

# EDEM

Escuela de Empresarios

Escuela de Empresarios

# Datos Maestros

## Stratio

Alfonso Fernández Revenga

Curso \_\_\_\_ - Edición \_\_\_\_

Fecha 09/01/2021

# SOBRE MI

## Alfonso Fernández Revenga

<https://www.linkedin.com/in/alfonsofernandezrevenga/>

[alfonsofernandez@stratio.com](mailto:alfonsofernandez@stratio.com)



**EDEM**  
Escuela de Empresarios



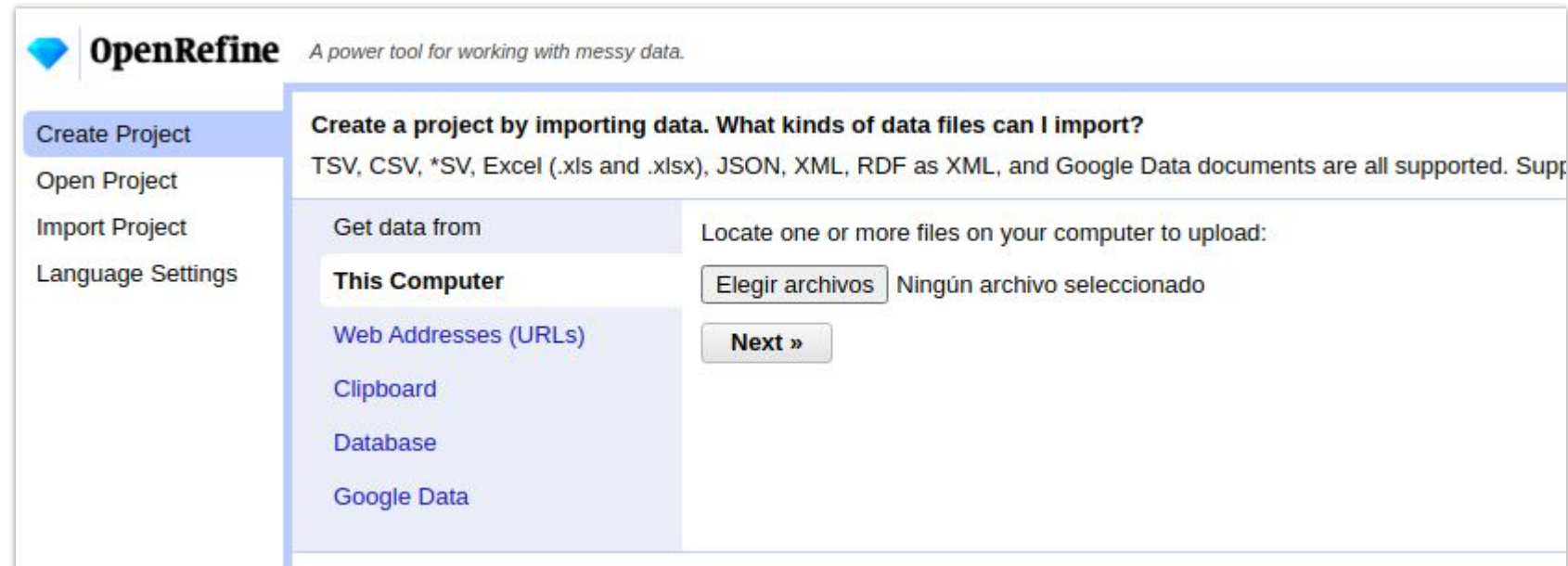
# Caso práctico

## Open Refine

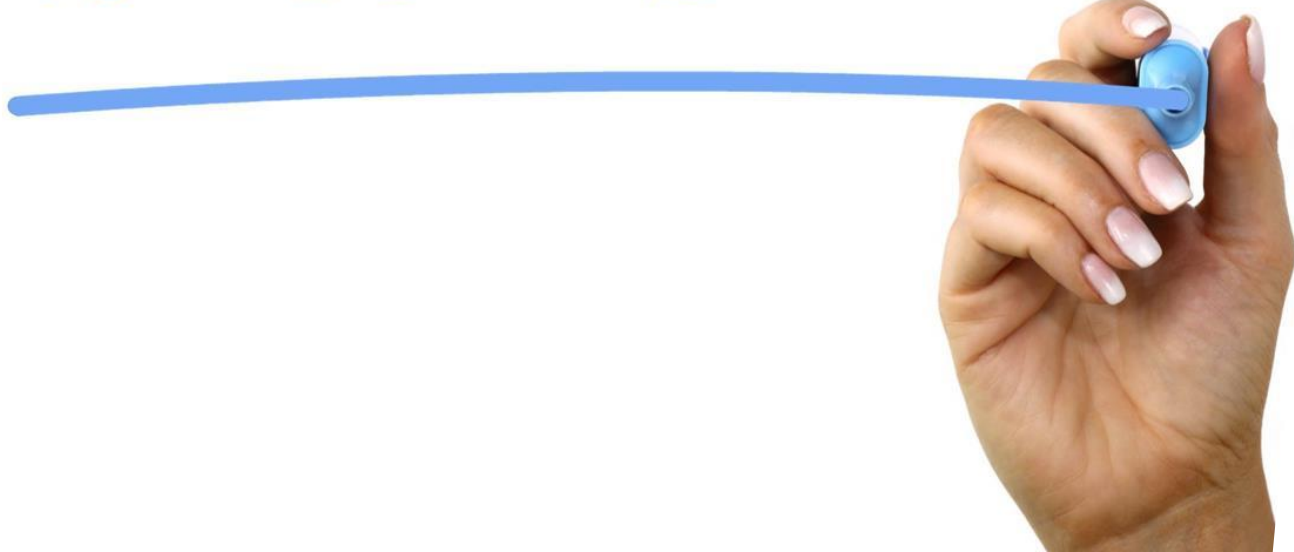
OpenRefine: <https://openrefine.org/download.html>

Librería de matching: <http://okfnlabs.org/reconcile-csv/>

Datasets: <https://tinyurl.com/yxg5tl9k>



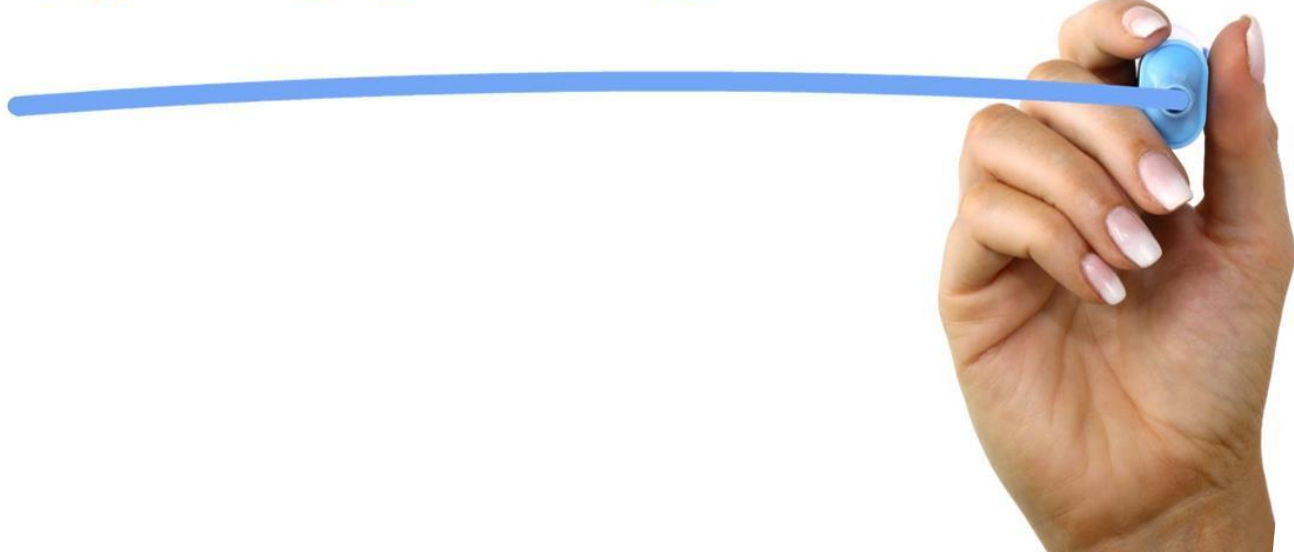
# INDEX



## Índice

1. Introducción a MDM
2. Arquitectura MDM
3. Valor de Master data
4. Gestión de proyectos MDM
5. Conclusiones
6. Herramientas

# INDEX



## Índice

1. **Introducción a MDM**
2. Arquitectura MDM
3. Valor de Master data
4. Gestión de proyectos MDM
5. Conclusiones
6. Herramientas

# Introducción a MDM

## Historias...

Un cliente con una tarjeta de crédito **se muda** de 2847 North 9th St. a 1001 11th St. North. El cliente cambió su dirección de facturación de inmediato, pero no recibió una factura durante varios meses. Un día, el cliente recibió una llamada telefónica **amenazante** del departamento de facturación de la tarjeta de crédito preguntando por qué **no se había pagado la factura**. El cliente verifica que tiene la nueva dirección y el departamento de facturación verifica que la dirección registrada sea 1001 11th St. North. El cliente solicita una copia de la factura para liquidar la cuenta.

Después de dos semanas más sin factura, el cliente vuelve a llamar y descubre que la cuenta ha sido entregada a una agencia de morosidad. Esta vez, el cliente descubre que aunque la **dirección en el archivo** era 1001 11th St. North, la **dirección de facturación** aparece como **101** 11th St. North. Después de varias llamadas telefónicas y cartas entre abogados, la factura finalmente se resuelve y la compañía de la tarjeta de crédito **ha perdido un cliente de por vida**. .... ¿Qué opináis?

# Introducción a MDM

## Historias...

En este caso, el dato maestro era preciso, pero la copia tenía fallos. Los datos maestros deben ser **correctos** y **consistentes**. Incluso si los datos maestros no tienen errores, pocas organizaciones tienen sólo un conjunto de datos maestros. Muchas compañías crecen a través de **fusiones y adquisiciones**, y cada compañía que adquiere la organización matriz viene con su propio maestro de clientes, productos, etc.

Esto no sería malo si pudiera unir los nuevos datos maestros con los datos maestros actuales, pero a menos que la compañía adquirida esté en un negocio completamente diferente en un país lejano, hay una muy gran posibilidad de que aparezcan algunos **clientes y productos en ambos conjuntos de datos maestros**, generalmente con diferentes formatos y diferentes claves de base de datos.



