UNIVERSIDAD DE BURGOS
ESCUELA POLITÉCNICA SUPERIOR
Máster en Ingeniería en Informática

Trabajo final del Máster Ing.Informática:

El gobierno del dato en el internet de las cosas

Presentado por Mikel Villanueva Gutiérrez
en julio de 2025
Tutora: Raquel Redondo Guevara

UNIVERSIDAD DE BURGOS

ESCUELA POLITÉCNICA SUPERIOR

Máster en Ingeniería en Informática

Dña. Raquel Redondo Guevara, profesora del departamento de Digitalización, área de Ciencia de la Computación e Inteligencia Artificial.

Expone:

Que el alumno D. Mikel Villanueva Gutiérrez, con DNI 72833075X ha realizado el Trabajo Fin de Máster del Máster en Ingeniería Informática titulado “El gobierno del dato en el internet de las cosas”.

Y que dicho trabajo ha sido realizado por el alumno bajo la dirección de la que suscribe, en virtud de lo cual se autoriza su presentación y defensa.

En Burgos, 7 de julio de 2025

Vº. Bº. de la Tutora:

Dña. Raquel Redondo Guevara

El gobierno de los datos es un aspecto crítico de las organizaciones modernas, ya que garantiza que los datos se gestionan de forma eficaz, segura y ética en diversos ámbitos. Sin embargo, la complejidad de los entornos de datos en evolución, las crecientes presiones normativas, la importancia cada vez mayor de la toma de decisiones basada en datos y las nuevas posibilidades brindadas por herramientas que aprovechan las capacidades de la inteligencia artificial han puesto de manifiesto la necesidad de un marco completo y adaptable para la gestión del gobierno de datos. Esta situación se ve doblemente agrandada en el contexto del internet de las cosas, por sus características de contexto descentralizado, heterogéneo y masivo.

Por lo tanto se propone la definición de un marco sólido para la gobernanza de datos que sea completo y útil; que integre principios clave como la privacidad, la calidad y la seguridad. Que también se incorporen buenas prácticas de la industria, se analicen las distintas herramientas disponibles en el mercado y se contemplen las regulaciones existentes.

También poder desarrollar unas normas para orientar la toma de decisión al crear políticas y procesos específicos. Creando guías para definir requisitos y orientar las acciones de implementación a realizar.

Descriptores

Data gobernance, Internet of Things, calidad del dato, Apache NiFi, OpenMetadata, IoT-Lite

Data governance is a critical aspect of modern organizations, ensuring that data is managed effectively, securely and ethically in a variety of settings. However, the complexity of evolving data environments, increasing regulatory pressures, the growing importance of data-driven decision making and the new posibilities brought forward by tools harnessing the strengths of artificial intelligence have highlighted the need for a comprehensive and adaptable framework for managing data governance. This situation is doubly magnified in the context of the Internet of Things, due to the characteristics of its decentralized, heterogeneous and massive context.

It is therefore proposed to define a solid framework for data governance that is complete and useful; that integrates key principles such as privacy, quality and security. It should also incorporate industry best practices, analyze the different tools available in the market and take into account existing regulations.

Also to be able to develop standards to guide decision making by creating specific policies and processes. Creating guidelines to define requirements and guide the implementation actions to be carried out.

Keywords

Data governance, Internet of Things, data quality, Apache NiFi, OpenMetadata, IoT-Lite

Índice de contenido

Índice de ilustraciones....2
Índice de tablas....2
I - Objetivos....3
1. ¿Qué nos aporta el gobierno del dato?....3
2. Objetivo principal....3
3. Objetivos específicos....3
4. Metodología....4
II - Alcance....5
1. Contexto....5
1.1. Los pilares del gobierno del dato....5
2. Actores implicados....5
III - Equipo....6
1. Consejo de gobierno del dato....6
2. Grupo de coordinación....6
3. Grupos de áreas....6
4. Roles....6
4.1. Sponsor....6
4.2. Líder del gobierno del dato....7
4.3. Propietario del dato....7
4.4. Data steward....7
4.5. Interesados....7
4.6. Custodios....7
5. Organización de los roles en los equipos 7
IV - Formación de data stewards....9
V - Ontología....10
1. Ontología dedicada a campos específicos 10
2. Ontology Modeling for Intelligent Domotic Environments (DogOnt)....10
3. IoT-Lite....11
4. Web of Things Thing Description (WoT-TD)....11
5. FIESTA-IoT....11
6. Machine-to-Machine Measurement (M3) 11
7. Comparativa....11
8. Funcionamiento de IoT-Lite....12
9. Caso práctico....13
VI - Políticas....15
1. Políticas de calidad....15
2. Políticas de privacidad....16
3. Políticas de seguridad....18
4. Políticas de la vida del dato....18
5. Políticas éticas....19
6. Políticas de definiciones....19
VII - Gestión de prácticas y políticas....20
1. Auditorías....20
VIII - Calidad del dato....22
1. Selección de las reglas....22
2. Requisitos....23
IX - Regulaciones....24
X - Especificaciones....25
1. UNE 0077:2023....25
2. UNE 0078:2023....26
3. UNE 0079:2023....26
4. ISO 8000....27
5. ISO/IEC 11179....27
6. ISO/IEC 38505....27
XI - Clasificación del dato....28
1. Personal....28
2. Sensible....29
3. Interna....30
4. Abierta....30
XII - Titularidad del dato....31
XIII - Herramientas....32
1. Herramientas ETL....32
1.1. Adeptia Connect....32
1.2. Apache NiFi....33
1.3. IBM InfoSphere Information Server 33
1.4. Oracle Data Integration Cloud Service....33
1.5. Comparativa....33
1.6. Usando Apache NiFi....34
1.7. Prueba de concepto....35
1.7.A. Instalación....35
1.7.B. Ejecución....35
1.7.C. Extracción de datos....36
1.7.D. Enrutado....37
1.7.E. Procesado....40
1.7.F. Unión....41
1.7.G. Guardado....42
1.7.H. Ejecutando la prueba de concepto....43
2. Herramientas de calidad del dato....44
2.1. Alternativas....44
2.1.A. Suites completas....44
2.1.B. Herramientas dedicadas....44
2.2. Herramientas de data profiling....45
2.3. Apache NiFi y la calidad del dato. 45
2.4. Incluyendo calidad del dato en la prueba de concepto....45
3. Catálogo de datos....52
3.1. Amudsen....53
3.2. Collibra....53
3.3. OpenMetadata....54
3.4. DataHub....54
3.5. Comparativa....55
3.6. Utilizando OpenMetadata....55
3.6.A. Administradores....56
3.6.B. Usuarios....57
3.6.C. Instalación de la herramienta 58
C.1. Problemas encontrados....58
C.2. Requisitos....58

C.3. Instalación.....58
4. Glosario de términos.....61
XIV - Monitorización.....62
1. Retos de la monitorización en el IoT....62
2. Buenas prácticas.....62
3. El futuro de la monitorización.....63
XV - Cultura del gobierno del dato.....64
XVI - Conclusiones y líneas futuras.....65
1. Líneas de trabajo futuro.....65
2. Futuro del gobierno del dato.....66
XVII - Referencias.....67
Bibliografía.....67

Índice de ilustraciones

Ilustración 1: Gráfico de grupos y roles.....8
Ilustración 2: Diagrama modelado Netatmo..13
Ilustración 3: Pantalla de inicio NiFi.....35
Ilustración 4: Procesador GetFile.....36
Ilustración 5: Configuración procesador GetFile NiFi.....36
Ilustración 6: Configuración enrutado por atributo NiFi.....37
Ilustración 7: Enrutado inicial NiFi.....38
Ilustración 8: Enrutado complejo NiFi.....38
Ilustración 9: Enrutado a procesador transformador NiFi.....39
Ilustración 10: Propiedades transformación unidades.....40
Ilustración 11: Flujo de procesado para dos orígenes.....41
Ilustración 12: Flujo completo ETL.....42
Ilustración 13: Salidas ETL.....43
Ilustración 14: Validación del dato en el flujo de ETL.....46
Ilustración 15: Enrutamiento de registros inválidos.....47
Ilustración 16: Final del flujo con monitorización.....50
Ilustración 17: Flujo completo con elementos de calidad del dato.....51
Ilustración 18: Página de login de OpenMetadata.....58
Ilustración 19: Panel de administrador OpenMetadata.....59
Ilustración 20: Panel lateral OpenMetadata. .59

Índice de tablas

Tabla 1: Relación grupos-roles.....7
Tabla 2: Certificaciones para Data Stewards...9
Tabla 3: Ontología de campos específicos....10
Tabla 4: Comparativa ontología.....12
Tabla 5: Ejemplos de políticas de calidad.....14
Tabla 6: Ejemplos de políticas de privacidad 16
Tabla 7: Ejemplos de políticas de seguridad. 17
Tabla 8: Ejemplos de políticas de la vida del dato.....18
Tabla 9: Ejemplos de políticas éticas.....18
Tabla 10: Ejemplos de políticas de definiciones.....18
Tabla 11: Dimensiones de la calidad del dato 21
Tabla 12: Unidades de indicadores de calidad del dato.....22
Tabla 13: Correlación dimensiones de la calidad del dato con unidades.....22
Tabla 14: Procesos UNE 0077:2023.....24
Tabla 15: Procesos UNE 0079:2023.....25
Tabla 16: Requisitos información personal...27
Tabla 17: Requisitos información sensible....28
Tabla 18: Requisitos información interna.....29
Tabla 19: Requisitos información abierta.....29
Tabla 20: Comparativa herramientas ETL.....32
Tabla 21: Directorios Apache NiFi.....34
Tabla 22: Estructura entradas NiFi.....39
Tabla 23: Operaciones necesarias según origen 39
Tabla 24: Propiedades para transformación de eliminar columna.....39
Tabla 25: Propiedad para nombrar salidas.....41
Tabla 26: Acciones de calidad del dato Apache NiFi.....44
Tabla 27: Condiciones para validez del dato. 45
Tabla 28: Esquemas de validación del dato...45
Tabla 29: Log de errores tras ejecución.....48
Tabla 30: Características Amudsen.....52
Tabla 31: Características Collibra.....52
Tabla 32: Características OpenMetadata.....53
Tabla 33: Características DataHub.....53
Tabla 34: Comparativa herramientas de catálogo de datos.....54

I - Objetivos

Los objetivos se separan en dos tipos[1]:

- Por un lado, se identifican los objetivos que la gobernanza del dato aporta a las organizaciones, es decir qué aspectos son los que se verán mejorados en nuestra organización tras la aplicación de un framework de gobierno del dato.
- Por otro lado, también es necesario definir los objetivos que el proyecto en sí ha de cumplir, qué aspectos de el gobierno del dato se van a tratar.

1. ¿Qué nos aporta el gobierno del dato?

Los beneficios de aplicar un marco de gobierno del dato a cualquier organización son claros e inmediatos, con un efecto directo en la eficiencia, seguridad y desarrollo.

Entre los objetivos que se buscan se encuentran los siguientes:

1. Mejora de los datos: de forma que no solamente se obtienen datos de mayor calidad, sino que son más confiables.
    1. Utilizamos estándares que nos ayudan a medir y cumplir con la calidad
    2. Se monitorizan los datos de forma continua para garantizar el cumplimiento
2. Cumplimiento de las normativas pertinentes, y mejora de la seguridad en general
3. Optimización del uso de los datos, consiguiendo que las operaciones diarias sean más rápidas
4. Mejora de la organización, dotando a todo el mundo de una visión unificada del significado y uso de los datos

2. Objetivo principal

El objetivo principal es el diseño y especificación de un framework para implementar el gobierno del dato en una organización que trabaja en el ámbito de el Internet of Things. Este debe de cubrir todos los aspectos incluyendo la organización de los equipos, las directrices a seguir para diseñar políticas y procesos a implementar, gestión del dato, cumplimiento con las regulaciones y las herramientas a utilizar.

3. Objetivos específicos

- Identificar y categorizar los actores principales en un proyecto de este ámbito, definiendo equipos, roles y sus relaciones
- Investigar estándares actuales para el uso de un lenguaje común en el ambiente de el IoT
- Definir los procesos a seguir para definir políticas de gobierno del dato, y los procedimientos a seguir para gestionar dichas políticas
- Definir los procesos a seguir para especificar reglas y requisitos de calidad del dato
- Analizar las regulaciones pertinentes al área de gobierno del dato
    - Utilizar dicha regulación para especificar la clasificación del dato
- Definir el tipo de herramientas necesarias
    - Por cada tipo realizar un estudio de mercado para poder realizar una comparativa y de

dicha comparativa poder realizar una selección

4. Metodología

- Estudio de el estado del arte, estándares y buenas prácticas para la implementación del gobierno del dato
- Estudio y comparativa de estándares ontológicos para definiciones de IoT
- Revisión de pautas comunes para la especificación de políticas y procesos
- Análisis de regulaciones que sea necesario tener en cuenta
- Estudio de mercado de las herramientas existentes
- Implementación de una prueba de concepto que utilice una de las herramientas seleccionadas

II - Alcance

1. Contexto

El gobierno del dato es un aspecto clave en cualquier organización, pero tiene un carácter especialmente importante en el ámbito del IoT (Internet de las Cosas). No solo se parte desde una situación en la que los dispositivos IoT son cada vez más en número, sino que cuenta con bastantes retos que son claves desde una perspectiva de gobierno del dato. Por ejemplo, la generación masiva de datos que estos dispositivos conllevan, o la heterogeneidad de las fuentes. A esto se le añaden agravantes como la necesidad de procesamiento de los datos en tiempo real y la necesidad de una mayor seguridad.

1.1. Los pilares del gobierno del dato

1. La calidad del dato: La calidad de los datos se refiere a los atributos de los datos que determinan en qué medida satisfacen las necesidades de sus usuarios.. Los datos de alta calidad son esenciales para tomar decisiones informadas, impulsar el análisis y respaldar los procesos operativos en diversas funciones.
2. Seguridad y privacidad: La seguridad y la privacidad de los datos garantizan su conformidad con la legislación pertinente y su protección frente al acceso no autorizado, la alteración, la destrucción o el uso indebido.
3. Catalogación y metadatos: La catalogación de datos y la gestión de metadatos permiten a las organizaciones organizar, documentar y describir sus activos de datos. Proporcionando información sobre la finalidad y el uso previstos de los datos, lo que facilita a los usuarios la comprensión y el aprovechamiento de los datos.
4. Gobernanza organizativa: Establece estructuras, funciones y responsabilidades de las actividades relacionadas con los datos.

2. Actores implicados

El gobierno del dato ha de implicar no sólo a los actores directos del proyecto, sino que es algo en lo que toda la organización ha de aportar y concienciarse. Sin embargo, sí que hay ciertos departamentos que son de vital importancia:

- Dirección, dado que han de mostrar su apoyo para que el proyecto cuente con los recursos
- Departamento de las tecnologías de la información, por su conocimiento técnico
- Especial mención a el área de la seguridad como expertos de la protección de datos
- Departamento legal, por su conocimiento de las leyes y normativas pertinentes
- Responsables de operaciones y producto, por su conocimiento del negocio y el uso de los datos en el día a día
- Equipos de desarrollo, ya que son los encargados de construir las soluciones que hacen uso de los datos
- De la misma manera, se definirá un equipo específico del gobierno del dato en un apartado posterior. Estas serán las personas principales encargadas de llevar a cabo el proyecto. Para ello se definirán roles específicos y se acotarán las responsabilidades de cada uno.

III - Equipo

El equipo de gobierno del dato se puede separar en 3 grupos principales[2].:

1. El consejo de gobernanza del dato, que tomará las decisiones a alto nivel
2. El grupo de coordinación
3. Un grupo reducido por cada área de negocio

1. Consejo de gobierno del dato

El consejo ha de ser formado por miembros de la organización con capacidades para asignarle tanto recursos como los fondos necesarios y miembros que vayan a ser los encargados de tomar las decisiones sobre políticas y procesos. Como tercer grupo minoritario también se contará con gente relacionada a los procesos de negocio, aunque no sean miembros principales aportarán su conocimiento sobre el negocio en momentos en los que se necesite.

Sus principales responsabilidades son las siguientes[3]:

- Definir los estándares (conjunto de normativas y criterios establecidos para garantizar la consistencia, calidad y seguridad en la gestión, almacenamiento y uso de los datos dentro de una organización). Como puede ser para las definiciones, formatos, o nomenclaturas
- Definir las qué se ha de monitorizar y qué acciones hay que seguir en caso de fallos
- Definir las políticas de acceso a datos y seguridad
- En aspectos generales, el rol principal consiste en decidir cuales son los objetivos y metodologías principales a utilizar.

2. Grupo de coordinación

Este grupo contará con miembros que también participan en el consejo, quienes se encargan de definir las políticas a seguir, han de definir los requisitos específicos que se han de lograr para lograr los objetivos a alto nivel decididos en el consejo. Funcionará como un intermediario entre la dirección de alto nivel del consejo y los grupos de áreas específicos, marcando las acciones que los grupos de áreas han de realizar y retransmitiendo al consejo los resultados y comentarios de las áreas al consejo.

3. Grupos de áreas

Formando parte de cada área de negocio, este grupo une a ciertos miembros de el grupo de coordinación con interesados y personal técnico de las áreas de negocio. Estos serán los encargados de implementar las órdenes de trabajo específicas que lleven al cumplimiento de los requisitos definidos.

4. Roles

Para conformar estos grupos definiremos varios roles, cada persona implicada en el proyecto contará con un rol, y formará parte de uno o varios grupos [4].

4.1. Sponsor

Será el líder del proyecto, encargado de asignar los suficientes fondos y recursos que se acuer-

den en el consejo. Toma carácter de guía de la estrategia que ha de seguir el proyecto.

4.2. Líder del gobierno del dato

Es el responsable máximo de que se lleve adelante la implementación del proyecto. Por lo tanto sus responsabilidades son de coordinar y guiar al equipo, como jefe de proyecto.

4.3. Propietario del dato

El propietario del dato es quien tiene autoridad para realizar las decisiones sobre las definiciones de términos de negocios y requerimientos que el negocio tiene sobre el dato. Términos de negocio en este caso hace referencia a términos como “expediente” o “usuario”, que más allá de su definición de diccionario tienen un sentido específico en el ámbito de un negocio en particular.

4.4. Data steward

Los data steward son los encargados de trasladar las decisiones de alto nivel a acciones específicas y actúan como sus representantes en la reuniones del día a día. Son personas de los diferentes departamentos que trabajan bajo la coordinación del propietario del dato.

4.5. Interesados

Son los representantes de los afectados por las políticas y acciones que se definen y realizan a causa del proyecto. Por un lado se encargan de transmitir la información pertinente a los usuarios y consumidores de los datos, y también se encargan de recoger todas las dudas, comentarios y opiniones de éstos.

4.6. Custodios

Se trata de un rol técnico que consiste en el mantenimiento de los datos, por lo tanto la implementación técnica de todas acciones definidas por los data stewards, así como de los arreglos y mejoras que se necesiten según se especifique en la monitorización.

5. Organización de los roles en los equipos

Ya definidos tanto los grupos como los roles, se especifica la organización de la siguiente manera

Tabla 1: Relación grupos-roles

Grupo	Roles
Consejo del gobierno del dato	Sponsor, líder de gobierno del dato, propietario del dato Personas claves de entre los interesados
Grupo de coordinación	Líder de gobierno del dato, propietario del dato, data stewards, personas claves de entre los interesados
Grupos de áreas	Data stewards ,custodios, interesados

Las relaciones entre los grupos se muestran en el siguiente diagrama

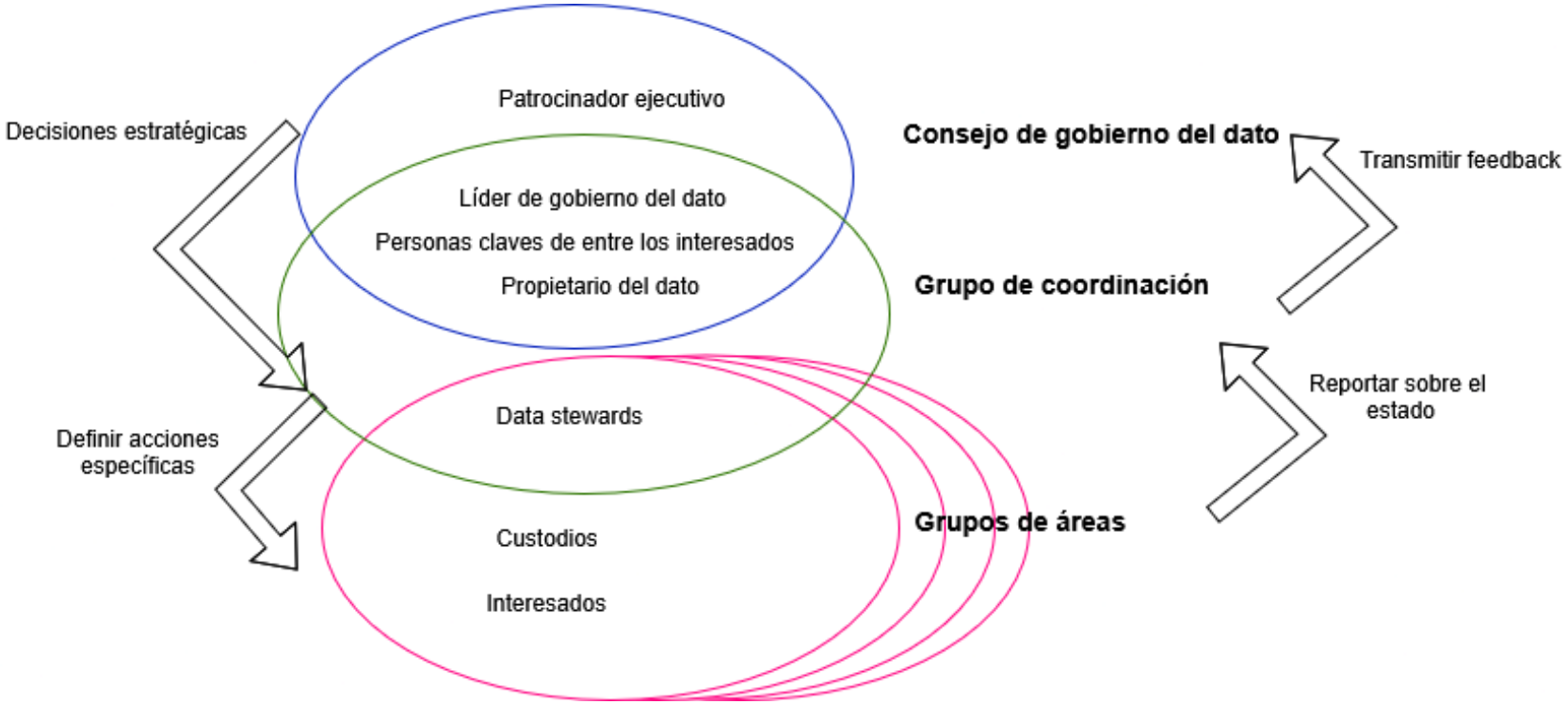

Ilustración 1: Gráfico de grupos y roles

IV - Formación de data stewards

Los data stewards son un perfil con alto conocimiento de ingeniería de datos, para poder realizar la función de data steward se requerirá alguna de las siguientes formaciones:

- Grado en Ingeniería Informática
- Grado en Ingeniería en Tecnología de Telecomunicación
- Grado en Matemáticas
- Grado en Ciencia de Datos
- Algún otro grado equivalente

Además de los estudios de grado hay ciertos posgrados que también ayudarán en la formación:

- Máster en Big Data
- Máster en Ciberseguridad
- Máster en Ingeniería del Software

Es importante que la organización busque políticas para fomentar este tipo de estudios, como ayudas económicas o conciliación de horarios. Sin embargo, no es realista esperar que los data stewards inviertan tanto tiempo, recursos y energías en formaciones tan extensas, ni se puede permitir ninguna organización los recursos ni tiempo completos que requieren. En España sí que existen varias organizaciones que ofrecen diferentes certificaciones en el ámbito del gobierno del dato, aunque ninguna de ellas es una certificación estandarizada, todas ellas son propias de las instituciones que las imparten. Estas sí que pueden considerarse como una buena opción para compaginar con el trabajo, dado sus precios más reducidos y duración asumible. Por lo tanto se ofrecerá a los data stewards la opción de cursar estos diferentes cursos y obtener sus correspondientes certificaciones:

Tabla 2: Certificaciones para Data Stewards

Certificación	Ofrecido por	Idioma	Virtual/presencial	Duración
Data Stewardship y gestión de datos de investigación	Sociedad Española de Documentación e Información Científica (SEDIC)	Castellano	Virtual	60h
Certified Data Steward	eLearningCurve	Inglés	Virtual	20h
Curso de Talend Data Stewardship	Talend	Castellano	Virtual o presencial (Bilbao, Madrid, Málaga o Barcelona)	14h

De la misma forma que se sugieren estos cursos, existirá la opción de los data stewards de solicitar otros alternativos que crean pertinentes, bajo la aprobación de su jefe de proyecto.

