. Fijándonos en la implementación práctica del Capítulo 6 podemos observar el análisis del conjunto de datos Cliente en las Figuras 6.5 y 6.6 donde encontramos datos duplicados que pueden afectar negativamente a la empresa, en concreto un 12,24 % (2.497 registros).

- 7. Facilidad de uso y mantenimiento y 9. Calidad de presentación.

  Estas dimensiones las hemos seleccionado ya que los conjuntos de datos pueden ser utilizados dentro de la empresa por distintos empleados, con roles diferentes. Estudiar los datos definiendo los usuarios que pueden acceder o no a dichos datos es algo clave para el buen funcionamiento de la empresa.

  Con respecto a estas dos dimensiones, en este proyecto no se han podido llevar a cabo ya que desconocemos el tipo de usuarios que van a manejar los datos y por tanto no podemos definir la mejor forma de presentar la información. Sin embargo, a modo de ejemplo podemos establecer una serie de reglas a la hora de acceder a los datos que podrían llegar a implementarse. Suponiendo que hubiera dos roles, rol contable y rol fabrica: el rol contable debería tener permisos únicamente para acceder a los datos referentes a temas de contabilidad/finanzas: facturas, compras, ventas, declaraciones de IVA, etc.; mientras que el rol fabrica debería tener acceso a toda la información referente a los productos, pedidos, entregas, gestión del almacén, etc.

3.4. La Gestión de la calidad de los datos

La gestión de la calidad de los datos se define como: *"principio empresarial que requiere una combinación de las personas, los procesos y las tecnologías adecuadas, con el objetivo común de mejorar las medidas de la calidad de los datos que más importan a una organización empresarial"* [17]

Esta gestión define cómo se debe trabajar con los datos y establece unas normas para su tratamiento [32]:

- Establece cómo se debe gestionar la obtención y tratamiento de los datos.
- Define las responsabilidades de cada uno de los roles implicados en el proceso.
- Determina cuáles son los parámetros de calidad, aquellas variables que van a definir si existe calidad o no en los datos.
- Implementa el conjunto de herramientas técnicas que van a servir para medir y asegurar la calidad de los datos.

David Loshin, experto en calidad de datos, propone un proceso al que llama *Data Preparation*, donde expone cómo debería realizarse la gestión de la calidad de los datos, que tiene como objetivo garantizar una mejora constante de la calidad de dichos datos en una empresa [68].

Este ciclo se divide en tres etapas (Ver Figura 3.2) que conforman un proceso continuo a través de las cuales se van analizando y gestionando los datos [60].

- Comienza con la **identificación y medición del efecto** de los datos en la empresa, un proceso en el que se busca encontrar los datos y describir los conjuntos de datos para conocer y comprender su utilidad antes de destinarlos a un contexto específico.
- Seguido de una **limpieza y consolidación** de estos datos en un mismo lugar donde se eliminan los datos defectuosos y/o extraños, los datos que estén vacíos o incompletos se completan, se crean patrones estándar de los datos y se validan mediante una prueba de error durante el proceso.
- Llegando a la tercera etapa donde se **definen una serie de reglas, se fijan unos objetivos de rendimiento y se implementan procesos de mejora de la calidad**, repitiéndose continuamente las fases en el mismo orden. El análisis y gestión de datos en estas fases nos permite encontrar los datos erróneos antes de que se introduzcan en los sistemas

Figura 3.2: Elaboración propia: Fases del Data Preparation

Ejemplo práctico Cogiendo como muestra el supuesto práctico vamos a identificar las etapas del proceso *Data Preparation* con datos reales. Este proceso se lleva a cabo de forma semiautomática a través de los catálogos de datos, concretamente con la herramienta de perfiles de datos.

- La primera etapa "identificación y medición del efecto de los datos en la empresa" , corresponde a un estudio de los datos del sistema antiguo, los cuales deben ser aportados por la empresa, y se relaciona con el estudio de la segunda dimensión de la calidad de datos, "Fundamentos de la Integridad de los datos". Identificamos que existen conjuntos de datos específicos que se pueden dividir en distintas clases: Clientes, Proveedores, Productos, Plan de Cuentas, donde cada conjunto tiene una estructura determinada.

- Dentro de cada conjunto de datos aplicamos la segunda etapa "limpieza y consolidación de los datos", relacionada con las dimensiones de "Duplicación" y "Exactitud / Precisión", donde se eliminan aquellos datos que hemos catalogado como Incorrectos y que por lo tanto se deben eliminar. Existen datos, por ejemplo en el conjunto Producto donde en un momento dado se creó un registro de un producto pero que al final no se completó y quedó con sus campos vacíos, o ese producto se dejo de fabricar por lo que su información está desactualizada, por lo que se debería eliminar o actualizar la información.

- Por último, en la tercera etapa definimos una serie de normas o reglas que deben ir asociadas a los distintos datos, como por ejemplo que el nombre de un Cliente, Proveedor o Producto no pueda estar en blanco, o si un Proveedor o un Cliente registrado llevase un tiempo determinado, por ejemplo de tres años, sin tener movimientos de Compra / Venta hacia la empresa, ese registro debe ser eliminado.

Este proceso, en nuestro caso, ya teníamos parte hecho antes de empezar a analizarlo ya que la empresa que nos proporcionó los datos ya tenía, como hemos mencionado en el ejemplo práctico de la Sección 3.3, una estructura definida para los datos (Ver Tablas 1.1 a 1.4) por lo que la primera fase no haría falta volverla a realizar.

En caso de no haber realizado la primera fase, se deben estudiar los datos de forma que nos indique el efecto, tanto positivo como negativo, que tendrían sobre la empresa. Este paso lo debe realizar la empresa misma ya que son los que mejor conocen los riesgos que podrían conllevar los datos. Para la segunda etapa, se realiza a través de los catálogos de datos que con ayuda de los perfiles de datos nos permite analizar los datos en busca de datos redundantes o imprecisos, este paso lo realizamos en el Capítulo 6. La última etapa, definir las normas que deben tener asociadas los datos también es un paso que la empresa nos dio de antemano y que podemos ver definidas en el Capítulo 6, estas reglas pueden ser aplicadas por sentido común, por ejemplo, que no puedan existir dos clientes iguales ya que conllevaría descuadres a la hora de analizar las finanzas, sin embargo, la empresa es quién toma la decisión sobre que normas poner y cuales no.

Es imprescindible tener un enfoque global, proactivo y colaborativo con respecto a los datos. Este enfoque se consigue con la realización de una buena gestión de los datos y garantiza la verificación del proceso *Data Preparation*. Además, este enfoque se debe tener en cuenta todos los ámbitos de la empresa y los diferentes equipos de trabajo.

Las tres fases de este ciclo están estrechamente relacionadas con los componentes de la calidad de los datos definidos en el apartado 3.2.1: Monitoreo de Datos, perfiles de datos y Corrección de Datos.

El Monitoreo nos permite establecer los estándares a los que someteremos los datos en la primera etapa del ciclo. Además, gracias a los perfiles de datos, que nos proporcionan la capacidad para analizar esos volúmenes de datos, en esta primera fase del *Data Preparation* podremos averiguar si los datos cumplen o no con los estándares definidos. Una vez identificados los datos erróneos, en la fase de Corrección de Datos, podremos corregir las anomalías que impiden que los datos cumplan con los estándares o lo que es lo mismo, pasarán por un proceso de limpieza y consolidación en el que se eliminarán esos errores y anomalías para que, en la siguiente fase del ciclo, se pueda trabajar con datos verídicos, relevantes y útiles.

Por otra parte, debemos señalar que la Organización Internacional de Estandarización (ISO) ha publicado la norma de calidad de datos, ISO 8000, en el año 2011. Está compuesta por 4 partes principales: conceptos generales sobre la calidad de datos, procesos de gestión de calidad de datos, aspectos de intercambio de Datos Maestros entre organizaciones (atendiendo a ciertos aspectos de calidad) e información de ingeniería [41].

La primera parte **"Conceptos generales sobre la calidad de datos"** consiste en una introducción al estándar, sus relaciones entre otros conceptos de calidad de datos, se definen los principios en los que se basa la calidad de datos, conceptos fundamentales, un vocabulario específico de calidad de datos, la arquitectura de la familia ISO 8000. En los **"procesos de gestión de calidad de datos"** correspondiente a la segunda parte se especifica cómo crear, almacenar y transferir los datos que dan soporte de manera rentable y oportuna a los procesos de negocio y esta parte a su vez se divide en tres subpartes, Visión General, Modelo de Referencias de Procesos y Modelo de Evaluación [31].

En cuanto a la parte de **"Datos Maestros"**, explica como se debe abordar el intercambio entre datos maestros, Gestión de Datos Maestros o *Master Data Management*, proponiendo un formato específico y exponiendo los requisitos que debe cumplir este intercambio para aportar mayor precisión y completitud a los datos maestros [30]. Por último, nos encontramos la parte de **"Información de ingeniería"** donde se encuentra la norma ISO 8000-311, que hace alusión a la "Orientación para la aplicación de la calidad de los datos del producto para la forma PDQ-S (Product Data Quality for Shapes)", dónde PDQ-S se refiere a la norma ISO 10303-59 que dicta, en el contexto de calidad de los datos, que se debe cumplir con cuatro estándares que son: consistencia, completitud, integridad e idoneidad para su propósito de los datos del producto [50].

3.5. Beneficios

Una buena calidad de datos en la empresa puede marcar una gran diferencia frente a la competencia. Como hemos visto una buena gestión de esta calidad puede traer consigo una serie de beneficios, en general, como es ahorrar en tiempo y dinero lo que conlleva un aumento en la productividad.

Los beneficios que puede aportar la calidad de datos en la empresa pueden ser [29]:

- Una **mayor satisfacción por parte de los clientes**. Un mejor control sobre los datos, asegurando su calidad, ofrece mejores soluciones, más rápidas y eficaces, además de ser más comprensibles para los clientes, aportando confianza en la empresa.
- Se **reducen los riesgos** en cada proyecto. Contando con datos de calidad se toman decisiones más acertadas, por lo que se minimizan los riegos de daños que no pueden evitarse.
- Se hace un mejor uso de la infraestructura tecnológica y sistemas para explotar su información, con lo que se **ahorra en tiempo y recursos**. Esto mejora la eficiencia en las operaciones, por ejemplo, evitando tener información duplicada en las distintas áreas de negocio que acceden a la información.
- Con el estudio de los datos y añadiéndolos calidad, se obtiene **información de confianza, precisa y (casi) libre de errores**, ayudando a tomar decisiones críticas lo más acertadas posibles.
- **La empresa crea un estándar** sobre cómo se debe manejar la información y facilitando su uso correcto de los datos.
- Agregando calidad a los datos podemos asegurar **sacar todo el potencial que se esconde en los datos**, por lo que su utilización supone un mayor beneficio en los proyectos.

3.6. Problemas que enfrenta la calidad

Introducir calidad de datos en una empresa por primera vez presenta una serie de dificultades, sobre todo al inicio del proceso, ya que es una tarea ardua y que requiere una buena implementación para sacar el máximo valor a los datos.

Algunos de los retos que puede presentar este proceso son los siguientes [22]:

- **Grandes volúmenes de datos** que deben estudiar y clasificar, además de su variedad de tipos (datos estructurados, semiestructurados y no estructurados).
- La existencia de **datos duplicados u obsoletos**, lo que dificulta la comunicación entre las distintas áreas de la empresa.

- La **gran diversidad de fuentes de datos** desde dónde se puede acceder a los datos.
- Un **tiempo elevado para analizar** los datos obtenidos y hacerlos de calidad.
- Los **requisitos de cumplimiento y seguridad** de los datos provienen de diversas fuentes y pueden incluir requisitos corporativos además de los mandatos de la industria y el gobierno, como HIPAA o Estándares de seguridad de datos de PCI (PCI DSS). El incumplimiento de estas reglas puede dar lugar a fuertes multas y, quizás incluso más costoso, a la pérdida de la confianza del cliente. A menudo, los requisitos descritos por mandatos como HIPAA y PCI constituyen un caso sólido para implementar un programa integral de gestión de la calidad de los datos [54].
- **Humano**. Los errores por parte del factor humano son comunes, como las equivocaciones a la hora de actualizar o clasificar algún dato.

3.7. Calidad de datos en Big Data

Las tecnologías Big Data permiten a las empresas captar grandes volúmenes de datos con facilidad, algo que, en la actualidad, es imprescindible debido a la gran cantidad de datos que se generan día tras día. Sin embargo, cuanto mayor sea el número de datos existentes, más difícil resulta manejarlos y gestionarlos de forma óptima. Es por este motivo, que hemos decidido incorporar este apartado, en el que analizaremos brevemente la implementación de la calidad en los datos en el ámbito de Big Data, ya que la reciente utilización de este tipo de tecnologías en las empresas cada vez es mayor y por lo tanto, cada vez adquiere una posición más relevante en la gestión y utilización de datos dentro de estas organizaciones.

Big Data ofrece un punto de referencia donde las empresas pueden identificar los problemas de manera más fácil, ya que al tratar con volúmenes de información mayores se pueden abordar mejor un problema específico, encontrando la solución más certera y de forma más rápida. Por lo tanto, en este tipo de empresas que apuestan por este "gran almacén de datos", la gestión y control de la calidad en sus datos se vuelve esencial, crítica y decisiva para progresar y destacar entre la competencia, y es así porque contar con información de calidad marca la diferencia en cuestión de elegir opciones decisivas para la empresa con una mayor certeza y obtención de un mayor beneficio económico, administrativo y humano.

Para definir y aplicar la calidad en los datos dentro del ámbito de Big Data es necesario conocer de antemano las características especiales que tiene este tipo de tecnología y que, entre otras cosas, provocan inconvenientes a la hora de capturar datos reales [6].

Estas características son las denominadas *5Vs*:

- **Volumen**. Hace referencia a la gran cantidad de volúmenes de datos que se pueden capturar. No se ha establecido una cifra exacta que defina lo que es o no Big Data, pero suele estar a partir de volúmenes medidos en Terabytes.

- **Velocidad**. Se almacena y recopila un número elevado de datos a una gran velocidad, lo que provoca que los datos pasen a un estado desfasado rápidamente y pierdan el valor que puedan aportar.

- **Variedad**. Los datos almacenados pueden ser de múltiples tipos y formatos, estructurados o no estructurados. Se debe tener especial cuidado a qué datos se pretende dar más énfasis.

- **Veracidad**. Los datos deben ser precisos y representar de forma exacta la realidad o, al menos, aproximarse lo máximo posible a lo que representan. Muestra el grado de confianza que tienen los datos para poder ser usados.

- **Valor**. Los datos deben aportar el valor necesario para poder ser usados en las empresas y presentar soluciones óptimas que permitan sobresalir sobre otras empresas.

Estas características presentan nuevos desafíos a los que se debe enfrentar la calidad de datos en un entorno Big Data [6]:

- Se **aumentan el número de fuentes** de las que se pueden obtener datos, lo que conlleva un aumento considerable en el volumen de datos a tratar y por tanto aumenta la diversidad y disparidad entre ellos.

- Los **datos evolucionan y cambian a un ritmo más elevado**, por lo que tratarlos y tenerlos listos en el momento que se necesiten es más complicado y se requiere aumentar los requisitos en el procesamiento de los datos.

- **No existen estándares** para manejar la calidad de datos en el ámbito del big data, lo que dificulta el estudio de estos datos.

Teniendo como referencia las dimensiones de la calidad, mencionadas en la sección 3.3, en Big Data se establecen unos estándares que deben cumplir con los requisitos establecidos en las características de la calidad de los datos, redefiniendo esos conceptos atendiendo a las necesidades reales y actuales de los negocios.

Las dimensiones se redefinen de la siguiente manera [33]:

Disponibilidad. Se relaciona con la sexta dimensión de la calidad de los datos, '6.Puntualidad y disponibilidad'.Se compone de dos elementos asociados:

- Accesibilidad. Indica si existe una interfaz que facilite el acceso a los datos y si pueden hacerse públicos y fáciles de adquirir.

- Oportunidad. Determina un periodo de tiempo límite para la entrega del dato al usuario, así como para su actualización.

Usabilidad. Relacionada con la séptima dimensión de la calidad de los datos, 'Facilidad de uso y mantenimiento'. Se compone de un elemento asociado:

- Credibilidad. Los datos provienen de múltiples fuentes cada una con sus propias peculiaridades, por lo que se debe establecer un rango de valores para los datos en los que se puedan consideran "aceptables".

Confiabilidad. Se relaciona con la segunda, cuarta y quinta dimensión de la calidad de los datos, 'Fundamentos de la integridad de los datos', 'Exactitud/Precisión' y 'Coherencia y Sincronización'. Se compone de cuatro elementos asociados:

- Exactitud. Los datos deben ser lo más precisos posibles, ya que deben reflejar una representación real de la información que contiene y no generar ambigüedades.
- Consistencia. Los datos deben ser iguales que su representación original, siendo consistentes y verificables.
- Integridad. Se debe seguir un criterio que establezca integridad y contenido a los datos y estos sean claros.
- Completitud. Los datos deben ser completos evitando que existan datos deficientes que afecten al valor de los datos.

Pertinencia. Relacionado con la primera dimensión de calidad de datos, 'Especificación de datos'. Se compone de un elemento asociado:

- Convivencia. Los datos deben tener una serie de "etiquetas" que los identifique, y permita encontrar datos incluso cuando sólo existe una relación parcial con el tema a tratar.

Calidad de presentación. Se relaciona con la novena dimensión de calidad de los datos, 'Calidad de presentación'. Se compone de un elemento asociado.

- Legibilidad. Los datos deben ser claros y fácilmente entendidos por todos los usuarios y satisfacer sus necesidades.

La calidad de los datos y Big Data son dos términos que deben ir unidos en empresas donde la recolección de datos se realiza de forma masiva y se quiere que esta información aporte valor y tenga un impacto positivo dentro de la empresa. Por ello, para que no existan problemas graves dentro de una empresa como pudiera ser procesos desactualizados, una desventaja competitiva frente a otras empresas de un mismo sector o una insatisfacción por parte de los clientes, debemos dotar de calidad a todos aquellos datos que se recogen a partir de las tecnologías Big Data.

Capítulo 4

Los metadatos y los perfiles de datos

En este capítulo se trata de dos conceptos que están directamente asociados con la calidad de datos, los metadatos y los perfiles de datos. Hemos considerado hacer un apartado que agrupe a estos dos conceptos clave debido a la estrecha vinculación que tienen con la calidad de los datos y el valor que aportan.

4.1. Metadatos

En este apartado nos centramos en uno de los aspectos más delicados y a la vez más influyentes con respecto a la calidad de datos, los metadatos, que de forma literal significa "datos sobre datos" o dicho de otro, los metadatos se encuentran dentro de los datos. En otras palabras, los datos son la suma de la información general que conforma el dato y de sus metadatos, es decir, información específica que se aporta al dato para que éste llegue a un nivel de concreción mayor. Éstos proporcionan una imagen completa de los datos y ayudan a comprenderlos en su totalidad y sin ambigüedad, ya que contienen una descripción detallada de los datos y aportan el contexto necesario para aportar un significado más completo de la información [65].

Para ilustrarlo podemos coger como ejemplo una fotografía. Al realizar la fotografía la cámara captura una serie de **información general** como la velocidad de obturación, el modelo de la cámara o la velocidad ISO, entro otros muchos. A posteriori, el mismo fotógrafo puede incluir una serie de **metadatos específicos** que ayuden en la identificación única del dato o fotografía en este caso, como podrían ser el nombre del autor,la descripción de la imagen o palabras clave que ayuden en su posterior búsqueda [46].

Los metadatos ofrecen información detallada de los datos, lo que permite a las empresas determinar si un conjunto de datos es apropiado para un fin específico, así como una manera de recuperar ese conjunto de datos identificados, la forma de procesarlos y de qué manera pueden ser utilizarlos. También, proveen a los usuarios de un inventario estandarizado de todos los datos que existen en la empresa y permite a los usuarios encontrar y verificar si esos datos son apropiados para sus necesidades y dónde pueden localizarlos.