# Introducción a MDM

## Historias...

Si ambas compañías usan el **Número de DNI/Pasaporte** el **Número de Seguridad Social** como el identificador del cliente, descubrir qué registros de clientes son para el mismo cliente es un **problema directo**; pero eso rara vez sucede.

En la mayoría de los casos, el software que crea los registros maestros asigna los números de cliente (IDs internos), por lo que las posibilidades de que el mismo cliente o el mismo producto tengan el mismo identificador en ambas bases de datos son bastante remotas. Los maestros de productos pueden ser aún más difíciles de conciliar si se compran piezas equivalentes de diferentes proveedores con diferentes números de proveedor.



# Introducción a MDM

## Experiencia personal

Una compañía telefónica me llama por teléfono para hacer una acción comercial: **ofrecerme una tarifa** super barata para hacer el traspaso de mi línea a esta compañía.

Hasta ahí, nada raro... salvo porque yo **ya era cliente de esa compañía** desde hace años. Aprovechando la llamada, le digo que la oferta me interesa, pero que ya soy cliente.

Respuesta: “Esta oferta es sólo para **nuevos clientes**”

Sensación: Tratan mejor a los nuevos clientes que a los antiguos.

Acción: amenacé con cambiarme, y me hicieron la oferta. Pero la imagen de la compañía se degradó.



# Introducción a MDM

## Tipos de datos en las compañías

1. **Datos no estructurados:** datos encontrados en correos electrónicos, documentos técnicos, artículos de revistas, portales de intranet corporativos, especificaciones de productos, material publicitario y archivos PDF.
2. **Datos transaccionales:** datos sobre eventos comerciales (a menudo relacionados con transacciones del sistema, como ventas, entregas, facturas, tickets de incidencias, reclamos y otras interacciones monetarias y no monetarias) que tienen un significado histórico o son necesarios para el análisis de otros sistemas. Los datos transaccionales son transacciones a nivel de unidad que utilizan entidades de datos maestros. A diferencia de los datos maestros, las transacciones son inherentemente temporales e instantáneas por naturaleza.
3. **Metadatos:** datos sobre otros datos. Puede residir en un repositorio formal o en varias otras formas, como documentos XML, definiciones de informes, descripciones de columnas en una base de datos, archivos de registro, conexiones y archivos de configuración.

# Introducción a MDM

## Tipos de datos en las compañías

4. **Datos jerárquicos:** datos que almacenan las relaciones entre otros datos. Puede almacenarse como parte de un sistema de contabilidad o por separado como descripciones de las relaciones del mundo real, como las estructuras organizativas de la empresa o las líneas de productos. Los datos jerárquicos a veces se consideran un dominio importante MDM porque es fundamental para comprender y, a veces, descubrir las relaciones entre los datos maestros.
5. **Datos de referencia:** un tipo especial de datos maestros utilizados para clasificar otros datos o para relacionar datos con información más allá de los límites de la empresa. Los datos de referencia se pueden compartir entre objetos de datos maestros o transaccionales (por ejemplo, países, monedas, zonas horarias, condiciones de pago, etc.)

# Introducción a MDM

## Tipos de datos en las compañías

6. **Datos maestros:** los datos centrales dentro de la empresa que describen objetos alrededor de los cuales se llevan a cabo negocios. Por lo general, cambia con poca frecuencia y puede incluir datos de referencia que son necesarios para operar el negocio. Los datos maestros no son de naturaleza transaccional, pero describen transacciones. Los sustantivos críticos de una empresa que cubren datos maestros generalmente se dividen en cuatro dominios y las categorizaciones adicionales dentro de esos dominios se denominan áreas temáticas, subdominios o tipos de entidad.

Los típicos dominios de datos maestros:

**Clientes:** Dentro del dominio del cliente, hay subdominios de clientes, empleados y vendedores.

**Productos:** Dentro del dominio de productos, hay subdominios de productos, partes, tiendas y activos.

**Ubicaciones:** Dentro del dominio de ubicaciones, hay subdominios de ubicación de oficina y división geográfica.

# Introducción a MDM

## Contexto

Las **grandes organizaciones**, con multitud de procesos y sistemas comerciales para procesar transacciones, a menudo se enfrentan al desafío de no tener una "**Fuente única de verdad**" para sus Datos maestros. Los **objetivos** de Arquitectura de sistemas y Arquitectura de datos parecen ser divergentes y tácticos en lugar de **coherentes y estratégicos**.

Los datos son un **activo empresarial** utilizado para tomar decisiones comerciales estratégicas. Muy a menudo la **precisión, integridad, accesibilidad y seguridad** de los datos **impiden** la toma de decisiones de negocio efectivas

Las organizaciones están dotadas de conjuntos de datos aislados (**silos**) que no se aprovechan de manera óptima para que la suma de las partes resulte un conjunto global.

# Introducción a MDM

## Definición

- **Master Data** is the consistent and uniform set of identifiers and extended attributes that describes **the core entities of the enterprise** including customers, prospects, citizens, suppliers, sites, hierarchies and chart of accounts (*sic*).
- **Master data management (MDM)** is a technology-enabled discipline in which business and IT work together to ensure the uniformity, accuracy, stewardship, semantic consistency and accountability of the enterprise's official shared master data assets.

- Source Gartner



# Introducción a MDM

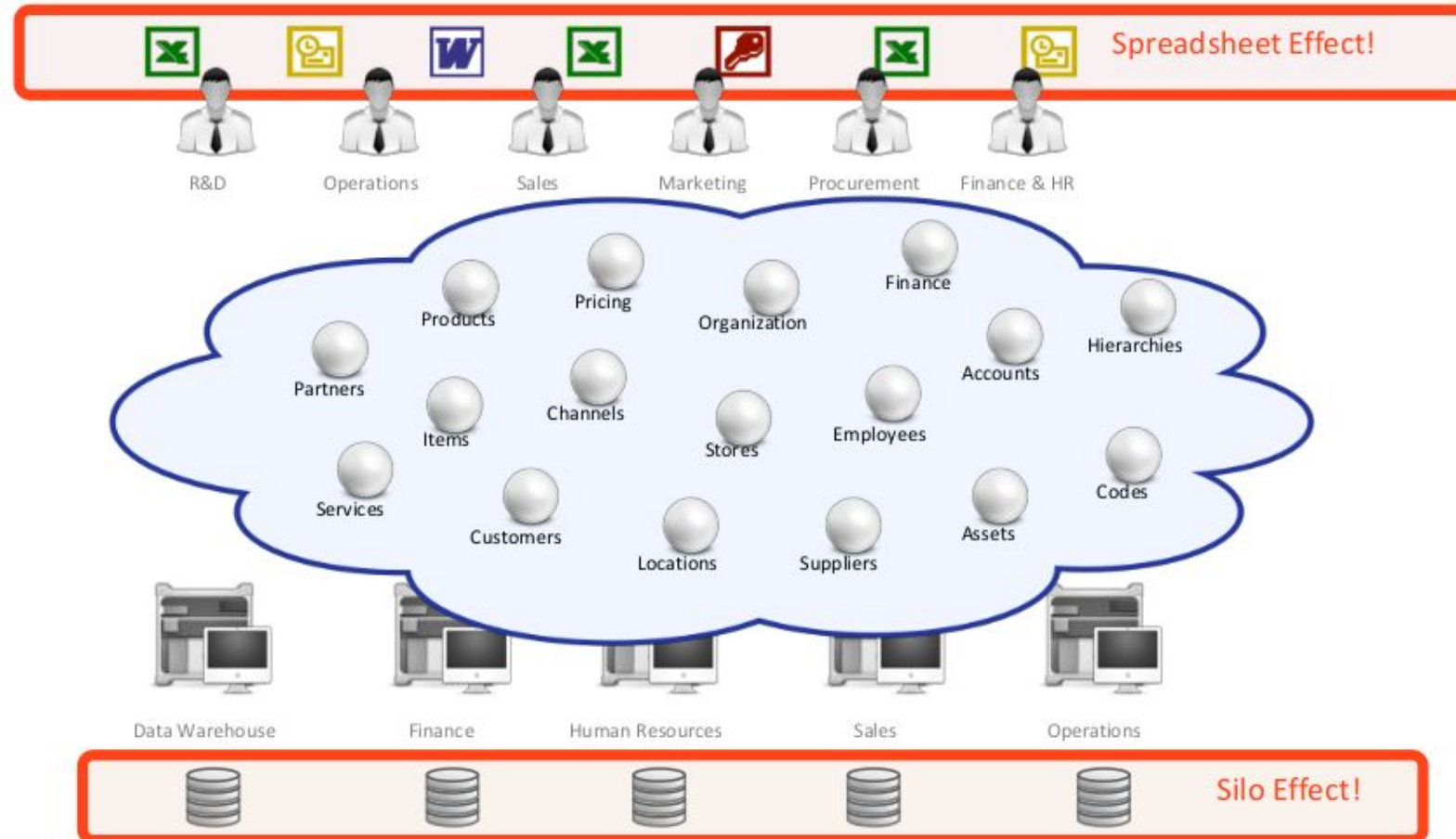
## Definición

MDM es una **capacidad de negocio** que permite a una organización **identificar** primero **datos maestros** y luego aprovechar los datos maestros para **mejorar los procesos y las decisiones de negocio**.

- Identificar datos maestros: MDM define y/o deriva la "**versión**" **más confiable y única de datos** empresariales importantes (por ejemplo, proveedor, cliente, producto, empleado, activo, material, ubicación, etc.).
- **Aprovechar los datos maestros** para mejorar los procesos y las decisiones empresariales: MDM incorpora esta versión maestra de los datos dentro de los procesos de negocio (ventas, marketing, finanzas, soporte, etc.) que **proporcionarán un beneficio directo** a los empleados, clientes, socios u otras partes interesadas relevantes dentro de una organización.
- Los datos maestros por sí solos proporcionan poco valor: Es necesario anticiparse a **cómo los datos serán consumidos** por otras aplicaciones o sistemas dentro del contexto de un proceso de negocio para maximizar su valor.