V - Ontología

De cara utilizar un lenguaje común, es una buena práctica no intentar partir de 0, sino utilizar estándares ya existentes como base. Existen marcos que definen los conceptos del IoT de una manera estándar. Para ello, se ha realizado un estudio de diferentes estándares ontológicos para ver sus diferentes características y poder decidirse por uno de ellos.

1. Ontología dedicada a campos específicos

No se consideran como soluciones a las necesidades en el contexto de IoT en general, dado que no tienen uso directo fuera de sus campos. Pero pueden ser tenidos en cuenta Algunos ejemplos a continuación:

Tabla 3: Ontología de campos específicos

Nombre	Dominio	Documentación
Agronomy Ontology (AgrO)	Agronomía	https://bigdata.cgiar.org/resources/agronomy-ontology/
Brick	Construcción	https://brickschema.org/
ExtruOnt	Industria , concretamente máquinas de extrusión	https://arxiv.org/pdf/2401.11848
SAREF	Es un grupo de varios proyectos cada uno con su dominio: • Automovilismo • Construcción • Ciudades • Salud • Energía • Medio ambiente • Industria • Aguas	• https://www.etsi.org/deliver/etsi_ts/103400_103499/10341007/01.01.01_60/ts_10341007v010101p.pdf • https://www.etsi.org/deliver/etsi_ts/103400_103499/10341003/01.01.02_60/ts_10341003v010102p.pdf • https://www.etsi.org/deliver/etsi_ts/103400_103499/10341004/01.01.02_60/ts_10341004v010102p.pdf • https://www.etsi.org/deliver/etsi_ts/103400_103499/10341001/01.01.02_60/ts_10341001v010102p.pdf • https://www.etsi.org/deliver/etsi_ts/103400_103499/10341002/01.01.02_60/ts_10341002v010102p.pdf • https://www.etsi.org/deliver/etsi_ts/103400_103499/10341005/01.01.02_60/ts_10341005v010102p.pdf • https://www.etsi.org/deliver/etsi_ts/103400_103499/10341010/01.01.01_60/ts_10341010v010101p.pdf

2. Ontology Modeling for Intelligent Domotic Environments (DogOnt)

Es un sistema ontológico principalmente orientado al entorno de la domótica, pero que también es utilizado como estándar genérico.

Alcance	Orientado a redes locales de dispositivos IoT
Documentación	https://iot-ontologies.github.io/dogont/
Licencia	Apache License, Version 2.0

3. IoT-Lite

IoT-Lite es un sistema de representación de recursos de IoT con el objetivo de ser de amplio alcance y al mismo tiempo ligero. Por lo que puede ser de utilidad cuando se quieren mezclar conceptos de diferentes ámbitos de las IoT.

Alcance	Lo más amplio posible, aplicable a cualquier ámbito
Documentación	https://www.w3.org/submissions/2015/SUBM-iot-lite-20151126/
Licencia	Creative Commons Attribution 3.0 Unported License

4. Web of Things Thing Description (WoT-TD)

Otro estándar orientado a uso general para cualquier ámbito de las IoT, con el objetivo específico de integrar los dispositivos IoT en un contexto Web of Things. Es decir dispositivos accesibles desde internet.

Alcance	General, Web of Things
Documentación	https://wot-td-ontology.github.io/wot-thing-description/
Licencia	W3C Software and Document License

5. FIESTA-IoT

Centrado en la privacidad, y en soportar los frameworks de pruebas de los dispositivos IoT. Pero sin dejar de lado el modelar las IoT de forma amplia y genérica.

Alcance	General, aunque con especial énfasis en privacidad y testbeds.
Documentación	https://ieeexplore.ieee.org/abstract/document/7845470
Licencia	Copyright EU H2020 FIESTA-IoT

6. Machine-to-Machine Measurement (M3)

Creado con el objetivo de utilizar un lenguaje unificado que compartan todos los sensores de dispositivos de todos los ámbitos.

Alcance	General
Documentación	http://www.eurecom.fr/fr/publication/4553/download/cm-publi-4553.pdf
Licencia	GNU GPLv3 license

7. Comparativa

Para realizar una selección de qué estándar utilizar, se ha decidido utilizar 4 cualidades. No todas necesariamente de obligatorio cumplimiento al 100%:

- Aplicable a un ámbito general: en este contexto significa que no está diseñado para un ámbito específico, y se considerará mejor según más tipos de elementos, propiedades y valores sea capaz de modelar. En este caso sólo DogOnt no es del todo aplicable a todos lo dominios, aunque sigue su desarrollo para que en algún momento lo sea. Y por otro lado M3 tiene un diseño que modela ciertas unidades, sensores y dominios de una forma muy rígida, por lo que podría darse el caso de que se necesitara modelar algún concepto que no lo contemple este sistema.
- Suficientemente desarrollado: implica que está listo para ser utilizado en aplicaciones productivas ya. En el caso de los estándares propuestos, todos están ya lo suficientemente desarrollados, exceptuando a DogOnt que sigue en el desarrollo previamente mencionado

de pasar de un dominio específico a un ámbito genérico.

- Licencia permisiva: se valorará que cuente con una licencia que no limite su uso, ni condicione ser parte de ninguna organización
- Utilizado en el marco europeo: la cualidad más subjetiva de las 4, aunque sí se considera útil el saber que proyectos y empresas en un marco europeo lo están ya utilizando

Tabla 4: Comparativa ontología

	Aplicable a un ámbito general	**Suficientemente desarrollado**	**Licencia permisiva**	**Utilizado en el marco europeo**
DogOnt	En proceso de expandirse del ámbito de la domótica	En proceso de desarrollo de su segunda versión	Sí	No extendido
IoT-Lite	Sí	Sí	Sí	Sí, desarrollado como parte de la iniciativa EU FP7
WoT-TD	Sí	Sí	No	Presencia en entidades europeas
FIESTA-IoT	Sí	Sí	Sólo bajo previa petición	Sí
M3	Sí, aunque tiene una estructura muy rígida	Sí	Sí	Sí

Como conclusión, IoT-Lite es la que mejor balancea las 4 cualidades que se buscan, por lo que sería nuestra primera opción. De carácter abierto a todos los dominios y sin limitaciones con ningún dispositivo que nos vayamos a encontrar, cuenta con una licencia que no nos limita en ningún aspecto y sabemos que ha sido desarrollado y se está utilizando en un marco como el nuestro.

8. Funcionamiento de IoT-Lite

IoT-Lite es una extensión de Semantic Sensor Network Ontology (SSN) y utiliza qu-taxo (herramienta de taxonomía) para definir unidades y cantidades[5]. Categoriza los 3 elementos en 3 clases fundamentales:

- Objetos, entidades IoT
  - Pueden tener ubicaciones (puntos en el espacio físico)
  - Pueden tener atributos y estos esta asociados con dispositivos
- Sistemas, abstracción que representa infraestructura. Puede tener componentes y al mismo tiempo subsistemas (que a su vez son de tipo sistema)
  - Estos componentes o dispositivos a su vez pueden ser de diferentes tipos: dispositivos sensores, dispositivos de etiqueta (como puede ser un código QR o chip RFID)
  - Los dispositivos sensores son a su vez también elementos de tipo "sensor" que pueden contener sub-sensores. Lo que estos dispositivos captan se define mediante objetos de tipo unidad y cantidad
  - Los dispositivos también pueden tener cobertura (define su rango de actuación) y esta se define mediante utilizando lo siguiente
    - Puntos, que representan puntos en el espacio físico
    - El sub-tipo de el elemento cobertura nos indica como utilizar los puntos para

dibujar el alcance

- Círculo
- Rectángulo
- Polígono

- Servicios, que son proporcionados por dispositivos IoT

9. Caso práctico

Para modelar un caso práctico, se ha seleccionado un sistema IoT que muestre bien muchas de las cualidades que IoT-Lite permite modelar. Se ha seleccionado el sistema de sensores para hogar de Netatmo; compuesto por dos dispositivos físicos unidos por el mismo sistema, uno orientado a estar en el interior y el otro a estar en el exterior. Se ha decidido no modelar el dispositivo exterior del todo, ya que es casi idéntico al interior, cuenta con los mismos sensores, sólo que alguno cuenta con un mayor rango. Cada uno de estos dispositivos así mismo tiene varios sistemas con sensores, para la temperatura, humedad, calidad del aire, presión y sonido. Cada uno de los dispositivos obviamente cuenta con una ubicación distinta, pero ambos están expuestos al usuario conjuntamente mediante un servicio ofrecido por HomeKit.

En el siguiente diagrama se modela el sistema:

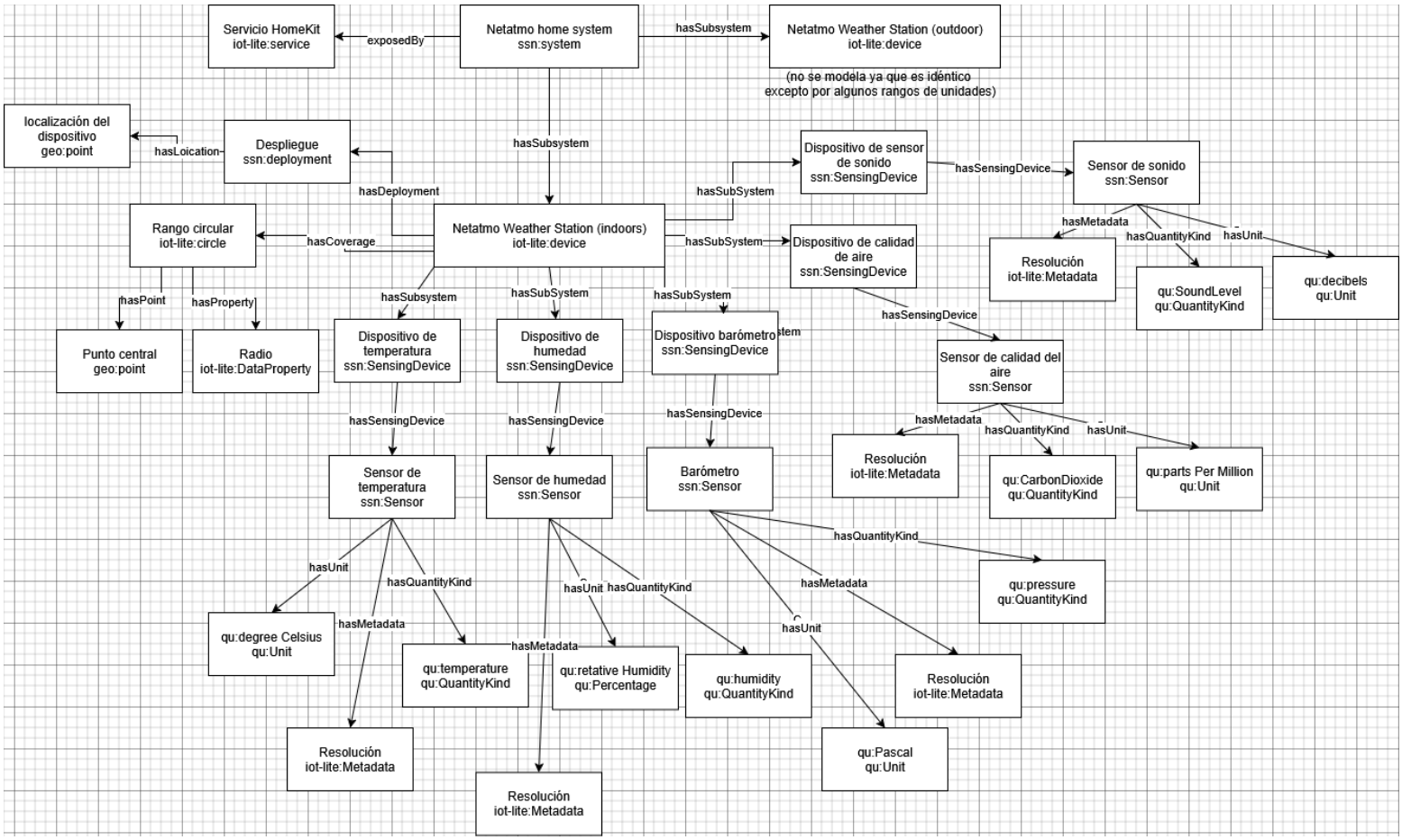

Ilustración 2: Diagrama modelado Netatmo

VI - Políticas

Las políticas de gobernanza de datos son un conjunto de normas, directrices y prácticas que definen cómo se gestionan, protegen y utilizan los datos dentro de una organización. Estas políticas establecen el marco para la toma de decisiones en torno a los datos, y su propósito es alinear las prácticas de gestión de datos con los objetivos de las organizaciones y los requisitos exigidos por las regulaciones.

Un ejemplo podría ser “Cuando se utilicen datos con fines analíticos, los datos personales o sensibles han de ser anonimizados o seudonimizados”.

A continuación definimos las políticas que se han definido, organizadas en las siguientes categorías [6]:

- Políticas de calidad
- Políticas de privacidad
- Políticas de seguridad
- Políticas de la vida del dato
- Políticas éticas
- Políticas de definiciones

1. Políticas de calidad

El reto principal de la calidad en el contexto de IoT viene dada la alta cantidad de datos y los heterogeneidad. Para ello es vital definir políticas que por un lado minimicen los posibles problemas de origen, y también realizar validaciones periódicas. A continuación se definen algunas de las políticas de este aspecto:

Tabla 5: Ejemplos de políticas de calidad

Política	Responsable	Momento de aplicación
Comparación de la información de los sensores con datos de referencia. Por ejemplo si tenemos sensores de lluvia comprobar que coincida con información meteorológica de referencia.	Proceso automatizado, los márgenes de diferencias serán marcados por el data steward, mientras que la implantación será de mano del custodio.	Periódica, diario como proceso automatizado fuera del horario laboral.
Validación de las unidades utilizadas por los nuevos orígenes de datos.	El custodio es el encargado de validar que cada nuevo origen que añada al sistema utilice las mismas unidades que los demás. Dichos estándares son los especificados por el data steward. Indican por ejemplo que los pesos han de especificarse en gramos, el tiempo en mili-segundos y la temperatura en grados Celsius. En caso de que el origen no cumpla por defecto, el custodio habrá de implementar la conversión de los datos antes de empezar a almacenarlos	Al momento de incorporar un nuevo origen de datos al sistema

Inspección visual de la información. No de forma exhaustiva, pero sí que ha de combinarse con las validaciones automáticas para poder encontrar casos que no se estén validando bien	El encargado es el data steward del área, ya que cuenta tanto con los conocimientos del negocio como de los estándares que se han marcado.	Periódica, semanal
Calibración de los sensores	Personal técnico	Periódico, según las especificaciones del fabricante
Implementación de redundancia en los sensores. En los casos que sea posible, utilizar sensores redundantes para la validación del correcto funcionamiento de cada uno de ellos. No sólo para la validación, sino el poder continuar con la recolección de datos en caso de fallo de uno de los sensores.	Custodios del dato, personal técnico	Crear la redundancia en cuanto sea posible, después una vez esté implementada la validación ha de ser automática

2. Políticas de privacidad

En el aspecto de la privacidad, las pautas principales en el ámbito de la Unión Europea las marca el Reglamento General de Protección de Datos (GDPR) [7]. A la hora de identificar si algún dato es de carácter privado es necesario utilizar la definición que nos indica [8]. A grandes rasgos se trata de información relacionada a una persona física (es decir, no aplicable a que sea identificada o que pueda ser identificada mediante un punto o varios de información. Bajo esta definición queda claro que muchísima de la información que capturan los dispositivos IoT cae bajo esta categoría. Sí que hay información que no es de este tipo, como puede ser:

- Información meteorológica en general
- Sensores de maquinaria industrial, o sensores de información anónima como puede ser la humedad en una planta industrial
- Sensores de movimiento en lugares públicos que no capten ninguna información que identifique a las personas (por ejemplo para contar el número de posibles clientes que entra a una tienda)

Por el otro lado, ejemplos muy usados en IoT que sí son de tipo privado:

- Cualquier dispositivo médico que monitorice a una persona
- Máquinas de fichaje de entrada y salida del trabajo en una planta industrial
- Dispositivos que localicen a una persona en el espacio/tiempo

El GPR también contempla un tipo de información extra sensible, que hemos de tratar con especial cuidado, siendo esta información de tipo:

- Genética
- Biométrica
- De salud
- Información personal que pueda revelar raza u origen étnico, opiniones políticas religiosas o ideológicas o que revelen pertenencia a sindicatos.

Tabla 6: Ejemplos de políticas de privacidad

Política	Responsable	Momento de aplicación
Identificar información privada. Se ha de identificar todos los orígenes que nos nutran de información que se considere privada. Se deberá ejecutar en cuanto sea posible para información ya capturada y orígenes de datos ya utilizados, y siempre que se añada un origen de datos nuevo. Especial hincapié en que no hay que tratar los nuevos orígenes como puntos independientes, sino que hay que considerarlos en contexto para poder identificar si la nueva información puede ser combinada con otra ya existente para des-anonimizarla.	Data steward	Inmediato, siempre en el momento de añadir nuevos orígenes de información
No compartir información privada con otras organizaciones. Esto ha de ser aceptado por el líder del proyecto, ya que tiene que guiar también al negocio. Es posible compartir esta información con otras organizaciones bajo ciertas condiciones, pero siempre la opción por defecto ha de ser que esta información no sea compartida con nadie.	Sponsor: responsable de alinear a la organización con la política Líder de gobierno del dato: responsable de argumentar su importancia Data steward: responsable de identificar la información privada	A la hora de solicitudes de información desde otras organizaciones
Notificación de la información utilizada. Es necesario notificar a la persona de la que se está capturando información. Esto implica, que las aplicaciones de nuestra organización informen al usuario de las mismas, se informen a los trabajadores y por último se informe a cualquier cliente.	Data steward: como supervisor Interesados: relacionados a las personas pertinentes Personal técnico, en caso de que se necesite alguna implementación de carácter técnico	Siempre que alguna de las personas afectadas interactúe con nosotros
Eliminación de la información bajo demanda de la persona a la que pertenezca esa información	Data steward: como supervisor Personal técnico, en caso de que se necesite alguna implementación de carácter técnico	Bajo demanda, cada vez que una persona solicite su eliminación
Cifrado de la información personal [9]. Como medida para minimizar los riesgos de una brecha de seguridad	Custodios Personal técnico	Continua
Creación de los casos de procesado de la información. Según la regulación, es necesario especificar qué uso se le da a cada dato, por lo que es necesario identificar todos los usos que se den y actualizar cada vez que se cree un uso nuevo para que se pueda ejercer la política de notificación a la persona afectada.	Data steward Custodios	Continua

3. Políticas de seguridad

No existe un sólo estándar universal de políticas de seguridad, y al igual que en los demás tipos de políticas hemos de tener en cuenta nuestro ámbito. Internet Society [10] por ejemplo distingue 3 tipos de ámbitos según su desarrollo en las regulaciones de la seguridad:

- Desarrollo en fases iniciales: hay interés en el IoT, pero no existe una regulación aún.
- Desarrollo en estado intermedio: hay regulaciones, pero no contemplan el IoT o no se ejercen dichas regulaciones
- Desarrollo avanzado: existen las regulaciones, son aplicables al IoT y hay cuerpos dedicados a ejercer su cumplimiento

Viendo estas tres categorías queda claro que cualquier país de la Unión Europea cae bajo la categoría de desarrollo avanzado, e Internet Society mismo sugiere unas acciones a realizar que pueden guiar nuestras políticas.

Tabla 7: Ejemplos de políticas de seguridad

Política	Responsable	Momento de aplicación
Revisión periódica de las regulaciones de seguridad. Que pueden incorporar requisitos para defenderse de nuevos riesgos y vulnerabilidades.	Líder de gobierno del dato Propietario del dato Data steward En general se trata de un equipo mixto que sea capaz de hacer cumplir lo acordado y que al mismo tiempo tenga algo de conocimiento específico del negocio	Periódica, con carácter mensual
Asignación de responsables para los diferentes sistemas	Líder de gobierno del dato Data steward Custodios	Cada vez que se realicen modificaciones en el equipo o en la propiedad del dato
Creación de protocolos en caso de una situación de violación de datos	Data steward y custodios asignados como responsables en la política anterior	Creación inmediata, con 2 revisiones anuales

4. Políticas de la vida del dato

Estas políticas se aplican en los diferentes puntos de la vida del dato, y aunque es cierto que existe cierto solapamiento con otras categorías, es importante que se reflejen en las políticas los retos que afectan especialmente al IoT[11]. Se hace especial hincapié en las etapas que son especialmente complejas en este ámbito; como por ejemplo la colección del dato, ya que vienen de fuentes, maneras y formatos mucho más heterogéneos que en la mayoría de otros ámbitos. Aunque por supuesto se han de considerar todas las fases.

Algunas fases del ciclo de vida que ya se cubren en anteriores políticas:

- Creación, colección y limpieza: se aplican las políticas de calidad del dato relacionadas con la estandarización y veracidad de la información
- Almacenamiento: se deben seguir las políticas de seguridad para garantizar su seguridad física (daño del lugar de almacenado) y de acceso (no haya accesos indebidos o hacking)
- Procesado: Aplicados en el contexto de la privacidad
- Retención: Se aplican las políticas de privacidad. Estas especifican los datos que no se pueden retener, por ejemplo alguien que solicita su eliminación según lo especificado en el GDPR

Como ejemplo de alguna política no considerada en el apartado anterior:

Tabla 8: Ejemplos de políticas de la vida del dato

Política	Responsable	Momento de aplicación
Purga de datos innecesarios. Ha de ser una decisión tomada entre los data stewards correspondientes y los custodios, para realizar la decisión de qué datos han dejado de ser relevantes para el negocio y pueden ser eliminados.	Data steward Custodios	Periódica de carácter anual

5. Políticas éticas

En lo relacionado a la ética, principalmente atañe a un cumplimiento correcto de la privacidad y seguridad anteriormente especificadas. Sin embargo es correcto también la revisión periódica de los usos y captura de los datos para revaluar estos procesos, tanto de cara a tener un trato más justo hacia las personas a las que atañe la información como para intentar adelantarnos a posibles regulaciones futuras.

Tabla 9: Ejemplos de políticas éticas

Política	Responsable	Momento de aplicación
Reuniones periódicas para la creación de directrices éticas que expandan los requisitos mínimos exigidos legalmente.	Líder de gobierno del dato Data steward	Periódica de carácter anual

6. Políticas de definiciones

Implican principalmente la creación de los estándares de nomenclatura que ha de seguir todo el mundo, y el uso de dicha nomenclatura por todos los usuarios de la información.