La información que nos proporcionan los metadatos es amplia y muy útil. Nos aportan conocimientos del origen del dato, qué pasos se han considerado para crearlo, los atributos que lo forman, qué proyección ofrece el dato, cómo se puede obtener una información completa de él, cuánto costaría temporal y económicamente manejar, encontrar y tratar el dato o incluso qué usuario puede aportar una copia del mismo. Además, los metadatos también describen, entre otras características, el contenido, la calidad, las condiciones y la disponibilidad de los datos [45].

A la hora de crear un documento web, por ejemplo, es imprescindible tener en cuenta qué metadatos analizar y añadirlos según cada caso. Por ejemplo añadiendo el origen del documento, es decir, la fuente de dónde ha sido sacada la información nos permite, en caso de una alerta por desactualización de dicha información, poder llegar al origen del dato pudiendo solucionar el problema rápidamente. Otro ejemplo serían, incluir distintos tipos de atributos del documento como la fecha de creación o de última actualización. Al especificar un mayor número de metadatos, es deducible que podemos encontrar de forma rápida su ubicación al introducir palabras clave en un buscador.

Como comprobaremos más adelante, los metadatos son de vital importancia para las empresas, ya que les permite conocer los datos en su totalidad, y es por este motivo por el cual se debe dedicar tiempo y esfuerzo a su estudio y comprensión. En este contexto, los metadatos son utilizados en diferentes áreas dentro de una empresa y para distintas tareas, como la administración de datos, ya que ayudan a los usuarios a la comprensión de la información que manejan y facilitan su localización y su evaluación.

A lo largo de este documento hemos hecho hincapié en que conocer y comprender los datos con los que se trabaja de forma completa y tenerlos localizados para acceder fácilmente a ellos, es imprescindible en cualquier empresa que trabaje con datos. En este sentido, la mejor forma de gestionarlos actualmente, y expecíficamente con la llegada del Big Data y el análisis de autoservicio, es utilizando una herramienta de gestión de la calidad llamada catálogo de datos, en la que profundizaremos con más detalle en el capítulo 5, que se ha convertido en un estándar para la gestión de los metadatos, ya que es la encargada de proporcionar el inventario estandarizado donde se almacenan y se organizan los metadatos.

Al igual que ocurre con una mala gestión de la calidad en los datos, una mala gestión de los metadatos trae consigo consecuencias fatídicas para la empresa, ya que se podrían estar manejando datos con falta de información, que esta esté obsoleta o, incluso, que exista una inexactitud en el linaje de los datos, provocando una falta de información entre las relaciones de los datos. En última instancia, podrían surgir Pantano de datos que imposibilitasen el uso de estos datos.

4.1.1. Clasificación

Los metadatos no son específicos de la calidad de datos, es decir, pueden estar presentes en otros ámbitos diferentes a la gestión de datos. Por este motivo, pueden clasificarse en diversas clases y tipos. En este apartado nos centraremos en su clasificación en función de tres criterios principales: según su función, variabilidad y contenido; basándonos

en los metadatos utilizados específicamente en la calidad de los datos. Para exponer esta clasificación nos guiaremos de un artículo realizado por la multinacional PowerData, especializada en gestión de datos [38].

Y estos criterios son:

- Según la **función** que aporte el metadato. Se diferencian tres tipos de metadatos por función [55]:
  - **Metadatos técnicos o estructurales.**
    - Aportan información sobre la organización de los datos y ayudan a los usuarios a definir cómo será la estructura, las tablas, columnas, índices y relaciones de los datos que manejen.
    - Este tipo de metadatos indica a los usuarios cómo se debe trabajar con ellos para obtener el máximo valor del dato.
  - **Metadatos de proceso o administrativos.**
    - Dan una visión del linaje que tienen los datos. Además, informa y define quién tiene permiso para acceder y utilizar el dato.
    - Aporta un histórico de los datos, facilitando la resolución de problemas en las consultas y dando acceso a información sobre los usuarios.
  - **Metadatos externos.**
    - Llamados metadatos empresariales, son aquellos que proporcionan una misma visión y en un mismo idioma sobre los datos para los trabajadores que los manejan y para los usuarios que utilizan los datos en cuestiones de negocio. Proporcionan una definición del dato desde un punto de vista del negocio.
    - Describen los aspectos de negocio del activo de datos, es decir, informa del valor que tiene para la empresa, su idoneidad para uno o varios propósitos o información sobre el cumplimiento normativo.
- Según la **variabilidad** del metadato. Se clasifica en dos tipos:
  - **Inmutables.** El metadato no cambia, sin importar a qué parte del dato se acceda. Un ejemplo sería la fecha de creación de un documento, siempre es la misma no cambia.
  - **Mutables.** El metadato puede ser modificado, es decir, un metadato mutable es aquel donde se modifica la parte específica del metadato, convirtiéndolo en uno nuevo. Como ejemplo sería la fecha de actualización de un documento, que varía en el tiempo.
- Según el **contenido** que ofrece el metadato, facilita la búsqueda y recuperación de la información. Se distinguen dos tipos.

- Metadatos que **detallan el dato** en sí, o lo que es lo mismo describen la naturaleza del dato, como su dominio, el tipo de dato, etc.
- Metadatos que **describen el contenido del dato** en sí, detallan el valor que presenta el dato como por ejemplo, un identificador único del dato o la longitud de dicho valor.

Ejemplo práctico Haciendo referencia al supuesto práctico nos encontramos con distintos tipos de metadatos. A nivel de función en nuestro caso contamos con metadatos empresariales, ya que son datos que expresan información a nivel de negocio donde por ejemplo, existe el campo Grupo Contable, donde dependiendo de si eres Cliente o Proveedor se estable un valor concreto, "GCC-CLT" para los clientes y "GCP-PVD" para los proveedores. En el caso de los metadatos según su variabilidad podemos distinguir como metadato inmutable el campo Nº Proveedor en el conjunto de datos Producto, donde una vez registrado un producto se estable el proveedor que les vende ese producto, como metadato mutable el campo Precio por unidad que puede variar en función del mercado. Por último, en el caso de los metadatos por contenido, podemos mostrar como metadato que detalla el dato en sí sería la categoría 'Producto' o 'Proveedor', mientras que metadatos que describen el contenido del dato sería el Identificador o Código del registro.

4.1.2. Funciones principales

A continuación, procedemos a enumerar las funciones principales que debería tener cualquier metadato. Esta lista se basa dos propuestas realizado por los autores Gayatri Ramachandran (2007), Ingeniero de software sénior en Google, y Kate Beard (1996), Ingeniera de software [45].

- **Búsqueda**. Los metadatos son esenciales en la localización e identificación de los conjuntos de datos requeridos en el momento preciso. Por ello, los metadatos deben proporcionar suficiente información que permita determinar si el dato existe o dónde localizarlo.
- **Recuperación**. Los metadatos aportan información de referencia para que, un usuario encuentre un dato o un conjunto de ellos en cualquier documento o proyecto a través de referencias, permitiéndoles identificar de nuevo el dato o conjunto de datos y además hacerlo de una forma rápida.
- **Transferencia**. Esta función permite que, en el proceso que se genera entre que se recupera el dato hasta que el usuario lo utiliza de nuevo, la información que contiene el metadato llegue de nuevo al usuario correctamente y siendo igualmente verídica, es decir, detalla la información de cómo acceder y/o utilizar el dato.
- **Evaluación**. Ofrecen información que ayude a los usuarios a determinar si los datos son útiles para sus necesidades o no.

- **Archivo y conservación**. Esta función está relacionada con la documentación del proceso de creación del dato en donde se recopilan aspectos como: quién ha creado el dato, con qué finalidad, para qué ámbito fue creado o cómo se debería utilizar ese metadato.
- **Interoperabilidad**. Se establecen estándares de metadatos que dan una visión unificada del conjunto de datos y protocolos para el intercambio de información, lo que ayuda por ejemplo, a la realización de búsquedas simultáneas en sistemas distribuidos.

Además de estas funciones, existen otras que aportan valor al estudio de los metadatos y, por lo tanto, es conveniente tenerlas en cuenta a la hora de realizar dicho estudio [73].

- Los metadatos favorecen la **gestión de cambios**, es decir, gracias a los metadatos es posible detectar cambios que se realicen sobre los datos y resolver los conflictos que surjan en tiempo real.
- Garantizan el **cumplimiento de las normativas**, aportando una mayor seguridad. Además, permiten tener especial cuidado con los datos que son sensibles para el negocio.
- **Facilita el acceso a los datos** aportando un acceso inteligente, es decir, que la obtención de esos datos se realice de una forma más fácil y eficiente, haciendo el muestreo más seguro y confiable.

4.1.3. Ciclo de vida

El ciclo de vida de un metadato es un proceso por el cual los metadatos van sufriendo diferentes transformaciones según van avanzando por las distintas etapas [38].

Figura 4.1: Elaboración propia: Ciclo de Vida de los metadatos

Como se puede observar en el gráfico (Figura 4.1), el ciclo de vida de un metadato se compone de tres etapas y es continuo, es decir, una vez que finaliza la tercera etapa, el metadato vuelve al inicio para comenzar de nuevo y en el mismo orden, todo el proceso.

La primera de estas etapas se llama **Creación** y es en la cual los metadatos son creados. Existen tres maneras diferentes por las que la información de los metadatos puede desarrollarse:

- De forma manual: es la menos utilizada ya que se trata de un proceso anticuado y complicado que depende únicamente del usuario que se encargue de la creación del metadato y del formato y tamaño que se requiera para crearlo.

- De forma automática: en la que el usuario no tiene que realizar ninguna tarea, sino que el software es el encargado de realizar todo el proceso de creación del metadato. Actualmente, es improbable que los algoritmos puedan extraer toda la información y valor al metadato por lo que esta forma no es del todo eficaz.

- Forma semiautomática: es la más aconsejable y utilizada por los usuarios ya que combina las poderosas tecnologías y los algoritmos que crean de manera automática los metadatos con la aprobación y supervisión manual del usuario, sólo en los casos en que sea necesaria su intervención.

La segunda etapa es la **Manipulación**, que consiste en los procesos y transformaciones que puedan sufrir los metadatos. Hace uso de procesos automáticos para tener control simultáneo de los cambios en los metadatos, lo que conlleva un mejor control sobre ellos y facilita su manejo a los usuarios. Por último, la etapa de Destrucción es la más complicada de manejar y se debe estudiar muy bien la manera en que se hará este proceso porque puede dar lugar a datos desactualizados o duplicados. Esta última etapa, consideramos que sería mejor llamarla etapa de **Actualización** ya que el metadato no se elimina en sí sino que consiste en modificar, de manera conjunta, el metadato proveniente de la etapa de Manipulación y sus datos asociados.

Ejemplo práctico En el caso del supuesto práctico, los datos que tenemos ya contaban con sus propios metadatos, por lo que este ciclo no empieza de cero. Sin embargo es un ciclo por el cuál pasan todos los metadatos una y otra vez, y con la herramienta de catálogo de datos el proceso se automatiza.

La primera etapa: Creación, se realiza de manera semiautomática para todos los metadatos, ya que los metadatos se crean por defecto según la información y el esquema del dato que el usuario ha definido al principio. Al insertar un nuevo registro añade todos los metadatos que pueda identificar y es el usuario quién decide si añadirlos o no. En la segunda etapa: Manipulación, correspondería a procesos de cambio simultaneo. Tomamos como ejemplo el identificador (campo 'No') del conjunto Proveedor. Si ese campo se modificase, gracias a los procesos automáticos de cambios debería cambiarse en todas las referencias que tenga en otros conjunto, como en el conjunto Producto se modificaría simultáneamente el campo Proveedor. En su última etapa: Actualización, los metadatos quedan registrados con las modificaciones que se han realizado en la anterior etapa.

4.1.4. Ventajas y Desventajas

Comprender los beneficios e inconvenientes que supone el uso de metadatos es imprescindible para desarrollar una gestión de metadatos efectiva.

A continuación, detallamos un listado de las ventajas y desventajas más comunes y críticas que se deben tener en cuenta a la hora de trabajar específicamente con metadatos.

Ventajas

En la actualidad existen herramientas, que se mejoran y evolucionan rápidamente, para gestionar de manera eficiente los datos y además integran aprendizaje automático para obtener mejores resultados.

Estas herramientas permiten explotar los metadatos para aportar una serie de beneficios en la gestión de los datos dentro de las empresas, que detallamos en el siguiente listado, extraída del Blog de Anna Pérez, responsable de contenidos de OBD Business School [47]:

- El uso de metadatos **facilita de manera considerable la búsqueda de información** dentro de la empresa, ya que disponen de una clasificación basada en diferentes atributos.

- Los metadatos aportan información muy detallada, lo que proporciona al dato una mayor calidad y veracidad. Este hecho es muy valioso para las empresas ya que, a la hora de **tomar decisiones**, sean críticas o no, es más probable que sean **acertadas y beneficiosas** para la organización.

- Los metadatos aumentan la cantidad de información de los conjuntos de datos y esto facilita la definición y diseño de estrategias de mejora que, en el ámbito de la competitividad entre las empresas, repercute directa y positivamente en la **diferenciación de una empresa sobre las demás**, siendo más competente frente a otras organizaciones.

Desventajas

La información es poder. Saber buscar y analizar toda la información que ofrecen los metadatos puede ser crucial para el éxito de un negocio y es por este motivo por el cual es necesario tener muy en cuenta no solo las ventajas, sino también las desventajas que expondremos a continuación, aunque debemos advertir que en la época actual en la que vivimos, son problemas menores que no causan muchos contratiempos.

Pese a que existen múltiples beneficios en la manipulación y gestión de los metadatos, también traen consigo una serie de problemas que no pueden ser pasados por alto [37]:

- El uso de estos metadatos es muy **costoso a nivel económico y** la inversión de **tiempo** para tratarlos es muy alta.

- La **comprensión de los metadatos** puede llegar a ser **muy complicada**, ya que los usuarios no definen unos estándares para su uso y esto dificulta el entendimiento de los mismos.
- Los **metadatos pueden tener una interpretación subjetiva** dependiendo del punto de vista de los usuarios y el contexto en el que se utilicen. Cada usuario puede interpretar un mismo metadato a su conveniencia.
- **No existen límites** en cuanto al detalle que puede adquirir un metadato, por lo que pueden llegar a surgir problemas para encontrar el dato que se necesita de forma rápida. Esto quiere decir que la casi inexistencia de metadatos, así como el exceso de metadatos, puede ralentizar el proceso de obtención del dato. No contar con muchos metadatos puede dar lugar a datos muy similares y se hace difícil identificar el dato exacto que se quiere. De la misma manera, si se tiene un dato con muchos metadatos, estaría sobrecargado y podría contener demasiadas referencias e invalidar el dato para una tarea específica.
- Los metadatos pueden ser superfluos, lo que se convierte en un inconveniente a la hora de su búsqueda, es decir, existen metadatos que no aportan la información necesaria en un contexto especifíco o no está bien especificado para ese contexto, por lo que la búsqueda por ese metadato puede dar lugar a conjuntos de datos variados que no son exactamente lo que se busca. Sin embargo, se está avanzando en buscadores que sean capaces de realizar búsquedas basadas en ejemplos que ayuden en la localización de los metadatos en el momento preciso.

Estos son algunos de los inconvenientes generales que puede traer consigo el uso de metadatos, sin embargo, el contar con ellos es imprescindible para facilitar y potenciar tareas importantes en la empresa en donde el uso de datos sea vital. Por este motivo, también es beneficioso conocer algunos problemas derivados específicamente de su utilización, para poder combatirlos en la medida de lo posible como: la falta de compromiso por parte de los usuarios para definir y detallar de forma completa los registros de metadatos, al ser una tarea dificultosa y de gran dedicación, o problemas que surgen por una inconsistencia semántica, dando lugar a metadatos distintos que contienen la misma información o que están insuficientemente documentados.

4.2. Perfiles de datos

4.2.1. Introducción

Los **perfiles de datos o *Data Profiling*** son uno de los aspectos más importantes cuando se habla de calidad de datos.

Se definen como el *proceso de examinar, analizar y revisar los datos desde una fuente de origen y hacer un resumen útil de la información obtenida* [61]. Descrito de otra forma, los perfiles de datos *son procesos de examinación de los datos desde su origen, comprendiendo*

su estructura, contenido e interrelaciones, e identificando el valor potencial que puedan llegar a aportar los datos [63].

Antes de profundizar sobre este aspecto de la calidad de datos, es necesario entender la diferencia entre perfiles de datos y evaluación de datos, ya que son dos conceptos que se suelen confundir porque sus funciones están muy relacionadas entre sí, aunque no tratan los mismos aspectos [15].

Mientras que los perfiles de datos se centran examinar y clasificar los datos, la *evaluación de datos es un proceso donde se determina que valor tienen los datos, la importancia de aquello que se está midiendo y pone en disposición dos objetos para poder compararlos entre sí*, es decir, determina si los datos obtenidos en cada perfil de datos son valorados de forma equilibrada y correcta, si son significativos, si aportan valor a la empresa y si son un reflejo exacto del verdadero alcance de un tema concreto.

La elaboración de perfiles de datos implica conocer una serie de características para que la empresa [63] pueda integrar mejor esta herramienta. Los perfiles de datos hacen uso de las estadísticas, proporcionando información descriptiva como mínimos, máximos, recuentos o sumas. Hacen recopilaciones sobre el tipo de datos que se encuentra, su longitud o patrones comunes. Realizan un etiquetado más exhaustivo de los datos con palabras clave, descripciones o categorías. Son útiles en la realización de evaluaciones de la calidad o del riesgo de realizar uniones entre los datos. Ayudan en el descubrimiento de metadatos y evalúan la precisión que pueden llegar a tener. Además, los perfiles de datos pueden identificar a nivel técnico distribuciones, claves candidatas, claves foráneas candidatas, dependencias funcionales, dependencias de valor incrustado y realizar análisis entre tablas.

Motivación

Los perfiles de datos son herramientas que se integran generalmente en los catálogos de datos para encontrar la mejor solución a los datos que se recogen [67]. El añadir perfiles de datos mejora significativamente la calidad de los datos y por tanto la credibilidad en ellos, y esto es así porque ayudan al proceso de eliminación de duplicados y anomalías, así como a determinar qué información es útil para la toma de decisiones de la empresa.

El perfilado permite a una empresa predecir cuáles serán las mejores opciones a partir de las cuales poder elegir las más acertadas y, por lo tanto, contribuyen directamente al estado de este tipo de organizaciones. Ayudan a identificar con mayor rapidez algunos problemas que puedan surgir, incluso antes de que aparezcan y aportan información válida sobre la manera de abordarlos de forma óptima.

Además, con los perfiles se puede analizar con mayor precisión las distintas fuentes y bases de datos donde se originaron los datos y asegurarse de que todos estos datos cumplan con las medidas y reglas empresariales específicas de cada organización.

Gracias a esta herramienta también se puede comprender con mayor exhaustividad los datos y sus relaciones, permitiendo fijar objetivos a largo plazo, construir una estrategia con visión de futuro, y todo ello simplificando los procesos asociados a los datos.

4.2.2. Proceso de los perfiles de datos

Para poder elaborar un perfil de datos, es necesario que el conjunto de datos que vaya a formar parte de este perfil, pase por dos grandes etapas: el Análisis de la Fuentes de Datos y la Consolidación de perfiles de datos.

Primera Fase:

Originalmente y según ETL Tools [13] esta fase se llama Elaboración de perfiles de datos. En este trabajo hemos considerado más acertada denominarla "Análisis de las Fuentes de Datos" ya que, esta primera etapa es un paso previo de análisis de fuentes de datos, un análisis del estado en el que se encuentras los datos recopilados en sus fuentes, para después poder elaborar estos perfiles.

En esta etapa se examinan los datos en sus fuentes de origen con el fin de encontrar anomalías y problemas que puedan surgir en ellos. Se compone de procesos de limpieza, como la comprobación de las relaciones entre tablas, de las reglas de negocio y de la familiarización con las fuentes de datos [13]

El proceso del análisis de las fuentes de datos está conformado por 8 pasos que deben realizarse antes de pasar a la siguiente etapa. Estos pasos son:

1. La realización de un **manual o documento inicial** del proyecto en sí. Contiene información necesaria para evitar gasto de tiempo y esfuerzo innecesario, así como el alcance, tiempo del proyecto, las expectativas y requisitos.