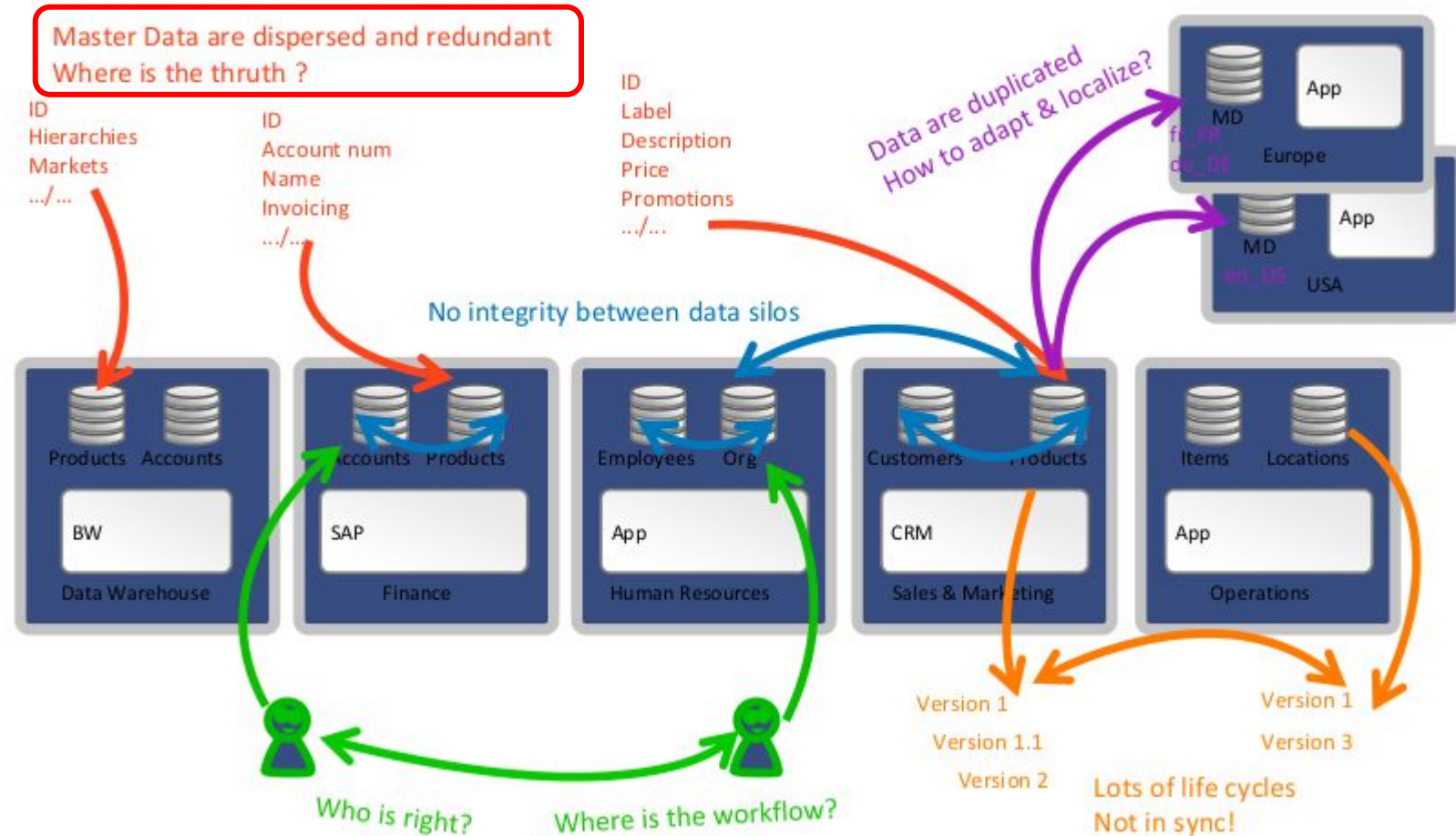
# Introducción a MDM

## Datos entre negocio e IT



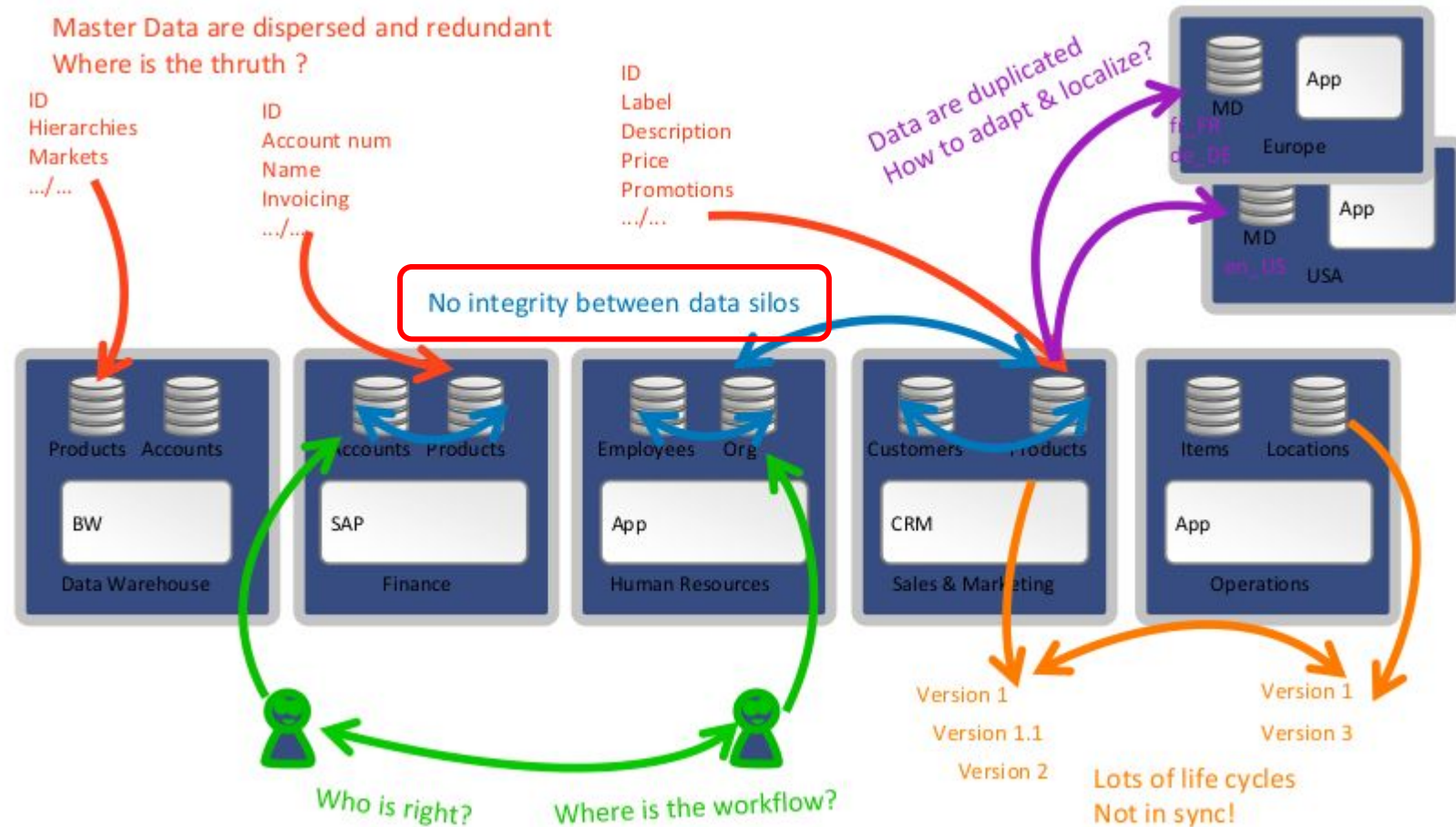
# Introducción a MDM

## Problemas



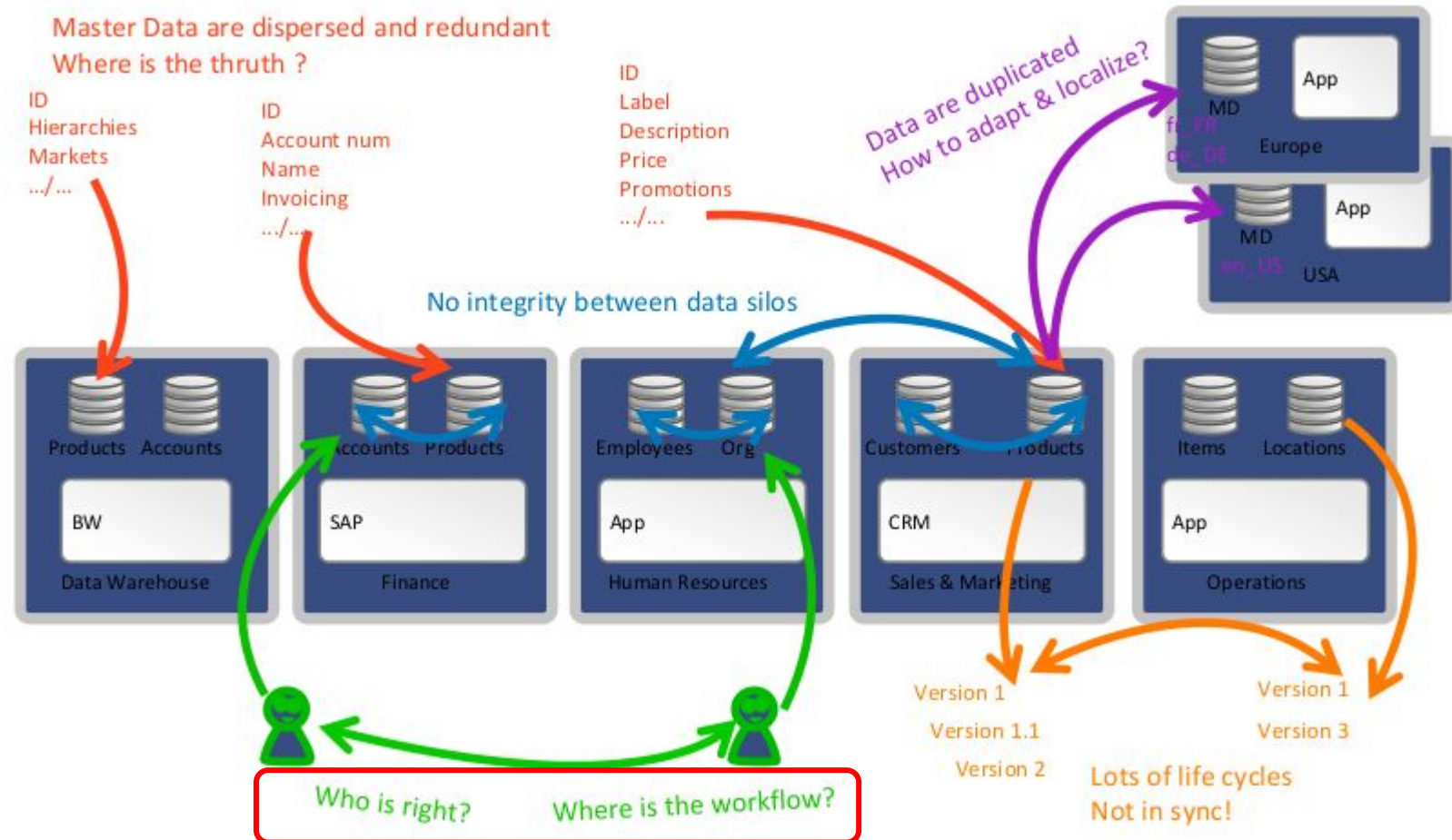
# Introducción a MDM

## Problemas



# Introducción a MDM

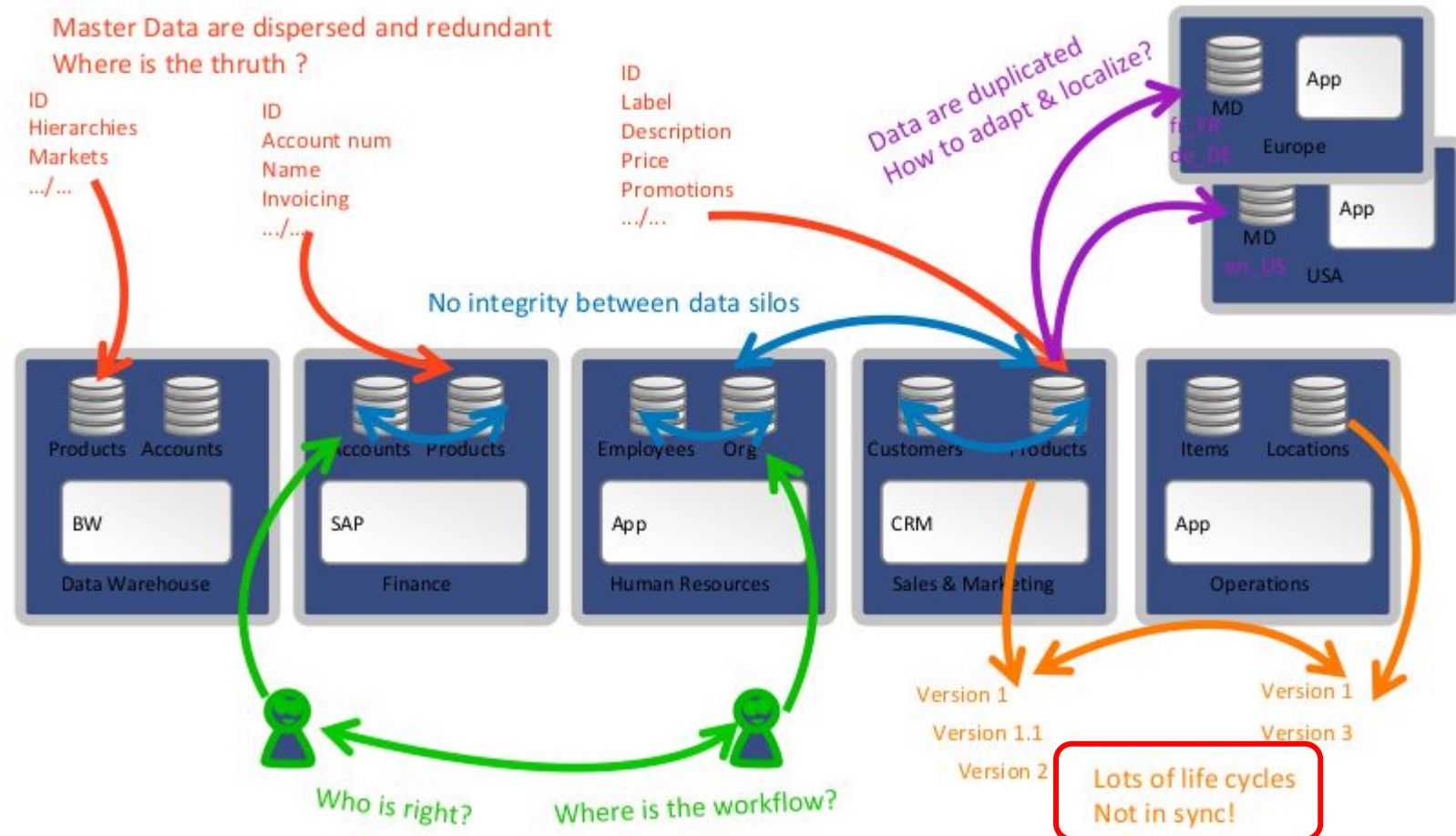
## Problemas





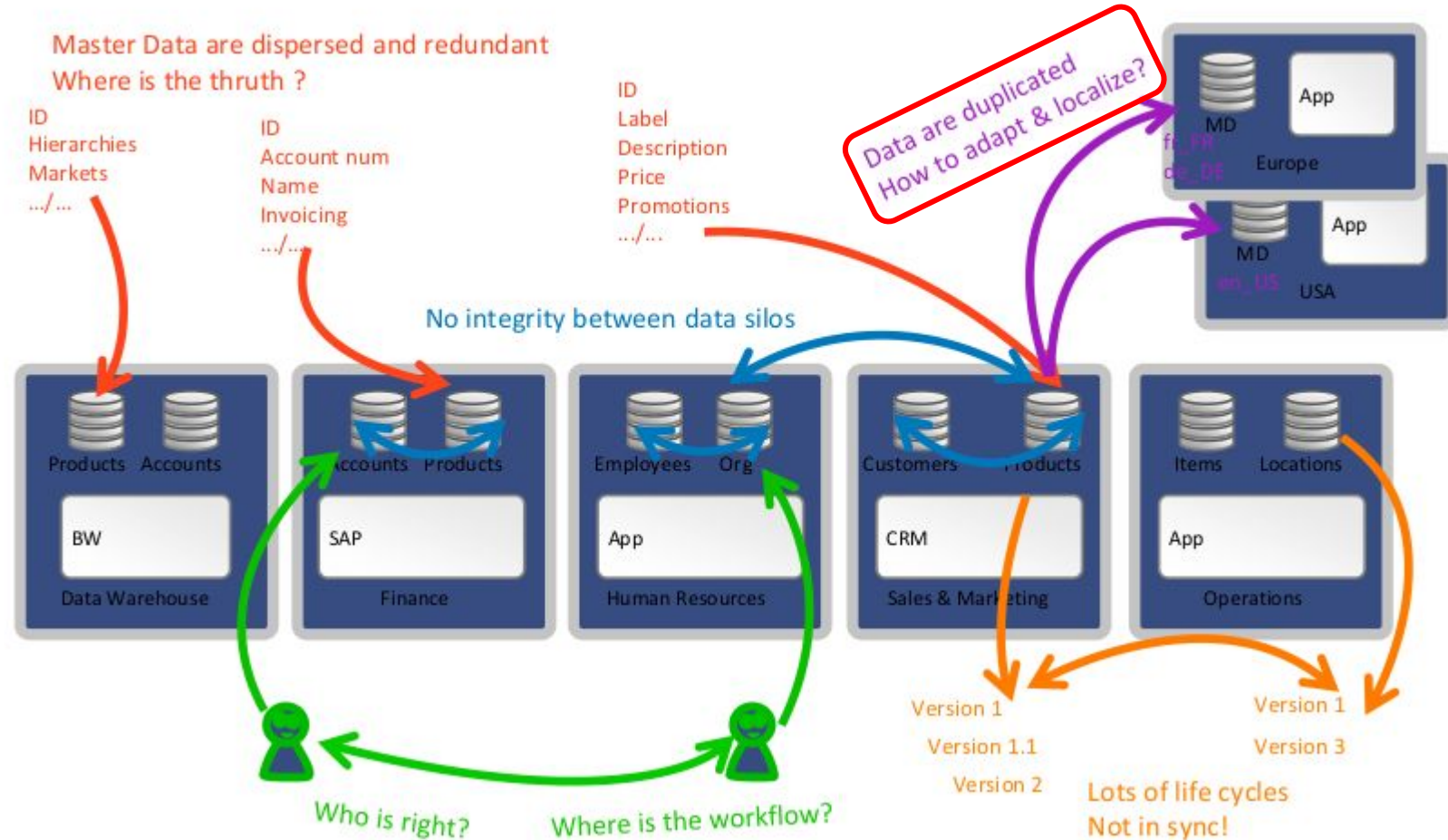
# Introducción a MDM

## Problemas



# Introducción a MDM

## Problemas





# Introducción a MDM

## Problemas

MDM como capacidad de negocio es **difícil de lograr** debido a:

- La **complejidad** de las alternativas de integración y arquitectura.
- La **falta de gobierno** de datos.
- **Procesos** existentes que impiden la captura de datos de **alta calidad**.
- **Costos** de implementación prohibitivos combinados con escaso alcance y priorización.

Tirando de ironía: esta solución que ayuda a habilitar una versión única de la verdad, no cuenta con una versión única de la verdad con respecto a su propia definición de mercado.

Hay muchas definiciones de lo que es MDM o de cómo hacerlo, y todas válidas.