Tabla 10: Ejemplos de políticas de definiciones

Política	Responsable	Momento de aplicación
Creación y mantenimiento de un catálogo de datos. En el se definirán las nomenclaturas a utilizar.	Data stewards	Continua
Utilización de las nomenclaturas especificadas en el catálogo de datos. No será permisible utilizar nomenclaturas que no estén definidas en este.	Custodios Interesados	Continua

VII - Gestión de prácticas y políticas

Para realizar una buena gestión de todo lo implementado y verificar su funcionamiento se realizarán reuniones bianuales. Estas reuniones las realizará el consejo de grupo de coordinación para evaluar el grado de implantación de las prácticas y políticas definidas y valorar su efectividad. Para ello los data stewards de cada área deberán realizar una auditoría sobre su área y compartir los resultados con el líder del gobierno del dato. Deberán cubrir los siguientes aspectos:

- Nivel de implantación del plan
  - Y los problemas encontrados que puedan estar retrasando su implantación
- Eficacia de las acciones tomadas
  - Se deberá poner especial énfasis en las más y menos eficaces
- Comentarios y sugerencias recibidas

De los puntos en los que se han notado carencias o problemas será necesario distinguir qué tipo de solución es suficiente:

- Si se considera solucionable mediante medios técnicos o formas de implantación, el responsable de dicha acción seguirá siendo el encargado de realizar lo necesario. Para ello puede solicitar orientación de otros miembros del grupo o usar alguno de los casos de éxito como referencia
- En caso de ser un problema que requiere más recursos o de un cambio en la estrategia del proyecto, será necesario realizar un informe con las deficiencias y necesidades. Este será redactado por el líder del dato y llevado a discusión al consejo del dato

De la misma manera, el consejo del dato se reunirá también con la misma cadencia, después de la anterior reunión. En ella se tratará el avance de el proyecto y se discutirán las mejoras que requieran de cambios a nivel estratégico o de recursos. En caso de llegar a un acuerdo el líder del dato se encargará de llevar de vuelta estas decisiones a los data stewards.

1. Auditorías

Las auditorías son una de las herramientas clave para valorar la correcta implementación y uso del framework. Es una forma de garantizar que se cumplan los objetivos marcados, observando qué aspectos están progresando adecuadamente y cuales no han progresado o no están funcionando como deberían. Es el origen del que surgirán todas las acciones de mejora continua.

Las auditorías pueden realizarse en cuatro fases [12]:

1. Preparación: implica la definición de objetivos, alcance y plan de acción
2. Participación de los interesados: definición de los roles, identificación de todos los interesados
3. Ejecución: Se obtiene la información y se analiza
4. Resultados: Se presentan los resultados

De cara a realizar una buena auditoría también es necesario definir que preguntas son clave realizar durante las auditorías, The Institute of Internal Auditors sugiere varias para su uso en el ámbito de el gobierno del dato [13]:

- ¿Saben las personas afectadas que se está recopilando su información?
- ¿Se está inventariando toda la información que recopilamos?
- ¿Se están cumpliendo nuestros estándares para la calidad?
- ¿Son las soluciones técnicas correctas y adecuadas?
- ¿Es la seguridad actual suficiente?
- ¿Son las prácticas de gobierno del dato utilizadas suficientemente maduras?
- ¿Se está formando a los trabajadores en el aspecto ético de el gobierno del dato?

VIII - Calidad del dato

De cara a la calidad del dato se tratarán dos aspectos principales. El primero consiste en el proceso de selección de reglas, es decir las guías que servirán para seleccionar los indicadores de calidad del dato. La segunda parte consiste en el proceso de selección de los requisitos, que implica definir qué expectativas se busca cumplir en cada uno de los casos. Estos requisitos se definirán en indicadores; que pueden ser usados tanto como métrica para valorar como de bien se están cumpliendo los requisitos y, en un contexto con herramientas de calidad del dato que lo permitan, poder establecer avisos automáticos en caso de sobrepasar los umbrales que se establezcan.

1. Selección de las reglas

Para realizar un buen proceso de selección de reglas, se puede seguir un enfoque orientado a las dimensiones de la calidad [14], que pueden categorizarse como 60 dimensiones diferentes. Cada una de ellas representa una cualidad específica de los datos, y actúa de guía para especificar qué reglas se pueden aplicar en cada caso.

Algunas de las dimensiones son las siguientes:

Tabla 11: Dimensiones de la calidad del dato

Dimensión	Definición
Accesibilidad	Facilidad para consultar u obtener el dato
Exactitud	La cercanía de los valores de los datos a valores reales
Coherencia	La facilidad con la que dos datasets pueden combinarse
Credibilidad	El grado en el que se puede confiar en que un dato sea cierto
Equivalencia	El grado en el que dos datos de dos sets de datos diferentes son iguales conceptualmente
Naturalidad	El grado en el que los datos se alinean con los objetos reales que representan
Relevancia	El grado en el que el dato cumple con las expectativas del consumidor del dato
Redundancia	Grado en el que cada registro ocurre más de una vez

Para poder relacionar un dato con sus dimensiones relevantes, se puede seguir un proceso de 3 pasos

1. Determinar qué dimensiones son relevantes para el dato
    1. Para la selección de posibles dimensiones en [15] se propone una correlación de dimensiones con categorías de datos a las que son aplicables
    2. Por lo que es buena idea realizar el proceso inverso, analizar qué categoría de dato es con el que se está trabajando, y realizar un listado inicial de las dimensiones que le pueden ser aplicadas
2. Determinar de cada una de las dimensiones, si contribuyen a algún objetivo de negocio
    1. En este paso se progresa de una lista inicial de dimensiones que son aplicables a dimensiones que aportan valor. Las dimensiones que se considere contribuyen suficiente serán las que se utilizarán
3. Priorizar las dimensiones
    1. Una vez seleccionadas las dimensiones que se van a trabajar se priorizan en base a el

ratio de coste-beneficio esperado

2. Toda acción conlleva sus costes, por lo que la prioridad ha de ser con los recursos que se tienen maximizar los beneficios a obtener

Una vez se ha formado la lista de dimensiones, el siguiente paso es la creación de indicadores que nos ayuden a cumplir con los requisitos que se busquen.

2. Requisitos

Los requisitos representan el cumplimiento de indicadores, cada indicador está formado por un valor y una unidad, siendo las unidades posibles las siguientes:

Tabla 12: Unidades de indicadores de calidad del dato

Unidad	Significado
Porcentaje	Representa la fracción de casos que han cumplido cierta condición
Numérico	Un valor numérico absoluto
Calificación	Es la percepción de las personas, en al escala que sea conveniente
Booleano	Valor binario sí/no
Duración	Especificada en una unidad de tiempo
Historia	Para valores complejos que no se pueden expresar con un sólo número, sino que que se explican en texto

De estas unidades,en [15] se realiza una vinculación de las diferentes dimensiones con sus unidades. Siguiendo ese ejemplo las dimensiones de ejemplo anteriores se regirían usando las siguientes unidades:

Tabla 13: Correlación dimensiones de la calidad del dato con unidades

Dimensión	Unidad
Accesibilidad	Calificación
Exactitud	Porcentaje
Coherencia	Historia
Credibilidad	Calificación
Equivalencia	Porcentaje
Naturalidad	Calificación
Relevancia	Historia
Redundancia	Numérico

Algunos ejemplos utilizando estas dimensiones:

- Accesibilidad se podría valorar en qué calificación promedio le dan los consumidores a la facilidad con la que han accedido a la información que buscaban
  - El indicador podría ser si se valor sobre 5, mantener esa calificación igual o por encima del 4/5
- La redundancia se mide en el número de veces que se guardan los mismos registros lógicos en dos lugares de almacenamiento diferente
  - El indicador en este caso puede ser comparar el número obtenido con el número de copias que se quieren tener por seguridad
  - Un valor mayor es incorrecto, ya que se está causando un gasto de recursos innecesario
  - Un valor inferior es incorrecto, ya que hay datos a riesgo de perderse

IX - Regulaciones

La ley principal en España que regula el dato es la "Ley Orgánica de Protección de Datos Personales y garantía de los derechos digitales"[17], que así mismo se trata de una adaptación al ámbito nacional de la ley de la Unión Europea "Reglamento General de Protección de Datos"[7] (GDPR por sus siglas en inglés).

Este marco define muchos de los aspectos que ya hemos tratado, como son:

- Qué se considera un dato privado
- Necesidad de información la persona sobre qué datos suyos se utilizan y cómo se utilizan.
- Necesidad de consentimiento para recopilar información no estrictamente necesaria para el funcionamiento básico de los servicios aportados
- Limitaciones a la hora de compartir datos personales con otras organizaciones
- El derecho al olvido, el poder de las personas a exigir que sus datos sean eliminados
- Aplicación de medidas de seguridad suficientes, como puede ser la anonimización o seudonimización de los datos
- Sanciones en caso de incumplimiento
- Exenciones, como puede ser para cumplimiento de otra ley, uso personal o seguridad nacional
- Al ser una ley se trata por lo tanto de algo con un cumplimiento obligatorio, no pautas a usar de referencia sino unos estándares mínimos que han de ser cumplidos en todos los casos menos en las exenciones contempladas.

Sin embargo, no es la única ley que afecta al gobierno del dato, otro caso es el Reglamento - UE – 2024/1689 [16] que prohíbe ciertas actividades dañinas llevadas a cabo con inteligencia artificial (IA). Dos artículos en particular son directamente relacionados el gobierno del dato:

- Artículo 31: prohíbe el uso de sistemas de IA para llevar a cabo una puntuación ciudadana, para evitar discriminaciones. Esto indica que hay que tener mucho cuidado al recopilar y agregar información personal que pueda considerarse como una puntuación ciudadana.
- Artículo 38: prohíbe el uso de sistemas de IA para la identificación biométrica remota en tiempo real de personas físicas. Este artículo afecta en especial a el ámbito del IoT ya que la captura de esta información de la que se alimentaría el sistema de IA suele hacerse mediante dispositivos IoT.

A diferencia de la regulación GDPR, no es que estos datos tengan especial protección, sino que su captura está completamente prohibida a no ser que se trate de alguna de las excepciones explícitamente citadas en la ley, como puede ser el artículo 25 que añade una excepción para actividades estrictas de investigación y desarrollo científico.

X - ESPECIFICACIONES

Las especificaciones a diferencia de las leyes, no son de aplicación obligatoria, pero siempre es una buena práctica el utilizar estándares de referencia. En España los estándares de referencia son las especificaciones UNE [18] (Una Norma Española). En particular en el ambiente de gobierno del dato las aplicables son las siguientes:

- UNE 0077:2023 Gobierno del dato
- UNE 0078:2023 Gestión del dato
- UNE 0079:2023 Gestión de la calidad del dato

En el ámbito internacional las especificaciones más conocidas son las ISO, específicamente las que afectan al gobierno del dato:

- ISO 8000
- ISO/IEC 11179
- ISO/IEC 38505

1. UNE 0077:2023

Es una especificación a nivel estratégico, diseñada para guiar la toma de decisiones a la hora de definir los objetivos del proyecto de gestión del dato. Por lo tanto se tendrá en cuenta tanto los objetivos a nivel general del proyecto, como las definiciones de las políticas que habrán de ser implementadas. Esto implica que es del dominio de todos los roles del consejo del gobierno del dato.

Esta especificación está dividida en 5 procesos [19].

Tabla 14: Procesos UNE 0077:2023

Proceso	Carácterísticas
Establecimiento de la estrategia del dato	Es el primer paso, en el que se identifican las capacidades de la organización, los datos usados por la misma, los contextos en los que se usan y se elabora un plan general de gobierno del dato
Establecimiento de políticas, buenas prácticas y procedimientos del dato	Es el paso en el que se definen las políticas de gobierno del dato que se van a implementar, las regulaciones y estándares que nos afectan.
Establecimiento de estructuras organizativas para el gobierno, gestión y uso del dato	Se analiza la estructura de la organización y como el uso del dato está dividido entre las diferentes partes. También se define el equipo, roles y responsabilidades del proyecto de gobierno del dato
Optimización de los riesgos del dato	Identificación, evaluación, mitigación de los riegos del dato, definición de las políticas de seguridad y privacidad
Optimización del valor del dato	Consiste en la maximización del valor de los datos, alineándolos los objetivos estratégicos de la organización y realizando una buena gestión en su uso

2. UNE 0078:2023

Mientras que UNE 0077:2023 era de carácter más estratégico, esta resalta más los procesos relacionados con las acciones que se van a llevar a cabo para realizarlos. Es decir por ejemplo de una política que se ha decidido implementar, que acciones va a realizar su responsable para que esta se cumpla. Esta especificación por lo tanto tendrán que tenerla más presente los integrantes de los grupos de áreas.

La especificación define 13 procesos a tener en cuenta [20], cada uno relacionado con un aspecto técnico de la gestión del dato:

1. Procesamiento del dato
2. Gestión de la infraestructura tecnológica
3. Gestión de requisitos del dato
4. Gestión de la configuración del dato
5. Gestión de datos histórico
6. Gestión de seguridad del dato
7. Gestión del metadato
8. Gestión de la arquitectura y diseño del dato
9. Compartición, intermediación e integración del dato
10. Gestión del dato maestro
11. Gestión de recursos humanos
12. Gestión del ciclo de vida del dato
13. Análisis del dato

3. UNE 0079:2023

Este estándar define 4 procesos que hemos incorporado en nuestro proyecto [21], todos ellos dedicados a la mejora continua del dato, al ser una mezcla de procesos estratégicos y técnicos algunas veces serán más del ámbito de unos roles u otros del equipo:

Tabla 15: Procesos UNE 0079:2023

Proceso	Carácterísticas	Tipo
Planificación de calidad del dato	Es la planificación llevada a cabo para cumplir con el resto de los procesos. Se especifican objetivos, responsabilidades y las acciones a seguir	Estratégico
Control y monitorización de calidad del dato	Es la monitorización continua del dato	Técnico
Aseguramiento de calidad del dato	Se definen los procesos para identificar los datos que no estén cumpliendo con los requisitos mínimos decididos	Estratégico
Mejora de calidad del dato	Se interviene en caso de que no se estén cumpliendo los mínimos de calidad	Técnico

Los procesos identificados como estratégicos van a ser responsabilidad principalmente de los data stewards, mientras que los custodios serán los principales responsables de los técnicos.

4. ISO 8000

Se trata de un estádar internacional para la calidad del dato, estructurado en cuatro partes [22]:

1. ISO 8000-1, ISO 8000-2 e ISO 8000-8: Conceptos generales de la calidad de los datos
2. ISO 8000-6x: Procesos de gestión de la calidad de los datos
    1. ISO 8000-60: Visión general de los procesos de mejora continua
    2. ISO 8000-61: Modelo de referencia para la implementación de la calidad de datos
    3. ISO 8000-62: Evalua la madurez de la gestión de calidad
3. ISO 8000-100 a ISO 8000-150: Aspectos relacionados con el intercambio de datos maestros entre organizaciones
4. ISO 8000-311: aplicación de la calidad de los datos de producto

5. ISO/IEC 11179

Es el estándar internacional que documenta la estandarización de los metadatos. Define los conceptos a seguir al modelar los metadatos y como se han de compartir. Cuenta con 6 partes [23]:

1. Ofrece una visión de alto nivel, define el concepto de metadatos
2. Define procesos para clasificar objetos y esquemas
3. Define un registro de metadatos
4. Especifica requisitos y recomendaciones para crear metadatos
5. Especifica como nombrar e identificar elementos
6. Define como registrar dichos elementos y asignarles un identificador

6. ISO/IEC 38505

Este estándar pone énfasis en los principios de gobierno del dato a nivel organizativo, con el objetivo de alinear el gobierno del dato con los objetivos de negocio. Se guía por 6 principios [24]:

1. Responsabilidad: Funciones y responsabilidades de cada persona
2. Estrategia: Alinear prácticas de gobernanza con objetivos estratégicos de la organización
3. Adquisición: Obtener los datos de forma que cumpla con los requisitos éticos, legales y organizativos
4. Rendimiento: Supervisión del rendimiento para ver si se cumplen los objetivos
5. Conformidad: Cumplir requisitos legales y organizativos (en el sentido de los requisitos impuestos por la ley y los requsitos impuestos por la organización)
6. Comportamiento humano: Los aspectos humanos del gobierno del dato y fomentar una cultura de gobierno del dato

XI - Clasificación del dato

Expandiendo los conceptos tratados en el apartado de políticas de privacidad, dividiremos los datos en 4 categorías, dependiendo de su grado de privacidad y de si se trata de información interna de la organización. La privacidad se definirá según la regulación GDPR [7], mientras que la información interna será por decisión de los encargados del negocio.

1. Personal

Se considerará personal la información que cumpla las siguientes carácterísticas, siguiendo su definición en el Artículo 4 (1) del GDPR [25]

- Es información relacionada a una persona física
  - Es decir, excluye a las personas jurídicas
- Que identifican a dicha persona
  - De manera directa
  - De manera indirecta, mediante la unión de varios datos
  - En particular hace especial hincapié en identificación mediante
    - Nombre
    - Número de identificación
    - Localización
    - Identificador online
    - Direcciones IP
    - O la unión de datos que sean específicos a una persona, como puede ser características físicas, fisiológicas, genéticas, mentales, económicas, culturales, o sociales
- Información que es tanto objetiva, como subjetiva (por ejemplo, opiniones o pareceres). Aunque sí es cierto que en el ámbito de las IoTs la mayoría va a ser información objetivo, como puede ser la recogida mediante diferentes sensores
- Es relacionada a una persona viva
  - Se considera que el derecho se aplica desde el nacimiento hasta la muerte de una persona

La ley hace referencia a "cualquier información", por lo que ha de interpretarse de la manera más amplia posible, en caso de duda se considerará personal.

Una vez identificada la información que cumple dichas condiciones se le asignan diferentes requisitos:

Tabla 16: Requisitos información personal

Contexto	Requisito
Tipo de almacenado	Siempre en servidor, nunca en almacenamiento extraíble (USBs, discos duros portátiles)
Localización física	Siempre dentro de la organización
Encripción	Siempre
Control de acceso	Estricto

Destrucción de la información	Bajo solicitud
Prevención de pérdida	Sí
Puede ser compartida con terceros	No
Logging	Sólo si se anonimiza previamente

2. Sensible

Es un nivel por encima de la información personal, que como su nombre indica se trata de datos que pueden ser más sensibles.

Se considera que es información sensible cuando:

- Cumple las condiciones para ser información personal
- Y encima es de una de las siguientes categorías:
- Genética
  - El Artículo 4 (13) del GDPR [25] la define como información de características genéticas que ha heredado o desarrollado una persona, en particular la información obtenida mediante el análisis de una muestra biológica
- Biométrica
  - El Artículo 4 (14) del GDPR [25] la define como información relacionada a las características físicas, fisiológicas o de comportamiento de una persona
- De salud
  - El Artículo 4 (15) del GDPR [25] la define como información relacionada a la situación de salud física o mental de una persona, incluyendo las prestaciones de servicios de salud
- Información personal que pueda revelar raza u origen étnico, opiniones políticas religiosas o ideológicas o que revelen pertenencia a sindicatos.

Por lo tanto se establecen una condiciones aun más restrictivas que con la información personal no sensible:

Tabla 17: Requisitos información sensible

Contexto	Requisito
Tipo de almacenado	Siempre en servidor, nunca en almacenamiento extraíble (USBs, discos duros portátiles)
Localización física	Siempre dentro de la organización
Encripción	Siempre
Control de acceso	Muy estricto
Destrucción de la información	Bajo solicitud
Prevención de pérdida	Sí
Puede ser compartida con terceros	No
Logging	Nunca

3. Interna

Se trata de información que no es personal, pero que por decisiones de negocio no puede ser accedida por organizaciones externas de forma libre. Su clasificación se hace mediante la decisión de los responsables de negocio.

Su seguridad no compromete el cumplimiento de las regulaciones, pero sí que ha de protegerse para no afectar negativamente a la organización. Por lo tanto tendría ciertos requisitos, aunque menos estrictos que en el caso de información personal.

Tabla 18: Requisitos información interna

Contexto	Requisito
Tipo de almacenado	En servidor o equipos profesionales de los trabajadores En almacenamiento extraíble sólo en caso de estar propiamente cifrada
Localización física	El almacenamiento principal se realizará tanto en servidores propios de la organización como en datacenters contratados
Encripción	Necesaria sólo en caso de usar medios de almacenados extraíbles
Control de acceso	Medio
Destrucción de la información	Según necesidades de negocio
Prevención de pérdida	Sí
Puede ser compartida con terceros	Sí, pero bajo permiso del propietario del dato
Logging	Permitido

4. Abierta

La información abierta o pública es la que ni es personal ni se considera que necesariamente ha de permanecer dentro de la organización. Su clasificación se hace mediante la decisión de los responsables de negocio.

Al ser abierta, los requisitos son los más laxos de todos.

Tabla 19: Requisitos información abierta

Contexto	Requisito
Tipo de almacenado	Cualquiera
Localización física	El almacenamiento principal se realizará tanto en servidores propios de la organización como en datacenters contratados
Encripción	No necesaria
Control de acceso	Medio
Destrucción de la información	Según necesidades de negocio
Prevención de pérdida	Sí, aunque se considera que puede ser compartida, no significa que su pérdida no sea un contratiempo
Puede ser compartida con terceros	Sí
Logging	Permitido

XII - Titularidad del dato

La titularidad del dato es un tema de especial complejidad en el ámbito del IoT, y no es un debate que tenga un consenso concreto a nivel global ni regional [26].

Para discutir sobre la titularidad del dato, es importante distinguir 3 tipos de actores:

- Los generadores del dato, que en el contexto de sensores y cualquier otra información generada por máquinas sería la organización
- Las personas relacionadas a información, en caso de que la información sea de carácter personal. Es decir que un dato o conjunto de datos identifique a una persona segun lo definido en el apartado anterior
- Otros participantes, pueden ser los fabricantes de dispositivos IoT que utilizan sus propias plataformas para transmitir o almacenar los datos

Dado que no existe un framework estandarizado para la titularidad, se han de definir unas pautas a seguir para guiarnos:

1. Aunque la titularidad puede ser nuestra, hay ciertos derechos que las personas tienen sobre su información personal. Por lo que se ha de hacer un cumplimiento estricto de los derechos para con estas personas:
    1. Esto implica informar del uso y obtener el consentimiento cuando es necesario
    2. El derecho a las personas a obtener la información relacionada a ellas
    3. El derecho a solicitar que dicha información sea borrada, por lo que tiene que haber mecanismo para eliminar todos los datos que la componen
2. Identificar hasta qué punto pueden tener fabricantes y otros terceros a los datos que recogemos
    1. Priorizar siempre en la manera de los posible sensores que no utilicen las propias redes del fabricante, o de fabricante que garanticen no almacenar nada
    2. Analizar el tipo de dato que puede ser almacenado por dicho fabricante, valorar si puede identificar a personas físicas o si es considerado interno de la organización
    3. Analizar el tipo de uso que se va a dar en caso de que sí se recopile información, valorar si es aceptable, por ejemplo, utilizar datos que nosotros consideramos "abiertos" para mejora de sus servicios
3. Identificar si hay más puntos entre el punto de captura del dato y el lugar de almacenado final en el que pueda estar implicado un tercero
4. Reevaluar los contextos y condiciones bajo las que se comparte información con otras organizaciones

XIII - Herramientas

A continuación se exploran herramientas que permiten implantar y mantener una gobernanza de datos eficaz. Las herramientas a tratar van a ser de tipo ETL, las herramientas de calidad de datos y los catálogos de datos.