2. Escoger las **herramientas analíticas y estadísticas** adecuadas para las necesidades del proyecto. Con estas herramientas se pretende comprender la calidad de los datos mostrando qué datos son más importantes, aportan mayor valor o son usados más frecuentemente.

3. Analizar detalladamente las **distintas fuentes de datos** que se tienen disponibles.

4. Comprender y **determinar el alcance** al que pueden llegar los datos, sin perder el valor que aportan.

5. Estudiar los **distintos patrones**, junto a sus variaciones, y formatos de los datos que se encuentren en la base de datos.

6. **Identificar anomalías** en los conjuntos de datos. Estas anomalías pueden ser: codificación múltiple, valores redundantes, duplicados, nulos, perdidos, etc.

7. Verificar las **relaciones entre los datos** y comprobar si esa relación influye en la extracción de datos.

8. Hacer un **análisis de las reglas de negocio** que puedan llegar a afectar al proyecto.

Además de los pasos que se han de seguir en el análisis, este cuenta con una serie de *técnicas consideradas mejores prácticas para el análisis de la calidad y los perfiles de*

datos [63], es decir, se aplican estas técnicas al proceso para conseguir un mejor análisis de las fuentes de datos y la calidad.

Estas técnicas son:

- **Recuento y porcentaje**. Detecta claves principales, valores distintos en cada columna que ayudan en los procesos de inserción y actualización.
- **Porcentaje de valores cero / en blanco / nulos**. Identifica anomalías en los datos, como falta de algún dato o datos mal identificados.
- **Longitud de cadena mínima / máxima / promedio**. Identifica qué tipos y el tamaño de los datos que se encuentran en la base de datos. Con este análisis se permite saber el tamaño máximo que puede alcanzar un atributo y por lo tanto mejorar el rendimiento en general.
- **Integridad de la clave**. Hace uso de la técnica del porcentaje de valores cero / en blanco / nulos para garantizar que las claves estén siempre actualizadas y presentes en los datos. Con esta técnica se identifican claves huérfanas, datos con referencias perdidas o desactualizadas, con lo que mejoran los procesos ETL.
- **Cardinalidad**. Estudia los tipos de relaciones ya sean uno a uno, uno a muchos o muchos a muchos, entre los distintos conjuntos de datos relacionados.
- **Distribuciones de patrones y frecuencias**. Hace comprobaciones en los formatos de los distintos campos de los datos y verifica que esos formatos sean válidos para evitar complicaciones a futuro.

En esta primera fase se comienza a crear el perfil de datos ya que gracias al análisis se conforma un conjunto de datos con unas mismas características que después, en la segunda etapa se consolidarán para elaborar oficialmente un perfil de datos.

Segunda Fase:

La Consolidación de perfiles de datos o como originalmente se denomina "Creación de los perfiles de datos", consiste en el estudio, análisis y creación de resúmenes de datos útiles que gracias a la etapa anterior hemos conseguido obtener [67].

En este trabajo hemos creído conveniente denominar a esta etapa como "consolidación" en vez de "creación" puesto que el perfil de datos ya ha comenzado a crearse en la fase anterior, y es en esta segunda etapa donde se formaliza el perfil consolidando el conjunto de datos que lo componen.

Gracias a esta parte del proceso, podemos adquirir una visión de conjunto que facilita la identificación de problemas, riesgos y tendencias generales relativas a la calidad de datos.

Ralph Kimball, padre de la arquitectura de almacenamiento de datos, propone un conjunto de 4 fases para la consolidación de perfiles de datos [63]. Estas fases consisten en:

1. El inicio se basa en el **análisis de perfiles de datos**, con el fin de descubrir si los datos recopilados pueden influir positivamente y ser adecuadlos para el análisis. Dependiendo de este paso, se decide si seguir con un proyecto o no.

2. **Detectar y solucionar los problemas derivados de la calidad de datos** en el origen antes de que ser se derivan a su destino final.

3. Solucionar los problemas que puedan surgir en los datos desde el origen hasta el destino, mediante el **uso de herramientas ETL (*Extract Transform Load* o Extraer Transformar Cargar)**.

4. **Distinguir las distintas reglas de negocio** existentes, estructuras jerárquicas y relaciones adicionales. Este paso se usar para ajustar el proceso ETL y proporcionar mejores resultados.

4.2.3. Tipos de perfiles de datos

Los perfiles de datos se clasifican según las técnicas o procesos de elaboración de perfiles que se utilicen, dando a lugar a tres categorías.

En cualquiera de las categorías, los objetivos son los mismos: mejorar la calidad de los datos y alcanzar una mejor comprensión de ellos [62].

Proceso de Descubrimiento de estructuras

Este tipo de perfiles se centra en analizar la estructura del dato comprobando que sean correctos y no tengan ningún tipo de anomalía. El estudio de la estructura sirve para comprender al dato en sí, y verificar que sea consistente y tenga un formato adecuado.

Hace uso de la estadística básica para dar información sobre el valor que aportan los datos.

Proceso de Descubrimiento de contenido

Se encarga del estudio del dato de forma individual con el fin de descubrir inconsistencias, se centra en la calidad de los datos. Con este estudio es más fácil comprobar cuál y dónde se ha producido el error y si es un error sistémico, encontrarlo inmediatamente.

Proceso de Descubrimiento de relaciones

Se centra en el análisis de las relaciones que hay entre los distintos conjuntos de datos. Este análisis permite una mayor reutilización de los datos y fijar un único lugar donde se unen las distintas fuentes de datos.

4.2.4. Tipos de análisis

La función más importante de los perfiles de datos es que son capaces de analizar distintos conjuntos de datos y conseguir una perspectiva más completa sobre ellos.

En este sentido, no existe un único análisis, sino que se pueden unir varios tipos de análisis para un mismo conjunto de datos, dependiendo de las necesidades [14].

Los distintos tipos de análisis que se conocen son los siguientes:

- **Análisis de la exhaustividad**

  Determina la frecuencia con la que un campo de un atributo determinado aparece completado, en blanco o como valor nulo en la lista de resultados.

- **Análisis de Distribución de valores**

  Determina cuál es la distribución de los distintos registros a través de diferentes valores que pueden tener un mismo atributo.

- **Análisis de unicidad**

  Determina para un mismo atributo cuantos valores únicos o distintos puede llegar a tener, o dicho de otro modo es una medida para detectar duplicidades.

- **Análisis de patrones**

  Determina qué formatos tiene un mismo atributo, así como la distribución de registros a través de un formato u otro.

- **Análisis de rango**

  Determina, haciendo uso de la estadística, los valores mínimos, máximo y medio que pueda tener un atributo concreto.

4.2.5. Ámbitos o Escenarios propicios

En este apartado señalaremos distintos tipos de escenarios donde el uso de perfiles de datos es casi imprescindible para sacar el mejor partido de la situación y los mayores beneficios [14].

Estos escenarios son:

- **Iniciativas de calidad de datos**. El principal objetivo de estos proyectos es poder corregir posibles anomalías y prevenir que aparezcan en el futuro. Con los perfiles de datos se puede identificar en que partes del sistema se encuentran los problemas de calidad de mayor impacto para la empresa y con ello poder solucionar el problema con la mayor rapidez y eficacia.

- **Proyectos de migración de datos**. Estos proyectos consisten en el traspaso de grandes cantidades de datos de un lugar de origen a otro de destino. Los perfiles de datos ayudan a minimizar los riesgos que pueden surgir durante el traslado, identificando los problemas antes de realizar la migración. Nuestro supuesto práctico se corresponde con este tipo de escenario, por lo que usar perfiles de datos es adecuado para cumplir el objetivo.

- **Iniciativas de *Data Warehousing* e inteligencia empresarial**. Son dos proyectos que tienen la necesidad de recopilar datos de múltiples sistemas diferentes para ofrecer las mejores decisiones de negocio a través del análisis de estos conjuntos

de datos históricos, actuales y predictivos. Los perfiles de datos ayudan a la identificación de tres tipos de problemas típicos que pueden surgir en estos proyectos: los relacionados con la calidad de los datos en el punto de origen, los que surgen si no se identifican adecuadamente los atributos y que pueden corregirse en el proceso ETL, y otros problemas que aparecen de las nuevas reglas de negocio que se den a conocer.

4.2.6. Beneficios

A lo largo de este apartado sobre perfiles de datos, hemos podido averiguar cuáles son algunas ventajas que aporta el uso de estos perfiles gracias a la revisión y análisis de su conceptualización, del cómo, dónde y porqué se decide utilizarlos o en qué situaciones y escenarios es recomendable incluirlos en los proyectos.

En este subapartado recogeremos e indicaremos algunos beneficios principales que trae consigo la utilización de perfiles de datos en los proyectos de una empresa, ya que, como hemos podido comprobar, incluirlos de forma automática mejora considerablemente el rendimiento de una empresa en varios ámbitos de actuación (economía del tiempo, toma de decisiones, calidad y veracidad en sus datos, etc.)

Según un artículo de *Harvard Business Review*, escrito por Thomas C. Redman, el costo anual para las empresas estadounidenses por contar con una mala calidad de datos serían unos 3,1 billones de dólares[1]. Estas cifras invitan a la empresa a reconsiderar la implementación de herramientas de calidad de datos donde se incluyen los perfiles de datos, ya que éstos **mejoran la calidad y credibilidad de los datos**, analizando los conjuntos de datos y eliminando duplicaciones y anomalías que se detecten en ellos.

Por lo tanto, utilizar perfiles de datos beneficia en [64]:

- **Toma de decisiones**, ya que al comprender mejor los conjuntos de datos se pueden analizar ciertas informaciones y utilizarlas para "predecir el futuro", acertando en mayor porcentaje en la elección de las decisiones de la empresa.
- **Facilita la resolución de problemas**, aumentando la velocidad y agilidad para sobreponerse a los errores encontrados.
- **Permite obtener una clasificación organizada** de los distintos conjuntos de datos, que vienen de diversas fuentes de origen. Los perfiles de datos analizan estas fuentes asegurándose de que los datos cumplan con el estándar y las reglas de negocio.

[1] `https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year`

Capítulo 5

Catálogos de datos

5.1. Introducción

Según Gartner, empresa consultora y de investigación de las tecnologías de la información, **un catálogo de datos** se define como una herramienta que: *"mantiene un inventario de activos de datos a través del descubrimiento, descripción y organización de conjuntos de datos. Un catálogo proporciona contexto para permitir que los analistas de datos, científicos de datos, administradores de datos y otros consumidores de datos encuentren y comprendan un conjunto de datos relevante con el fin de extraer valor comercial"* [58]. A esta definición se añade que los catálogos de datos automatizan la administración de metadatos y la hacen colaborativa.

De manera más sencilla, un catálogo de datos es una colección de metadatos, que combina herramientas de búsqueda, con administración de datos, lo que ayuda a los analistas y otros usuarios de datos a encontrar, de forma más fácil, los datos que necesitan. En general, sirven como un inventario de los datos disponibles y proporciona información para evaluar los datos de aptitud para usos previos [56].

Este tipo de herramientas son esenciales para aquellas empresas que se basan principalmente en datos. Entre sus principales usos se encuentran el de facilitar la identificación, comprensión y colaboración de los datos. Además, son herramientas imprescindibles para la administración, conservación y gobierno de datos. Son útiles desde un punto de vista estratégico, ya que son buenos en tareas como la gestión de activos de datos y mejoran notablemente la calidad y productividad de los análisis [72].

Durante el presente trabajo hemos destacado la gran importancia que tienen los datos, su utilidad en distintas áreas y cómo pueden proporcionar valor a distintos tipos de usuarios. En este sentido, los catálogos de datos son comunes a múltiples tipos de usuarios, siendo capaces de proporcionar el mejor resultado a cada usuario de manera independiente a pesar de que estos usuarios tiene intereses y necesidades diversas.

A continuación, se muestra una breve descripción de los distintos tipos de usuarios que pueden existir y qué problemas pueden llegar a enfrentar [59].

Ingenieros de datos o *Data Engineers*

Los ingenieros son aquellos usuarios encargados del diseño y construcción de los datos,

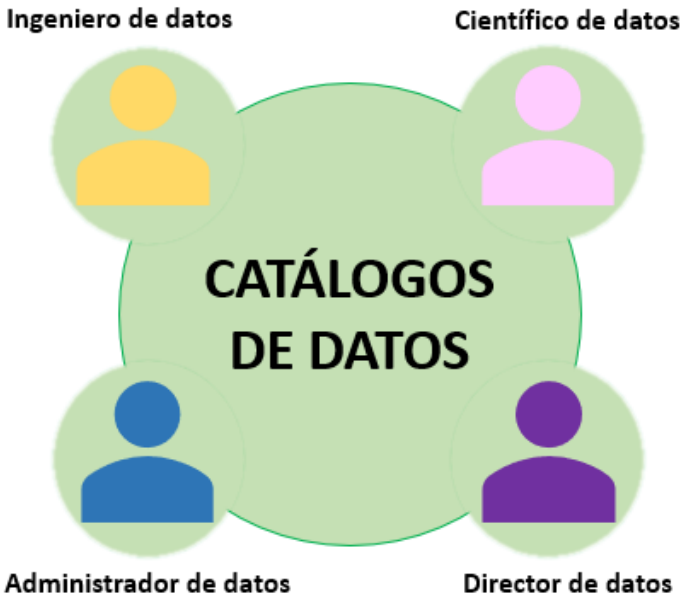

Figura 5.1: Elaboración propia: Usuarios para catálogos de datos

desarrollan su arquitectura, los procesos y son los encargados del rendimiento y calidad de estos datos.

Necesitan los datos para comprender a fondo cómo pueden afectar los cambios en conjunto al sistema.

Ejemplos: saber el impacto que puede provocar en la aplicación realizar un cambio en el esquema de los datos o las diferencias entre dos estructuras de datos.

Científicos de datos o *Data Scientists*

Los científicos de datos son aquellos usuarios encargados del análisis de los grandes volúmenes de datos, su principal función es organizar y analizar los datos recopilados para así sacar el máximo valor a los datos.

Necesitan tener un acceso a los datos rápido y sencillo, además de la calidad que aporta cada dato.

Ejemplos: saber dónde pueden encontrar y explotar datos concretos o como acceder de la forma mas sencilla a un Lago de datos o *Data Lake* concreto.

Administradores de datos o *Data stewards*

Un administrador de datos es aquel usuario encargado de la gestión de los recursos de datos, su definición, políticas, procedimientos, estándares, así como el ciclo de vida de los datos. Necesitan tener en cuenta toda la parte administrativa de los datos y sus procesos.

Ejemplos: saber si las medidas que se toman son verdaderamente acertadas y si mejoran la calidad de los datos, o si están definidos los estándares para los datos claves de la empresa.

Directores de datos o *Chief Data Officers*

Un director de datos es el encargado de gestionar y manejar a los demás usuarios y

las tareas que realiza cada uno dentro de la empresa.

No están vinculados directamente a los catálogos de datos, sin embargo, deben tener en cuenta cosas como, por ejemplo, quién accede a qué información o las políticas de retención que se han definido para todos los datos.

5.1.1. Por qué surgen

Las bases de datos y lagos de datos son estructuras potentes para almacenar, actualizar, consultar, buscar y procesar datos, sin embargo, sin una herramienta adecuada, es imposible sacar el máximo partido y valor a todos los datos recogidos y dificulta la compartición de estos datos con los usuarios.

Antes de la existencia de los catálogos de datos había herramientas que se encargaban de la extracción de los metadatos y daban la posibilidad de operar con ellos. Pero estas herramientas cuentan con una serie de limitaciones que hacen que sean poco útiles para ser usadas en las empresas basadas en datos, y más con las nuevas tecnologías que van apareciendo, como el Big Data.

Algunas de estas limitaciones pueden ser que se requiere demasiada experiencia técnica, lo que imposibilita el uso de estos datos a usuarios que no sean técnicos. Se utilizaban métodos manuales, que dificultan el manejo de datos en empresas con múltiples bases de datos. Los enfoques que ofrecen no son útiles para usuarios como los científicos de datos, que trabajan con aprendizaje automático. También se encuentran dificultades en la estrategia en la que se basan, ya que no facilita el gobierno de datos.

Todas las limitaciones mencionadas son eliminadas con la incorporación de los catálogos de datos. Estos permiten saber qué datos existen, cómo y dónde encontrar las mejores fuentes de datos, quién es el responsable de cada dato o conjunto de datos y además, aportan medidas de seguridad para proteger los datos, e incluyen funciones de gestión de metadatos, dando una mayor libertad al manejarlos.

Un ejemplo que muestra la importancia y utilidad de los catálogos de datos puede ser el de la plataforma Collibra de Estados Unidos, en una situación tan actual como puede ser el COVID-19 [34]. Tal situación ha demostrado la importancia de tener los datos necesarios para poder actuar de forma ágil y eficientemente, y que éstos sean de calidad y estén los más actualizados posible, especialmente para la toma de decisiones importantes.

Un gran sistema de salud en el área metropolitana de Boston utilizó el catálogo de datos de Collibra para realizar un seguimiento de los pacientes, el número de camas, los médicos, las enfermeras y el diagnóstico durante la pandemia. La oficina de datos de este hospital utilizó datos para orientar todas sus decisiones, como la cantidad de camas disponibles, los miembros del personal que fueron reubicados en estos lugares de interés improvisados y la cadena de suministro de EPP (Equipos de Protección Personal). También monitorearon la propagación regional del virus a diario y la propagación interna para determinar dónde trasladar a los pacientes y evitar el desbordamiento en un hospital individual. Utilizaron Collibra Data Catalog [9] para acceder fácilmente a estos datos e informes, de modo que pudieran tomar decisiones rápidas e informadas.

Además de este caso concreto en Boston, se dieron muchos otros en distintos hospitales e incluso en compañías de seguros, lo que provocó que Collibra crease su propio catálogo de datos para todo lo referente al COVID-19, que reúne conjuntos de datos de alta calidad de organizaciones como la Organización Mundial de la Salud, la Universidad de Oxford y el Centro de Ciencia e Ingeniería de Sistemas (CSSE) de la Universidad Johns Hopkins. Este catálogo de datos COVID-19 es público y está diseñado para ayudar a epidemiólogos, investigadores, científicos de datos y analistas que trabajan para hospitales, empresas de atención médica y agencias públicas a encontrar y acceder fácilmente a los datos que necesitan para tomar decisiones rápidas y que impacten positivamente.

A modo de resumen, el uso de catálogos de datos como herramientas de ayuda con los datos es esencial en cualquier empresa que considere que sus datos son fundamentales para conseguir el éxito. Como hemos ido señalando, los catálogos de datos son imprescindibles porque te permiten confiar en los datos y comprenderlos, lo que implica sacar el máximo valor de estos. Además, se obtiene una vista unificada de todos los datos de la empresa, lo que permite encontrar fácilmente los datos necesarios entre las múltiples fuentes de datos y en cada situación, lo que reduce el tiempo de búsqueda, ayudando también al cumplimiento normativo.

5.2. Desafíos a los que se enfrentan

Big Data, servicios en la nube o una mayor regulación en torno a la privacidad de los datos, son temas de actualidad donde se mueven grandes volúmenes de datos. Esto hace que se dificulte el encuentro y el acceso a los datos requeridos, lo que se convierte en un inconveniente para la Gobernanza de los Datos o *Data Governance* y las empresas.

Como hemos estado observando y analizando, es una tarea fundamental el comprender los datos que se captan, conocer de qué tipo son, quién los trata, la finalidad de su utilidad y las políticas de seguridad que se encargan de proteger esos datos. Otro hecho fundamental es que se debe tratar con datos lo suficientemente sencillos para que no existan complicaciones en su utilización.