# Introducción a MDM

## Debate

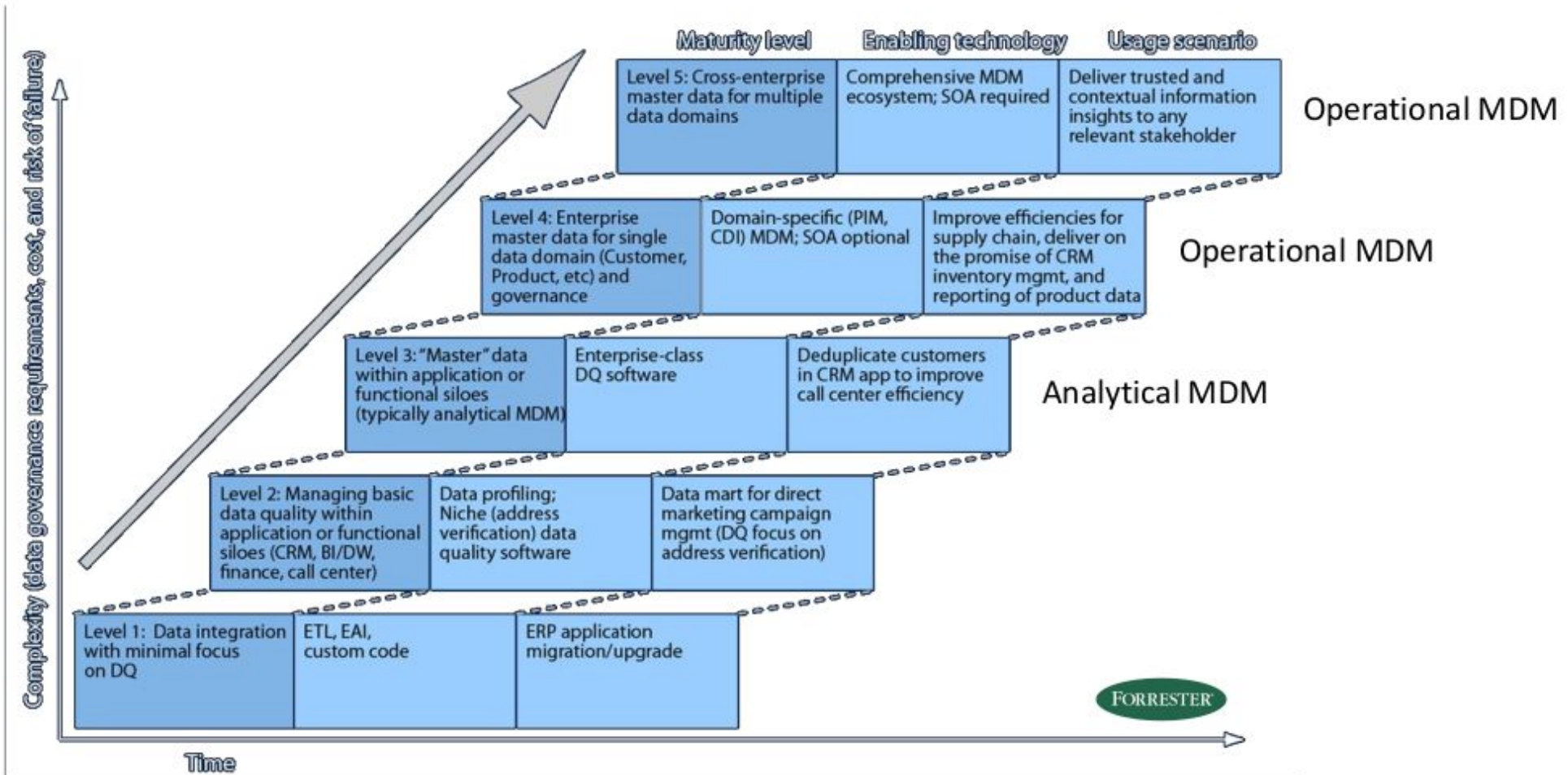
En vuestras empresas:

¿Sabéis dónde están los datos “buenos” que necesitáis?

¿Os fiáis de ellos?

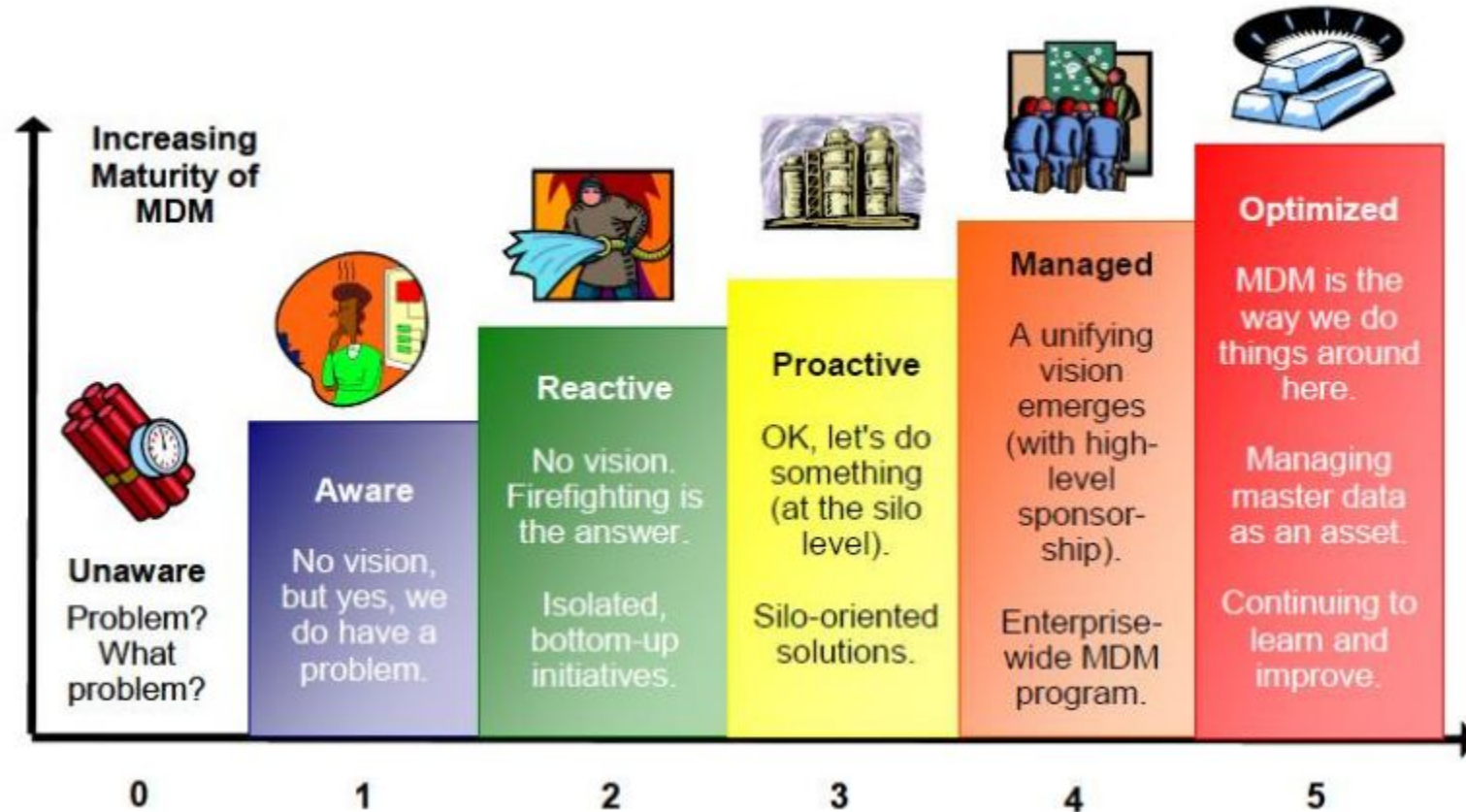
# Introducción a MDM

## Modelo de madurez de MDM (Forrester)

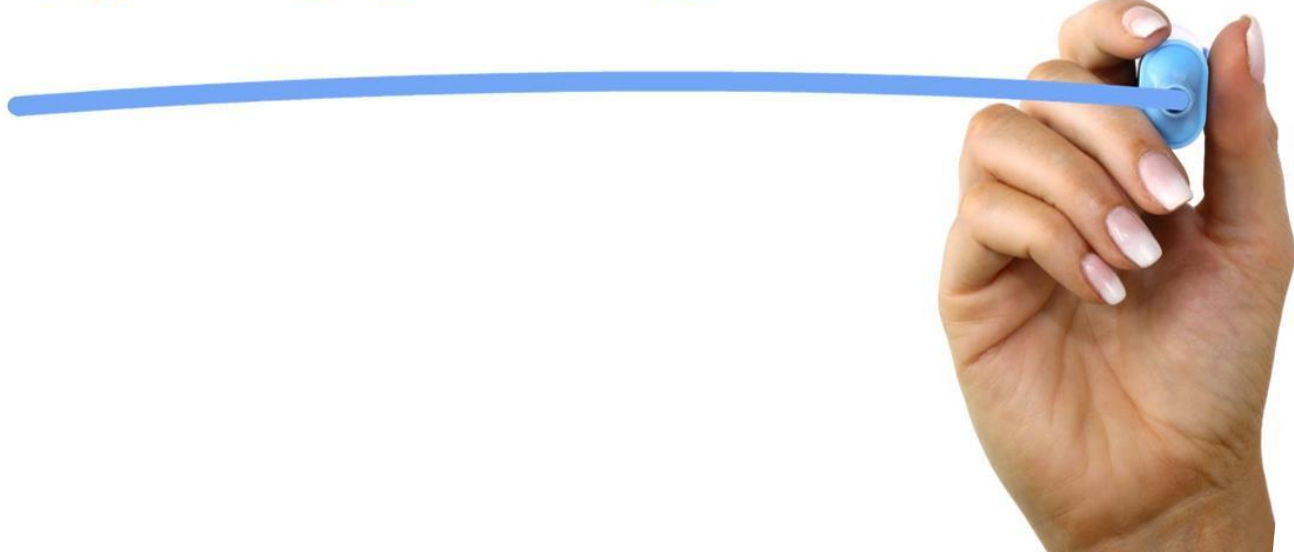


# Introducción a MDM

## Modelo de madurez de MDM (Gartner)



# INDEX



## Índice

1. Introducción a MDM
2. **Arquitectura MDM**
3. Valor de Master data
4. Gestión de proyectos MDM
5. Conclusiones
6. Herramientas

# Arquitectura MDM

## Ejemplo



Occupation = Ski Instructor



Address = Pontresina, Switzerland



Purchased €500 in outdoor gear in 2015



Member of Loyalty Program since 2010



Top Finisher in Engadin Ski Marathon 2010-2015



Stefan Krauss  
Age = 31



100% of purchases online



Prefers Text Message



# Arquitectura MDM

## Ejemplo



Occupation = Banker



Address = Zurich, Switzerland



Purchased €3.500 in outdoor gear in 2015



Member of Loyalty Program since 1990



**Stefan Krauss**  
Age = 62



75% of spending is while on holiday



Football Fan



Prefers Physical Mail



100% of spending in store



# Arquitectura MDM

## Transaction Data vs Master Data

Customer	Date	Product	Code	Price	Quantity	Location
Stefan Kraus	1/2/2017	Scarpa Telemark Ski Boot	SC1279	€250	1	St. Moritz, CH
Donna Burbank	1/5/2017	Scarpa Telemark Ski Boot	SCU1289	\$150	1	Boulder, CO
Stefan Kraus	1/2/2017	North Face Down Jacket	NF8392	€450	1	Zurich, CH
Stefan Kraus	1/2/2017	Garmin Sports Watch	GM29384	€200	2	Zurich, CH
Wendy Hu	3/4/2017	Prana Yoga Pant	PN82734	\$51	5	New York, NY
Joe Smith	4/1/2017	Garmin Sports Watch	GM29384	\$150	1	Albany, NY