Cada una de estas herramientas cuenta con un objetivo concreto en el ámbito de la gobernanza de datos:

- Las herramientas ETL facilitan la integración y transformación de fuentes de datos dispares, garantizando que los datos se consoliden y estructuren
- Las herramientas de calidad de datos están diseñadas para ayudar a las organizaciones a mantener datos de alta calidad, que son fundamentales para tomar decisiones informadas, cumplir los requisitos normativos e impulsar el rendimiento empresarial
- Los catálogos de datos actúan como repositorios que proporcionan una visión centralizada y organizada de la información de una organización

1. Herramientas ETL

ETL, de sus siglas en inglés "Extract Transform Load" ("Extraer Transformar Cargar") es un proceso vital en el mundo de las IoT. Consiste principalmente en reunir datos de distintas fuentes, realizar un proceso para que queden uniformes y almacenarlos donde se considere oportuno. Es el único modo de tratar con orígenes de datos heterogéneos, que se dan tanto en los IoT, y poder utilizar dichos datos de forma conjunta posteriormente.

El procesado ha de ser guiado por un data steward que será quién especifique que forma final ha de tener los datos, es decir, qué estándares y condiciones tienen que cumplir. Por lo tanto será el trabajo de personas de perfil más técnico, como son los custodios el adaptar cada uno de los orígenes. Para ello se realizan los siguientes pasos:

1. Identificar todos los orígenes que han de ser agrupados
2. Identificar las salidas de cada uno de ellos
3. Definir las operaciones que han de realizarse sobre cada uno de ellos para transformarlos a lo que el steward ha especificado
4. Crear las entradas y operaciones identificadas y definidas anteriormente en la herramienta de ETL

1.1. Adeptia Connect

Adeptia Connect es una herramienta orientada a maximizar la facilidad de uso. Ofrece opciones para usuarios avanzados y con conocimientos más técnicos, pero busca ofrecer la mayoría de su funcionalidad como "no code" es decir puramente mediante su interfaz gráfica.

Es de licencia comercial, y su uso requiere de un contrato activo con Adeptia, no ofrecen opción de self-hosting. Toda la gestión se hace desde su plataforma, y sí ofrece algunos beneficios como el trabajo colaborativo de varias personas al mismo tiempo.

Ofrece procesamiento a tiempo real, además de algunas herramientas de monitorización.

1.2. Apache NiFi

Apache NiFi es una herramienta open-source, con soporte para procesamiento en vivo y opciones de procesado dinámico. De esta forma es posible realizar operaciones diferentes a los mismos orígenes dependiendo de condiciones establecidas.

Al ser open-source, sí dan la opción de gestionarla uno mismo, y aunque puede que no sea tan avanzada en el aspecto de la interfaz de usuario, sí que ofrece la mayoría de su funcionalidad directamente desde la interfaz web. Por lo que es posible funcionar correctamente con un equipo mixto de personas técnicas que se encarguen de el despliegue y funcionamiento del sistema, mientras que los menos técnicos se encargan de el uso de la herramienta.

1.3. IBM InfoSphere Information Server

Es una herramienta multiusos de IBM, que cuenta con su sección de ETL, aunque también busca ser una herramienta general para Data Warehousing y Business Intelligence, ofreciano la habilidad de gestionar la calidad del dato, análisis y auditoría.

Está optimizada para su uso conjunto con otras herramientas del ecosistema IBM.

Puede ser utilizado desplegándolo en la organización mismo o en la nube.

1.4. Oracle Data Integration Cloud Service

Se trata de una plataforma completa, similar a la herramienta anterior de IBM, que busca unificar diferentes soluciones para todos los aspectos del gobierno del dato, Business Intelligence o migración de datos.

También se integra con las herramientas del ecosistema Oracle.

Es una herramienta orientada a su uso puramente en la nube, ya que la estrategia de Oracle ha sido la de desarrollar todo su ecosistema en la nube.

1.5. Comparativa

Realizamos la comparativa, principalmente valorando el tipo de licencia, precio, opciones para despliegues en la propia organización o en la nube e integraciones con herramientas del propio ecosistema.

Tabla 20: Comparativa herramientas ETL

Herramienta	Licencia	Precio	Uso "in-premise"	Hosting en la nube	Ecosistema
Adeptia Connect	Comercial	Bajo acuerdo	Sí	Sí, nube propia o de terceros	No cuenta con ecosistema propio
Apache NiFi	Apache License, Version 2.0	Gratuito, a coste de realizar el mantenimiento nosotros	Sí	Sí, nube de terceros	Ecosistema Apache
IBM InfoSphere Information Server	Comercial	Bajo acuerdo	Sí	Sí, nube propia o de terceros	Ecosistema IBM
Oracle Data Integration Cloud Service	Comercial	Bajo acuerdo	No	Sí, nube propia	Ecosistema Oracle

La elección de herramienta dependerá bastante de las capacidades técnicas del equipo, para

equipos con pocos recursos de este tipo o que consideren que no pueden dedicarle los recursos necesarios, podrían considerar que la selección más correcta es alguna de las 3 que ofrecen soporte. Mientras que si nos consideramos habilitados para realizar esa gestión, la opción de Apache puede ser más adecuada, que es la que consideramos para nuestro caso. La elección de herramienta dependerá bastante de las capacidades técnicas del equipo, para equipos con pocos recursos de este tipo o que consideren que no pueden dedicarle los recursos necesarios, podrían considerar que la selección más correcta es alguna de las 3 que ofrecen soporte. Mientras que si nos consideramos habilitados para realizar esa gestión, la opción de Apache puede ser más adecuada, que es la que consideramos para nuestro caso.

La opción de integrarlos con servicios en la nube lo consideramos un punto a favor, pero sin embargo lo que se considera crítico es la opción de hosting en las premisas de la organización. Es la forma más sencilla de garantizar que ciertos datos no salen en ningún momento de nuestra organización.

Por lo tanto, se selecciona Apache NiFi como la solución más correcta en este caso, con la condición de que será necesario dedicar ciertos recursos a su gestión.

1.6. Usando Apache NiFi

NiFi funciona mediante lo que denomina "processors" que son las unidades que definen orígenes, transformaciones y salidas.

Un flujo de uso tal como se define en el manual [27] podría ser el siguiente, dividido para mostrar las 3 fases del proceso Extract-Transform-Load:

1. Extracción: Crear un nuevo processor de tipo como de "data ingestion"
    1. Estos son los processors que formarán la entrada de todos nuestros datos
2. Transformación: procesadores que distribuyen los datos y los transforman
    1. Crear los procesadores dedicados a mover los datos
        1. De tipo "Routing" o "Mediator", es decir, que enrutan de un punto a otro
        2. De tipo "Attribute extraction", para extraer atributos concretos
        3. De tipo "Splitting" o "Merging", para separar atributos del mismo origen en varios, o de varios a un solo punto
    2. Crear los procesadores que transforman los datos en sí, de tipo "Data Transformation"
        1. Estos son los que realizan operaciones sobre los datos, como puede ser convertir de una unidad a otra
3. Carga:
    1. Procesadores que envían los datos ya transformados a su destino
        1. Tipo "Sending" para envío de datos a el servidor donde configuremos
        2. Tipo "Database Access" si modificamos directamente en base de datos

Se definirá también el "run duration" o el tiempo que se dedicará a la ejecución de los procesadores cada vez que se pongan en marcha.

Finalmente definimos el momento de ejecución de los procesadores, en la pestaña de "scheduling". La configuración principal es la estrategia de ejecución "Scheduling Strategy" que puede ser de 3 tipos:

- "Timer driven": ejecución continua cada X tiempo, para ejecución a tiempo real
- "Event driven": sólo se ejecuta bajo ciertas condiciones especificadas

- "CRON driven": ejecuciones planificadas por fecha y hora. Por ejemplo, todas las semanas el viernes a las 2pm

Otras configuraciones más técnicas para la ejecución son las de "concurrent tasks" y "execution". La primera especifica cuantos hilos puede ocupar la ejecución, mientras que la segunda especifica el nodo específico en el que se ejecutan. Estas opciones no son definidas por el data steward, sino que se han de crear unas directrices por parte del equipo técnico que mantiene la herramienta para no sobrecargarla y hacer un uso óptimo de la misma.

1.7. Prueba de concepto

De cara a crear una pequeño prueba de concepto para mostrar el uso de apache NiFI. Es vital antes de decantarse totalmente por una herramienta poder crear una prueba de concepto para validar la viabilidad de la herramienta, comprobar que cumple con los casos de usos que se van a dar y cuenta con todas las herramientas que se necesitan. Se recogerán los pasos seguidos para realizar cada uno de los pasos.

1.7.A. Instalación

Apache NiFi es una herramienta portátil por lo que no requiere de instalación como tal. Eso sí, es necesario tener instalada una versión de Java superior a 8. Una vez Java instalado y configurada la variable de entorno JAVA_HOME, el proceso de instalación es tan sencillo como extraer el archivo zip que Apache NiFi ofrece en su página.

Sí que de cara a documentar lo realizado, se ha editado la configuración en "./conf/nifi.properties" para que los diferentes directorios donde se guarda el trabajo queden controlados por el control de cambios, ya que por defecto se guarda mezclado con los archivos propios del funcionamiento de la herramienta. Se trata de las siguientes configuraciones:

Tabla 21: Directorios Apache NiFi

Directorio	Contenido
./database_repository/	Base de datos interna de NiFi
./flowfile_repository/	Datos que fluyen por el sistema de NiFi
./content_repository/	El contenido de los "flowfile" anteriores
./provenance_repository/	Es un log de las procedencias de los diferentes flujos
./flow_repository/	Las definiciones de los flujos
./assets/	Contenido creado por el usuario, como pueden ser procesadores creados
./nar_repository/	Metadatos de los contenidos

También es posible la instalación de NiFi como servicio, aunque sólo en sistemas Mac y Linux. De cualquier manera, al tratarse de una prueba de concepto, incluso existiendo la opción lo más adecuado es la instalación portable que NiFi ofrece por defecto.

1.7.B. Ejecución

Uno de los posibles pasos es la modificación del usuario por defecto, que se puede modificar con el siguiente comando, aunque por el momento se dejará el auto-generado.

```
./bin/nifi.sh set-single-user-credentials
```

El usuario generado lo podremos ver en "./logs/nifi-app.log", junto con el estado actual de la ejecución. El estado no se muestra en la ventana de la terminal, y es necesario consultarlo en el log.

Una vez arrancada la herramienta, se puede navegar a https://localhost:8443/nifi para acceder a la herramienta. En los navegadores modernos será necesario aceptar la conexión a una página no segura antes de continuar.

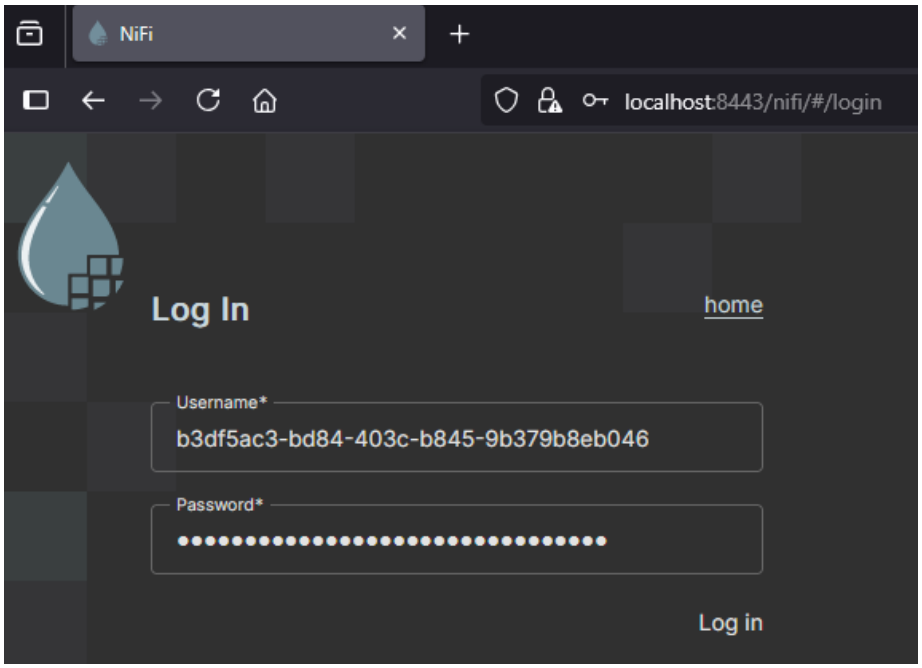

Ilustración 3: Pantalla de inicio NiFi

1.7.C. Extracción de datos

Para recoger los primeros datos, deberemos crear un procesador. El tipo más simple por el que podemos empezar es el de tipo "GetFile"

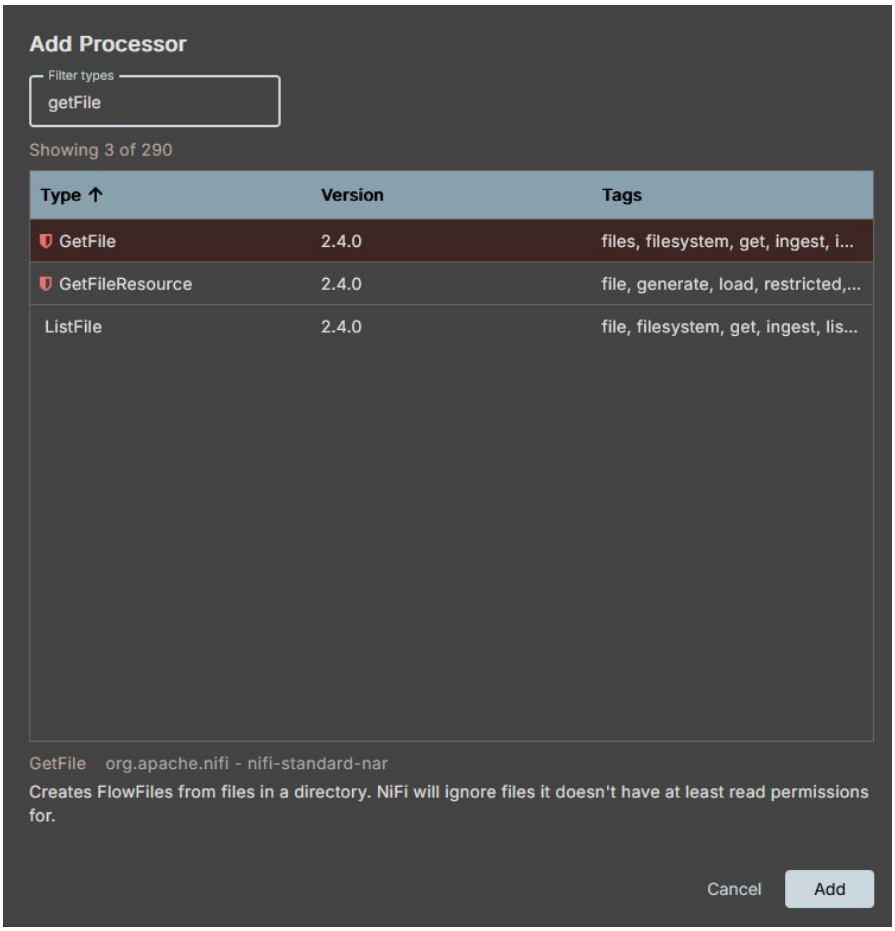

Ilustración 4: Procesador GetFile

Una vez generado podemos darle un nombre y configurar sus propiedades. Le podemos indicar el directorio donde va a leer los archivos, y el filtro que va a usar para decidir qué archivos recojer. En el siguiente caso, por ejemplo, recogerá todos los archivos ".csv" del directorio especificado.

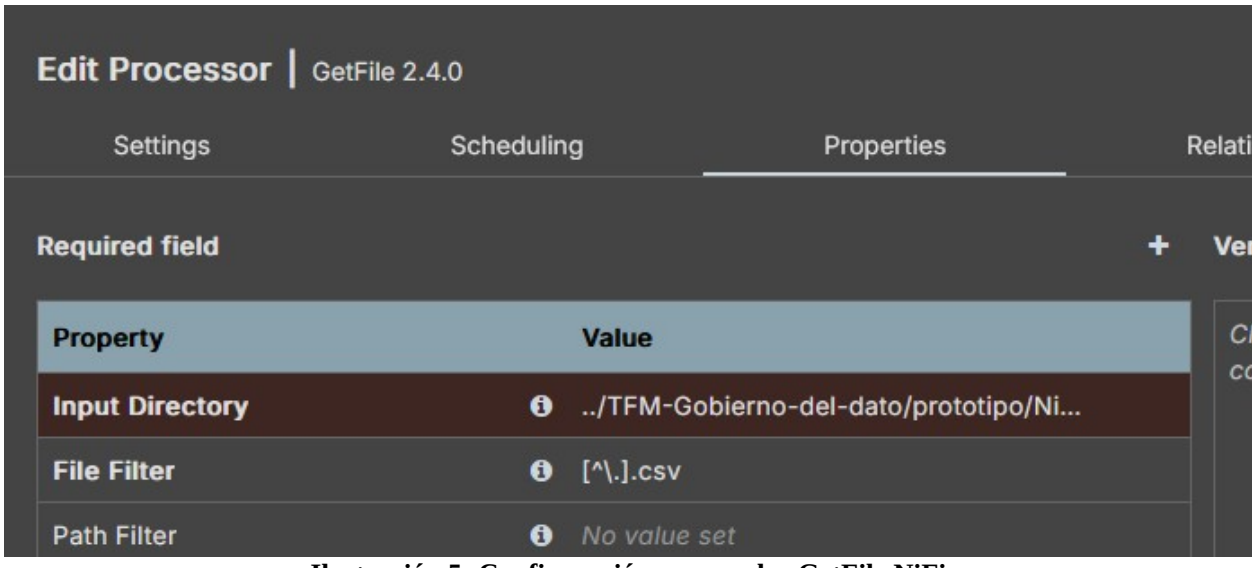

Ilustración 5: Configuración procesador GetFile NiFi

1.7.D. Enrutado

Para mostrar las capacidades del enrutado de NiFI, se crean varias situaciones que pueden solucionar con esta herramienta.

Se propone el caso 1, simulando dos orígenes que cuenten con entradas compatibles. Esto significa que se quieren enrutar a el mismo destino. Para ello se crea un procesador de tipo "RouteOnAttribute". Se modifica el anterior procesador GetFile, para que obtenga un sólo archivo csv, este simulará uno de los orígenes. Se crea un segundo archivo con la misma estructura pero diferentes valores. El archivo se encuentra en el mismo directorio, pero para simular un origen diferente se crea otro procesador GetFile que consuma ese único archivo. Siendo ambos compatibles, el objetivo es enrutarlos al mismo lugar, por lo que en el procesador de enrutado, se controlarán estos dos orígenes para un enrutado en particular:

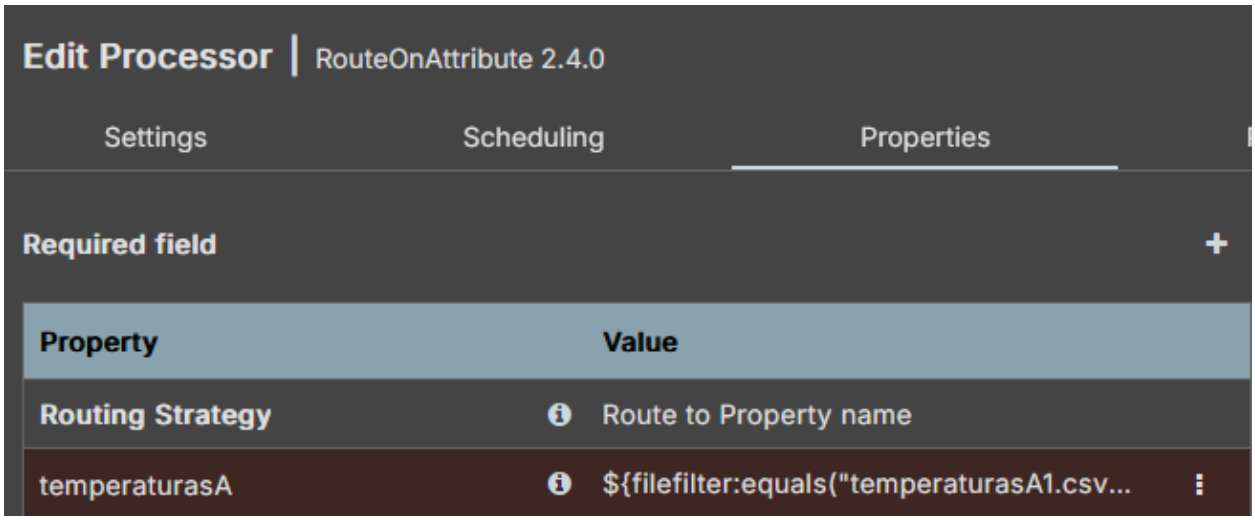

Ilustración 6: Configuración enrutado por atributo NiFi

El siguiente paso consiste en unir todos los procesadores, mediante la interfaz visual de NiFi.