Ahora bien, nos encontramos con una serie de desafíos a los que se deben enfrentar los catálogos de datos para poder encontrar y acceder a los datos correctos. Entre estos desafíos destacamos los siguientes [59]:

- Se estima dedicar una cantidad considerable de tiempo y esfuerzo en la búsqueda y captación de los datos.
- Si no se lleva un control correcto y apropiado, los lagos de datos que se tienen pueden llegar a convertirse en pantanos de datos.
- La inexistencia de un vocabulario de negocio común para que todo tipo de usuarios pueda entender e interpretar los datos de manera correcta.
- La aparición de datos oscuros dificulta el trabajo de los usuarios por la comprensión de la estructura y la variedad de estos datos.

- Se debe llevar un riguroso control en el linaje de los datos, donde se evalúa la procedencia, calidad y confiabilidad de los datos.

- *"Todos los datos son iguales"*. Se refiere a que a priori no se puede distinguir entre un dato crítico para la tarea y un dato estándar.

- Se debe hacer un esfuerzo extra para documentar datos preparados manualmente y Datos ad-hoc.

5.3. Cualidades que deben tener

Los catálogos de datos son herramientas que sirven para analizar y sacar el máximo valor a los datos, por eso es necesario conocer qué cualidades debe tener un catálogo para ser considerado como una herramienta válida y de calidad.

Estas herramientas deben ofrecer, en primer lugar, una vista unificada de todos los datos de los que se compone, para que resulte más accesible a los distintos usuarios. Por otro lado, debe evitar que surjan pantanos de datos, llevando un control apropiado de los conjuntos de datos. Además, debe elevar la confianza y seguridad de sus datos, mejorando la satisfacción de los clientes y conseguir un control adecuado para incrementar la productividad y eficacia operativa. Para los usuarios una de las tareas más significativas que pueda ofrecer un catálogo es la economía del tiempo, es decir, que el tiempo dedicado a comprender y asumir la información requerida, se reduzca considerablemente [57].

Por otra parte, los catálogos deben aportar cualidades como añadir opciones flexibles en cuanto a la búsqueda y filtrado de los datos, donde los usuarios puedan decidir en todo momento, y además se les permita añadir etiquetas o información relevante según el criterio del usuario. Deben poder recopilar todo tipo de metadatos de múltiples fuentes de información ya sean bases de datos o sistemas locales. Debe automatizar, dentro de lo posible, todas aquellas tareas que se realicen de manera manual, y para ello se hace uso de las nuevas tecnologías, como la inteligencia artificial o las técnicas de aprendizaje automático. Es necesario que tengan capacidades de nivel empresarial referentes a la administración de identidades y accesos, y capacidades básicas a través de API REST[1].

5.4. Características y funciones

Las empresas que están en continuo cambio necesitan tener una gran capacidad de adaptación a esos nuevos cambios tecnológicos y de negocio. Los catálogos de datos, al igual que las empresas, deben ser capaces de adaptarse y gestionar estos cambios de manera rápida y eficaz. Por ello deben [11]:

[1]API REST: Arquitectura software que transmite la información mediante el protocolo HTTP, sin llevar con sigo información entre las peticiones

- **Proporcionar capacidades necesarias para su construcción inicial**, para un mantenimiento continuo y que sean lo suficientemente fáciles de usar para usuarios que necesiten encontrar y trabajar con los datos rápidamente.
- Dotar de **seguridad a sus datos**: algunas de estas funciones pueden ser control de accesos, saber quién accedió a los datos, además de tener una capacidad de auditoría y encriptación.

Los catálogos de datos pueden tener capacidades variadas y únicas ya que se crean con un fin específico. Sin embargo, todos los catálogos o la mayoría de los catálogos de calidad tienen una serie de características comunes y claves y que son imprescindibles:

- **Gran capacidad de implantación** dentro de la empresa, para evitar el máximo de inconvenientes posibles.
- Deben tener una serie de **conectores y herramientas** que ayuden a construir un único lugar de confianza.
- Implementar un **catálogo de metadatos** amplio que muestre la conectividad entre ellos.
- Impulsar la **automatización con el aprendizaje automático** para ganar velocidad y agilidad, evitando conectar las fuentes de datos de forma manual.
- Llevar un **control sobre el linaje de los datos**, haciendo que exista una mejor comprensión entre los diferentes tipos de datos y sus fuentes.
- Ofrecer **gobernanza y privacidad de datos** integrados, lo que conlleva aportar contexto empresarial a los datos.
- Capacidad de **colaboración con fuentes y componentes externos** que mejoran la búsqueda y muestra de resultados.
- Incorporar una **herramienta de perfil de datos** para la evaluación de la calidad de datos.
- Por último, una de las características más importantes es la **búsqueda poderosa y multifacética** que tienen para explorar los distintos conjuntos de datos de forma instantánea.

5.5. Criterios de elección

Una vez que tenemos el conocimiento suficiente sobre el concepto, características y funciones de los catálogos de datos y sus aportaciones a la empresa, es imprescindible escoger un catálogo que se adapte a las necesidades de cada organización.

La lista de criterios de elección de catálogos de datos que ofrecemos a continuación sirve de guía en este proceso y está estrechamente relacionada con las funciones y características descritas en el subapartado anterior. No todos los criterios son igual de importantes y necesarios para la empresa, por ello se deberán priorizar aquellos que aporten un mayor valor a la toma de decisiones y a la empresa en general.

Criterios de elección de catálogos de datos [2]:

1. **Catalogación de conjuntos de datos.**

   Un catálogo de datos debe incluir procesos automatizados para descubrir nuevos conjuntos de datos, ya sea al comienzo de la creación del catálogo o a lo largo de su vida.

   Debe admitir aprendizaje automático para recopilar metadatos, añadir la inferencia semántica y proporcionar un etiquetado automático. Con todo ello se pretende extraer el máximo valor de la automatización y disminuir el trabajo manual por parte de los usuarios.

2. **Catalogación de operaciones de datos.**

   Se debe tener una lista catalogada de todas aquellas operaciones necesarias en la preparación de los datos y además permitir la asociación de estas operaciones a los conjuntos de datos oportunos.

   Estas operaciones consisten en procesos para mejorar, enriquecer, formatear y combinar los datos.

3. **Búsqueda.**

   El criterio de búsqueda es imprescindible en cualquier catálogo de datos. Este criterio es el que busca cualquier usuario de catálogos, sobre todo si se trata con usuarios no técnicos, para cubrir cualquiera de sus necesidades. La búsqueda debe permitir buscar por distintos filtros, como puede ser, por facetas, palabras clave o términos empresariales.

   Además de la búsqueda por filtro, un catálogo de datos debe ser capaz de mostrar los resultados de la forma que mejor convenga al usuario, como es búsqueda por relevancia o por frecuencia de uso.

4. **Recomendaciones.**

   La capacidad de realizar recomendaciones a los usuarios basándose en sus experiencias anteriores puede resultar en procesos de búsqueda más rápidos y mejorar la calidad de las coincidencias entre los resultados de la búsqueda. Permite establecer conexiones entre los conjuntos de datos y las operaciones de preparación de datos y flujos de trabajo.

   Para las recomendaciones se hace uso de los metadatos proporcionados por el historial de uso además de la utilización del aprendizaje automático.

5. **Evaluación del conjunto de datos.**

   Los catálogos pueden proporcionar una opción de evaluación del conjunto de datos antes de tener que descargarse los datos en sí. Es decir, los catálogos pueden escogerse según el nivel de evaluación de la idoneidad de los conjuntos de datos a la hora de hacer una elección sobre cuáles escoger.

   Entre algunas de las características que ofrece la evaluación se encuentra tener una vista previa de los conjuntos de datos, visualizar sus perfiles de datos, obtener clasificaciones de otros usuarios, permitir la lectura de opiniones y anotaciones de los usuarios respecto a los conjuntos de datos y obtener información sobre la calidad que tienen esos datos.

6. **Acceso a los datos.**

   Este es otro de los criterios básicos, junto con la evaluación de los datos. Los catálogos deben proporcionar una buena experiencia al usuario, desde el momento en el que realizan la búsqueda del dato hasta que, finalmente, lo obtienen y para ello los conjuntos de datos deben ser accesibles directamente desde el catálogo.

   El acceso a los conjuntos de datos debe tener en cuenta la protección de datos para asegurar la seguridad, privacidad y cumplimiento de datos confidenciales.

7. **Metadatos de uso.**

   La recopilación de metadatos de uso de los datos es básica, ya que de los metadatos dependen otras funciones como la evaluación de los conjuntos de datos o las recomendaciones inteligentes.

8. **Valoración de datos.**

   Una buena experiencia para los usuarios a la hora de considerar con qué datos trabajar puede traducirse en tener datos cuantificables sobre su valor. A partir de la frecuencia de uso de los datos y los casos de uso analíticos se puede realizar una estimación del valor del conjunto de datos, que permita a los usuarios observar que datos son más adecuados para su utilización.

9. **Catálogo de metadatos.**

   Se recogen y recopilan metadatos pero, debemos tener en cuenta que, éstos tienen que seguir un orden que permita extraer información de estos metadatos, así como su linaje, cuáles son los datos que contienen, en qué conjuntos de datos están y otras informaciones. Para conseguir todo ello se necesitan catálogos de metadatos.

10. **Seguridad.**

    La seguridad de los datos es la primera de las cuatro capacidades esenciales de gobierno de datos.

    Los catálogos de datos deben tener la capacidad de trabajar en entornos seguros que incluyan procesos de autenticación y autorización por parte del usuario. Debe

existir una distinción entre los distintos tipos de usuarios que puede haber y sus distintas tareas.

Se debe considerar los niveles en los que se puede imponer las distintas restricciones de seguridad, ya sea a nivel de conjunto de datos, nivel de registro / fila, a nivel de columna / campo, o a nivel del valor del dato.

11. **Linaje.**

    Un catálogo de datos debe ser capaz de hacer un análisis completo y profundo del linaje de los datos que contiene. Es necesario conocer la fuente original del dato para poder generar informes que den la confianza necesaria en los datos.

    El linaje también se puede utilizar en la gestión de cambios, los análisis de impactos o para hallar la solución ante diversos problemas y la manera de resolverlos.

12. **Cumplimiento de las normativas.**

    Un problema muy común al tratar con datos es el que gira en torno a la protección de datos. Debemos tener muy en cuenta y no cometer errores con este tipo de regulaciones. El RGPD (Reglamento General de Protección de Datos) es un reglamento que regula este aspecto de los datos, siendo de los más relevantes, aunque existen otras regulaciones más específicas de la industria como HIPAA [53] y Dodd-Frank [20].

    Los catálogos, por tanto, deben centrarse en gestionar la PII (Información Personal Identificable) que protege la privacidad de los datos y admitir el cumplimiento de las normativas.

13. **Calidad en los datos.**

    Al igual que la seguridad, la calidad de los datos es una capacidad esencial para la gobernanza de datos. Cada vez es más compleja ya que han ido surgiendo conceptos como Big Data o lagos de datos.

    Los catálogos de datos no son los que proporcionan la calidad a los datos, sino que se limitan a la administración de esta Calidad. Para ello, utiliza algoritmos inteligentes que faciliten el encuentro de conflictos de datos e identifican si existe deficiencia en la calidad.

    Tener gestionada la calidad de los datos ayuda a los analistas a evaluar y escoger conjuntos de datos de manera más exacta.

14. **Curación de datos.**

    Los encargados de hacer que el catálogo de datos sea de utilidad y valioso dentro de la empresa son los llamados curadores de datos (*data curators*).

    Algunas de las acciones más importantes que realizan estos curadores de datos son las siguientes: evalúan la riqueza de las capacidades de curación junto a la capacidad de agregar conjuntos de datos, ocultarlos o eliminarlos. Son los encargados

de añadir anotaciones en la creación de metadatos y agregar términos y etiquetas de búsqueda. Identifican quiénes son los administradores y pymes[2], etiquetan los datos sensibles a la seguridad y el cumplimiento de la normativa, además de compartir consejos y técnicas, y fomentan el *crowdsourcing* (colaboración abierta distribuida) de metadatos.

15. **Socialización.**

    Una característica interesante de los catálogos es la capacidad que tienen para "socializar", es decir, que sean capaces de compartir información con componentes externos.

    Algunas de las capacidades son el *crowdsourcing* de metadatos, las características de colaboración, la publicación de valoraciones y opiniones de los usuarios y la captura de los comentarios de los usuarios.

16. **Integración e interoperabilidad.**

    Los catálogos de datos no pueden funcionar de forma aislada. Durante todo el ciclo de vida del análisis de datos (*data analytics*), desde la definición del problema, hasta la visualización de los datos, la herramienta debe funcionar sin inconvenientes.

    Debe ser una herramienta transparente, sin importar las opciones de preparación y análisis de datos.

17. **Despliegue.**

    Se debe tener en cuenta cómo encajará el catálogo dentro de la infraestructura técnica actual y futura de la empresa. Averiguar si cuenta con opciones para realizar implementaciones locales, en la nube u ofrece una solución hibrida.

18. **Servicios que puede ofrecer.**

    Es conveniente tener en cuenta si, además de los servicios básicos, el catálogo ofrece servicios extra. Por ejemplo disponer de servicios de consultoría ya que, a la larga, tener al alcance este tipo de servicios puede facilitar la resolución de tareas difíciles.

    Es necesario ver si cuentan con una serie de tutoriales de iniciación sobre la herramienta (o similar) para los usuarios que no tengan conocimientos básicos. Especialmente en el caso de los curadores de datos, ya que son los encargados de la captación de los datos para su posterior análisis, y por tanto necesitan profundizar más en los datos.

19. **Precios.**

    Según el presupuesto que se tenga para el proyecto, se debe escoger la herramienta de catálogos de datos que mejor se adapte a la empresa y a sus necesidades.

[2]Pyme: empresa pequeña o mediana en cuanto a volumen de ingresos, valor del patrimonio y número e trabajadores.

20. **Hoja de ruta del proveedor.**

    La hoja de ruta es un manual en el que se informa sobre las intencionalidades del proveedor en la integración de futuras características y funciones en el catálogo de datos. Tener en cuenta estos planes permite ver las consecuencias si se aumenta la interoperabilidad con diversas herramientas de preparación y análisis. Además, es conveniente considerar a cuántos orígenes de datos se puede conectar actualmente y si a futuro podrá aumentarse el número, es decir, ser más escalable.

Criterios elección catálogo de datos	
1. Catalogación de conjunto de datos.	11. Linaje.
2. Catalogación de operaciones de datos.	12. Cumplimiento de las normativas.
3. Búsqueda.	13. Calidad en los datos
4. Recomendaciones.	14. Curación de datos.
5. Evaluación del conjunto de datos.	15. Socialización.
6. Acceso a los datos.	16. Integración e interoperabilidad.
7. Metadatos de uso.	17. Despliegue.
8. Valoración de datos.	18. Servicios que puede ofrecer.
9. Catálogo de metadatos.	19. Precios.
10. Seguridad.	20. Hoja de ruta del proveedor.

Tabla 5.1: Criterios elección catálogo de datos.

5.6. Beneficios

Existen múltiples ventajas de las que se puede beneficiar la empresa con incluir la herramienta de catálogo de datos en la organización, pero implementarlo por primera vez puede ser realmente costoso, especialmente en lo que a tiempo y esfuerzo se refiere, ya que clasificar y estudiar los datos se debe realizar de acuerdo con las necesidades de la empresa y este proceso se ha de desarrollar de manera exhaustiva.

Pero estas posibles dificultades o desventajas que a priori se podrían identificar, si se lleva un riguroso control en la implementación, no deberían suponer grandes consecuencias negativas. Por el contrario, la utilización de estos catálogos sí puede aportar una serie de beneficios a medio y largo plazo que contrarrestan los costos económicos, temporales y de esfuerzo.

Con el uso de catálogos las empresas pueden **ahorrar en tiempo y dinero**, ya que se gana en productividad con un mejor monitoreo de los datos y la eliminación de redundancias. Se obtiene una **mejor comprensión de los datos** con la introducción de un contexto único y mejorado. Al disponer de un mayor y mejor control sobre los datos

se **minimizan considerablemente los riesgos** para la empresa. Además, se **aumenta la seguridad** en el control de acceso a datos, haciendo visible quién, cómo y por qué un usuario hace uso de ese dato o conjunto de datos. El **punto de vista unificado** que aportan los catálogos es de ayuda a la hora de aplicar las leyes de protección de datos necesarias para su tratamiento. Se hace una **mejor gestión de los datos**, con lo que aumenta la calidad de éstos y por tanto la toma de decisiones es más precisa. Por último, debemos señalar que uno de los beneficios que más llama la atención a las empresas es que la implementación y uso de catálogos de datos son una **ventaja competitiva** con respecto a otras empresas [26] [55].

5.7. Estado del arte

En el estado del arte se estudian diferentes ejemplos de herramientas prácticas de catálogos de datos. Esta sección está dedicado a la identificación de algunas de las herramientas de catálogo de datos más conocidas y utilizadas en el ámbito empresarial, indicando sus características más relevantes.

Comenzamos destacando algunos ejemplos de catálogos que utilizan páginas web muy conocidas en Internet y que son específicos para sus objetivos: LinkedIn usa DataHub, Airbnb usa Dataportal, la plataforma Netflix tiene integrada la herramienta Metacat o la empresa inmobilaria WeWork utiliza el catálogo de datos Marquez [12].

A continuación, destacamos algunas de las herramientas de catálogos de datos más conocidas y utilizadas por empresas grandes e importantes a nivel internacional [51]:

1. **COLLIBRA CATALOG**

   Collibra Data Catalog hace uso de sus capacidades de gobernanza de datos para garantizar a sus usuarios que siempre accederán a los datos más confiables disponibles. Les permite comprender y descubrir los conjuntos de datos más importantes para sus necesidades y que generan un mayor valor a la empresa, de una manera más rápida [9].

   Como observamos en la Figura 5.2 Collibra Catalog ofrece una vista unificada de todos los datos indicando el linaje de esos datos. En el menú que aparece a la izquierda en la interfaz aparecen varias opciones que nos permite navegar por los datos, realizar distintos análisis estadísticos, y verificar si los datos cumplen las políticas de privacidad.

   Usa algoritmos de aprendizaje automático para encontrar los datos, lo que hace las búsquedas más eficientes y poderosas, y te muestra opciones de conjuntos de datos dentro de un mismo contexto que ayude al usuario a manejar los datos. Además, tiene una funcionalidad que permite comparar datos similares gracias a la puntuación de datos, ayudando al usuario a elegir el conjunto que mejor se adapte a sus necesidades. Destacar también que la herramienta cuenta con la capacidad de subir tus datos a la nube.