### Transaction Data

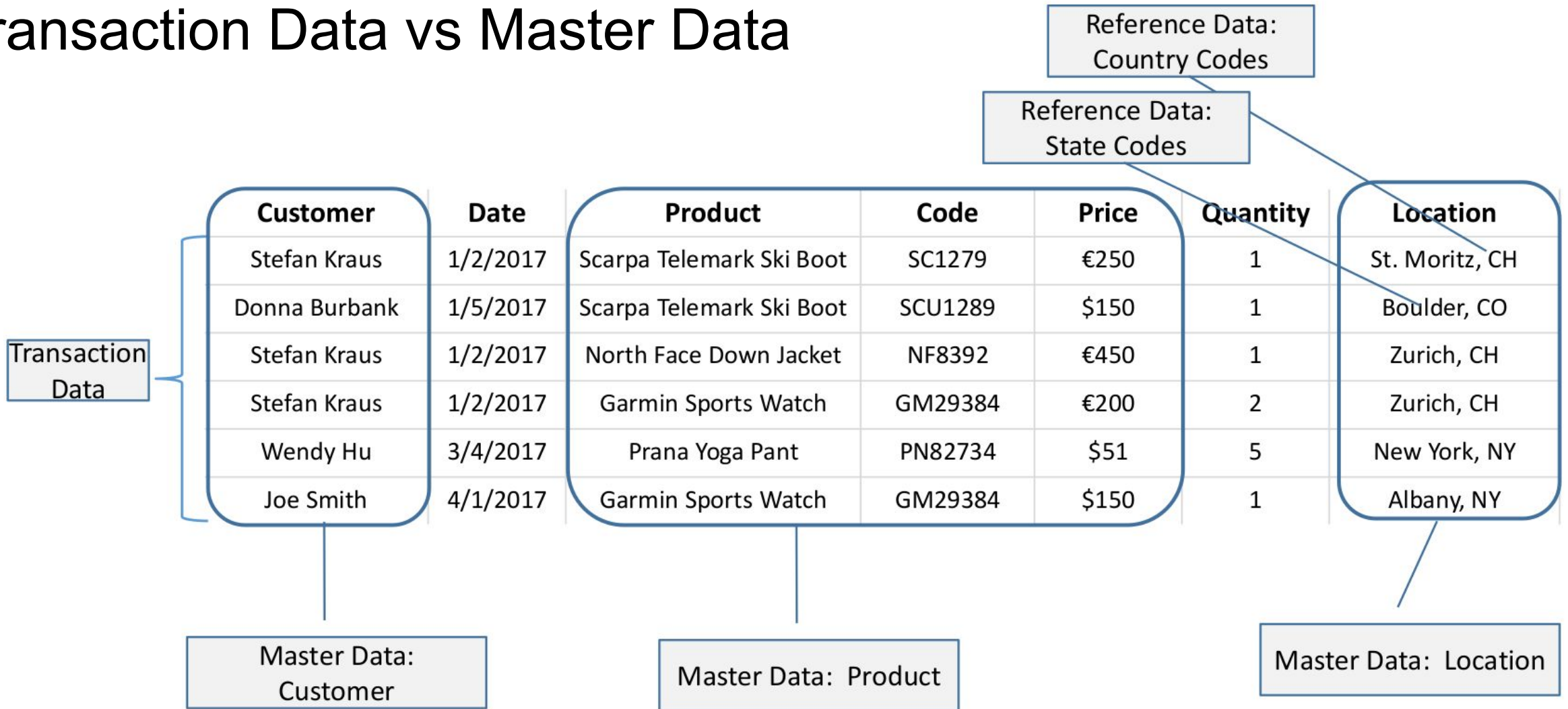
- Describes an action (verb): E.g. “buy”
- May include measurements about the action: (Who, When, What, How Many, Where, How Much, etc.)
- E.g. Stefan Kraus, 1/2/2017/, Scarpa Telemark Ski Boot, St. Moritz, CH, €250

### Master Data

- Describes the key entities (nouns), e.g. Customer, Product, Location
- Provides attributes & context for these nouns
- e.g. Wendy Hu, age 25, female, resident of New York, NY, Customer since 2005, preferred customer card, etc.

# Arquitectura MDM

## Transaction Data vs Master Data



# Arquitectura MDM

## MDM analítico/informacional

MDM analítico, se enfoca en proporcionar una **visión empresarial unidireccional** de la información a través de la gestión de jerarquía controlada por versión y capacidades de modelado dimensional

- Por ejemplo, las familias de productos, los canales de ventas y las regiones de ventas son vistas comunes administradas en estos entornos.

Muchos clientes **comienzan** su viaje de MDM con MDM analítico

- El MDM analítico es más fácil de abordar y es un primer paso recomendado porque se trata principalmente de los datos.
- Porque su naturaleza direccional introduce mucho menos riesgo y complejidad que intentar sincronizar bidireccionalmente los datos maestros con la aplicación de producción crítica.

El MDM analítico generalmente corresponde con el tercer nivel del modelo de madurez MDM de Forrester.

# Arquitectura MDM

## MDM operacional

MDM operacional se enfoca en consolidar datos de fuentes de datos dispares en un entorno analítico reconciliado (generalmente un data warehouse o un datastore operacional) para informes y análisis.

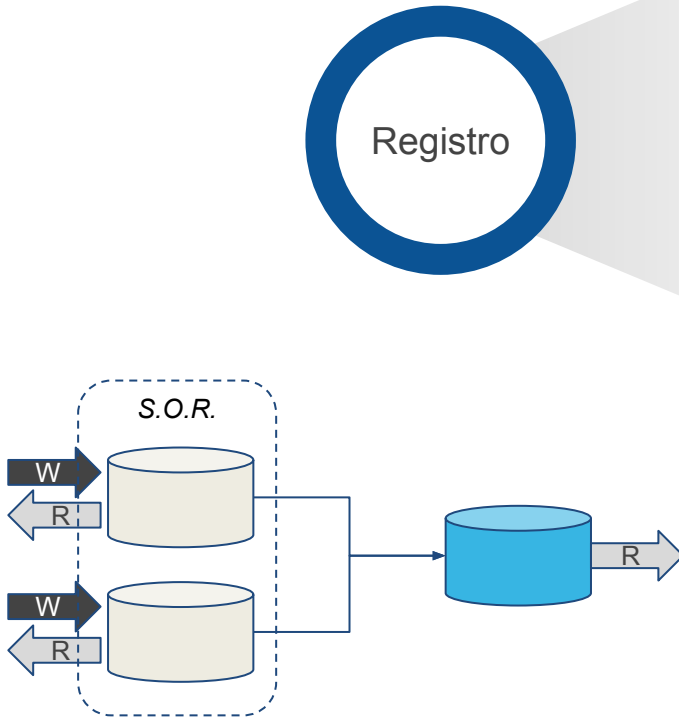
Sincroniza bidireccionalmente datos maestros confiables en tiempo real, en entornos de información heterogéneos.

- Requiere la necesidad mucho más desafiante de sincronizar los procesos comerciales y los datos.
- El MDM operativo generalmente corresponde con los niveles cuatro y cinco del modelo de madurez MDM.

# Arquitectura MDM

## Tipos de arquitecturas

- Este estilo de implementación de MDM se utiliza principalmente para detectar duplicados en los datos de diversos sistemas origen mediante la ejecución de algoritmos de limpieza y matching.
- Asigna identificadores únicos a los registros coincidentes, con el fin de ayudar a identificar una única versión de la información.
- Este estilo no envía datos de vuelta a los sistemas origen, por lo que los cambios en los datos maestros deben realizarse sobre los sistemas origen existentes.
- En este caso, el MDM no es generalmente el system-of-record para los atributos de los datos maestros del dominio en cuestión, si no que las fuentes de datos operacionales actúan como system-of-record.

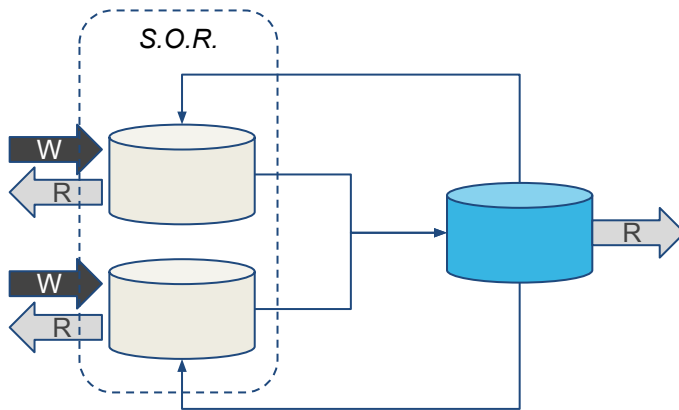


# Arquitectura MDM

## Tipos de arquitecturas



- En este estilo los datos maestros se consolidan generalmente en el MDM desde múltiples fuentes para crear una visión única de la información (conocido como “golden record”).
- Los datos almacenados en el MDM se utilizan principalmente con fines de reporting y análisis.
- Sin embargo, cualquier actualización realizada sobre los datos maestros debe aplicarse posteriormente sobre las fuentes de origen.
- En este caso, el MDM no es generalmente el system-of-record autorizado para los atributos de datos maestros del dominio en cuestión, si no que las fuentes de datos operacionales actúan como system-of-record.

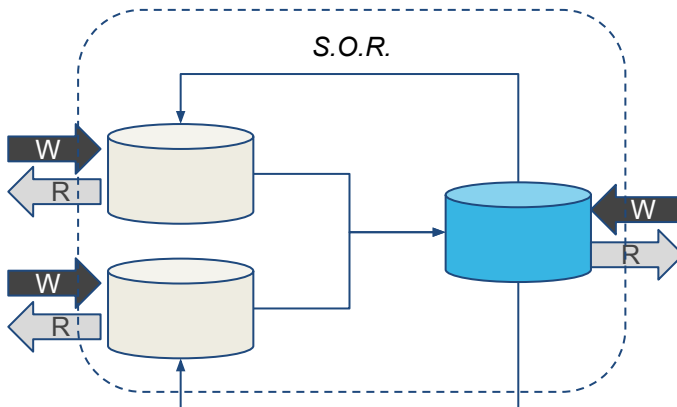


# Arquitectura MDM

## Tipos de arquitecturas



- El estilo de Convivencia permite construir un “golden record” de la misma manera que el estilo de Consolidación, pero sus datos maestros se almacenan en el sistema central de MDM y se actualizan sobre los sistemas de origen.
- Este estilo puede ser más costoso de implementar que el estilo de Consolidación, ya que los cambios en los datos maestros pueden ocurrir tanto en el sistema MDM como en los sistemas origen.
- Por último, en este estilo, la solución MDM generalmente es sólo el system-of-record autorizado para una parte de los atributos de datos maestros para el dominio en cuestión. Las fuentes de datos operacionales actúan como system-of-records para el resto de atributos.