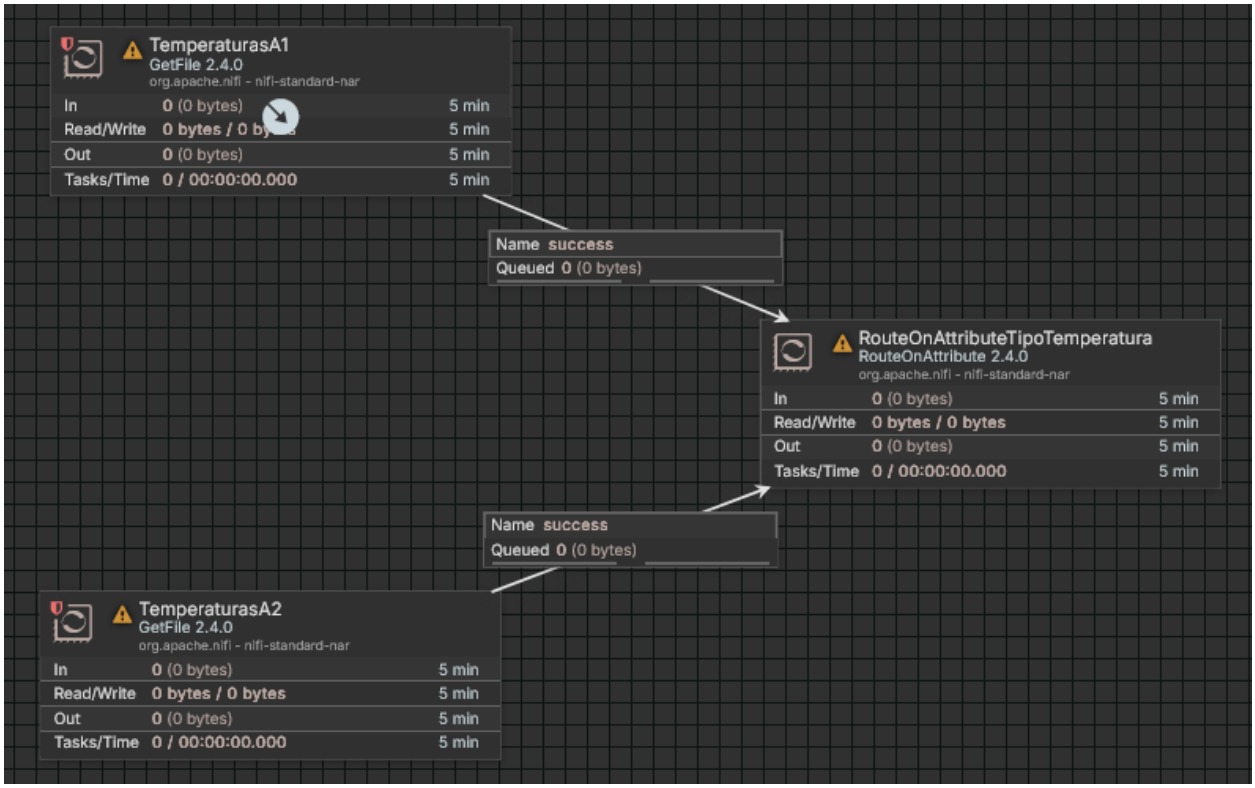

Ilustración 7: Enrutado inicial NiFi

A continuación se imagina un caso 2, con otro origen, que no sigue la misma estructura, sino que aunque sí que ha de acabar uniéndose a los anteriores deberá de tener un procesado diferente. Para ello se definirá otro procesador GetFile. Y de la misma manera se definirá una nueva propiedad en el enrutador para las entradas de tipo B. De esta forma la entrada queda de la siguiente manera:

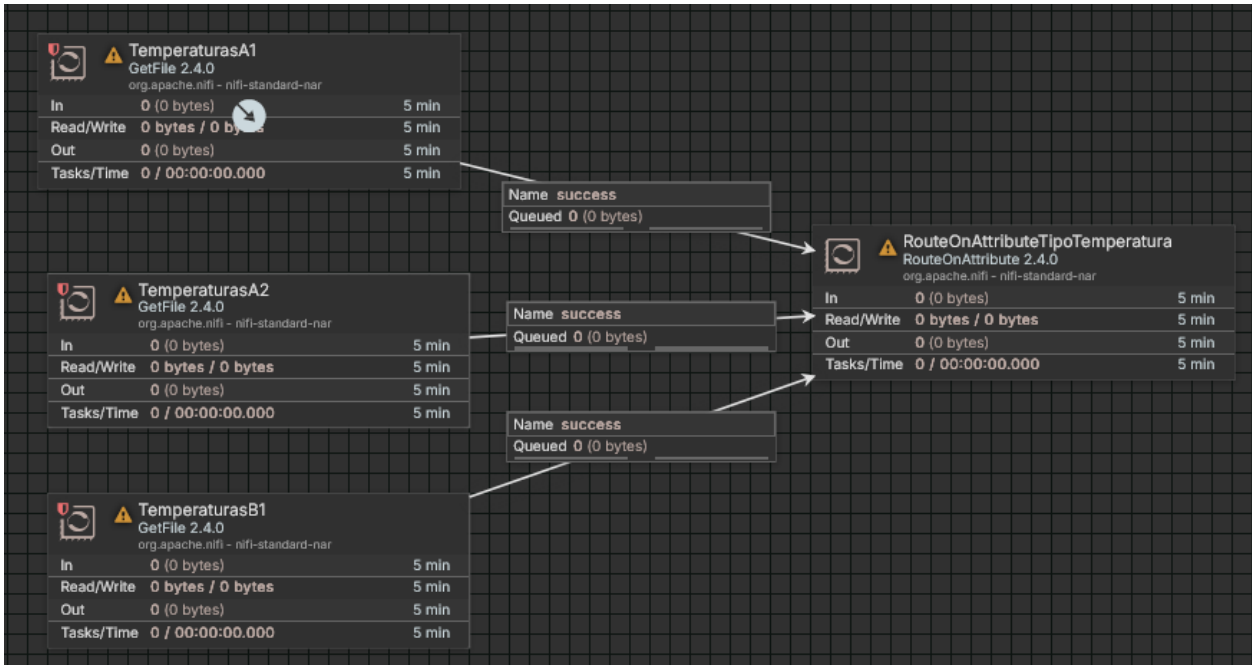

Ilustración 8: Enrutado complejo NiFi

Las propiedades de salida del enrutador tendrán que utilizarse a la hora de generar las conexiones al siguiente paso, que consiste en un procesador de tipo "SplitRecord" para separar las columnas de tipo A. Para ello se crearán 2 servicios de tipo RecordReader y RecordWriter, utilizando las plantillas que existen en NiFi para CSVs. Utilizaremos estos servicios el procesador para las temperaturas de tipo A.

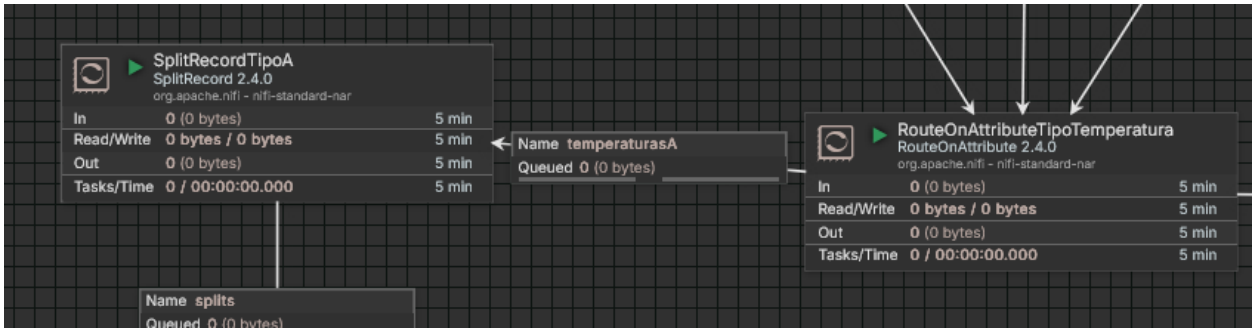

Ilustración 9: Enrutado a procesador transformador NiFi

Estos procesadores lo que harán será separar los archivos SCV en filas separadas, ignorando la primera fila que consta de los títulos.

1.7.E. Procesado

Para los casos propuestos, se ha decidido simular un caso de tener que estandarizar orígenes heterogéneos. Hasta ahora se ha tratado cómo capturar y enrutar los datos, pero no cómo realizar la equivalencia. En la siguiente tabla se muestran los campos de ambos orígenes y el estándar que se busca:

Tabla 22: Estructura entradas NiFi

	Columnas	Unidades
A	"id ciudad", "nombre ciudad", "temperatura", "fecha"	Temperatura en Celsius
B	"id ciudad", "temperatura", "fecha"	Temperatura en Fahrenheit
Objetivo	"id ciudad", "temperatura", "fecha"	Temperatura en Celsius

Por lo que para la casuística especificada las operaciones a realizar son las siguientes:

Tabla 23: Operaciones necesarias según origen

Origen	Operaciones
A	Eliminación de la columna "nombre ciudad"
B	Conversión de la columna "temperatura" a Celsius

La eliminación de la columna puede hacerse mediante un procesador "ReplaceText", para ello se asignan las diferentes propiedades:

Tabla 24: Propiedades para transformación de eliminar columna

Propiedad	Valor	Significado
Search Value	^(.*,)(.*,)(.*,)(.*)	Es una expresión regular que separa una fila de 4 columnas con separadores "," en 4 secciones.
Replacement Value	$1$3$4	Significa que la fila será cambiada por las secciones 1,3 y 4 especificadas antes, es decir, todo menos la sección 2

Para la modificación de la temperatura es necesario crear un sistema un poco más complejo. Para empezar, es necesario definir un esquema en el RecordReader. Por lo que se crea un nuevo servicio de este tipo con el siguiente esquema que se corresponde a la estructura del el CSV de tipo B:

```
{
	"type": "record",
	"name": "TemperaturaB",
	"fields": [
		{"name": "id", "type": "string"},
		{"name": "temperatura","type": "double"},
		{"name": "fecha","type": "string"}
	]
}
```

Una vez especificado el RecordReader se puede hacer referencia a las columnas, por lo que en las propiedades del procesador UpdateRecord se puede realizar un "Literal value" (Existe el RecordPathValue para actualizar los valores con otros valores de otras entradas del esquema) y realizar la operación de conversión de Fahrenheit a Celsius.

${field.value:toDecimal():minus(32):multiply(5):divide(9)}

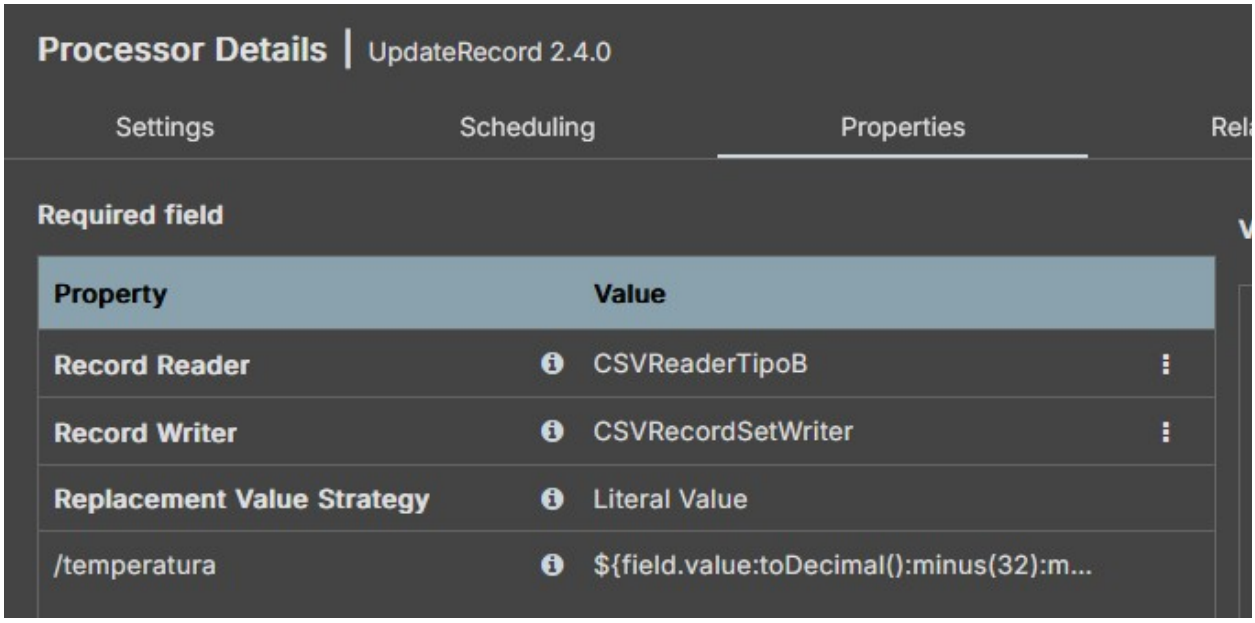

Ilustración 10: Propiedades transformación unidades

1.7.F. Unión

El siguiente paso una vez contamos con registros que son perfectamente compatibles los unos con los otros, el siguiente paso consiste en unirlos para poder consumirlos juntos.

Sería posible concatenarlos seguidos de cada uno de los orígenes, pero ya que ahora sí son perfectamente compatibles se pueden unir registro por registro (en vez de cada archivo todo junto), haciendo que datos de uno y otro puedan entremezclarse. Esto ya se había logrado con los de tipo A, pero también se realizará para los de tipo B

Para ello se utiliza un procesador de tipo "SplitRecord" que separará cada registro, de la misma forma que en el otro caso. Y posteriormente se unirán usando un procesador "MergeContent", de modo que es posible hacer que se vayan guardando los registro por ejemplo de 500 en 500 (para el caso de la prueba de concepto, en situaciones reales se manejarían muchos más registros de golpe). De esta forma se realizarán los dos casos: uno que divide en líneas y procesa cada línea, y otro que procesa todo los registros del archivo de golpe y luego divide.

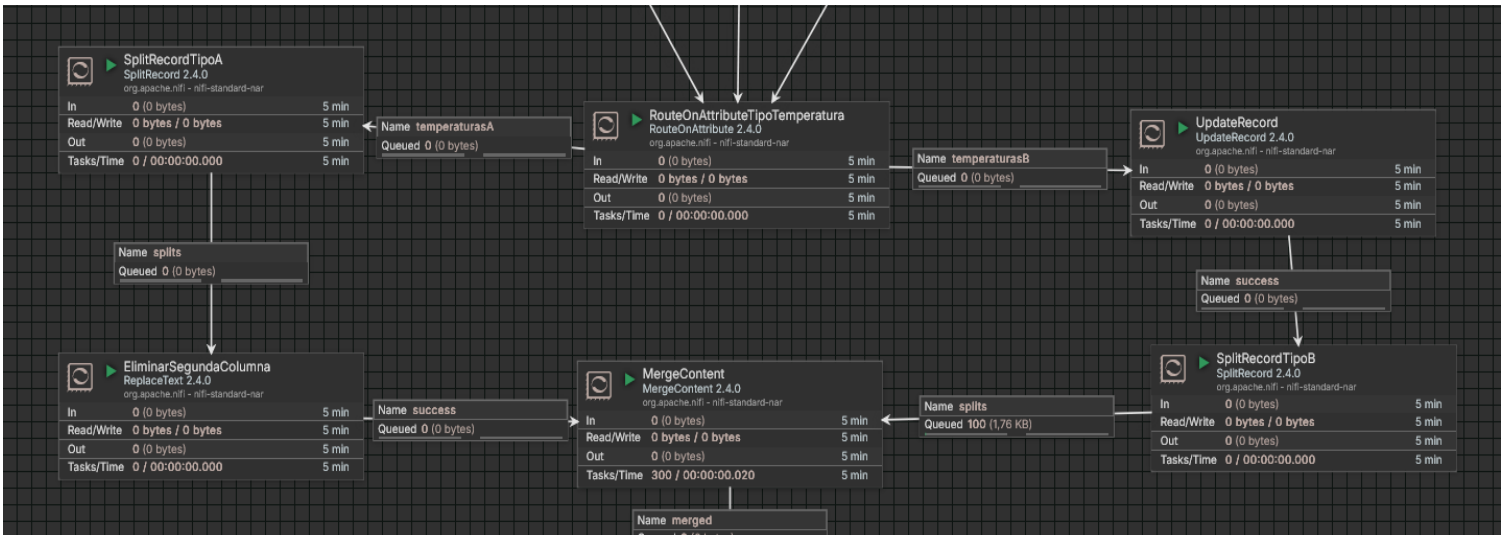

Ilustración 11: Flujo de procesado para dos orígenes

1.7.G. Guardado

Finalmente se sacarán los resultados del flujo. Para ello se realizan dos pasos. Primero se le asigna un nombre a el archivo que se va a generar, para evitar conflictos se utiliza la fecha-hora actual hasta milisegundos. Se utiliza un procesador de tipo "UpdateAttribute" con la siguiente propiedad.

Tabla 25: Propiedad para nombrar salidas

Propiedad	Valor
filename	${now():format("yyyy-MM-dd-HH-mm-ss.SSS")}.csv

Y la salida de este procesador se conecta a un procesador de tipo "PutFile", en el que se configura un directorio de salida. Sólo se configura el directorio de salida, y nada más, por lo que utiliza el atributo "filename" del recurso que le llegara, por eso para evitar colisiones era necesario el procesador anterior.

1.7.H. Ejecutando la prueba de concepto

Una vez realizado todo el flujo, y comprobado que no queda ningún procesador con errores, se puede comenzar a arrancarlos. Lo correcto en ste caso es dejar todos menos los 3 de entrada funcionando, y ejecutar de forma discreta las entradas para simular entradas de datos. A continuación se muestra el flujo completo:

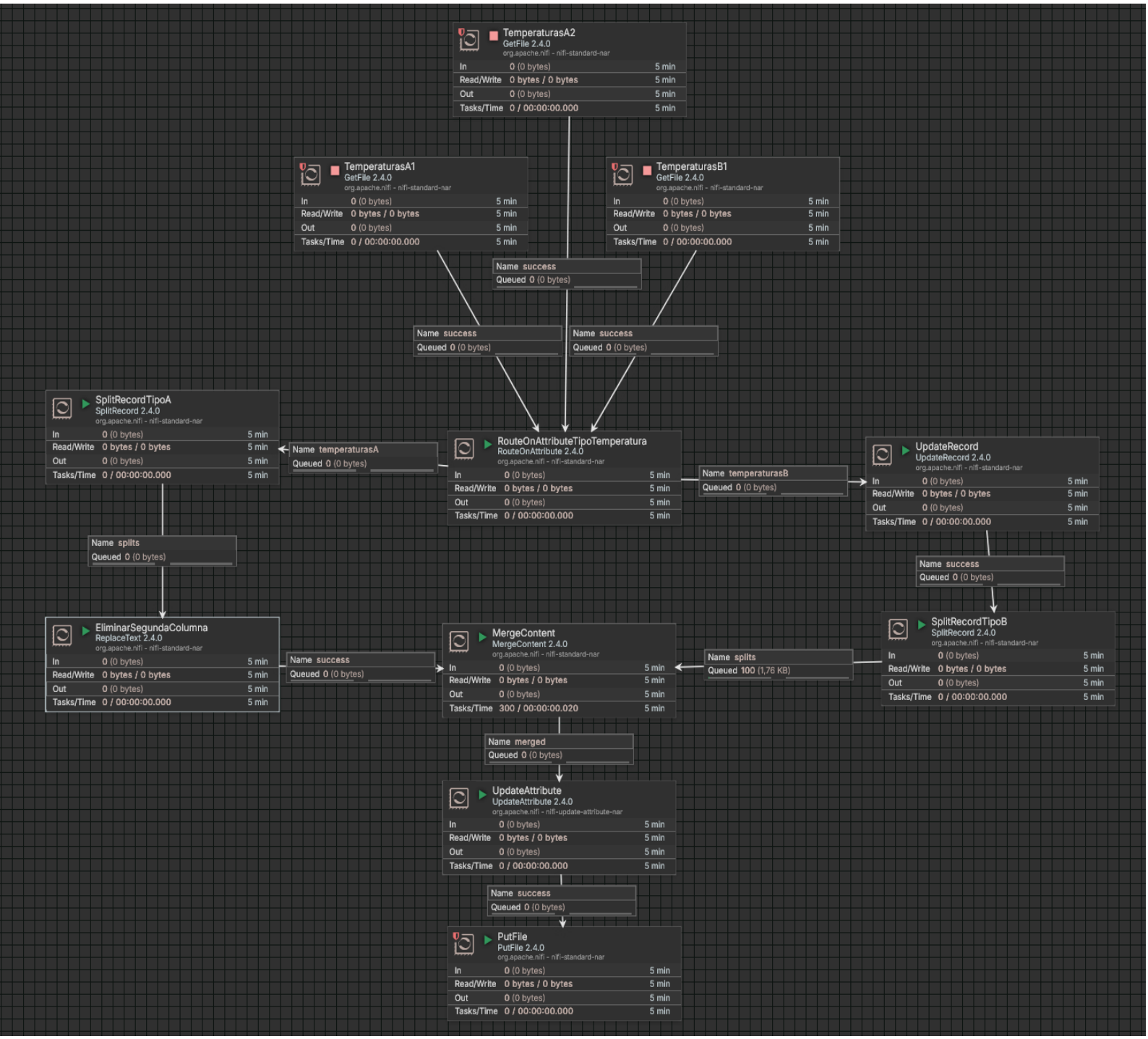

Ilustración 12: Flujo completo ETL

Si todo ha ido correctamente, se habrán grabado los resultados en el directorio indicado:

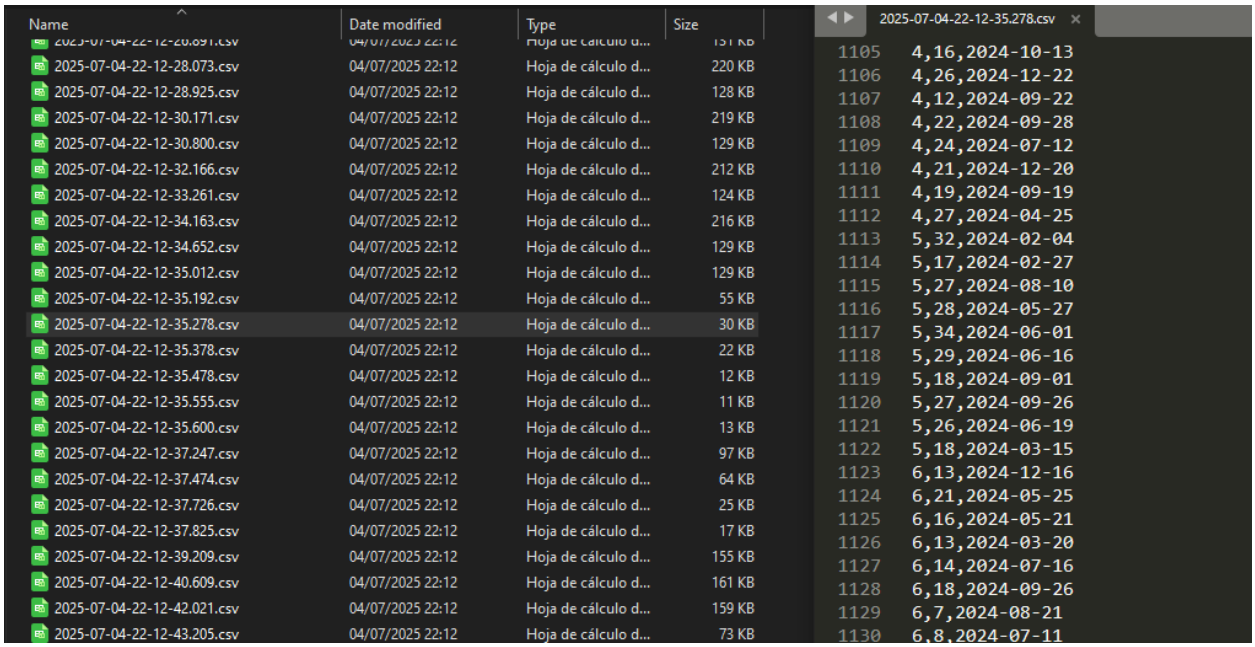

Ilustración 13: Salidas ETL

2. Herramientas de calidad del dato

De cara a utilizar una herramienta para gestionar la calidad del dato, ya que estamos utilizando Apache NiFi como herramienta ETL, optaremos por utilizarla también por sus cualidades de gestión de la calidad del dato para evitar redundancias con dos herramientas que se solapan. Esto se debe a que a pesar de que no se trate de una herramienta dedicada a la calidad del dato, cuenta con muchas funcionalidades que sí permiten la implementación de ciertos aspectos de la calidad del dato mediante sus capacidades de procesamiento, transformación y control de flujos. De todas formas también se considerarán otras herramientas por si en algún futuro se decidiera dejar Apache NiFi de lado y se necesitara alguna alternativa.

2.1. Alternativas

Existen tanto alternativas orientadas puramente a la calidad del dato como suites completas que intentan cubrir todos los aspectos del gobierno del dato. A continuación se mostrarán alternativas de ambos tipos.

2.1.A. Suites completas

Una opción en caso de decidirse por la opción de optar por ecosistemas completos serían los ecosistemas de las otras herramientas ETL consideradas. En caso por ejemplo de pivotar a el ecosistema de IBM, también nos ofrecería capacidades de calidad del dato mediante sus propias herramientas. Es también el caso con Oracle Data Integration Cloud Service, que como se ha especificado antes, busca ser una herramienta que cubra todas las necesidades de este ámbito.

2.1.B. Herramientas dedicadas

- OpenRefine, antes "Google Refine" es una herramienta dedicada exclusivamente a la calidad de la información, con especial énfasis en la eficiencia del análisis. Por lo que es una herramienta que destaca por ejemplo a la hora de identificar patrones en la información. Aunque es más limitada en el aspecto de la gestión de errores.
- Data Ladder, herramienta muy orientada a la limpieza de los datos. Sin embargo su muy limitado alcance y falta de documentación para sus características más avanzadas son un factor grande en su contra.

- Talend es una herramienta que potencia sus cualidades de análisis utilizando machine learning. Cuanta con varias rutinas de limpieza y de-duplicado muy eficientes, aunque su gran desventaja es la complejidad a la hora de usarse por parte de usuarios en perfiles poco técnicos.

2.2. Herramientas de data profiling

Las herramientas de data profiling son herramientas que automatizan y sistematizan el proceso de análisis del dato, Se podrían considerar un paso previo a utilizar las herramientas de calidad del dato, ganando conocimiento sobre datasets, patrones en los datos y descubrir perspectivas que quizás no se hubieran considerado sin el uso de estas herramientas.