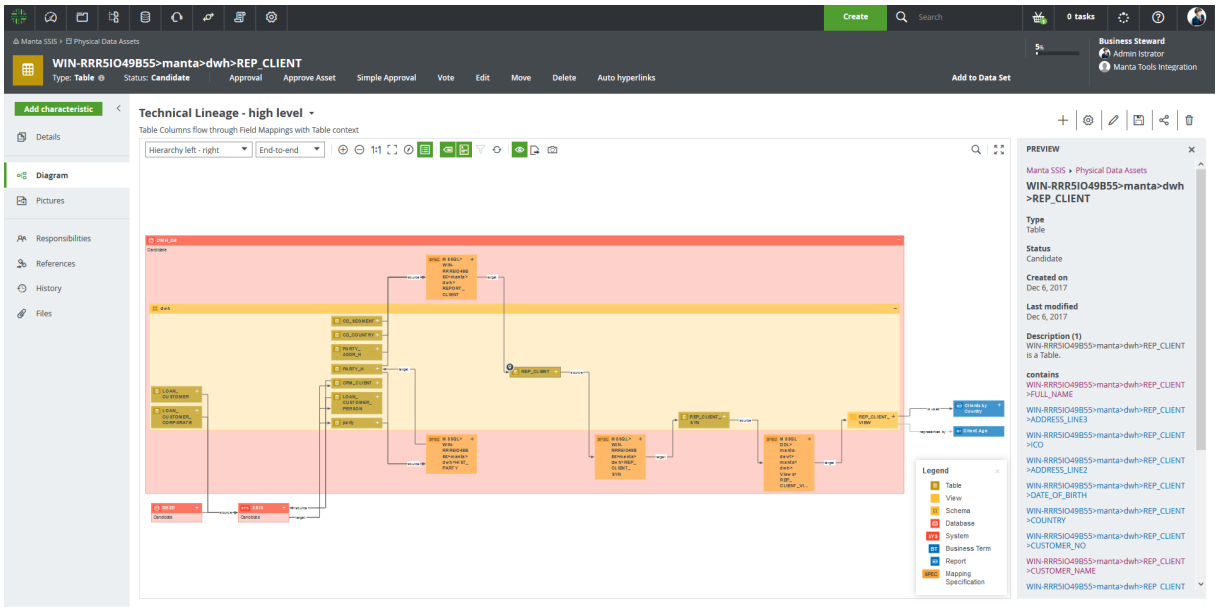

Figura 5.2: Interfaz Collibra Data Catalog

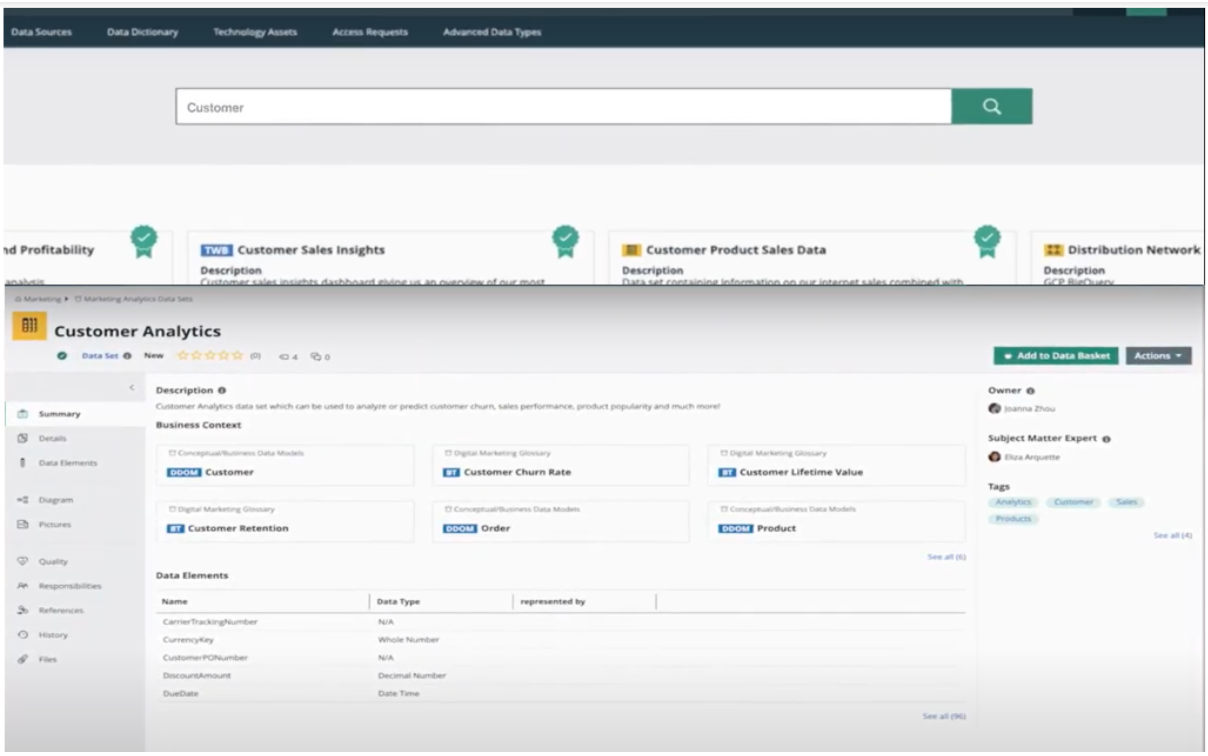

Figura 5.3: Interfaz Collibra Data Catalog - Búsqueda

2. WATSON KNOWLEDGE CATALOG

Watson Knowledge Catalog es una herramienta de catálogo de datos desarrollada

por IBM, que está integrada con una plataforma de gobernanza de datos para empresas y basada en Cloud. Permite a los usuarios, de manera más sencilla y rápida, encontrar, preparar, comprender y utilizar los datos y establece políticas de acceso a los datos garantizando que los datos sean accesibles por los usuarios correctos [28].

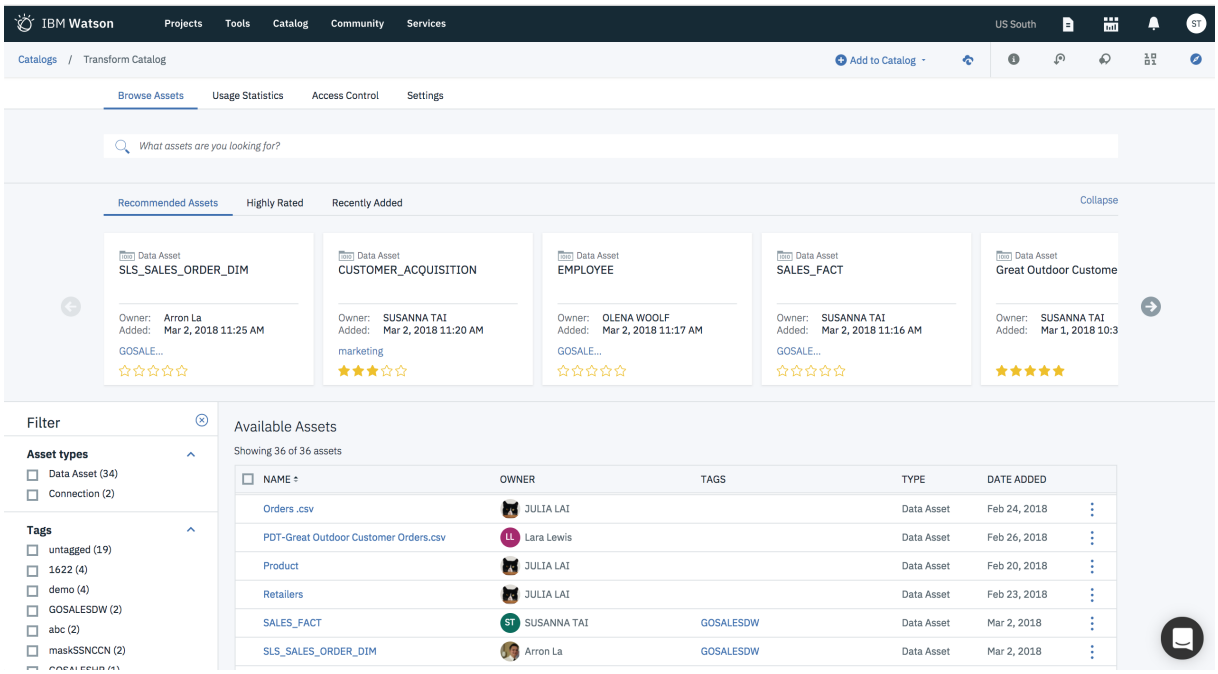

Figura 5.4: Interfaz Watson Knowledge Catalog

Se establece como una única fuente de datos para cualquier tipo de usuario, los cuales pueden acceder a los datos ellos mismos contando con datos fiables. Hace uso de la Inteligencia Artificial para una búsqueda más potente y personalizada. Además, trata datos tanto estructurados como no estructurados con Watson Natural Language Understanding[3]. Permite crear un glosario que introduzca un vocabulario común a todos los usuarios.

3. **ORACLE CLOUD INFRASTRUCTURE DATA CATALOG**

Oracle Cloud Infrastructure Data Catalog ofrece un servicio de gestión de metadatos que ayuda a los usuarios de datos a gestionar los datos y respaldar la gobernanza de datos. Diseñado por Oracle, proporciona un inventario de datos y gestión de lagos de datos y se acopla bien a los entornos de Oracle, como Oracle Cloud Infrastructure [42].

Como observamos en la Figura 5.5 la interfaz del catálogo de datos de Oracle permite realizar búsquedas a través de etiquetas o tag que son populares, es decir, nos permite encontrar datos dentro de un contexto específico en base a las búsquedas más populares realizadas por otros usuarios.

[3]`https://www.ibm.com/mx-es/cloud/watson-natural-language-understanding`

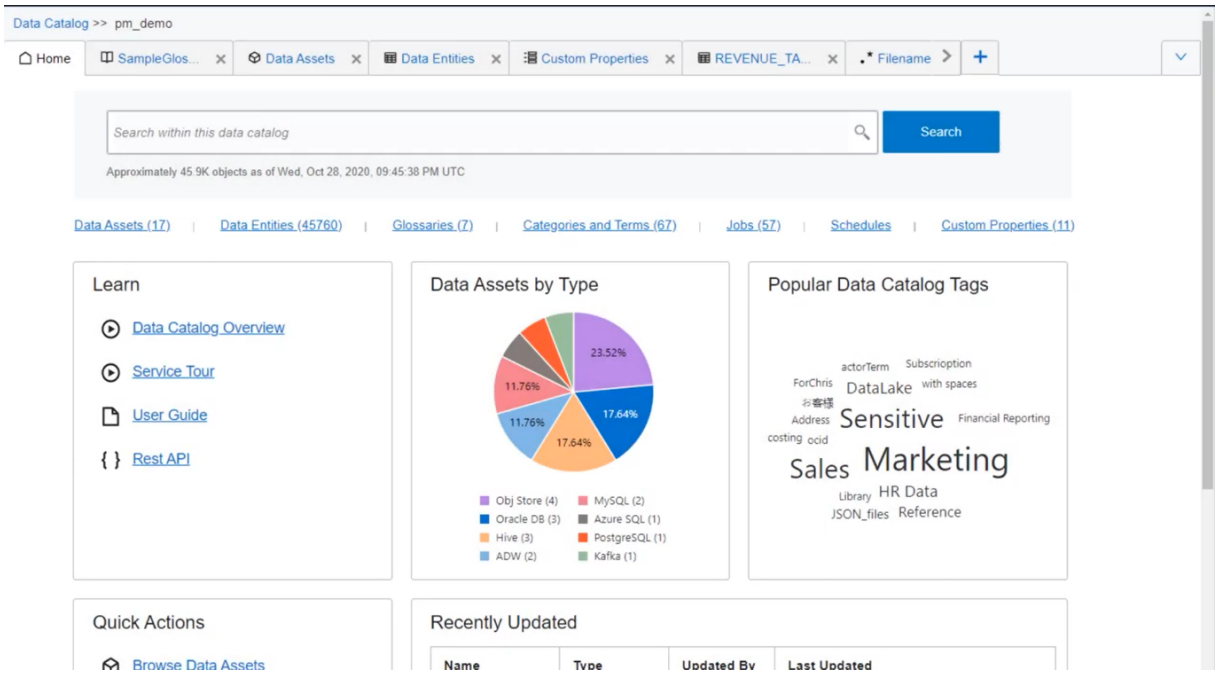

Figura 5.5: Interfaz Oracle Cloud Infrastructure Data Catalog

4. **QLIK CATALOG**

   Qlik Catalog se centra en la gestión de datos empresariales, simplificando y acelerando la forma de catalogar, gestionar, preparar y entregar sus datos a sus usuarios, de forma que sean fiables y procesables [44].

   Ofrece una único repositorio seguro de todos los datos donde permite realizar análisis y obtener información de cualquier fuente de datos empresarial de la que recopilen información. Agiliza la transformación de los datos sin procesar en activos listos para analizar, gracias a las herramientas automatizadas de preparación de datos y metadatos.

   En la Figura 5.6 podemos observar un único repositorio con los diferentes perfiles de datos. Se puede apreciar una de las funciones características de este catálogo que es enriquecer al catálogo, es decir, los usuarios pueden crear un catálogo de Smart Data donde documenten aspectos destacados de los datos como por ejemplo, diferenciar entre metadatos técnicos, administrativos (operativos) y externos (empresariales). Esto provoca que los datos sean comprendidos de forma más fácil, y trabajar con ellos sea un proceso sencillo.

5. **SAP DATA INTELLIGENCE**

   SAP Data Intelligence transforma la dispersión de los datos distribuidos en información vital y estratégica para las empresas. Utiliza inteligencia de datos para procesar estos datos distribuidos [10].

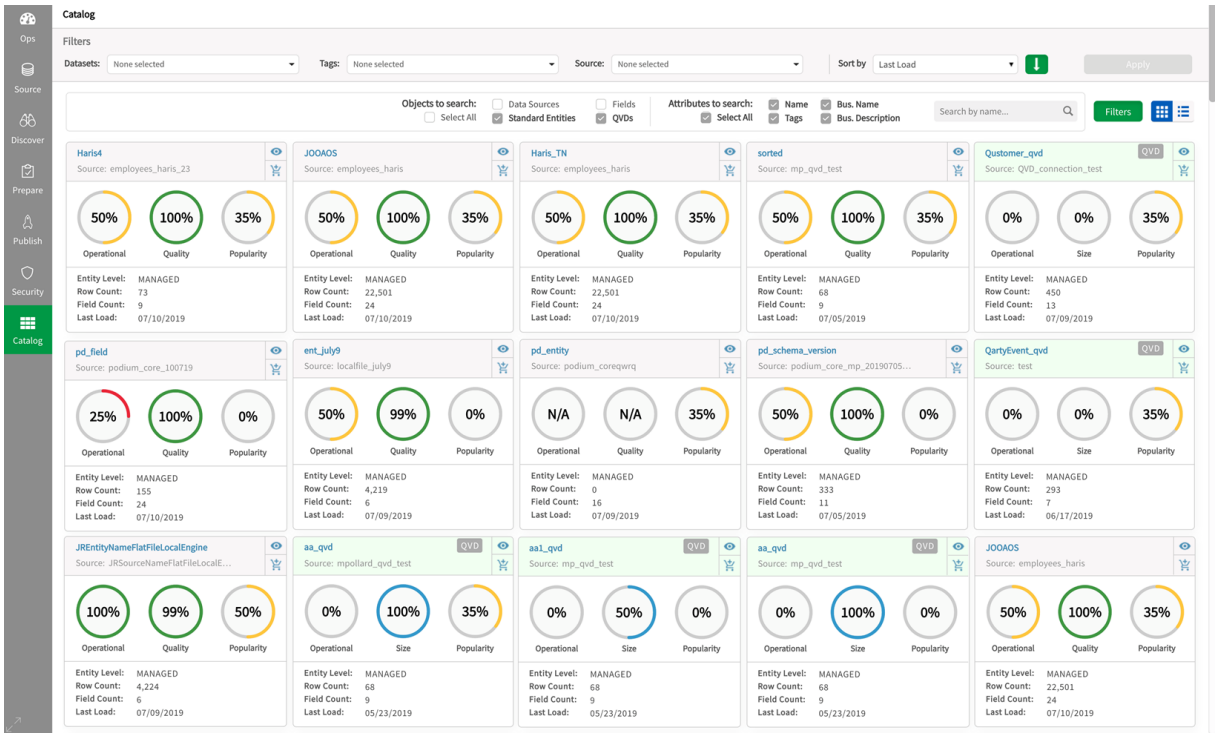

Figura 5.6: Interfaz Qlik Catalog

Los usuarios tendrán oportunidad de dotar a los datos de una mayor información creando etiquetas de metadatos, estableciendo clasificaciones o generando comentarios sobre los datos. Esta herramienta trabaja con volúmenes masivos de datos y permite trabajar desde cualquier tipo de dispositivo. Además, hace uso del aprendizaje automático para impulsar la búsqueda.

6. **CKAN**

   CKAN es un software de código abierto, gratuito y muy flexible, que se caracteriza por sus funciones como recolección de datos, búsqueda por facetas e interfaces para datos y metadatos. Los datos públicos pueden compartirse entre diferentes sitios dentro de CKAN [8].

   Es un catálogo valioso ya que proporciona colecciones de datos y metadatos, facilitando la búsqueda y ofreciendo una visión simplificada de los datos. En la Figura 5.8 podemos observar un conjunto de datos donde podemos explorar esos datos, viendo como se comportan.

7. **TALEND OPEN STUDIO FOR DATA QUALITY**

   Se trata de una herramienta de catálogos de datos creada por la empresa Talend, cuya principal función es la generación de perfiles de datos para identificar distintos problemas que puedan existir antes de la iniciación de un proyecto con un uso masivo de datos. Otras de sus funciones son: (1) Analizar distintas fuentes de datos

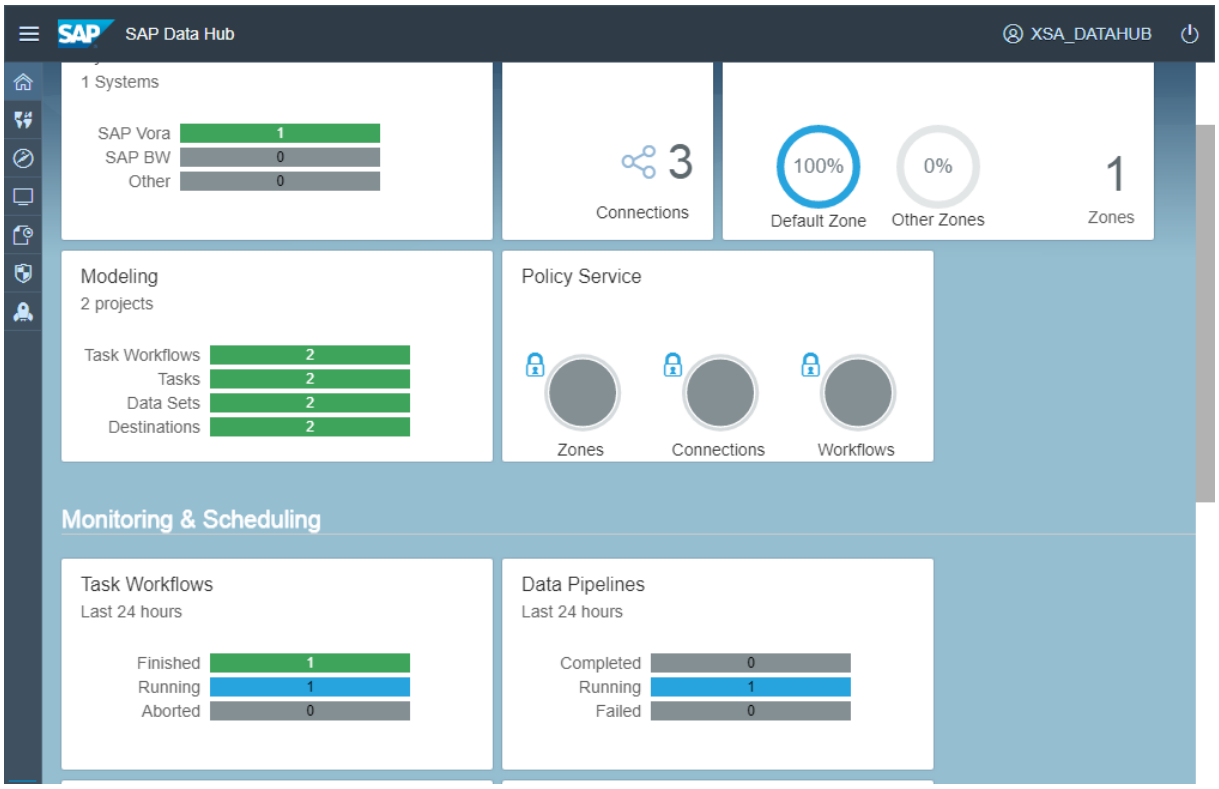

Figura 5.7: Interfaz SAP Data Intelligence

para captar los metadatos que contienen los datos y guardarlos en un repositorio de metadatos, que pueden ser utilizados como indicadores y métricas para los distintos análisis posteriores; (2) Mostrar resultados comparando los distintos datos y operaciones relacionadas con ellos, a través del estudio de patrones, cuya implementación da lugar a distintos indicadores; (3) La integración de una arquitectura funcional, un modelo arquitectónico que identifica las funciones del estudio, las interacciones y las correspondientes necesidades de TI (Tecnologías de la Información).

Gracias a Talend Studio podemos realizar una serie de análisis de perfiles de datos, a un conjunto de datos de gran volumen, que nos permita descubrir anomalías o problemas en los distintos datos, antes de exportar esos datos a la realidad, es decir, antes de que tengan una gran repercusión en el proyecto donde vayan a ser utilizados. Además, nos permite obtener una visión más amplia de todos los datos de una forma rápida y sencilla.