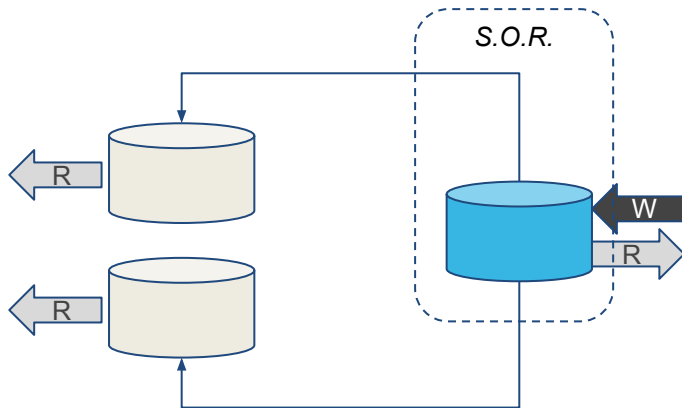


# Arquitectura MDM

## Tipos de arquitecturas



- El estilo Centralizado almacena y mantiene los atributos de datos maestros, utilizando algoritmos de limpieza, matching y enriquecimiento para mejorar los datos. Los datos mejorados pueden luego actualizarse en el sistema de origen correspondiente.
- El sistema MDM admite la consolidación de datos maestros, y los sistemas origen pueden suscribirse a las actualizaciones publicadas por el sistema central para obtener una consistencia total. Sin embargo, este estilo requiere intrusión en los sistemas origen para interacciones en ambos sentidos (sistemas origen-MDM y vice versa).
- En este estilo, el MDM es generalmente el system-of-record autorizado para todos los atributos de datos maestros.



# Arquitectura MDM

## Dudas a resolver para decidir la arquitectura

¿Qué requerimientos tienes?

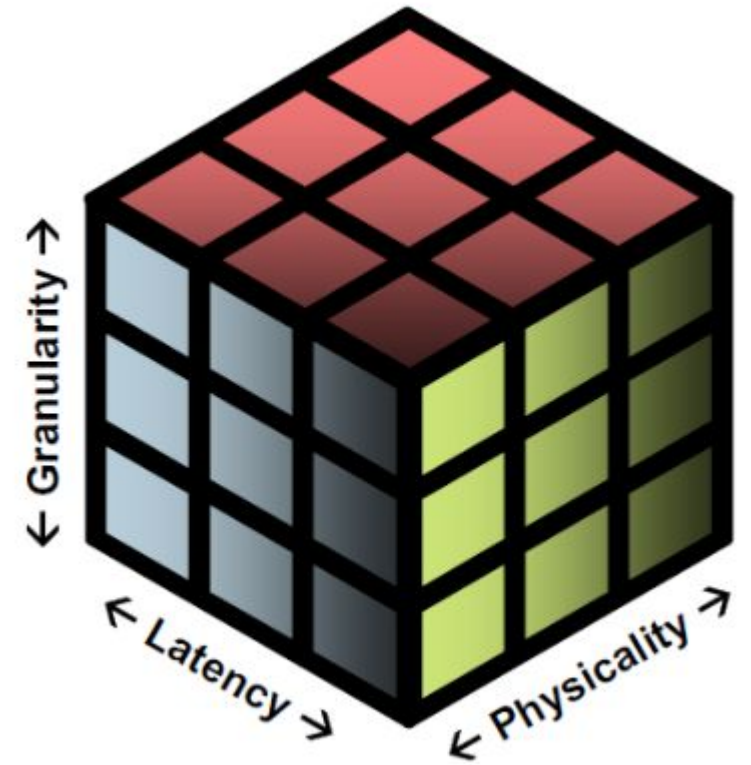
- Batch vs real-time
- Carga masiva vs registros individuales
- Almacenamiento físico vs en memoria

¿Hasta dónde puede llegar tu solución?

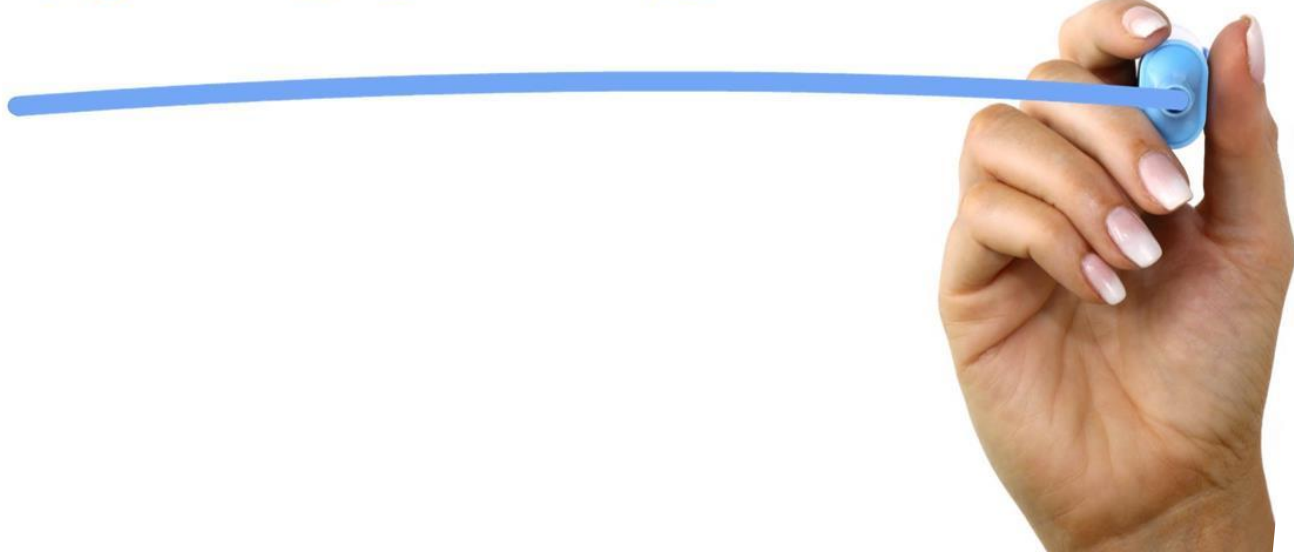
- Bifurcar ETLs, eventos de CDC, modelo federado (virtual), replicación de datos,...

¿Qué es mejor?

- Herramientas
- Arquitecturas de referencia



# INDEX



## Índice

1. Introducción a MDM
2. Arquitectura MDM
- 3. Valor de Master data**
4. Gestión de proyectos MDM
5. Conclusiones
6. Herramientas

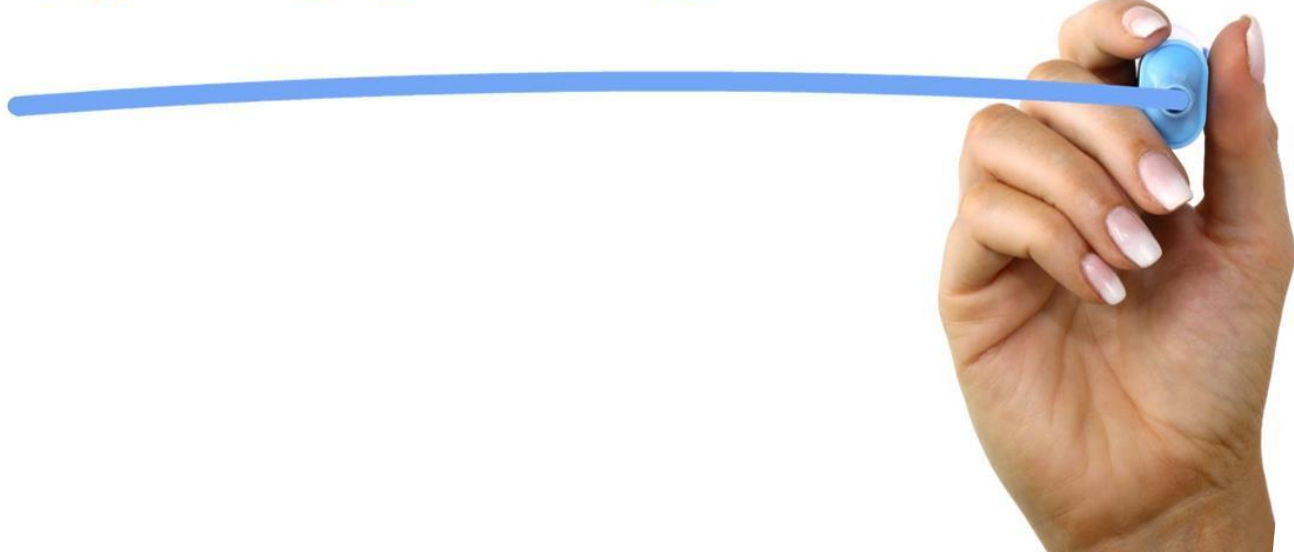
# Valor de Master Data

## Beneficios

Si bien la creación de un MDM puede ser un desafío **desalentador**, existen muchos **beneficios positivos** para el resultado final de tener un repositorio maestro común, que incluye:

- Una sola **factura consolidada**, que ahorra dinero y mejora la satisfacción del cliente.
- **Evitar** preocupaciones sobre el envío de la misma información publicitaria a un cliente **repetidas veces** desde múltiples listas de clientes, lo que desperdicia dinero e irrita al cliente
- Una visión **coherente** de los clientes en **toda la organización**, de esa manera los usuarios saben antes de pasar una cuenta de cliente a una agencia de morosidad si ese cliente debe o no dinero a otras partes de la organización o, lo que es más importante, si ese cliente es la **mayor fuente de ingresos** de otra división de negocios
- Una vista consolidada del stock de productos para eliminar el desperdicio de dinero y espacio en los y almacenes, así como evitar el riesgo de escasez que proviene de almacenar el mismo artículo con diferentes números de serie

# INDEX



## Índice

1. Introducción a MDM
2. Arquitectura MDM
3. Valor de Master data
- 4. Gestión de proyectos MDM**
5. Conclusiones
6. Herramientas

# Gestión de proyectos MDM

## Por dónde empezar

### Identificación

Identificar fuentes de datos maestros

Identificar los productores y consumidores de los datos maestros

Recopilar y analizar metadatos para sus datos maestros

Nombrar Data Stewards

### Definición

Implementar un programa y un comité de gobierno de datos

Desarrollar el modelo de datos maestros.

Elegir un conjunto de herramientas

Diseña la infraestructura

### Implementación

Generar y probar los datos maestros.

Modificar los sistemas productores y consumidores.

Implementar procesos de mantenimiento.

# Gestión de proyectos MDM

## Por dónde empezar

Identificar fuentes de datos maestros

Este paso suele ser un ejercicio muy revelador. Algunas compañías encuentran que tienen docenas de bases de datos que contienen datos de clientes que el departamento de TI no sabía que existían.