Por lo tanto se pueden considerar herramientas que una vez se ha establecido una buena gestión de la calidad del dato son capaces mejorar esa buena base.

Algunos ejemplos de este tipo de herramienta son Alteryx Designer, Trifacta Wrangler, OpenText Magellan Data Discovery y Oracle Data Profiling.

2.3. Apache NiFi y la calidad del dato

Apache NiFi nos ofrece soporte para diferentes acciones clave de cara a la calidad:

Tabla 26: Acciones de calidad del dato Apache NiFi

Validación	Es posible realizar una validación continua, ya sea mediante esquemas predefinidos o lógica programada
Transformación	La acción de la transformación en sí mejora la calidad, garantiza una estandarización y consistencia de los datos.
Procedencia	NiFi muestra lo que llama el "linaje" de los datos, es decir muestra todos los pasos por los que transcurre desde el origen hasta la salida
Establecimiento de perfiles	O "data profiling", es la capacidad de NiFi de analizar la distribución de datos en el sistema, identificando patrones o posibles anomalías.
Limpieza	Parte de las operaciones que los "processors" pueden realizar son de limpieza de datos, como puede ser eliminación de duplicados o corregir errores
Gestión de errores	Permite gestionar qué hacer en caso de errores, con políticas de reintentos o gestiones específicas. De esta forma se puede definir qué hacer al darse un fallo de validación o un error de conexión
Monitorización	Permite monitorizar la salud y rendimiento de todos los aspectos del sistema, mostrando métricas de tasas de error, velocidades de procesamiento...
Logging	Alta customización de los logs del sistema, para poder identificar los fallos y trazar sus puntos de error

2.4. Incluyendo calidad del dato en la prueba de concepto

En esta sección se mostrarán algunos ejemplo de como es el proceso de incluir algunos procesos de calidad del dato en la anterior prueba de concepto de Apache NiFi.

En primer lugar se valoran las mejoras a la calidad que se están haciendo ya, ya que el proceso de transformación mismo añade calidad al dato. Por lo tanto hay 3 mejoras que ya se están realizando en el proceso ETL actual

1. Unión de datos equivalentes. De los 3 orígenes simulados se ha pasado a un sólo flujo, mejorando el descubrimiento de los datos, ya no es necesario consultar 3 lugares, sino que uno solo es suficiente

2. Estandarización del esquema. Anteriormente se trabajaba con dos esquemas diferentes, duplicando el trabajo de quien necesitara consultarlos, añadiendo la posibilidad de causar errores al no aplicar el esquema adecuado a alguno de los orígenes. Después de pasar por el flujo establecido todos los datos tienen la misma estructura.
3. Estandarización de unidades. Similar al caso anterior, antes se contaba con dos unidades diferentes para la misma información (en el ejemplo, la temperatura). Causando los mismos problemas, trabajo añadido y posible causa de errores si no se convertían las unidades antes de realizar comparaciones.

Aunque aún hay puntos que pueden ser incluidos para mejorar aún más la calidad. Quizás el caso más claro es en el de la validación de datos. Ya se incluye cierta validación al comprobar las entradas contra el esquema del CSV especificado, pero es posible ir más allá y controlar por ejemplo datos incorrectos. Una opción ya que se está tratando con datos de temperatura por ejemplo podría ser controla datos que son claramente incorrectos, para identificar posibles sensores que estén funcionando incorrectamente. Se pueden establecer valores máximos y mínimos, para identificar datos sospechosos que se salgan de esos valores.

Para ello se puede utilizar un procesador de tipo "ValidateCSV", que compruebe que los valores se encuentren entre los que se han especificado. El procesador ha de ser introducido entre el punto en el que las filas han sido separadas y el punto en el que se vuelven a juntar. Es posible validar un conjunto entero, pero esto implicaría que se consideraría válido o inválido en su conjunto. Al realizarlo antes de la unión de los datos es posible validarlos de forma independiente y realizar diferentes acciones sobre cada uno de ellos.

El la siguiente tabla se especifican las columnas que tiene el CSV final, y las condiciones que se han de implementar para validar si son correctos o no.

Tabla 27: Condiciones para validez del dato

Columna	Condiciones
id	No puede ser nulo
temperatura	Consideramos que son valores sospechosos para temperaturas en ciudades españolas los menores a -20ºC y mayores a 55ºC
fecha	Ha de ser una fecha válida con el formato yyyy-MM-dd

Es posible evaluar el formato de la fecha en cualquier momento anterior en el que se utilice un RecordReader, especificando el formato que ha de seguir, pero realizándolo en este punto se puede centralizar el control de las validaciones en un solo punto. aunque no hay una sola forma de realizar las validaciones, por lo que se podría utilizar un sistema alternativo que validara la fecha al principio y los otros campos más tarde.

Para la validación se especifica el siguiente esquema:

NotNull(), DminMax(-30,55),ParseDate("yyyy-MM-dd")

Que se adhiere a la tabla especificada anteriormente:

Tabla 28: Esquemas de validación del dato

Esquema	Significado
NotNull()	Los valores nulos se considerarán inválidos para la primera columna
DMinMax(-30,55)	Sólo se considerarán válidos los valores de tipo decimal entre -30 y 55
ParseDate("yyyy-MM-dd")	Sólo se considerará válido el valor que cumpla el esquema de fecha especificado

Una vez definida la propiedad del esquema se conectará de la forma especificada anteriormente, de este modo los registros válidos llegarán al final del proceso:

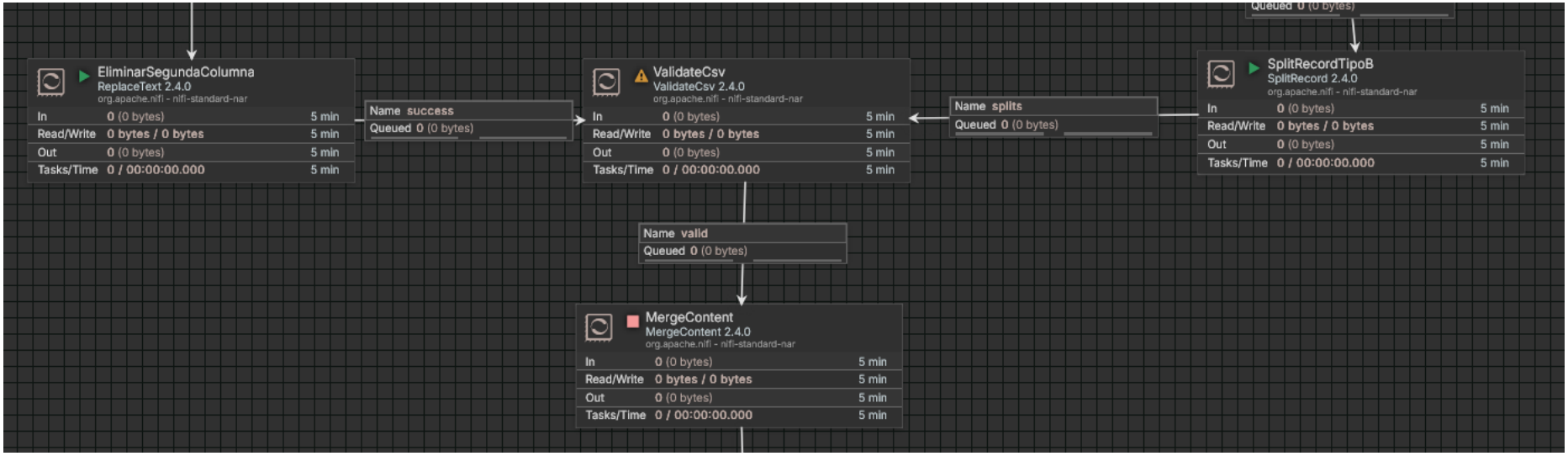

Ilustración 14: Validación del dato en el flujo de ETL

El siguiente paso una vez se ha definido el camino para los registros válidos es definir uno para los inválidos. Estos no deberían guardarse con los registros válidos, por lo que se aprovecha para utilizar otra de las funcionalidades de Apache NiFi, que es el logging. Para ello podemos usar un procesador e tipo "LogAttribute", que automáticamente escribirá en el log de la aplicación los atributos de los registros inválidos y la causa de ser inválidos. Este procesador se conectará a la salida "invalid" del ValidateCSV.

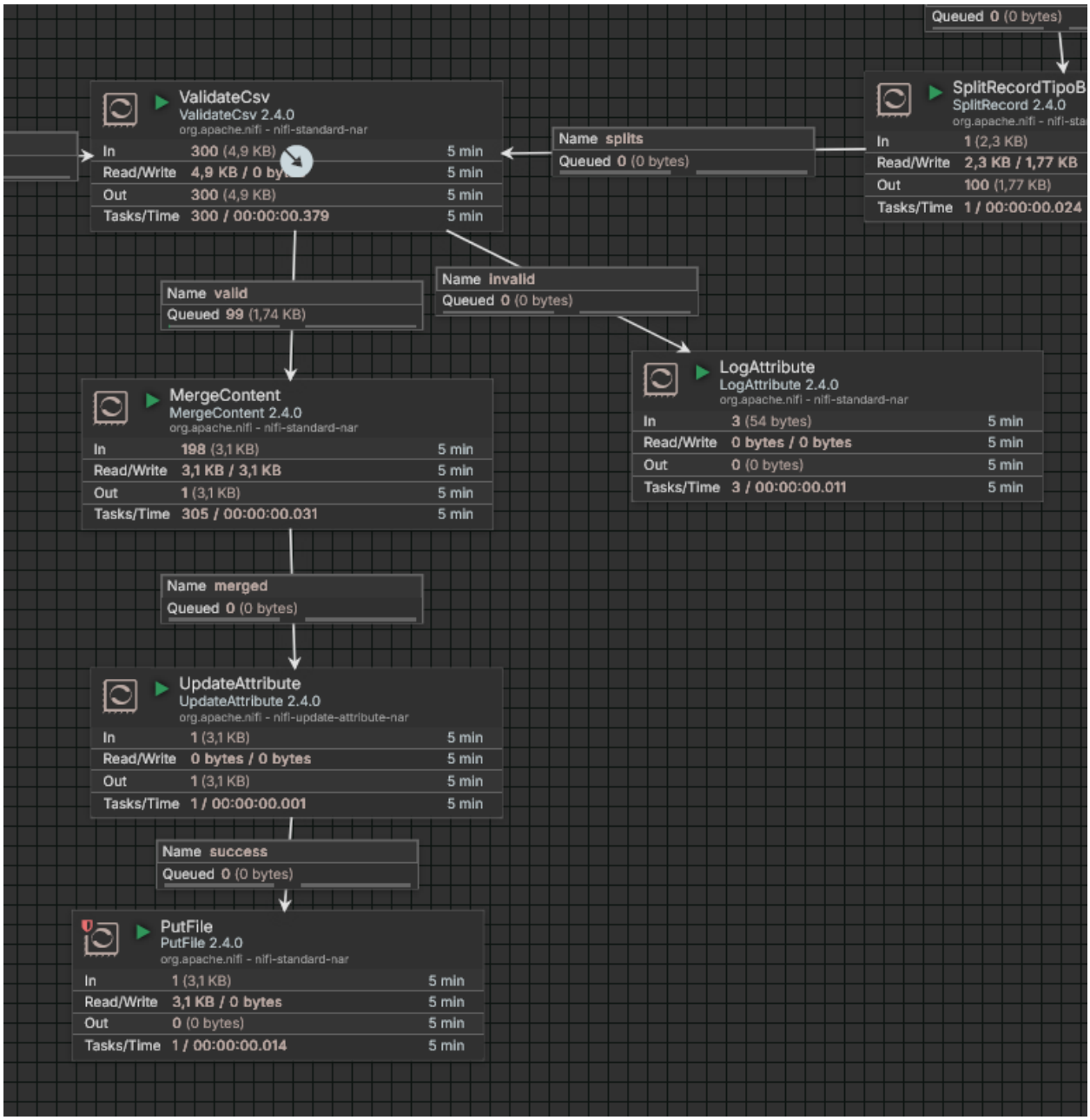

Ilustración 15: Enrutamiento de registros inválidos

Para realizar las pruebas se "esconden" 3 errores en los archivos de entrada, uno por cada uno de uno de los tipos especificados. Una vez ejecutado el flujo, se observa que aparecen mensajes en el log de la herramienta.

Tabla 29: Log de errores tras ejecución

Error	Mensaje
ID Nulo	---------- FlowFile Properties Key: 'entryDate' Value: 'Sat Jul 05 14:14:27 CEST 2025' Key: 'lineageStartDate' Value: 'Sat Jul 05 14:14:27 CEST 2025' Key: 'fileSize' Value: '15' FlowFile Attribute Map Content Key: 'RouteOnAttribute.Route' Value: 'temperaturasA' Key: 'absolute.path' Value: '….\recursos\csv-temperatura/' Key: 'file.creationTime' Value: '2025-07-04T10:49:52+0200' Key: 'file.lastAccessTime' Value: '2025-07-05T14:14:02+0200' Key: 'file.lastModifiedTime' Value: '2025-07-05T14:14:02+0200' Key: 'filename' Value: 'temperaturasA2.csv' Key: 'fragment.count' Value: '100' Key: 'fragment.identifier' Value: 'f7163921-955c-4672-b462-68bc3d269fa7' Key: 'fragment.index' Value: '47' Key: 'mime.type' Value: 'text/csv' Key: 'path' Value: '/' Key: 'record.count' Value: '1' Key: 'segment.original.filename' Value: 'temperaturasA2.csv' Key: 'uuid' Value: '2a86d1d9-eb9b-488d-b249-9cd13288ba54' Key: 'validation.error.message' Value: '**null value encountered** at {line=1, row=1, column=1}' ----------
Temperatura fuera de rango	---------- FlowFile Properties Key: 'entryDate' Value: 'Sat Jul 05 14:14:28 CEST 2025' Key: 'lineageStartDate' Value: 'Sat Jul 05 14:14:28 CEST 2025' Key: 'fileSize' Value: '23' FlowFile Attribute Map Content Key: 'RouteOnAttribute.Route' Value: 'temperaturasB' Key: 'absolute.path' Value: '….\NiFi\recursos\csv-temperatura/' Key: 'file.creationTime' Value: '2025-07-04T12:20:14+0200' Key: 'file.lastAccessTime' Value: '2025-07-05T14:14:18+0200' Key: 'file.lastModifiedTime' Value: '2025-07-05T14:14:18+0200' Key: 'filename' Value: 'temperaturasB1.csv' Key: 'fragment.count' Value: '100' Key: 'fragment.identifier' Value: '6068141d-2b5a-4efa-9925-f736be7fd3c1' Key: 'fragment.index' Value: '43' Key: 'mime.type' Value: 'text/csv' Key: 'path' Value: '/' Key: 'record.count' Value: '1' Key: 'segment.original.filename' Value: 'temperaturasB1.csv' Key: 'uuid' Value: 'ddd2c90c-0678-46da-b7e7-15e23197a015' Key: 'validation.error.message' Value: '**93,333336 does not lie between the min (-30,000000) and max (55,000000) values (inclusive)** at {line=1, row=1, column=2}' ----------

Formato de fecha incorrecta	---------- FlowFile Properties Key: 'entryDate' Value: 'Sat Jul 05 14:14:26 CEST 2025' Key: 'lineageStartDate' Value: 'Sat Jul 05 14:14:26 CEST 2025' Key: 'fileSize' Value: '16' FlowFile Attribute Map Content Key: 'RouteOnAttribute.Route' Value: 'temperaturasA' Key: 'absolute.path' Value: '...\NiFi\recursos\csv-temperatura/' Key: 'file.creationTime' Value: '2025-07-03T20:55:50+0200' Key: 'file.lastAccessTime' Value: '2025-07-05T14:13:48+0200' Key: 'file.lastModifiedTime' Value: '2025-07-05T14:13:48+0200' Key: 'filename' Value: 'temperaturasA1.csv' Key: 'fragment.count' Value: '100' Key: 'fragment.identifier' Value: 'ab77c18e-0ab3-49ce-afff-17ac9905c0e8' Key: 'fragment.index' Value: '84' Key: 'mime.type' Value: 'text/csv' Key: 'path' Value: '/' Key: 'record.count' Value: '1' Key: 'segment.original.filename' Value: 'temperaturasA1.csv' Key: 'uuid' Value: 'ce8af2d6-2022-4d61-9254-3ab6dbdf9fc1' Key: 'validation.error.message' Value: "**09-12-2024' could not be parsed as a Date** at {line=1, row=1, column=3}' ----------

Se observa también que el resto de registros llegan correctamente al archivo de salida de registros válidos.

Otro aspecto a contemplar es la monitorización que ofrece NiFi. Una cosa es controlar los errores, pero quizás también se quiere controlar cuando no hay actividad, para saber si ha habido un corte del funcionamiento.

Para ello ofrece los procesadores "MonitorActivity" que notifican la falta de actividad. Su funcionamiento consiste en notificar si se ha excedido una cantidad de tiempo desde la última vez que haya habido actividad. Es decir, que si se configura ese tiempo de forma adecuada (por ejemplo para datos que fluyen continuamente se puede especificar un tiempo bajo, mientras que para otros que se recogen cada hora se podría especificar un tiempo mayor a una hora para avisar que el flujo ha sido interrumpido por alguna razón). Estos procesadores tienen dos salidas que son útiles para este prueba de concepto, una es la salida de los datos normales, esta se conectará a el siguiente paso del flujo, mientras que las salida "inactive" es la que saltará cuando se haya cumplido ese periodo sin actividad.

Para el caso propuesto, se podría introducir el procesador después de la validación, y en caso de darse esa inactividad conectar esa salida con el procesador de logging. Aunque NiFi ofrece más opciones para notificar estos cambios, como puede ser enviar un email, o guardar registros de todos los errores en una base de datos. Por supuesto ninguna de estas soluciones es excluyente, por lo que se podría enviar un mail y escribir en el log en caso de interrupción.

El final del flujo por lo tanto podría quedar de la siguiente manera:

De esta forma guardamos en el log cuando la actividad se interrumpe y cuando se reanuda.

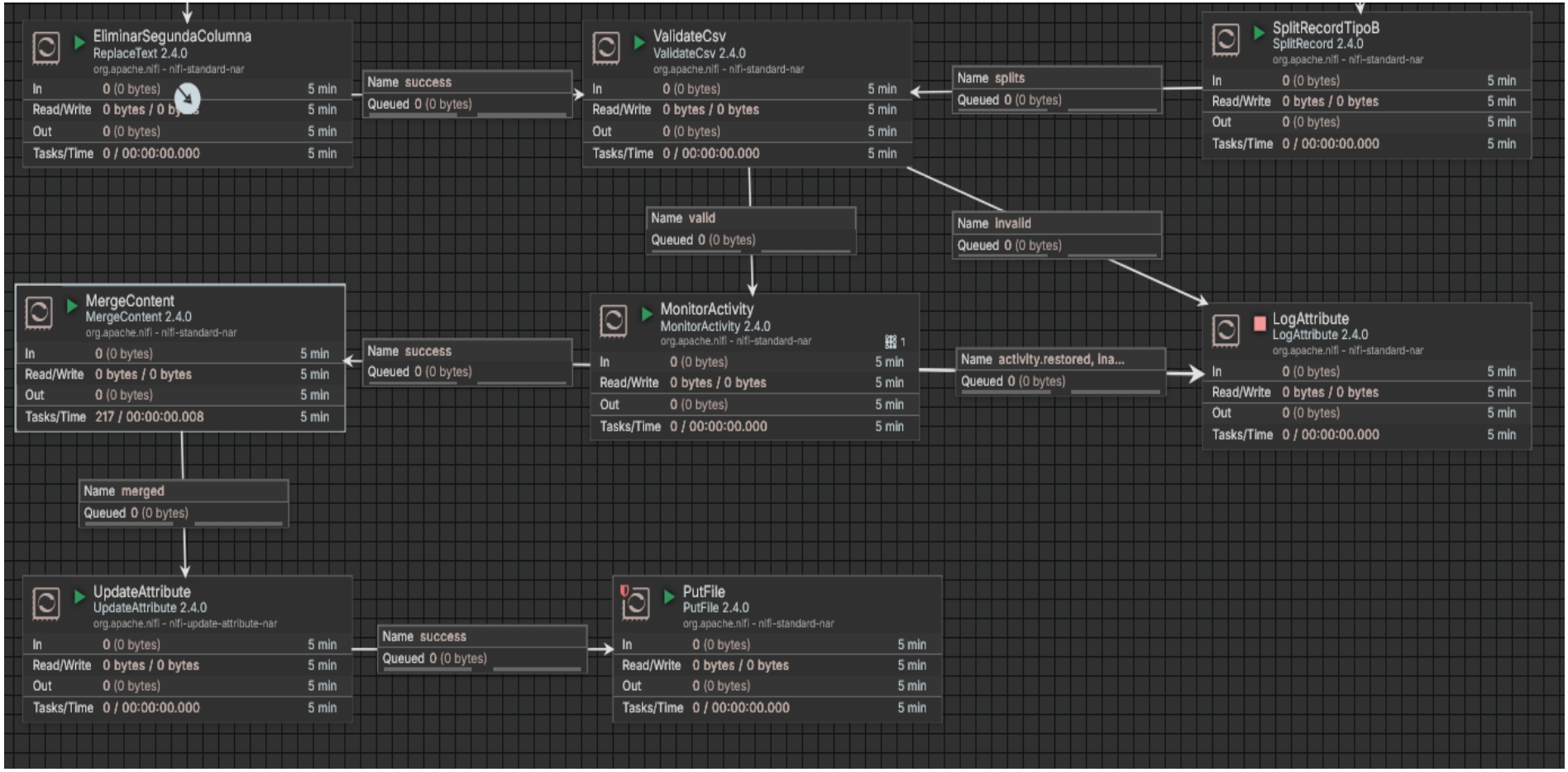

Ilustración 16: Final del flujo con monitorización

Una vez finalizada esta parte, se obtiene el gráfico completo de la prueba de concepto implementado en Apache NiFi. Desde la captura de datos, su procesado, validación, unión, monitorización y guardado.