En la Tabla 5.9, podemos observar una breve comparación de las seis de las herramientas de catálogos explicadas con anterioridad.

Para el supuesto práctico hemos utilizado una herramienta de software libre Talend Open Studio for Data Quality [66], en su última versión 7.3.1, que analizaremos con más detalle en el Capítulo 6 donde también conoceremos su aplicación práctica. Hemos optado por esta herramienta principalmente porque al ser de software libre es una herramienta

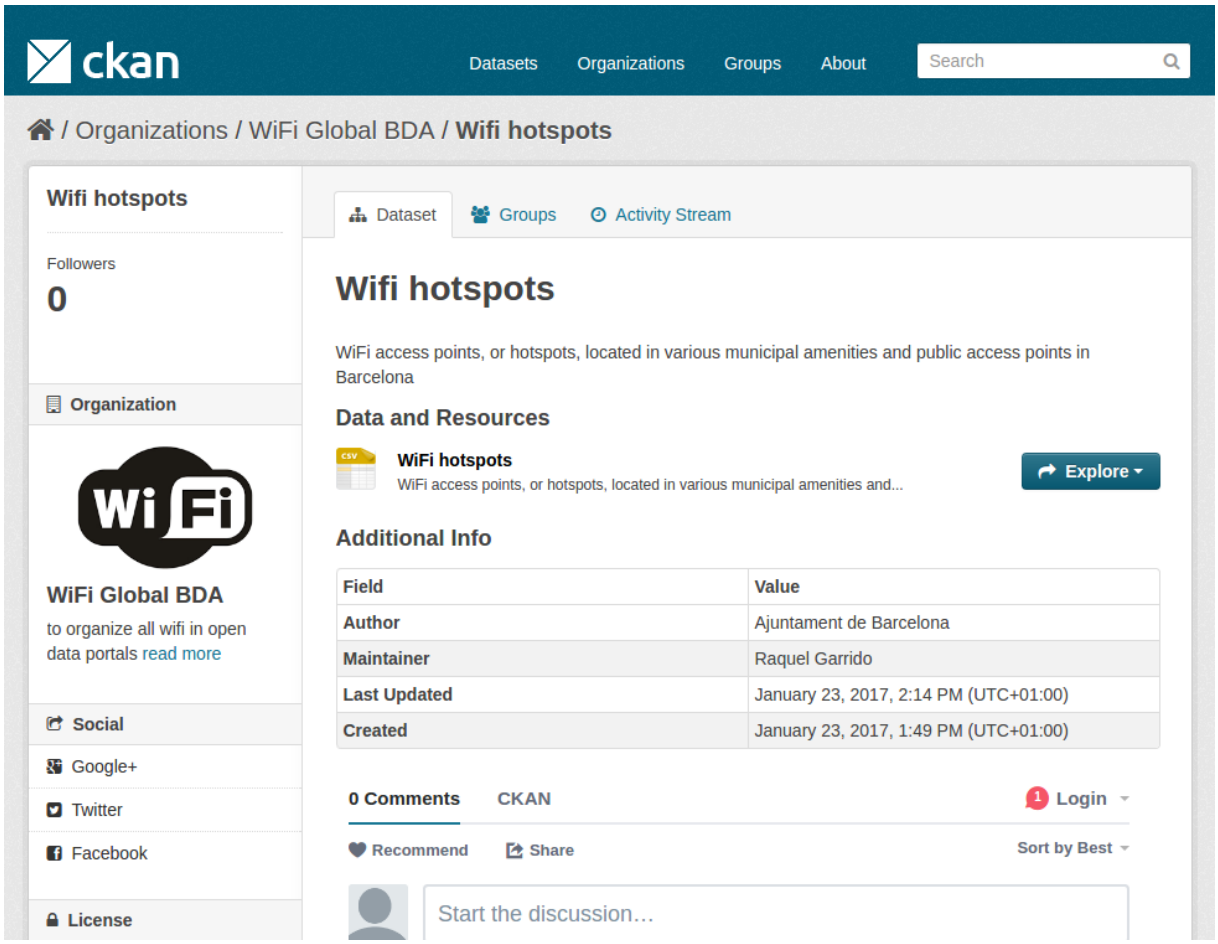

Figura 5.8: Interfaz CKAN

gratuita, y además nos permitía ilustrar de forma práctica todos los conceptos teóricos explicados en el trabajo.

Herramienta	Empresa	Metadatos	¿Gratuito?	Perfiles de Datos	Características
Collibra Catalog	Collibra	Sí	Pago	Si	• Repositorio de búsqueda, con el que comprender cómo y dónde se almacenan los datos y cómo utilizarlos. • Integra Gobernanza de datos. • Basado en la nube. • Pensada datos empresariales.
Watson Knowledge Catalog	IBM	Sí	Pago	No	• Descubrimiento de datos asistido por IA y modelos de aprendizaje automático. • Virtualización de datos en tiempo real.
Oracle Cloud Infrastructure Data Catalog	Oracle	Sí	Pago	No	• Diseñado específicamente para funcionar con el ecosistema de Oracle. • Glosario empresarial. • Especializado en lagos de datos.
Qlik Catalog (Qlik Data Catalyst)	Qlik	Sí	Pago	No	• Automatización ágil de Data Warehouse. • Crea lagos de datos. • Preparación de datos automatizados sin estar procesados.
SAP Data Intelligence	SAP	Sí	Pago	Sí	• Gestión de datos impulsada por IA. • Permite descubir datos de forma natural y aquellos que provienen de dispositivos basados em IoT.
CKAN	CKAN	Sí	Gratuita	No	• Software de código abierto, gratuito y muy flexible. • Funciones de recolección de datos, búsqueda por facetas e interfaces para datos y metadatos. • Permite conpartir sus datos públicos con otros sitios de ckan.
Talend Open Studio for Data Quality	talend	Sí	Gratuita	Sí	• Herramienta de gestión y desarrollo unificadas para integrar y procesar todos los datos dentro de un entorno visual fácil de usar. • Acceder y examinar los datos desde diferentes fuentes de datos. • Recopilar estadísticas e información sobre los datos.

Figura 5.9: Resumen de herramientas de catálogos de datos

Capítulo 6

Parte práctica

A lo largo de este trabajo hemos estado viendo distintos conceptos referentes a la calidad de datos, y cómo implementar, de una forma general, esa calidad dentro de las empresas. Nosotros como analistas de calidad de datos nos encargamos de ese proceso de implantación. En este capítulo nos apoyaremos de la herramienta de catálogo de datos **Talend Open Studio for Data Quality** [1], descrita en la sección 5.7, para mostrar el proceso de implantación desde un punto de vista práctico.

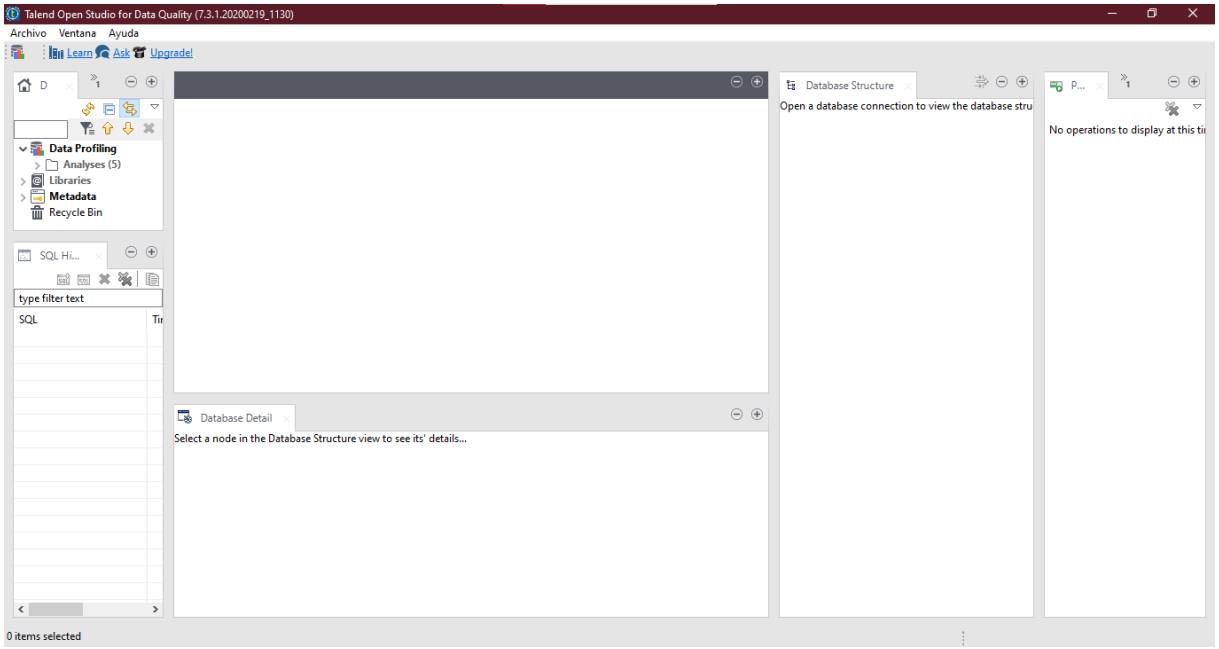

Figura 6.1: Interfaz Talend Studio

Partiendo del proyecto definido en la sección 1.2, el proceso de implantación de calidad de datos dentro de una empresa se forma de varios pasos, cuyo orden se establece de igual forma que la distribución de capítulos de este proyecto:

[1] `https://www.talend.com/es/products/data-quality/data-quality-open-studio/`

1. **Definir las dimensiones** con las que se va a trabajar en este proyecto. En la sección 3.3, ya expusimos las dimensiones que abordaremos y son: *2.Fundamentos de la integridad*, *3. Duplicación*, *4. Exactitud / Precisión*, *7. Facilidad de uso y mantenimiento* y *9. Calidad de presentación.*

2. **Proceso de gestión de la calidad: *Data Preparation***. Este ciclo nos permite aplicar calidad a los datos y tener un proceso de mejora constante, en la sección 3.4 definimos este paso.

3. **Identificar y definir los metadatos**. Conocer, en mayor medida, los distintos metadatos que contienen nuestros conjuntos de datos nos permite conocer en mayor grado nuestros datos, por lo que nos ayuda en su comprensión y nos permite aplicar calidad con mayor precisión. En las secciones 4.1.1 y 4.1.3, definimos los tipos de metadatos que podemos encontrar en nuestros conjuntos de datos y un ejemplo del ciclo de vida de un de ellos.

4. **Uso de catálogo de datos**. Los catálogos de datos como hemos visto, gracias a los perfíles de datos y los metadatos nos ayuda a organizar la información y tener una visión general de los datos. En este Capítulo nos centraremos en este paso.

Contamos con cuatro conjuntos de datos, los cuales para poder ser utilizados por Talend Studio, primero debemos conectarnos a una fuente de datos externa. En nuestro caso hemos utilizado como fuente de datos MariaDB junto a la herramienta de software libre para la gestión de bases de datos, phpMyAdmin. Los conjuntos de datos se podrían añadir directamente desde la herramienta, pero por familiarización con phpMyAdmin, escogimos ésta manera de crear la conexión con los datos.

Creamos una BBDD en phpMyAdmin, *data_ tfg*, donde definimos 4 tablas que contienen nuestros conjuntos de datos, siguiendo la estructura definida para cada conjunto en la sección 1.2. Estos conjuntos son datos reales sobre datos maestros de Clientes, Proveedores, Productos y Cuentas. Para estos datos la empresa establece una serie de reglas o requisitos que deben cumplirse para contar con datos verídicos y de calidad.

Algunos de estos requisitos son:

- El campo 'Nombre' en todos los conjuntos de datos no pueden estar vacío.
- No puede haber ningún registro duplicado en ningún conjunto de datos.
- No se permiten registros con el mismo nombre, aunque su identificador sea distinto.
- Los valores nulos o que no estén definidos no pueden existir dentro del conjunto.

Una vez creada nuestra fuente externa debemos realizar la conexión entre la fuente de datos y la herramienta Talend Studio. Con esta conexión podremos acceder a los diferentes conjuntos de datos y podremos ejecutar sobre ellos distintos análisis estadísticos sobre su calidad, lo que nos permitirá realizar un estudio que ayude a eliminar o corregir las posibles anomalías que puedan presentar.

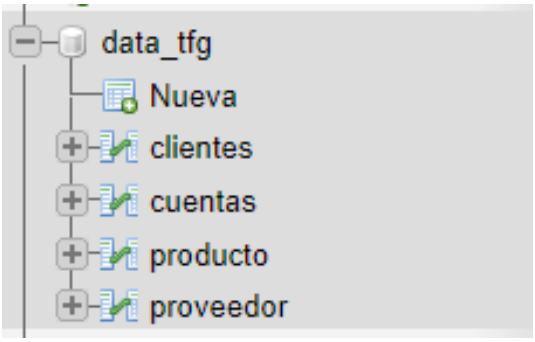

Figura 6.2: Creación de BBDD

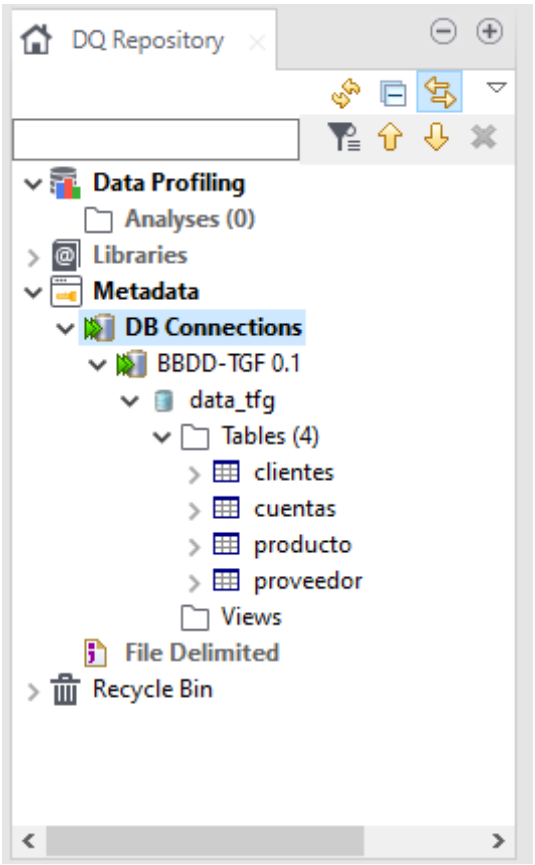

Figura 6.3: Conexión creada en Talend Studio

Ya tenemos preparado todo para poder realizar distintos tipos de análisis y hacer un estudio exhaustivo de los datos. Son lo que hemos llamado perfiles de datos, donde a partir de un conjunto de datos se realizan procesos de análisis que nos permita verificar que se cumplen los requisitos impuestos por la empresa. Esta herramienta de catálogos permite realizar el proceso de perfiles de datos (sección 4.2.2) de una forma sencilla y visual, gracias a los distintos análisis que puede realizar.

Esta herramienta distingue cinco grandes grupos de análisis:

1. **Análisis estructural.** Ofrece una visión general del contenido de la Base de Datos y calcula el número de tablas y filas por cada tabla que puede contener cada catálogo. Además, te permite saber el número de índices, claves primarias, etc., que tienes.

   Análisis que se pueden realizar dentro de este grupo:

   - Análisis de la Visión General de las Conexiones.
   - Análisis de la Vista General del Catálogo.
   - Análisis General de Esquemas.

2. **Análisis de tablas cruzadas.** Examina varias tablas, y te permite examinar la relación que existe entre dos tablas y descubrir a su vez las claves foráneas que tengan. En nuestro ejemplo, este tipo de análisis no lo podemos dar porque no tenemos tablas relacionadas que nos puedan servir de ejemplo.

   Análisis que se pueden realizar dentro de este grupo:

   - Análisis de Redundancia.

3. **Análisis de tabla.** Permite realizar análisis a nivel de registros, identifica filas idénticas o similares. Puede tratar a toda la fila como un elemento indivisible o elegir un conjunto de celdas para analizarlas. Permite aplicar reglas de calidad de datos creadas previamente por el usuario.

   Análisis que se pueden realizar dentro de este grupo:

   - Análisis de Reglas de Negocio.
   - Análisis de Coincidencias.
   - Análisis de Dependencia Funcional.
   - Análisis de Conjuntos de Columnas.

4. **Análisis de Columna.** Realiza el análisis en base a las columnas por las que se compone la tabla.

   Análisis que se pueden realizar dentro de este grupo:

   - Análisis Básico de Columnas.
   - Análisis de Valores Nominales.
   - Análisis de la Frecuencia del Patrón.
   - Análisis de Datos Discretos.
   - Análisis de Estadísticas de Resumen.

5. **Análisis de correlación.** Llamados también análisis exploratorios. Exploran la relación entre las columnas y nos ayuda en la identificación de problemas de calidad de datos.

   Análisis que se pueden realizar dentro de este grupo:

   - Análisis de Correlación Numérica.
   - Análisis de Correlación Temporal.
   - Análisis de Correlación Nominal.

A continuación (Figuras 6.4 a 6.9) mostraremos una serie de ejemplos de algunos de los análisis mencionados anteriormente. Todos los análisis vienen acompañados de una tabla resumen con los datos estadísticos obtenidos y un gráfico que ilustra esos datos, además te permite sacar los registros de cada estadística para comprobar cuales se ven afectados.

En nuestro caso el primer grupo de análisis no importa que tipo escoger, ya que al no tener más que una conexión y un catálogo nos saldría el mismo resultado en cualquiera de ellos. Como observamos en la Figura 6.4 tenemos un análisis completo de lo que contiene nuestra base de datos, los cuatro conjuntos de datos siendo las tablas, un total de 37.837 registros, 4 claves primarias una por cada tabla, etc.

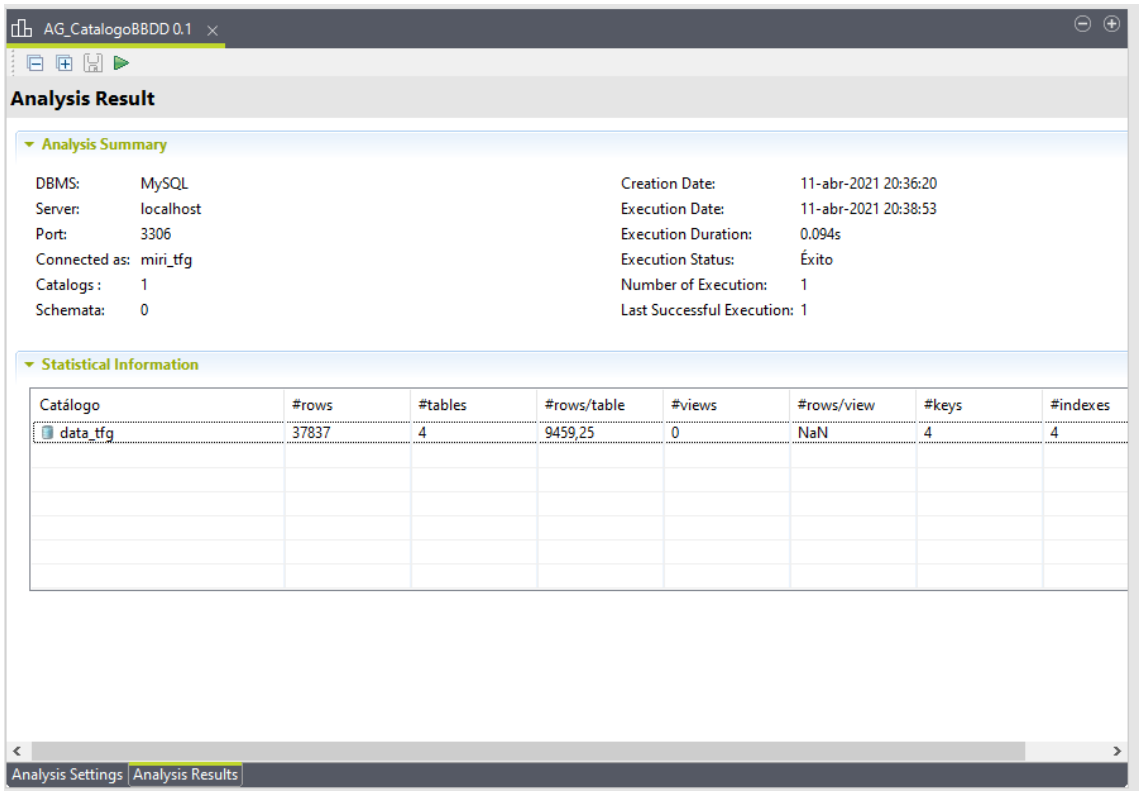

Figura 6.4: Análisis de la Vista General del Catálogo

El segundo grupo de análisis como hemos mencionado con nuestro ejemplo no podemos abordarlo ya que carecemos de tablas relacionadas que nos permita mostrarlo.