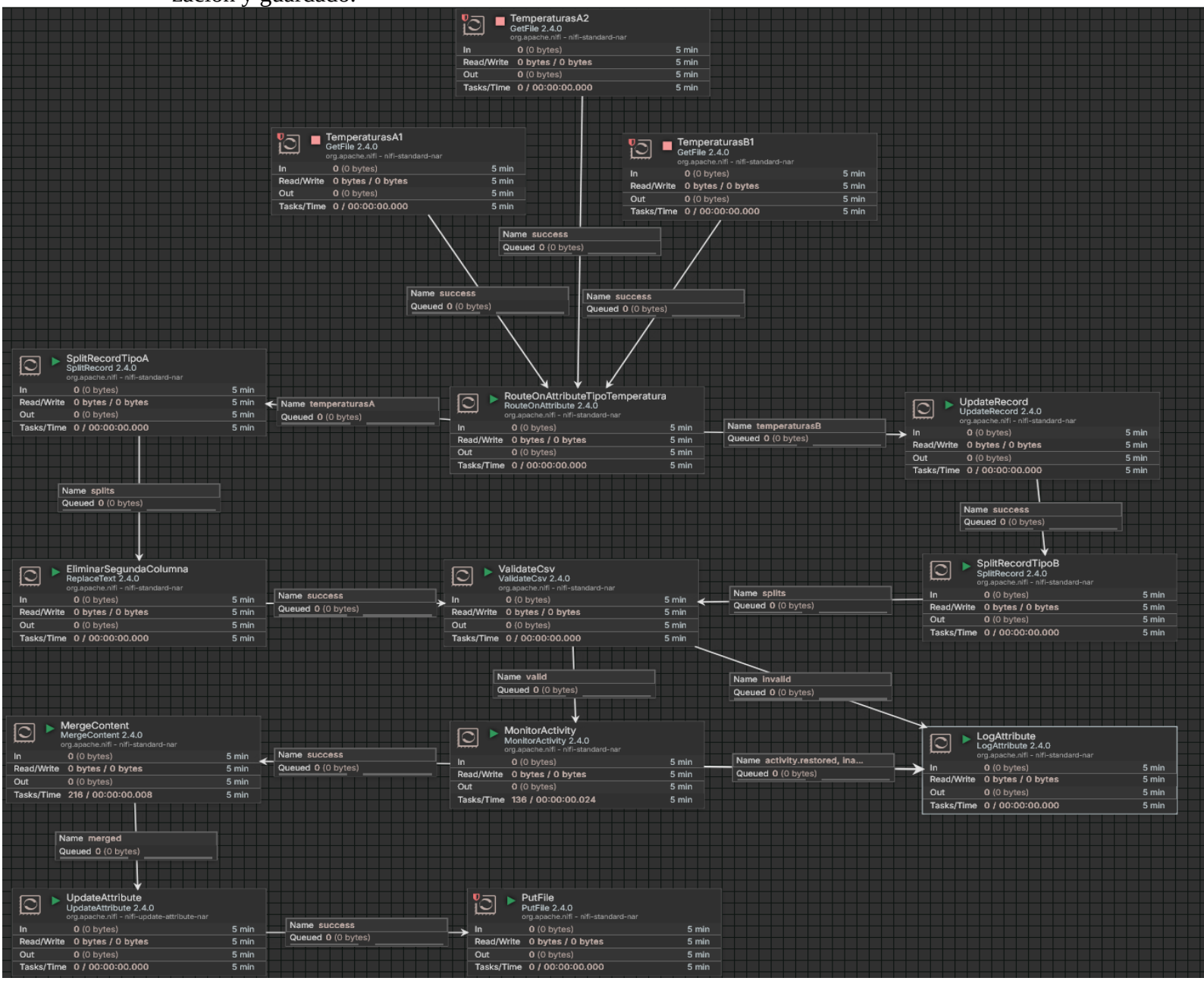

Ilustración 17: Flujo completo con elementos de calidad del dato

3. *Catálogo de datos*

Un catálogo de datos es un inventario de la riqueza del dato de una organización. No es un lugar para guardar los datos en sí, sino una recolección de metadatos sobre ellos. Es una forma de ayudar a los usuarios de dicha información a encontrar y acceder a ella de forma rápida y eficiente.

Estos metadatos pueden ser por ejemplo, la estructura de un dato, el tipo de dato (numérico, texto…), relaciones entre datos, orígenes...

No debe de ser sólo aplicable a datos estructurados, sino que debe de permitir informar a los usuarios sobre todo tipo de diferentes tipos, especialmente importante en el ámbito de IoT, donde se manejan decenas de formatos y estructuras.

Los dos usuarios principales de estos catálogos son por un lado los data stewards, encargados de organizar los datos de toda la organización realizando la tarea de mantenedores de el catálogo, y por otro lado los consumidores de la información, los cuales hacen uso de algún dato o conjunto de datos para servir a diferentes necesidades del negocio. Para estos segundos es de vital importancia el llegar a saber de qué pueden hacer uso, ya que muchas veces dada la gran cantidad de información es posible que ni siquiera se conozca. De la misma manera, y como hemos comentado antes sobre la gran heterogeneidad en el ámbito de las IoT, el catálogo especifica tanto de dónde puede recogerse esa información como la forma y estándares que tendrá.

Otra de las utilidades que nos brindan los catálogos de datos es el linaje, siendo este el seguimiento y la visualización del flujo y la transformación o modificación de los datos a medida que se mueven por los sistemas. Es decir, muestra la procedencia original de los datos (orígenes de datos), los cambios que sufren en cualquier momento del proceso, lugares donde quedan almacenados (ya sea de manera temporal o final) y ayuda a descubrir relaciones entre datos. Estas capacidades son aplicables de cara a mejorar la transparencia de la captura y uso de datos que hace una organización, y significa un gran apoyo a la hora de realizar auditorias.

También es una herramienta vital de cara a crear un lenguaje unificado. Este es el punto en el que los data stewards definen los datos, su nomenclatura, significado, clasificación y contexto. Y los demás consumidores han de usar el mismo lenguaje para así evitar posibles errores o confusiones.

3.1. Amudsen

Amudsen es una herramienta de catálogo de datos originalmente creada por Lyft y que ahora es de código abierto. Tiene un especial énfasis en la colaboración entre las personas usuarias, con capacidades para ver qué uso le dan las diferentes personas del equipo a las características de la herramienta como a qué tipo de datos se utilizan y de qué forma. Enfatiza el compartir con el resto del equipo anotaciones comentarios y valoraciones.

Tabla 30: Características Amudsen

Página web	https://www.amundsen.io/
Licencia	Apache-2.0 license
Proyecto	https://github.com/amundsen-io/amundsen

3.2. Collibra

Collibra es una de las soluciones líder en el mercado, ofreciando la mayoría de las características que son necesarias de una herramienta de este tipo para empresas. Es rápido en incluir nuevas tecnologías para mejorar sus funcionalidades, como es el caso de machine learning e inteligencia artificial, consiguiendo acelerar procesos de gestión de metadatos, clasificación y detección de anomalías.

Se trata de una solución completamente comercial, tanto en el sentido de no ser open source como de no contar con versión gratuita, todas sus funcionalidades requiren de un contrato con la empresa. Aunque sí es cierto que aunque sea sí que ofrecen dos modalidades, una puramente como servicio (es decir en su propia nube) y otra en las premisas de la organización usuaria.

Tabla 31: Características Collibra

Página web	https://www.collibra.com/
Licencia	Comercial
Proyecto	No es de código abierto

3.3. OpenMetadata

OpenMetadata es una herramienta de código abierto, una gran ventaja que ofrece es el poder realizar una demo directamente en su página web, de este modo es posible probar las funcionalidades incluso antes de hacer una primera instalación.

Ofrece la opción de ejecutar el catálogo directamente en cualquier servidor, o utilizar herramientas como Docker y Kubernetes o cualquier servicio cloud.

Se define a sí misma como la plataforma de metadatos con el mayor crecimiento en el mercado, y cuenta con amplia adopción en diferentes industrias.

Cuenta con herramientas de gestión y colaboración, y una amplia gama de conectores para todo tipo de entradas de datos.

Tabla 32: Características OpenMetadata

Página web	https://open-metadata.org/
Licencia	Apache-2.0 license
Proyecto	https://github.com/open-metadata/OpenMetadata

3.4. DataHub

DataHub es una solución mixta que cuenta con versiones gratuitas y open source (versión "Core") y comerciales (versión "Cloud"). Es una herramienta que tiene también gran presencia en la industria, y su versión Cloud está orientada a ofrecer servicios a empresas.

La versión Core cuenta con todas las funcionalidades básicas de un catálogo de datos maduro, pero hay ciertas características que son exclusivas de la versión Cloud, algunos ejemplos a continuación:

- Personalización de búsquedas para diferentes roles
- Automatizaciones con interfaz no-code
- Herramientas de IA para documentación
- Integraciones para ciertos datasources
- Seguimiento del flujo de información
- Observabilidad continua
- Evaluación de la calidad bajo demanda

Por lo que es cierto que la funcionalidad básica es gratuita, ofrece menos servicios que otras opciones que son puramente open source.

Tabla 33: Características DataHub

Página web	https://datahub.com/
Licencia	Apache-2.0 license / comercial
Proyecto	https://github.com/datahub-project/datahub

3.5. Comparativa

Para comparar las diferentes herramientas se valorará la permisividad de su licencia, tanto las opciones para realizar el mantenimiento nosotros mismo y por último las posibilidades de desplegarlo tanto en nuestras propias premisas como en la nube.

Tabla 34: Comparativa herramientas de catálogo de datos

Herramienta	Open source	Precio	Características completas	Opción de despliegue "on premise"	Opción en cloud
Amudsen	Sí	Gratuito	Sí	Sí	Sí
Collibra	No	Suscripción	Sí	Sí	Sólo en su propio servicio
OpenMetadata	Sí	Gratuito	Sí	Sí	Sí
DataHub	Mixto	Gratuito para "Core", suscripción para "Cloud"	Sólo en la versión "Cloud"	Sólo en la versión "Core"	En ambas versiones, aunque la versión "Cloud" ha de ser en su propia nube

En conclusión, se han considerado principalmente Amudsen u OpenMetadata, ya que ofrecen la mayor flexibilidad posible, con opciones de despliegues por nuestra cuenta o en diferentes proveedores cloud. Tal como se ha valorado en la sección de "Herramientas ETL", se considera a nuestra organización lo suficientemente avanzada para realizar la gestión de la herramienta, y de este modo no queda el servicio atado a una suscripción recurrente.

La versión open source de DataHub se ha considerado que no es lo suficientemente madura comparada con las otras opciones de código abierto, y contaba con demasiadas limitaciones que sólo estaban presentes en la versión comercial.

Aunque tanto Amudsen como OpenMetadata se han considerado opciones adecuadas, se le ha dado la ventaja a OpenMetadata. Esto se da debido a su mayor madurez, facilidad de uso y gestión y mayor utilización en la industria. También por ofrecer una plataforma de pruebas online en la que los data stewards pueden andar realizando pruebas previa a la instalación, haciendo posible el pivotar a la otra opción en caso de identificar puntos que sean insuficientes.

Un último punto a considerar, es que sea compatible con la herramienta de ETL que estamos utilizando, Apache NiFi. En este caso OpenMetadata lo es, por lo que podemos darle el visto bueno.

3.6. Utilizando OpenMetadata

Dividiremos la siguiente sección en 2 partes, dependiendo de qué tipo de perfil sea el responsable de los distintos aspectos de la herramienta:

- Administradores: perfiles técnicos, encargados de el buen funcionamiento de la herramient
- Usuarios: tanto data stewards que son los encargados de gestionar el dato como usuarios del dato. En general las acciones a realizar ya utilizando las prestaciones de gobernanza del dato que nos ofrece la herramienta

3.6.A. Administradores

La guía de administradores de OpenMetadata [28] divide las acciones básicas de administración de la plataforma en 3 secciones

1. Gestión de los usuarios y equipos
    - Creación de los equipos
        - OpenMetadata utiliza una estructura de la organización de 5 niveles: "Organization", "Business Unit", "Division", "Department" y "Group". Siendo Organization el nodo origen, cada nodo puede contener tanto nodos hijos de un tipo inferior o del mismo nivel (excepto Organization que no pueden ser del mismo nivel). Cada nodo puede tener varios nodos hijos, y en el caso de Department y Group pueden tener varios nodos padre también. Esto permite una forma flexible de adaptarse a casi cualquier organización.
        - Los datos se asignarán a nivel de "Group", no de ningún nodo superior.
    - Control de acceso
        - OpenMetadata permite el acceso a la plataforma mediante varios servicios Single Sign On, como puede ser Auth0
        - También permite definir políticas de acceso, estas así mismo puede ser de 3 tipos:
            - "Evaluate Deny", significa que se evalúan las condiciones y si se da alguna se niega el acceso
            - "Evaluate Allow", lo contrario, se evalúan y si se cumple alguna se otorga el acceso
            - "Disallow Access", nunca se tiene acceso
2. Añadir los usuarios
    - Desde el dashboard de OpenMetadata se pueden añadir los usuarios, se les asignan los datos básicos (nombre, correo electrónico, descripción) y su tipo dentro de la plataforma (Equipos a los que pertenece, roles y si se trata de un usuario administrador o no)
3. Configurar la entrada de metadatos
    - OpenMetadata ofrece servicios para la conexión a diferentes orígenes, en especial para nuestro caso, cuenta un conector para Apache NiFi. También es posible configurar más de una entrada; por lo que si contáramos con datos no considerados en el ETL, aunque sería una mala práctica, podríamos seguir añadiéndolos a nuestro catálogo
    - En este caso por lo tanto crearíamos un conector de tipo "Pipelines" y seleccionaríamos Apache NiFi
        - Una vez creado se puede probar la conectividad y se podrá observar el estado de la conexión
    - Dentro de cada conector se crearían los "Agents". En estos agentes se controla qué datos de los que expone de todos los que ofrece el origen a los que nos conectamos. De este modo podemos dar acceso a diferentes personas a diferentes recursos.
        - Para cada agente se configurará también el momento de ejecución, que puede ser manual bajo demanda o planificado en momentos concretos

3.6.B. Usuarios

Es más complicado definir unos pasos iniciales para los usuarios dada la amplia extensión de la aplicación y los diferentes tipos de usos que cada usuario va a darle a la aplicación. Un buen punto de comienzo es el manual del usuario de OpenMetadata [29] en el que se detallan las diferentes secciones de la interfaz web, por lo que será la primera lectura necesaria para los usuarios. De esta forma será posible orientarse inicialmente y empezar a hacer uso de la herramienta.

Algunas de las secciones tratadas en el mismo son las siguientes:

- Página de inicio, se trata de una página viva y modificable por cada usuario, con una estructura de "widgets" cada usuario puede seleccionar cuales quiere ver en su página de inicio y en qué posición en orden.
  - Widgets de actividad: como puede ser un listado de la actividad de algún equipo, las menciones que hayan hecho otros usuarios o anuncios de alguien de la organización.
  - Widgets de tareas pendientes, creadas por el usuario mismo u otro usuario
  - Widgets de seguimiento: listados de recursos a los que el usuario sigue o widgets de rendimiento de recursos completos o de la aplicación entera.
- Gestión de los recursos, realizada directamente por el usuario o realizando solicitudes. Es decir que si el usuario descubre algún punto que requiera trabajo, puede decidir informarlo el mismo, o en caso de no saber puede mencionar a otro usuario para subsanar el problema.
  - Asignar titularidad, o modificar dicha titularidad
    - Por defecto la titularidad de un recurso se propaga hacia abajo (por ejemplo de un esquema de una base de datos a las tablas que lo compongan) a no ser que se especifique lo contrario
  - Todos los recursos tienen ciertas características que se pueden configurar, como es la titularidad o el dominio, pero dependiendo de que tipo sean OpenMetadata ofrece unas pestañas de gestión u otras, para así sólo mostrar las que sean aplicables a el recurso en particular. Por ejemplo, un recurso de base de datos tendrá una pestaña para configurar sus esquemas, pero un recurso de tipo "pipeline" no tiene esquemas y por lo tanto no tendría sentido que tuviera esa pestaña. El manual de OpenMetadata especifica qué pestañas se aplican en cada caso [30]
  - Seguir un recurso, para luego poder verlo en la página principal y ser notificado de sus cambios
  - Eliminación de recursos
    - De forma lógica (OpenMetadata lo llama "soft delete"), de forma que el borrado sea reversible. A tener en cuenta que no cumple con regulaciones que requieran el eliminado real de datos
    - De forma física ("hard delete") para un borrado final que no es reversible
  - Descripciones de texto enriquecido mediante el uso de MarkDown
  - Creación de términos del glosario y etiquetas
  - Asignación de términos y etiquetas a recursos
  - Versionado de recursos
    - Pueden ser versiones menores (cambios no rompen compatibilidad) o mayores (sí se rompe compatibilidad)
  - Creación de propiedades a medida. Además de las propiedades por defecto para cada tipo de recurso, los usuarios podrán crear propiedades a medida para cubrir cualquier

caso de uso no contemplado por defecto. Estas propiedades se relacionarán con un tipo de recurso concreto, y estarán disponibles para todos los recursos de dicho tipo.

3.6.C. Instalación de la herramienta

En este apartado se documentarán los pasos seguidos para instalar la herramienta seleccionada de catálogo de datos en un equipo local.

C.1. Problemas encontrados

La solución inicial se ha intentado realizar sobre una máquina virtual. Y aunque sí es posible utilizar herramientas de virtualización como Docker dentro de una máquina virtual, no ha posido realizarse una instalación completa en este caso. Puede que por no contar con recursos suficientes para la ejecución de los contenedores sobre los recursos que una máquina virtual ya consume de por sí.

Después de realizar esta prueba, se ha optado por la instalación directamente en el sistema operativo actual del equipo, siendo este Windows 10. OpenMeetadata cuenta con soporte para la instalación de la herramienta en sistemas Windows que cuenten con el subsistema de Linux para Windows versión 2 (WSL2).

C.2. Requisitos

OpenMetadata requiere de una instalación de WSL2 en caso de trabajar sobre Windows, por lo que el primer paso es instalar esa herramienta. La distribución recomendada es Ubuntu, pero debería valer con cualquiera. Por lo que el primer paso ha sido instalar WSDL en su última versión y asegurarse de que cuenta con una imagen de Ubuntu.

El segundo paso para los rerequisitos es instalar Docker Desktop, que es un instalador normal para Windows.

Por último, se instalarán python3-pip y python3-venv dentro del entorno WSL.

```
mkdir openmetadata-docker && cd openmetadata-docker
```

C.3. Instalación

Una vez ya están los requisitos instalados, es el momento de comenzar la instalación. El primer paso consiste en obtener el archivo de Docker Compose, que es el que contiene la configuración del contenedor que se va a ejecutar. Se puede obtener descargándola directamente mediante un comando curl.

```
curl -sL -o docker-compose.yml https://github.com/open-metadata/OpenMetadata/releases/download/1.8.0-release/docker-compose.yml
```

Y de la misma manera, se puede arrancar el contenedor. Al tener el archivo previamente descargado, ya cuenta con la configuración de qué componentes ha de descargar y como tiene que arrancar.

```
docker compose -f docker-compose.yml up --detach
```

Llegado este momento se ha encontrado un problema, el docker-compose ofrecido por el equipo de OpenMetadata no es capaz de levantar la herramienta en el entorno actual. El script de inicio se encuentra con problemas , posiblemente porque el MySQL integrado no soporta una creación y ejecución en sistemas como puede ser Windows con la capa de compatibilidad de WSL. Inicialmente se ha intentado buscar soluciones modificando la configuración ofrecida, pero no se ha encontrado ninguna solución. Por lo tanto como solución alternativa se ha sugerido

utilizar la otra configuración que ofrece OpenMetadata, con PostgreSQL. Por lo que los comandos ejecutados han pasado a ser los siguientes:

curl -sL -o docker-compose-postgres.yml https://github.com/open-metadata/OpenMetadata/releases/download/1.8.0-release/docker-compose-postgres.yml

docker compose -f docker-compose-postgres.yml up --detach

Sin embargo no ha sido suficiente con esto, ya que la configuración ofrecida en este caso no ha funcionado de primeras tampoco. Ha sido necesario modificar el parámetro "volumes" del contenedor que corre PostgreSQL, y asignare un directorio del usuario al que se le han podido cambiar los permisos para poder realizar las tareas que requiere la ejecución.

Una vez realizado el cambio y terminada la ejecución estará ya corriendo OpenMetadata en local:

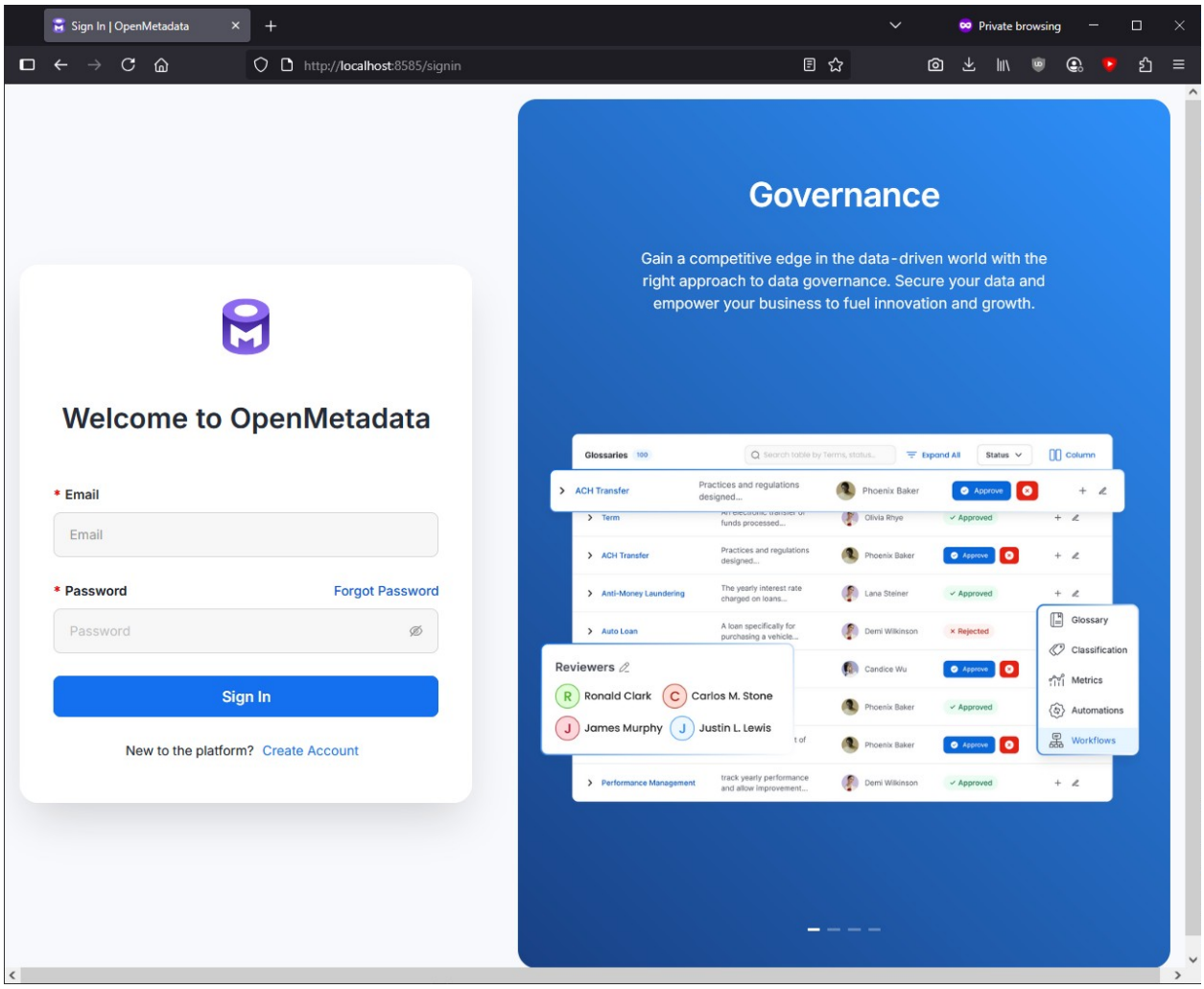

Ilustración 18: Página de login de OpenMetadata

Una vez dentro de la aplicación el primer administrador podrá crear el resto de usuarios desde la ventana de gestión de usuarios y equipos como se ha especificado en la sección anterior:

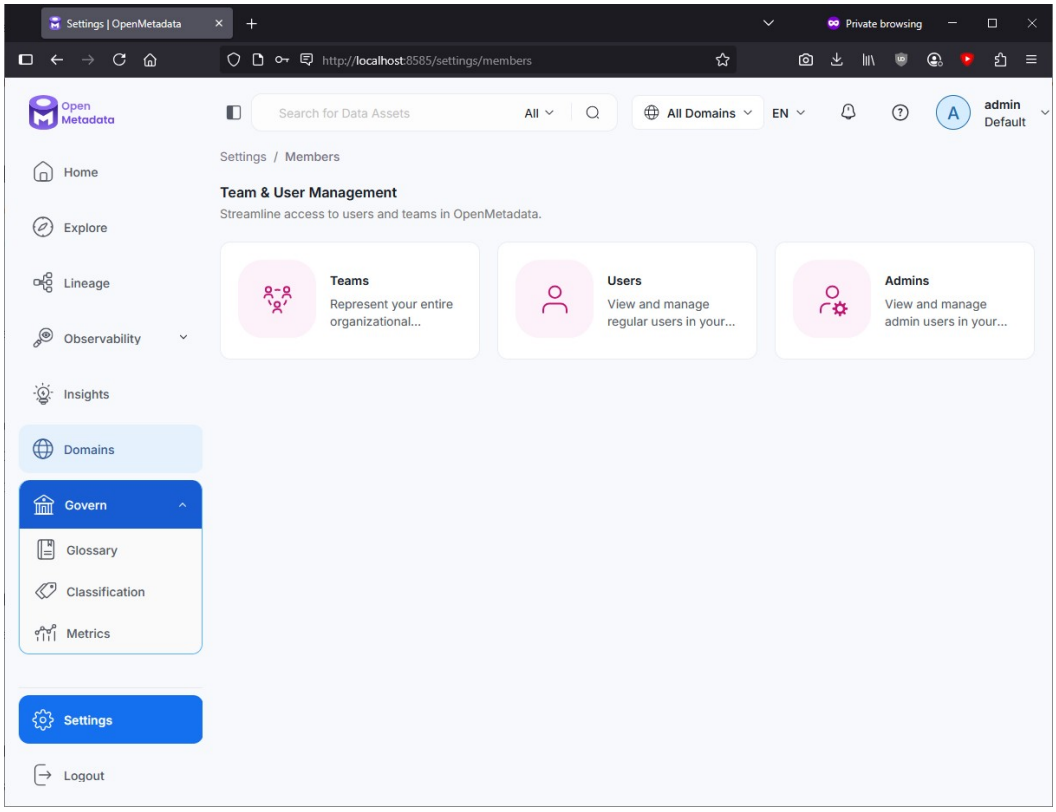

Ilustración 19: Panel de administrador OpenMetadata

Y una vez dados de alta los usuarios podrán empezar a explorar las diferentes secciones de la aplicación.