Con respecto al tercer grupo de análisis, vamos a realizar un análisis de Coincidencias, lo que nos permitirá evaluar el número de duplicados en sus datos, es decir, saber el número de grupos de datos similares en una tabla o un conjunto de columnas. La Figura 6.5 nos muestra un análisis de coincidencias, dentro de la tabla Clientes, entre sus registros. En los resultados podemos observar que existe un 12,24 % de los registros que están duplicados (*Matched Records*), y sabemos que hay que eliminarlos, ya que es uno de los requisitos que nos menciona la empresa. En la Figura 6.6 podemos observar un ejemplo de uno de los registros duplicados.

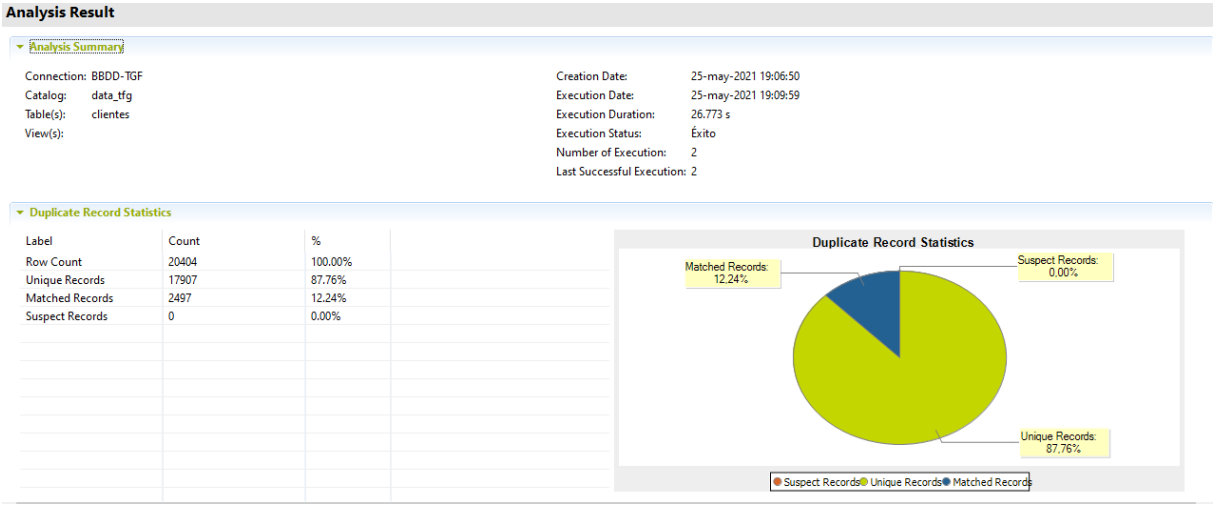

Figura 6.5: Análisis de Coincidencias

Figura 6.6: Análisis de Coincidencias

Dentro de este grupo también podemos realizar un análisis de conjuntos de columnas, que nos permitirá conocer el total de filas que tiene, cuantos valores de un conjunto de columnas son distintos, únicos o están duplicados. No son análisis a nivel de una única columna como es el caso del cuarto grupo de análisis. Para este análisis hemos utilizado el

conjunto de datos Cuentas, como observamos en la Figura 6.7 no existen registros que sean totalmente iguales en todos sus campos, esto es porque en el análisis cogemos también como columna a analizar 'Cuenta'que corresponde con la clave primaria del conjunto la cuál debe ser única para cada conjunto. Sin embargo, si nos fijamos en la Figura 6.8,donde en el análisis no incluimos la clave primaria, observamos como dentro del conjunto existen una serie de registros, un total de 2.668, que están duplicados.

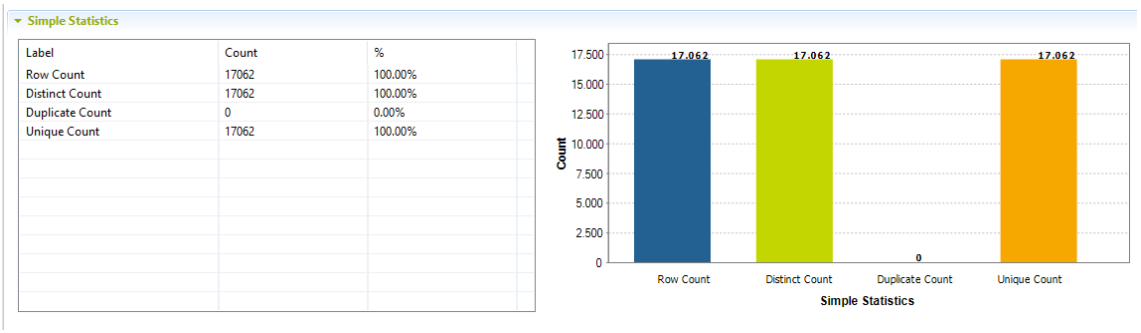

Figura 6.7: Análisis de Conjunto de Columnas

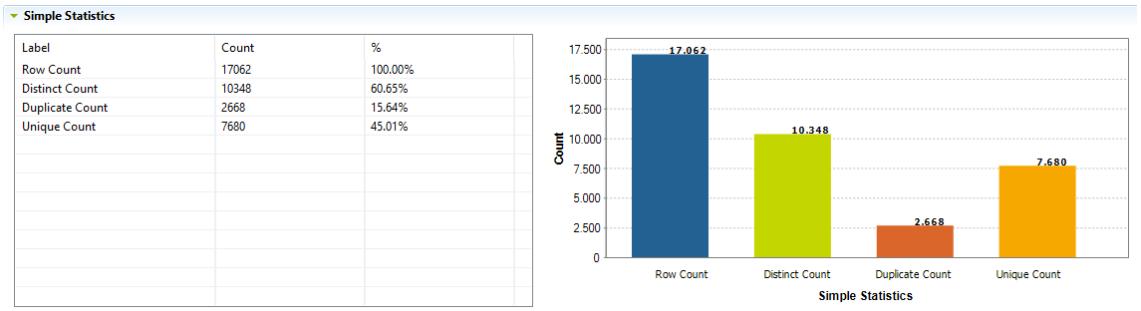

Figura 6.8: Análisis de Conjunto de Columnas 2

El cuarto grupo de análisis es uno de los más típicos y usados, ya que con ellos se examinan los datos individuales dentro de una columna, pudiendo comprobar si existen campos en blanco, nulos o existe algún patrón dentro del campo. En este grupo se identifican los tipos de análisis que se mencionan en la sección 4.2.4, pudiendo realizar un examen de cada uno de los tipos.

Para este análisis hemos utilizado la tabla Cuentas como referencia, en la Figura 6.9 vemos el análisis básico de columnas que, a diferencia del análisis anterior de conjunto de columnas realiza un análisis por cada columna individual, en concreto de las columnas 'Cuenta' y 'Valor' con algún valor nulo (*Null Count*), duplicado (*Duplicate Count*), en blanco (*Blank Count*) o único (*Unique Count*) que encuentre. La columna 'Cuenta' al ser clave primaria, como corresponde, no contiene ningún valor en blanco, nulo o duplicado, mientras que la columna 'Valor' si contiene valores duplicados (18,10 %). Por último, se muestra un gráfico de patrones de frecuencia, donde te muestra en la columna 'Valor', el número de registros que contienen un patrón en sus valores (este gráfico, en nuestro

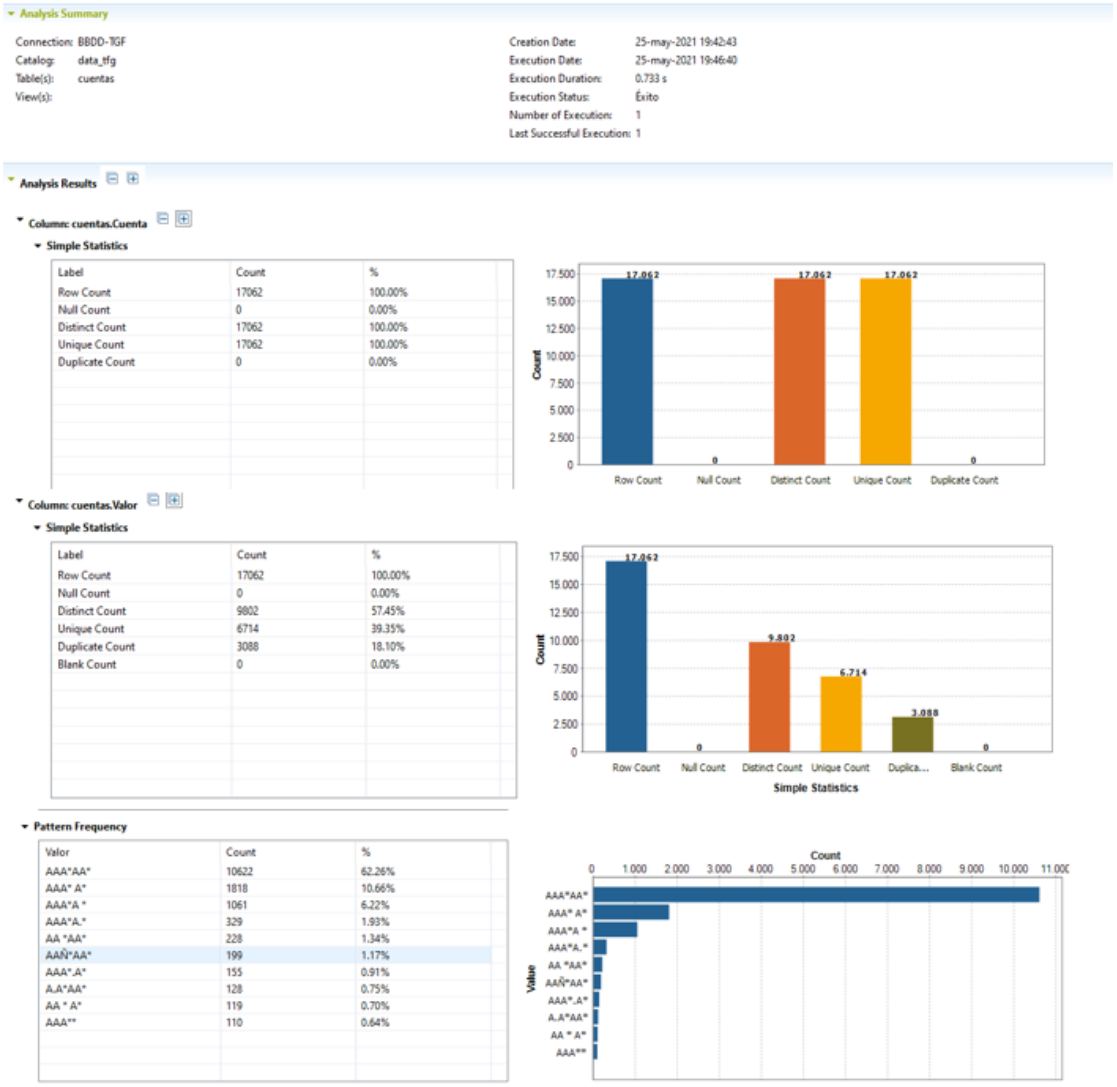

Figura 6.9: Análisis Básico de Columnas

gráfico, no es acertado ya que los nombres han sido anonimizados por lo que existen más patrones de los reales).

Otro análisis que podemos realizar es el Análisis de la Frecuencia del Patrón, dónde podemos identificar con que regularidad se repite un patrón en este caso en la columna Nombre del conjunto de datos proveedor. En nuestro ejemplo, nos salen unos resultados alejados de la realidad, ya que al estar los datos anonimizados se encuentran muchos más patrones. Un patrón sería AAA***AAA***AAA, dónde la A corresponde a una letra cualquiera, es decir, todos los registros que contengas tres letras, seguida de tres *, se considera un patrón. En la Figura 6.10 podemos observar dos tipos de análisis uno nos muestra la frecuencia general de todos los patrones que ha encontrado, junto al número de registros afectados y en detalle cuales son esos registros afectados (para el patrón

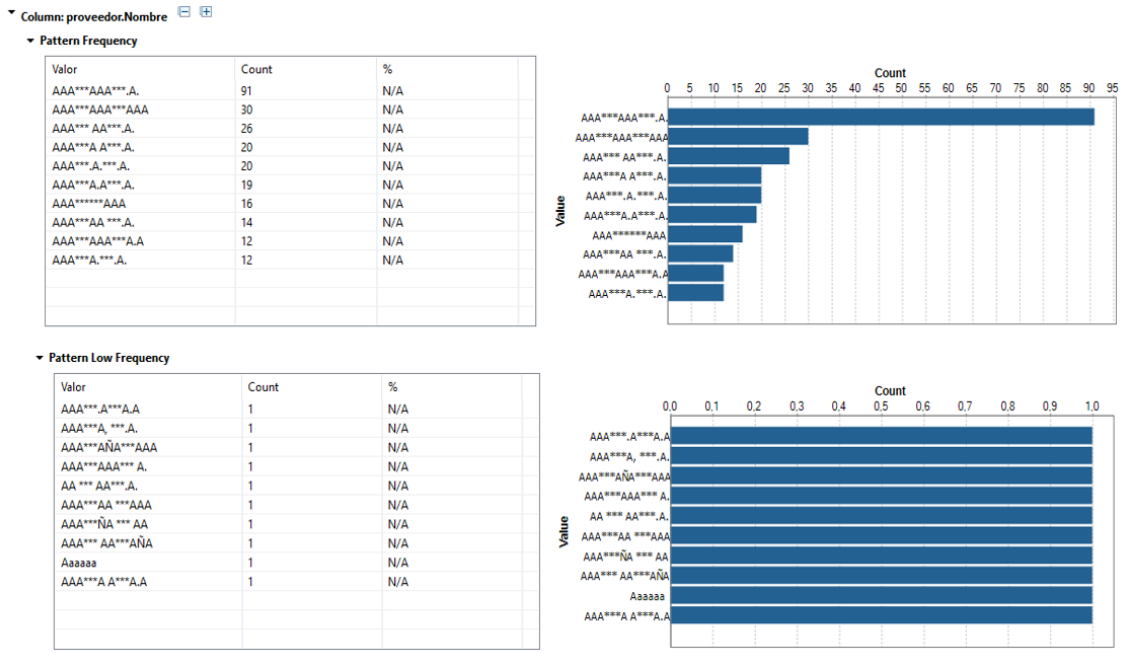

Figura 6.10: Análisis de la Frecuencia del Patrón

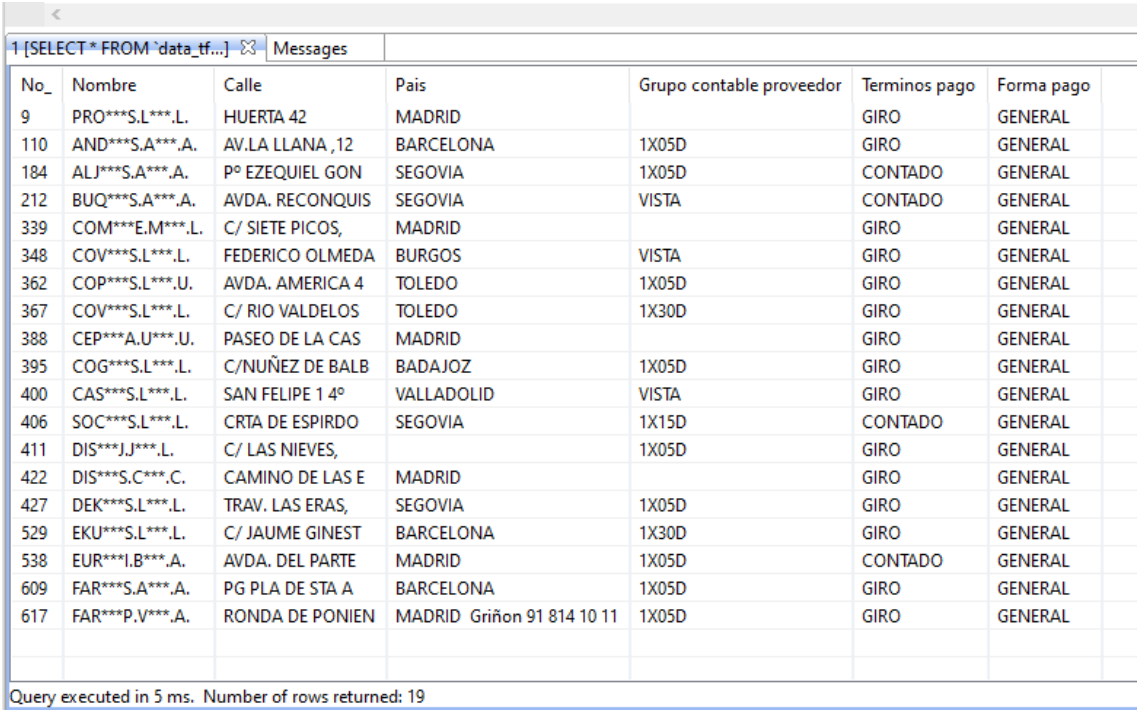

Figura 6.11: Análisis de la Frecuencia del Patrón Resultados

AAA***A.A***.A. ver Figura 6.11) y un análisis con los patrones de menor frecuencia, en este caso sólo se repiten una vez.

En el último grupo de análisis no hemos podido realizar ningún análisis ya que los datos que teníamos no son suficiente ya que no contamos con datos nominales que nos ayuden a mostrar resultados de este análisis. Con este análisis podríamos sacar gráficos en burbuja, para observar valores extremos, un diagrama parecido a un Gantt, para representar valores mínimos y máximos o un gráfico de red, para destacar las relaciones débiles entre los pares de datos, relacionando datos nominales con datos númericos, datos nominales con datos temporales o únicamente analizando datos nóminales, respectivamente.

Talend Studio nos permite realizar muchos tipos de análisis dependiendo de las necesidades que se requieran. Estos análisis nos permiten saber realmente con qué datos tratamos y nos ayuda a prevenir situaciones problemáticas provenientes de datos erróneos, lo que nos garantiza, que si tratamos los datos anómalos que nos muestran los análisis, trataremos con información verídica y libre de errores.

A través de este caso práctico hemos podido plasmar y analizar de forma práctica el contenido teórico descrito en el documento, lo más completo posible de acuerdo a los datos que nos facilitó la empresa. Los datos facilitados son suficientes para ilustrar y realizar el TFG de manera que la teoría pueda ser comprendida en mayor medida y nos sirve de inicio para realizar análisis de este tipo. Sin embargo, si esto se fuera a implementar para una empresa en un proyecto real necesitaríamos un mayor volumen de datos, mayor variabilidad, y unas reglas de negocio con respecto a los datos más descriptivas.

Capítulo 7

Conclusiones y trabajo futuro

En este último capítulo reflexionaremos sobre el trabajo realizado, exponiendo una serie de conclusiones donde se reflejará si los objetivos planteados han sido cumplidos y una breve reflexión personal.

7.1. Conclusiones

- El primer objetivo planteado en el proyecto es 'comprender qué es la calidad de datos y sus conceptos asociado' que ha sido abordado a lo largo de los capítulos: 3. Calidad de datos y 4. Los metadatos y los perfiles de datos. Estos capítulos contemplan de manera detallada estos conceptos.
- En el Capítulo 5. Catálogo de datos y en el 6. Ejemplo Práctico, conocemos y comprendemos las herramientas de catálogo de datos y cómo se utilizan para la gestión de la calidad de los datos, verificando así el segundo objetivo planteado en el proyecto 'Analizar herramientas de catálogos de datos para la gestión de la calidad de los datos'.
- El último objetivo 'Analizar la calidad de los datos y su repercusión en los entornos empresariales, se demuestra a lo largo de todo el trabajo y gracias a la herramienta de catálogo de datos, Talend Studio, hemos podido aplicar de forma práctica los conceptos teóricos que hemos ido desarrollando.

Al cumplir los tres objetivos específicos del proyecto, podemos decir que el objetivo general 'Demostrar los beneficios que trae consigo la Gestión de la calidad de los datos en las empresas' queda demostrado y concretamente lo podemos verificar gracias al Capítulo 6, donde aplicando calidad de datos a un conjunto de datos se reduce el número de anomalías, obteniendo mejores resultados y verificando que los datos son más eficaces, eficientes y reales.

Como ejemplo de lo que estamos mencionando, ponemos el conjunto de datos de Cuentas, qué inicialmente contenía un total de 17.086 registros. Utilizando el catálogo de datos definido comprobamos que muchas de las cuentas estaban duplicadas (ver Figura 6.9),

dando lugar a datos redundantes lo que ofrece información incorrecta. Con este estudio, detectamos las cuentas duplicadas y aplicamos calidad de datos sobre el conjunto, como consecuencia se redujeron los registros a un total 1.632 cuentas. Por lo que podemos comprobar que aplicar calidad a los datos resulta en conjuntos de datos más manejables y verídicos.

Como conclusión final del proyecto, podemos afirmar que implementar calidad en los datos dentro de una empresa es vital, ya que esta es la que determina si esos conjuntos de datos afectan de manera positiva o negativa a los objetivos de la empresa. Llevar una mala gestión de los datos, por tanto, puede provocar en grandes repercusiones económicas para la empresa e incluso causar una insatisfacción por parte de los clientes que conlleve una desventaja competitiva frente a otras empresas del mismo sector.

A modo de reflexión final, realizar este trabajo ha supuesto grandes beneficios personales a nivel profesional y personal. La calidad de datos es un tema muy frecuente e importante en cualquier tipo de empresa, sobre todo si se centran en datos, por lo que conocer este amplio concepto y a tanta profundidad me ha preparado para poder desempeñar tareas relacionadas a este ámbito. El realizar este trabajo también me ha llevado dificultades, ya que es un tema que durante la carrera no había abordado y tuve que hacer un esfuerzo extra en la búsqueda de bibliografía, sin embargo, me ha dado la oportunidad de conocer distintas empresas orientadas a datos, que en un futuro me resulten de ayuda para futuros proyectos.

7.2. Trabajo futuro

Este trabajo está centrado en una parte más teórica sobre calidad de datos y todos los conceptos asociados a ella, por lo que en un futuro se podría coger este proyecto y llevarlo a las empresas para implementarlo de forma práctica y real, como una guía para aplicar calidad de los datos en esas empresas desde el inicio.

Nos sirve como punto de inicio para aquellas empresas que no han implantado Calidad a sus Datos y quieran tratar con datos más óptimos y sacar beneficio de su utilización, ya que en este trabajo se recoge información que permite comprender y aclarar el proceso de gestión de la calidad de datos y ofrece herramientas para conseguirlo.

A nivel personal, este proyecto me abre las puertas a puestos de trabajo relacionados con datos, ya que he asimilado con bastante soltura estos conceptos y tengo una visión más acertada de la importancia de los datos y que éstos sean de calidad para sacar el mayor rendimiento de una empresa.

Glosario

Big Data Análisis masivo de datos. Una cuantía de datos, tan sumamente grande, que las aplicaciones de software de procesamiento de datos que tradicionalmente se venían usando no son capaces de capturar, tratar y poner en valor en un tiempo razonable. [70]. 35

Datos ad-hoc Término procedente del latín que significa "a propósito". Es uno de los principales tipos de investigación de mercado. Se caracteriza porque se recopilan los datos de la investigación en un período determinado, en función de las exigencias del cliente y de acuerdo al propósito de ésta. Con el ad hoc se buscan datos específicos sobre el objeto de estudio, es decir, no se utilizan datos ya existentes. Se busca responder a una necesidad específica con el objetivo de mejorar el producto o servicio en cuestión. [3]. 57

Datos Maestros Conjunto de información correspondiente a entidades que no se modifican una vez que las transacciones comerciales se han completado. [48] . 33

ETL ETL es un tipo de integración de datos que hace referencia a los tres pasos (extraer, transformar, cargar) que se utilizan para mezclar datos de múltiples fuentes. Se utiliza a menudo para construir un almacén de datos. Durante este proceso, los datos se toman (extraen) de un sistema de origen, se convierten (transforman) en un formato que se puede almacenar y se almacenan (cargan) en un data warehouse u otro sistema. Extraer, cargar, transformar (ELT) es un enfoque alterno pero relacionado diseñado para canalizar el procesamiento a la base de datos para mejorar el desempeño.[24]. 49

Gestión de Datos Maestros o *Master Data Management* Conjunto de metodologías, herramientas y procesos, necesarios para crear y mantener conjuntos precisos y consistentes de datos maestros. Se logra identificar la información más relevante de una empresa, creando una única fuente de datos, con lo que se consigue una mejora en los procesos empresariales. . 33

Gobernanza de los Datos o *Data Governance* Conjunto de procesos, funciones, políticas, normas y mediciones que garantizan el uso eficaz y eficiente de la información

con el fin de ayudar a cumplir los objetivos de la empresa. Define quién puede realizar las acciones, sobre qué datos, en qué situaciones y mediante qué métodos. [71]. 56

Historias de usuario Las historias de usuario son una herramienta con la que podremos tratar cualquier aspecto necesario para la creación de productos (utilizada especialmente el desarrollo de software). Son, esencialmente, pequeñas descripciones de los requerimientos de un cliente. [27]. 14

Lago de datos o *Data Lake* Entorno de datos compartido en su formato original (datos en bruto) que comprende múltiples repositorios, que aprovecha el poder de la gran tecnología de datos y lo combina con la agilidad del autoservicio. Dicho de otro modo, un lago de datos es un repositorio donde se almacenan grandes volúmenes de datos en bruto para utilizarlos en un futuro. Además, si se hace un buen uso de ello queda centralizado y ofrece a los usuarios transparencia a la hora de manejar los datos. [25]. 54

Linaje de los datos o *Data Lineage* Describe las fases que sufre cada dato, desde su origen, pasando por los movimientos, características y transformaciones que sufren a lo largo del tiempo. No solo esto, sino que además para entender el concepto en su totalidad se deberá añadir una serie de aspectos adicionales: saber quién usa cada dato, qué significa, cuándo se accede a la información, por qué se almacenan los datos y cómo se relacionan los elementos de los datos. [23]. 25

Pantano de datos Lagos de datos desestructurados, es decir, son repositorios que han llegado a ser lagos de datos por su gran volumen de datos pero que no ha seguido una buena gestión y mantenimiento de los datos, por lo que se dificultan el uso y recuperación de datos de una manera eficiente hasta el punto de hacerse imposibles de realizar. Algunos de los motivos por los que pueden surgir los pantanos son:Falta de metadatos, Recopilación de datos irrelevantes, Inexistencia de gobernanza de datos, que defina como tratar a los mismos, quién debe manejarles o donde se organizan, Falta de procesos automatizados o No tener una buena estrategia de limpieza de datos. [52] . 40

Bibliografía

[1] *¿Puede ser Agile la Docencia Universitaria? (UVagile).* http://uvadoc.uva.es/handle/10324/37210. (Visitado 12-04-2021).

[2] *20 Criteria You Should Use To Choose A Data Catalog.* https://www.topbots.com/choosing-a-data-catalog/. (Visitado 28-02-2021).

[3] *Ad hoc.* https://www.reasonwhy.es/diccionario/ad-hoc. (Visitado 10-04-2021).

[4] *Análisis de la calidad de datos en fuentes de la suite ABCD.* https://dspace.uclv.edu.cu/bitstream/handle/123456789/6792/TD_YordanAndreuAlvarez.pdf. (Visitado 04-04-2021).

[5] *Aplicar metodologías ágiles en sectores ajenos a las TIC.* https://pmi-mad.org/index.php?option=com_content&view=article&id=491:aplicar-metodologias-agiles-en-sectores-ajenos-a-las-tic&catid=137:articulos&Itemid=88. (Visitado 14-04-2021).

[6] *Big Data: ¿En qué consiste? Su importancia, desafíos y gobernabilidad.* https://www.powerdata.es/big-data. (Visitado 03-04-2021).

[7] *Calidad de Datos. Cómo impulsar tu negocio con los datos.* https://www.powerdata.es/calidad-de-datos. (Visitado 08-02-2021).

[8] *CKAN, la plataforma de portal de datos de código abierto líder en el mundo.* https://ckan.org/. (Visitado 15-04-2021).

[9] *Collibra Data Catalog.* https://www.collibra.com/data-catalog. (Visitado 15-04-2021).

[10] *Convierta el caos de datos en valor de datos con inteligencia de datos.* https://www.sap.com/spain/products/data-intelligence.html. (Visitado 15-04-2021).

[11] *data catalog.* https://searchdatamanagement.techtarget.com/definition/data-catalog. (Visitado 28-02-2021).

[12] *Data Discovery in 2020.* https://medium.com/bigeye/data-discovery-in-2020-8c85eed328bb. (Visitado 04-04-2021).

[13] *DATA PROFILING.*
https://www.etltools.org/data-profiling.html. (Visitado 10-04-2021).

[14] *Data profiling, el primer paso en calidad de datos.*
https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/data-profiling-el-primer-paso-en-calidad-de-datos. (Visitado 01-03-2021).

[15] *Data profiling no es lo mismo que evaluación de calidad de datos.*
https://blog.es.logicalis.com/analytics/data-profiling-no-es-lo-mismo-que-evaluacion-de-calidad-de-datos. (Visitado 04-04-2021).

[16] *Data Quality Assessment Framework.*
https://dsbb.imf.org/dqrs/DQAF. (Visitado 03-04-2021).

[17] *Data Quality Management: An Introduction.*
https://www.bmc.com/blogs/what-is-data-quality-management/. (Visitado 03-04-2021).

[18] *Datos Maestros: Definición y Tipología.*
https://prezi.com/omfh2k5ylqun/datos-maestros-definicion-y-tipologia/. (Visitado 05-05-2021).

[19] *Diferencia entre dato, información y conocimiento.*
https://www.estrategiaynegocios.net/opinion/977752-345/diferencia-entre-dato-informaci%C3%B3n-y-conocimiento. (Visitado 29-10-2020).

[20] *Dodd-Frank Act.*
https://www.history.com/topics/21st-century/dodd-frank-act. (Visitado 05-05-2021).

[21] *El conocimiento es poder.*
https://www.culturagenial.com/es/el-conocimiento-es-poder/. (Visitado 04-11-2020).

[22] *El data quality marca la diferencia.*
https://blogs.imf-formacion.com/blog/tecnologia/el-data-quality-marca-la-diferencia-201908/. (Visitado 03-02-2021).

[23] *Entendiendo lo que es Data Lineage.*
https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/entendiendo-lo-que-es-data-lineage?hs_amp=true. (Visitado 10-12-2020).

[24] *ETL Qué es y por qué es importante.*
https://www.sas.com/es_es/insights/data-management/what-is-etl.html. (Visitado 10-04-2021).

[25] Alex Gorelik. *The Enterprise Big Data Lake.* O'Reilly, 2019.

[26] *Guide to Data Catalog Tools and Architecture.*
https://www.xenonstack.com/insights/data-catalog/. (Visitado 25-02-2021).

[27] *Historias de Usuario Scrum: Definición, ejemplos y plantillas.* https://asesorias.com/empresas/modelos-plantillas/historias-usuario-scrum/. (Visitado 05-05-2021).

[28] *IBM Watson Knowledge Catalog.* https://www.ibm.com/es-es/cloud/watson-knowledge-catalog. (Visitado 15-04-2021).

[29] *Introducción a la Calidad de Datos: Definición, Control y Beneficios.* https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/bid/368784/introducci-n-a-la-calidad-de-datos-definici-n-control-y-beneficios. (Visitado 03-02-2021).

[30] *ISO 8000-1x0.* http://iso8000.es/normas-iso-8000-1x. (Visitado 03-04-2021).

[31] *ISO 8000-6x.* http://iso8000.es/normas-iso-8000-6x. (Visitado 03-04-2021).

[32] *LA CALIDAD DE LOS DATOS: QUÉ ES, CÓMO SE GARANTIZA Y QUÉ BENEFICIOS REPORTA A TU NEGOCIO.* https://www.roiscroll.com/blog/la-calidad-de-los-datos-que-es-como-se-garantiza-y-que-beneficios-reporta-a-tu-negocio. (Visitado 03-04-2021).

[33] *Los estándares de calidad en Big Data.* https://www.deustoformacion.com/blog/gestion-empresas/estandares-calidad-big-data. (Visitado 04-02-2021).

[34] *Making data-driven decisions during the COVID-19 pandemic.* https://www.collibra.com/blog/making-data-driven-decisions-during-the-covid-19-pandemic. (Visitado 22-02-2021).

[35] *Marcos de gestión de proyectos.* https://www.wrike.com/es/project-management-guide/marcos-de-gestion-de-proyectos/. (Visitado 24-03-2021).

[36] Danette McGilvray. *Executing Data Quality Projects Ten Steps to Quality Data and Trusted Information.* Morgan Kaufmann, 2008.

[37] *Metadatos.* https://es.calameo.com/read/000132245f2070a80f965. (Visitado 12-02-2021).

[38] *Metadatos, definición y características.* https://www.powerdata.es/metadatos. (Visitado 15-02-2021).

[39] *Metodologías ágiles ¿qué son y para qué sirven?* https://www.tithink.com/es/2018/10/16/metodologias-agiles-que-son-y-para-que-sirven/. (Visitado 11-04-2021).

[40] *Miguel A. Martínez-Prieto, Jorge Silvestre, Anibal Bregon, José Ignacio Farrán Hacia la consolidación de las aulas ágiles Actas de las XXVI Jornadas sobre Enseñanza Universitaria de la Informática Volumen 5, págs. 29-36, 2020.*
http://uvadoc.uva.es/bitstream/handle/10324/42479/uvagile.pdf?sequence=3&isAllowed=y. (Visitado 25-05-2021).

[41] *Normas ISO 8000.*
http://iso8000.es/normas-iso-8000. (Visitado 03-04-2021).

[42] *Oracle Cloud Infrastructure Data Catalog.*
https://www.oracle.com/es/big-data/data-catalog/. (Visitado 15-04-2021).

[43] *Principales indicadores para Calidad de Datos.*
https://www.grapheverywhere.com/principales-indicadores-para-calidad-de-datos/. (Visitado 08-02-2021).

[44] *Qlik Catalog™ | Enterprise Data Catalog.*
https://www.qlik.com/es-es/products/qlik-catalog. (Visitado 15-04-2021).

[45] *Qué son los Metadatos.*
https://www.geoidep.gob.pe/conoce-las-ides/metadatos/que-son-los-metadatos. (Visitado 15-02-2021).

[46] *Qué son los metadatos: definición, tipos y ejemplos.*
https://www.docunecta.com/blog/que-son-los-metadatos. (Visitado 12-02-2021).

[47] *Qué son los metadatos y por qué son tan importantes.*
https://www.obsbusiness.school/blog/que-son-los-metadatos-y-por-que-son-tan-importantes. (Visitado 12-02-2021).

[48] *Qué son y cuál es la importancia de los datos maestros.*
https://blog.stibosystems.lat/que-son-y-cual-es-la-importancia-de-los-datos-maestros. (Visitado 10-12-2020).

[49] Ken Schwaber y Jeff Sutherland. *La Guía Scrum. La Guía Definitiva de Scrum: Las Reglas del Juego.*

[50] *SO 8000-2:2020(en) Data quality — Part 2: Vocabulary [3.5.2].*
https://www.iso.org/obp/ui/#iso:std:iso:8000:-2:ed-4:v1:en. (Visitado 11-04-2021).

[51] *The 20 Best Data Catalog Tools and Software for 2021.*
https://solutionsreview.com/data-management/the-best-data-catalog-tools-and-software/. (Visitado 05-04-2021).

[52] *The difference between a data swamp and a data lake? 5 signs.*
https://www.information-age.com/data-swamp-data-lake-123481597/. (Visitado 10-12-2020).

[53] *The HIPAA Privacy Rule.*
https://www.hhs.gov/hipaa/for-professionals/privacy/index.html. (Visitado 05-05-2021).

[54] *Top Ten Data Quality Problems: Part I.*
http://ds.datasourceconsulting.com/blog/top-10-data-quality-problems-part-1. (Visitado 03-02-2021).

[55] *What is a data catalog?*
https://www.ibm.com/cloud/learn/data-catalog. (Visitado 25-02-2021).

[56] *What Is a Data Catalog?*
https://www.alation.com/blog/what-is-a-data-catalog/. (Visitado 18-02-2021).

[57] *What is a data catalog?*
https://www.collibra.com/blog/what-is-a-data-catalog. (Visitado 25-02-2021).

[58] *What is a Data Catalog, and Do You Need One?*
https://www.talend.com/resources/what-is-data-catalog/. (Visitado 18-02-2021).

[59] *What Is a Data Catalog and Why Do You Need One?*
https://www.oracle.com/big-data/what-is-a-data-catalog/. (Visitado 25-02-2021).

[60] *What is Data Preparation?*
https://www.talend.com/resources/what-is-data-preparation/?type=productspage. (Visitado 03-04-2021).

[61] *What is Data Profiling?*
https://www.alooma.com/blog/what-is-data-profiling. (Visitado 01-03-2021).

[62] *What is data profiling and how does it make big data easier?*
https://www.sas.com/es_es/insights/articles/data-management/what-is-data-profiling-and-how-does-it-make-big-data-easier.html. (Visitado 01-03-2021).

[63] *What Is Data Profiling? Process, Best Practices and Tools.*
https://panoply.io/analytics-stack-guide/data-profiling-best-practices/. (Visitado 01-03-2021).

[64] *What is Data Profiling? Tools and Examples.*
https://www.talend.com/resources/what-is-data-profiling/. (Visitado 09-04-2021).

[65] *What is Metadata—The Key to Unlocking the Value of Your Data Assets.*
https://atlan.com/what-is-metadata/. (Visitado 15-02-2021).

[66] *What is Talend Studio?*
https://help.talend.com/r/Se88elCF0HkomGAlfWThWA/1AYBDq4xfCACn_7ferVd8A. (Visitado 11-04-2021).

[67] *¿En qué consiste la creación de perfiles de datos?*
https://www.talend.com/es/resources/what-is-data-profiling/. (Visitado 10-04-2021).

[68] *¿Qué es Data Quality y por qué es importante?*
https://www.elternativa.com/blog/index.php/2019/04/03/que-es-data-quality-y-por-que-es-importante/. (Visitado 03-02-2021).

[69] *¿Qué es el procesamiento de datos?*
https://www.nextu.com/blog/que-es-el-procesamiento-de-datos/. (Visitado 02-04-2021).

[70] *¿Qué es la gobernanza de datos? ¿La necesito?*
https://www.masterbigdataucm.com/que-es-big-data/. (Visitado 03-04-2021).

[71] *¿Qué es la gobernanza de datos? ¿La necesito?*
https://www.talend.com/es/resources/what-is-data-governance/. (Visitado 07-01-2021).

[72] *¿Qué es un «catálogo de datos»?*
https://www.t-systemsblog.es/que-es-un-catalogo-de-datos/. (Visitado 22-02-2021).

[73] *¿Qué son los metadatos y por qué son importantes?*
https://blog.enzymeadvisinggroup.com/que-son-los-metadatos. (Visitado 15-02-2021).

[74] *¿Qué son los métodos agile?*
https://blog.selfbank.es/que-son-los-metodos-agile/. (Visitado 22-03-2021).