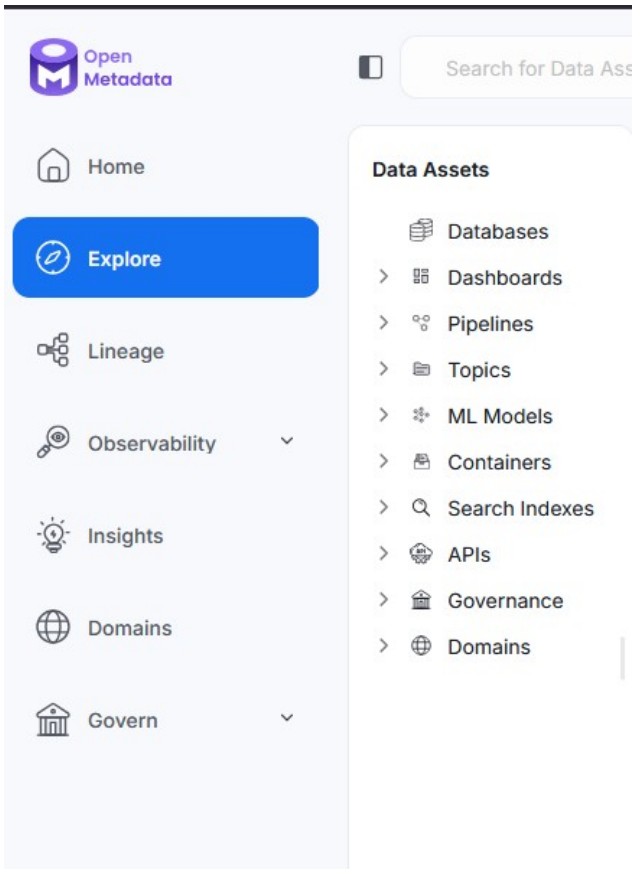

Ilustración 20: Panel lateral OpenMetadata

4. Glosario de términos

Un glosario de términos es una colección de definiciones, conceptos y terminologías que crean una entendimiento común de los datos. El glosario añade semántica o significado a los datos definiendo los conceptos relacionados con un ámbito específico. Un glosario bien definido ayuda a fomentar la colaboración en equipo mediante el uso de un lenguaje común. Los términos del glosario se pueden utilizar para etiquetar o marcar como metadatos adicionales de datos para describir y categorizar cosas.

Muchas de las herramientas de catálogo del dato contienen funcionalidades para crear y mantener un glosario de términos, y es el caso con la herramienta tratada anteriormente, OpenMetadata.

OpenMetadata permite crear varios glosarios, en caso de que se busque separar términos por dominios. A su vez estos glosarios están formados por términos. Término en este ambiente se refiere a “la terminología preferida para un concepto”, y se pueden relacionar entre ellos. De esta forma OpenMetadata crea mapas conceptuales mostrando estas relaciones, y esto fomenta el descubrimiento de los términos entre los usuarios.

Los términos en OpenMetadata tienen los siguientes campos:

- Nombre
- Descripción: es una descripción en lenguaje natural, que ha de ser entendida por los demás usuarios
- Etiquetas: para ayudar a buscar y relacionar términos
- Sinónimos: otros términos que se usan para el mismo concepto
- Hijos: Los términos pueden organizarse de forma jerárquica, creando una relación de más genérico a más específico. Por ejemplo en el contexto de IoT se podría definir “dispositivo sensor” como un término hijo de “dispositivo”
- Términos relacionados: Una lista de términos relacionados que ayudará a formar el mapa conceptual mencionado previamente
- Referencias: enlaces de internet que puedan ayudar a entender el término
- Términos mutuamente exclusivos: Una lista de términos que no se pueden aplicar al mismo tiempo que el termino que se está tratando.
- Revisores: Usuarios encargados de revisar y aceptar cualquier cambio sobre el término
- Activos: Activos que se hayan relacionado con el término

Además de incluirse en la mayoría de herramientas de catálogo de datos, también existen herramientas dedicadas de glosario de términos:

- erwin Data Literacy
- OvalEdge
- a.k.a. Taxonomy Manager

XIV - Monitorización

1. Retos de la monitorización en el IoT

Los retos que la monitorización tiene que superar en el contexto del IoT vienen de aspectos que ya hemos tratado en otras secciones:

- Heterogeneidad de los sistemas: un reto que está presente en todos los aspectos del IoT, sigue siendo igual de importante para la monitorización. Es imposible monitorizar correctamente todos estos sistemas sin un buen sistema de integración de todos ellos. Es por eso que se ha trabajado en utilizar herramientas que permitan esta estandarización y agregado de sistemas
- Problemas de conectividad: Al ser el IoT un ecosistema con muchísimos tipos de dispositivos separados en el espacio físico, es de gran importancia considerar que la tasa de fallos de conectividad será alta
- Generación masiva de datos: Es una de las ventajas del IoT que al mismo tiempo significa un reto, el IoT nos permite recopilar grandes cantidades de información que sistemas más centralizados son incapaces de alcanzar, pero esto supone un mayor coste a la hora de cualquier proceso que se de después de la captura del dato. Desde el transporte de altas cantidades, como el procesado de todos ellos y por supuesto la monitorización de todo ello
- Conformidad con las regulaciones: tratado principalmente en la sección de "Regulaciones", el IoT ofrece un reto extra no solo por la complejidad que aporta una mayor cantidad de datos, sino porque se dan muchas veces que son datos personales y la complicación que se da al considerar qué combinaciones de datos inicialmente no personales forman información personal

2. Buenas prácticas

La creación de los procesos de monitorización no sigue los mismos pasos, pero sí que se pueden establecer buenas prácticas que guíen las decisiones:

- Utilizar siempre la herramienta de monitorización centralizada. Es decir no crear puntos de monitorización separados, se han de utilizar siempre las herramientas seleccionadas en el proyecto y la monitorización debe ir siempre por esos canales. De esta forma toda persona que deba acceder a esa monitorización podrá tener acceso a ella y se podrán realizar comparativas de los diferentes puntos
- Identificar bases para las métricas, definir qué se considera el funcionamiento normal para cada una de las entradas y los rangos de valores para ese funcionamiento. Para ello es de vital importancia el haber establecido un glosario de términos, para que se vea claro qué conceptos son equivalentes (y por tanto se les puede aplicar los mismos rangos) y qué conceptos suenan igual pero no son equivalentes
- Crear alertas automáticas que avisen de cualquier problema. Identificar puntos críticos que requieran de intervención en caso de problemas, y utilizar las capacidades de nuestras herramientas para crear alertas. De esta forma dada una situación problemática las personas adecuadas estarán sobre aviso al momento
- Aplicar estándares de seguridad adecuados, de la misma forma que son necesarios para cualquier otro aspecto, en el ámbito de la monitorización es importante saber que datos se están moviendo a donde y cuando se trata de información personal o anonimizada.

3. El futuro de la monitorización

Prever las tendencias de cara al futuro es siempre complicado, y no es algo sobre lo que exista un consenso. Por ejemplo la plataforma de monitorización SigNoz contempla principalmente cuatro tendencias [31]:

- El auge de la inteligencia artificial y el machine learning, utilizando sus capacidades para análisis predictivos para poder adelantarnos a problemas antes de que estos de den
- Integración de la tecnología 5G mejorando la conectividad de todos los dispositivos IoT, reduciendo problemas de conectividad, mejorando el rendimiento de las comunicaciones tanto en el ancho de banda como en latencia. Mejorando de gran manera el rendimiento de cualquier procesado
- "Edge Computing" para un mayor procesado desde el principio, disminuyendo el trabajo a realizar por herramientas centralizadas que ya recibirán la información más estandarizada y procesada. De esta forma no se consumirán tantos recursos en un solo punto, sino que se repartirá parte de la carga por todo el sistema
- Blockchain para el aumento de la seguridad e integridad, creando registros transparentes e inmutables

XV - Cultura del gobierno del dato

No existe una sola manera de crear una buena cultura de gobierno del dato, aunque en general consiste en re-enfocar las responsabilidades que conlleva a los beneficios que nos aporta. No hacer énfasis en el trabajo que supone, sino en el trabajo que va a ser más sencillo y eficiente una vez se alcanza esta cultura.

Algunas recomendaciones para fomentar la buena cultura son las siguientes [32]:

- Cambiar las percepciones
  - Mostrar la gobernanza como una herramienta que ayuda, en vez de ser una imposición administrativa
- Fomentar prácticas proactivas
  - Incluir la gobernanza desde el principio de los workflows de trabajo, para maximizar sus ventajas
- Fomentar la colaboración
  - Creación de equipos de gobernanza para tratar las dudas y preocupaciones
  - Alinear la gobernanza con los objetivos de la organización
- Mostrar beneficios tangibles
  - Celebrando mejoras que ha traído la gobernanza
- Realizar formaciones regulares

XVI - Conclusiones y líneas futuras

Quizás la conclusión más clara es la enorme tarea que significa realizar una buena implantación del gobierno en cualquier organización. No sólo eso, sino la necesidad constante de valorar y reajustar procesos y acciones.

Sin ser una gran sorpresa, las regulaciones y su cumplimiento son una de las mayores preocupaciones, aunque también es cierto que las mejoras de eficiencia y rendimiento no son meros términos de marketing, tienen efectos reales y profundos. De esto son prueba la cantidad de herramientas disponibles y su madurez. No es difícil encontrar herramientas con un objetivo limitado como grandes suites que buscan ser la solución a todos los problemas de gobierno del dato. De la misma manera que se pueden encontrar con todo tipo de licencias, soluciones open source, soluciones comerciales, y soluciones orientadas por completo a la nube.

De cara a el aspecto IoT, se han notado las peculiaridades de este ámbito. Los problemas no son únicos de este ambiente, también se dan en entornos centralizados. Pero no es difícil notar la preocupación que causan aspectos como la heterogeneidad de los datos, que afectan desproporcionadamente al IoT.

Habiendo desarrollado los diferentes aspectos que son necesarios para la implantación de un marco de gobierno del dato, se ha obtenido una buena guía para los procesos y acciones a realizar para una buena implantación. Aunque bien es cierto que la implementación técnica de una pueba de concepto para todos ellos implicaría muchos desarrollos. De cara a hacer frente a este situación se ha optado por definir el marco entero pero realizar una prueba de concepto de una de las herramientas (Apache NiFi como herramienta ETL), en la siguiente sección se tratará los aspectos principales que sería interesante expandir y desarrollar en el futuro.

1. Líneas de trabajo futuro

De cara a desarrollos futuros, los más interesante sería continuar con el desarrollo práctico del marco desarrollado durante este trabajo. El objetivo podría ser desarrollar una prueba de concepto de todos los aspectos y finalmente conectarlos todos para un funcionamiento que cubra todo el ámbito técnico del marco.

Por un lado, se podría realizar un estudio exhaustivo de captura de datos. Es decir identificar las estructuras y formatos más comunes en el IoT y observar la facilidad con la que cada uno de ellos puede ser capturado. A ello se le podría sumar una sección de buenas prácticas para tratar con cada una de ellas.

Otro punto a explorar sería la calidad del dato, considerando más conjuntos de datos de prueba y valorando qué reglas sería correcto aplicar en cada caso y crear requisitos con sus indicadores para cada uno de los casos. Para finalizar esta parte se podrían implementar las comprobaciones y monitorización usando una herramienta de gobierno del dato.

También sería interesante valorar el impacto que introducir herramientas de machine learning o inteligencia artificial. Tanto de cara a maximizar el rendimiento como para aprovechar sus capacidades predictivas. Ya sea para un análisis más flexible de la situación del dato como herramienta que sea capaz de dar sugerencias de reglas y políticas de calidad del dato.

Finalmente, se podría realizar una prueba de concepto de catálogo de datos. Esto implicaría la gestión de diferentes perfiles, roles y grupos simulando usuarios reales que podrían hacer uso de esta herramienta. También se desarrollaría una prueba de concepto de glosario, utilizando buenas prácticas para las nomenclaturas y un sistema de etiquetado robusto. De la misma manera se incorporarían diferentes orígenes de datos, incluidos los implementados en los otras pruebas de concepto, haciendo que así quede todo el marco plasmado en un ejemplo práctico. O ir más allá e

intentar desarrollar una herramienta que integre los pilares del gobierno del dato, para no necesitar de herramientas comerciales de precios prohibitivos.

2. Futuro del gobierno del dato

El gobierno del dato cobra cada vez más importancia en la era de la inteligencia artificial. Uno de los aspectos más críticos de cualquier IA es los datos que se usan para entrenarla, es por eso que es crucial una buena implementación que asegure buenos resultados. Los beneficios que la gobernanza aporta a la IA son innumerables:

- La IA depende de datos de alta calidad, si los datos son inexactos, incompletos o sesgados, los modelos de IA pueden producir resultados incorrectos, engañosos o perjudiciales
- Se ha de garantizar que las prácticas de recopilación, almacenamiento, procesamiento e intercambio de datos se ajustan a la normativa sobre privacidad. De esta forma se asegura que los sistemas de IA no violan los derechos de privacidad de los usuarios ni operan fuera de los límites legales
- Se ha de documentar y rastrear el linaje de los datos utilizados para entrenar modelos de IA, lo que permite la transparencia sobre la procedencia de los datos y cómo se utilizaron. Esta trazabilidad facilita la rendición de cuentas de los sistemas de IA
- Los sistemas de IA trabajan a menudo con grandes cantidades de datos personales sensibles -registros sanitarios, información financiera o datos de comportamiento que requieren protecciones adecuadas
- Se han de incluir controles de auditoría, supervisión y acceso para garantizar que los datos no puedan manipularse o alterarse fácilmente. Al aplicar controles de integridad de los datos y mantener una supervisión rigurosa, se minimiza el riesgo de que agentes maliciosos afecten a los modelos de IA

XVII - Referencias

Bibliografía

[1] ¿Qué es la gobernanza de datos? | Definición, importancia y tipos | SAP. (n.d.). SAP. Accedido: 7 de julio de 2025. [Online] Disponible: https://www.sap.com/latinamerica/products/data-cloud/master-data-governance/what-is-data-governance.html
[2] EWSolutions. (2025, March 20). Data Governance Program Team structure. EWSolutions. Accedido: 7 de julio de 2025. [Online] Disponible: https://www.ewsolutions.com/data-governance-program-team-structure/
[3] Kazlow, D. (2024, September 18). The definitive guide to Data Governance Councils - The Data Governance. The Data Governance. Accedido: 7 de julio de 2025. [Online] Disponible: https://thedatagovernance.com/data-governance-council/
[4] Firican, G. (2021, November 22). The complete guide to data governance roles and responsibilities | LightsOnData. LightsOnData. Accedido: 7 de julio de 2025. [Online] Disponible: https://www.lightsondata.com/the-complete-guide-to-data-governance-roles-and-responsibilities/
[5] IoT-Lite Ontology. (n.d.). Accedido: 7 de julio de 2025. [Online] Disponible: https://www.w3.org/submissions/2015/SUBM-iot-lite-20151126/
[6] Atlan, T. (2024, December 18). Data Governance Policy: Comprehensive Guide with Examples for 2025. Atlan. Accedido: 7 de julio de 2025. [Online] Disponible: https://atlan.com/data-governance-policy/
[7] General Data Protection Regulation (GDPR) – Legal Text. (2024, April 22). General Data Protection Regulation (GDPR). Accedido: 7 de julio de 2025. [Online] Disponible: https://gdpr-info.eu/
[8] Personal Data - General Data Protection Regulation (GDPR). (2021, October 22). General Data Protection Regulation (GDPR). Accedido: 7 de julio de 2025. [Online] Disponible: https://gdpr-info.eu/issues/personal-data/
[9] Encryption - General Data Protection Regulation (GDPR). (2023, September 5). General Data Protection Regulation (GDPR). Accedido: 7 de julio de 2025. [Online] Disponible: https://gdpr-info.eu/issues/encryption/
[10] Internet Society (2020). Policy Toolkit on IoT Security and Privacy : Accedido: 7 de julio de 2025. [Online] Disponible: https://www.internetsociety.org/wp-content/uploads/2020/08/IoTtoolkit-August-2020.pdf
[11] Vinod V. Nair, R. Nanda Kishor. (2017). Getting the Most out of IoT with an Effective Data Lifecycle Management Strategy. Accedido: 7 de julio de 2025. [Online] Disponible: https://www.tcs.com/content/dam/global-tcs/en/pdfs/insights/whitepapers/getting-the-most-of-iot-data-effective-lifecycle-management-strategy.pdf
[12] Baker, E. (2024, September 26). Conducting a data governance audit. DataGovernancePlatforms.com. Accedido: 7 de julio de 2025. [Online] Disponible: https://www.datagovernanceplatforms.com/conducting-data-governance-audit/
[13] The Institute of Internal Auditors, Inc (2020). Data Governance - Providing assurance regarding data risk management : Accedido: 7 de julio de 2025. [Online] Disponible: https://www.theiia.org/globalassets/site/content/articles/industry-knowledge-brief/2020/data-governance/data-governance.pdf
[14] Black, A., Nederpelt, P. van. (2020). Dimensions of Data Quality Dimensions. Accedido: 7 de julio de 2025. [Online] Disponible: https://dama-nl.org/wp-content/uploads/2020/09/DDQ-Dimensions-of-Data-Quality-Research-Paper-version-1.2-d.d.-3-Sept-2020.pdf
[15] Black, A., Nederpelt, P. van. (2020). How to Select the RightDimensions of DataQuality. Accedido: 7 de julio de 2025. [Online] Disponible: https://dama-nl.org/wp-content/uploads/2020/11/How-to-Select-the-Right-Dimensions-of-Data-Quality-v1.1-d.d.-14-Nov-2020.pdf

[16] *Reglamento - UE - 2024/1689 - EN - EUR-LEX*. (n.d.). Accedido: 7 de julio de 2025. [Online] Disponible: https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:32024R1689
[17] BOE-A-2018-16673 Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales. (n.d.). Accedido: 7 de julio de 2025. [Online] Disponible: https://www.boe.es/buscar/doc.php?id=BOE-A-2018-16673
[18] datos.gob.es. (2025c, March 10). Especificaciones UNE – Gobierno, gestión y calidad del dato. datos.gob.es. Accedido: 7 de julio de 2025. [Online] Disponible: https://datos.gob.es/es/blog/especificaciones-une-gobierno-gestion-y-calidad-del-dato
[19] datos.gob.es. (2025d, June 6). Aplicación de la especificación UNE 0077:2023 a los datos abiertos. datos.gob.es. Accedido: 7 de julio de 2025. [Online] Disponible: https://datos.gob.es/es/blog/aplicacion-de-la-especificacion-une-00772023-los-datos-abiertos
[20] datos.gob.es. (2025e, June 6). Aplicación de las Especificación UNE 0078:2023 a los datos abiertos. datos.gob.es. Accedido: 7 de julio de 2025. [Online] Disponible: https://datos.gob.es/es/blog/aplicacion-de-las-especificacion-une-0078-2023-los-datos-abiertos
[21] datos.gob.es. (2025e, June 6). Aplicación de la Especificación UNE 0079:2023 de gestión de calidad a los datos abiertos. datos.gob.es. Accedido: 7 de julio de 2025. [Online] Disponible: https://datos.gob.es/es/blog/aplicacion-de-la-especificacion-une-00792023-de-gestion-de-calidad-los-datos-abiertos
[22] datos.gob.es. (2024, June 20). Normas Técnicas para alcanzar la Calidad del Dato. *datos.gob.es*. Accedido: 7 de julio de 2025. [Online] Disponible: https://datos.gob.es/es/blog/normas-tecnicas-para-alcanzar-la-calidad-del-dato
[23] R. K. Pon, D. J. Buttler (2008). METADATA REGISTRY, ISO/IEC 11179. Accedido: 7 de julio de 2025. [Online] Disponible: https://digital.library.unt.edu/ark:/67531/metadc926399/m2/1/high_res_d/973862.pdf
[24] Matta, I., & Matta, I. (2025, February 18). Data Governance Frameworks -The ISO 38505. Sogeti Labs. Accedido: 7 de julio de 2025. [Online] Disponible: https://labs.sogeti.com/data-governance-frameworks-the-iso-38505/
[25] Art. 4 GDPR – Definitions - General Data Protection Regulation (GDPR). (2018, March 29). General Data Protection Regulation (GDPR). Accedido: 7 de julio de 2025. [Online] Disponible: https://gdpr-info.eu/art-4-gdpr/
[26] Asswad, J., & Marx Gómez, J. (2021). Data Ownership: A Survey. Information, 12(11), 465. Accedido: 7 de julio de 2025. [Online] Disponible: https://doi.org/10.3390/info12110465
[27] Team, A. N. (n.d.). Apache NiFi User Guide. https://nifi.apache.org/docs/nifi-docs/html/user-guide.html
[28] Admin Guide | OpenMetadata Administration Documentation. (n.d.). Accedido: 7 de julio de 2025. [Online] Disponible: https://docs.open-metadata.org/latest/how-to-guides/admin-guide
[29] Guide for Data Users | OpenMetadata User Guide. (n.d.). Accedido: 7 de julio de 2025. [Online] Disponible: https://docs.open-metadata.org/latest/how-to-guides/guide-for-data-users
[30] Overview of Data Assets. (n.d.). Accedido: 7 de julio de 2025. [Online] Disponible: https://docs.open-metadata.org/latest/how-to-guides/guide-for-data-users/data-asset-tabs#data-asset-tabs
[31] Goswami, V. (2024, November 29). Essential Guide to IoT Monitoring - Benefits and Best Practices. SigNoz. Accedido: 7 de julio de 2025. [Online] Disponible: https://signoz.io/guides/iot-monitoring/
[32] https://www.secoda.co/authors/ainslie-eck. (2024, December 13). Embedding governance into culture: Practices that drive success. Accedido: 7 de julio de 2025. [Online] Disponible: https://www.secoda.co/blog/embedding-governance-into-culture-practices-that-drive-success

Impreso en Burgos el martes, 8 de julio de 2025