Identificar los productores y consumidores de los datos maestros

Este paso implica determinar qué aplicaciones producen los datos maestros del primer paso y es más difícil determinar qué aplicaciones usan los datos maestros. Dependiendo del enfoque que utilice para mantener los datos maestros, este paso podría no ser necesario si todos los cambios se detectan y manejan a nivel de la base de datos.

Recopilar y analizar metadatos para sus datos maestros

Para todas las fuentes identificadas en el paso uno, ¿cuáles son las entidades y atributos de los datos, y qué significan?  
Nombre del Atributo, Tipo de datos, Valores permitidos, Restricciones, Valores predeterminados, Dependencias

Nombrar Data Stewards

Estas deberían ser las personas con el conocimiento de los datos de origen actuales y con la capacidad de determinar cómo transformar los datos de origen en el formato de datos maestros.

# Gestión de proyectos MDM

## Por dónde empezar

Implementar un programa y un comité de gobierno de datos

Este grupo debe tener el conocimiento y la autoridad para tomar decisiones sobre cómo se mantienen los datos maestros, qué contienen, cuánto tiempo se conservan y cómo se autorizan y auditan los cambios. Cientos de decisiones deben tomarse en el curso de un proyecto de datos maestros, y si no hay un cuerpo y proceso de toma de decisiones bien definido, el proyecto puede fallar por temas políticos

Desarrollar el modelo de datos maestros.

Decidir cómo se verán los registros maestros, incluidos qué atributos se incluyen, qué tamaño y tipo de datos son, qué valores están permitidos, etc. Este paso también debe incluir mapeo entre el modelo de datos maestros y las fuentes de datos actuales.

Elegir un conjunto de herramientas

Se deberá comprar o crear herramientas para crear los datos maestros limpiando, transformando y fusionando los datos de origen. También necesitará una infraestructura para usar y mantener la lista maestra.

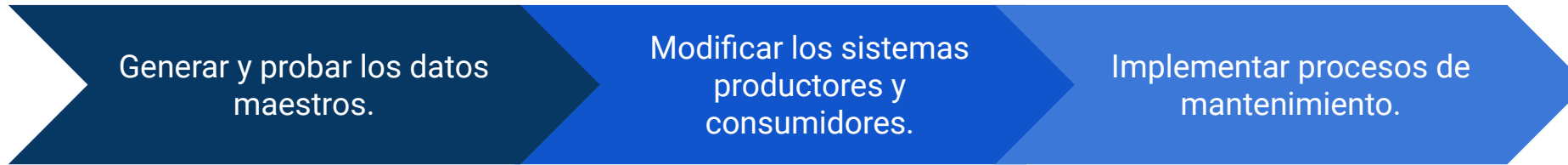
Diseñar la infraestructura

Una vez se tengan datos maestros limpios y consistentes, se deberá exponer a aplicaciones y proporcionar procesos para administrarlos y mantenerlos. La confiabilidad y la escalabilidad son consideraciones importantes para incluir en el diseño.



# Gestión de proyectos MDM

## Por dónde empezar



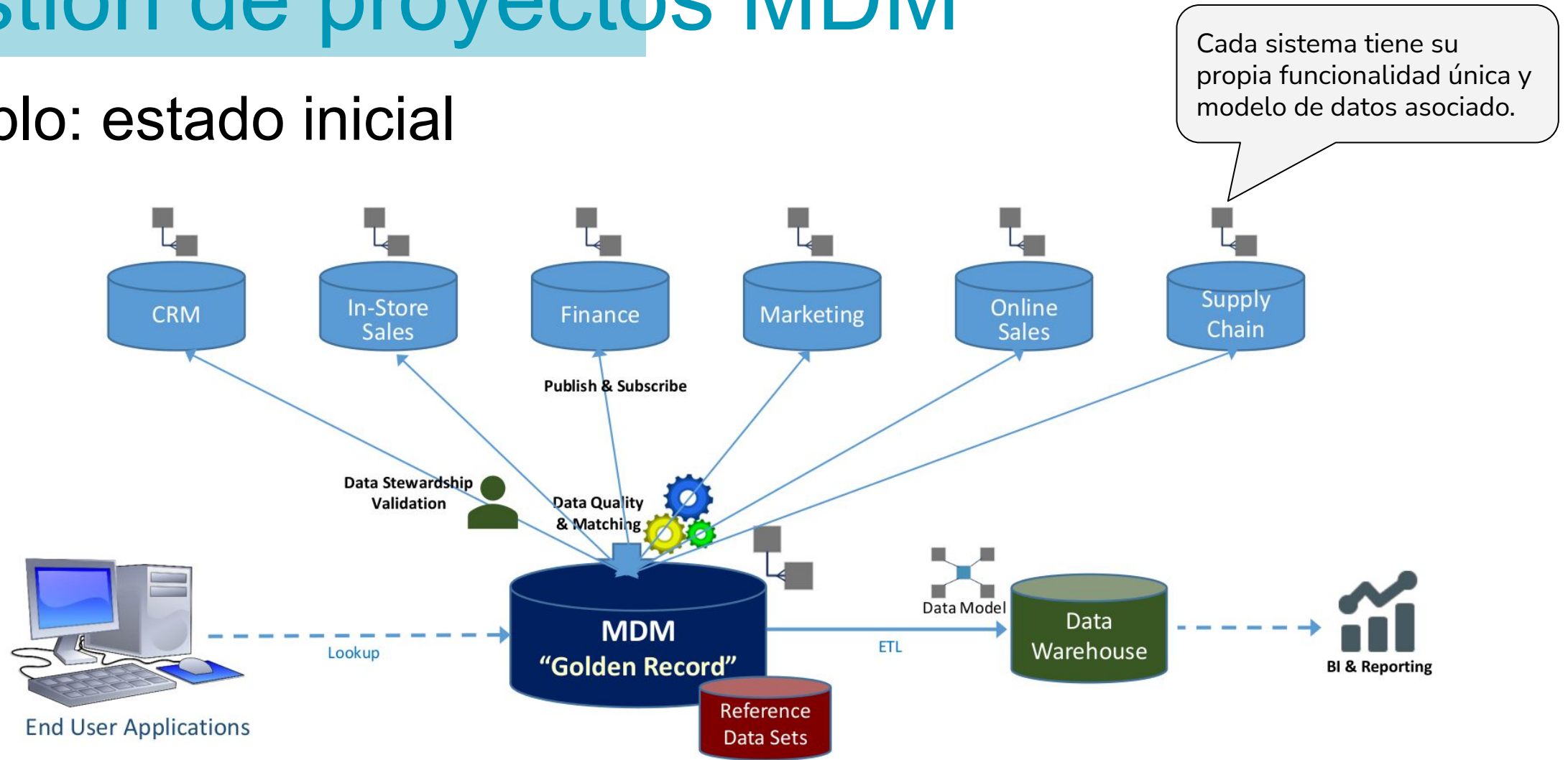
Este paso es donde utiliza las herramientas para fusionar los datos de origen en su lista de datos maestros. A menudo, es un proceso iterativo que requiere modificar las reglas y la para obtener la correspondencia correcta. Este proceso también requiere mucha inspección manual para garantizar que los resultados sean correctos y cumplan con los requisitos establecidos para el proyecto.

Dependiendo de cómo esté diseñada la implementación de MDM, es posible que deba cambiar los sistemas que producen, mantienen o consumen datos maestros para trabajar con la nueva fuente de datos maestros. Si los datos maestros se usan en un sistema separado, es posible que los sistemas origen no tengan que cambiar.

Cualquier implementación de MDM debe incorporar herramientas, procesos y personas para mantener la calidad de los datos. Todos los datos deben tener un data steward que sea responsable de garantizar la calidad de los datos maestros.

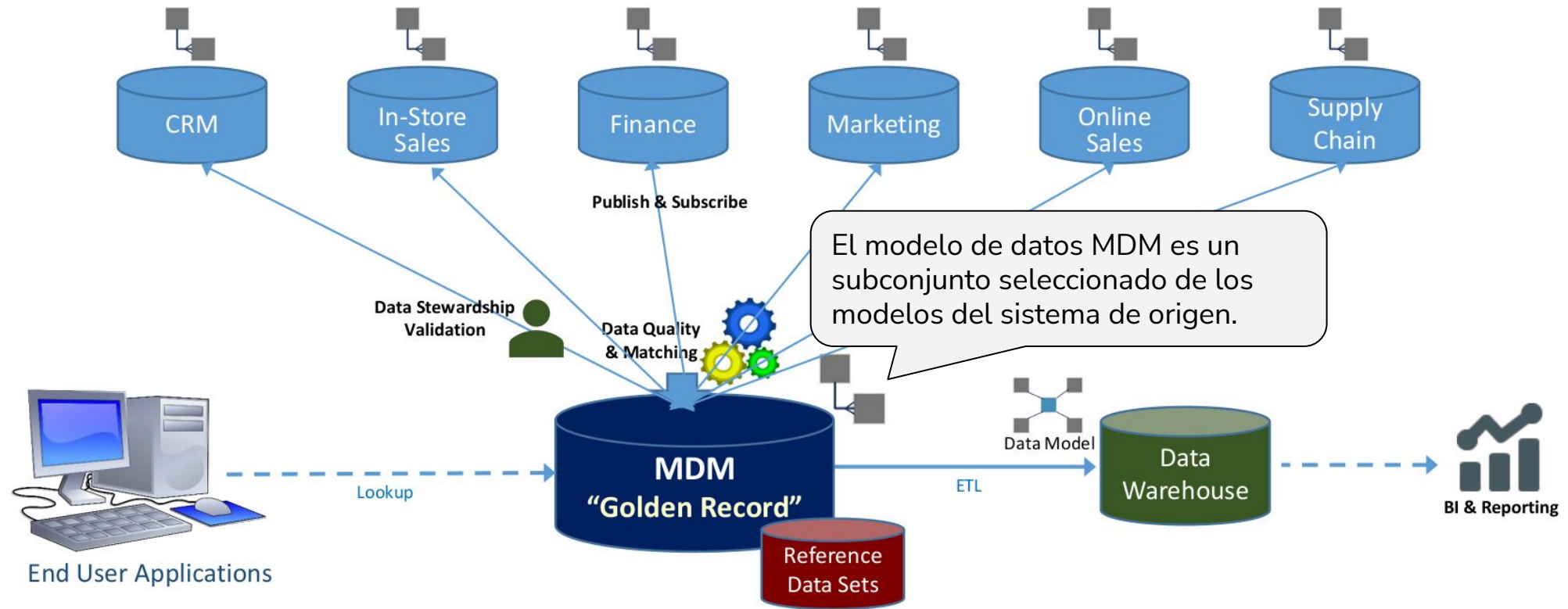
# Gestión de proyectos MDM

## Ejemplo: estado inicial



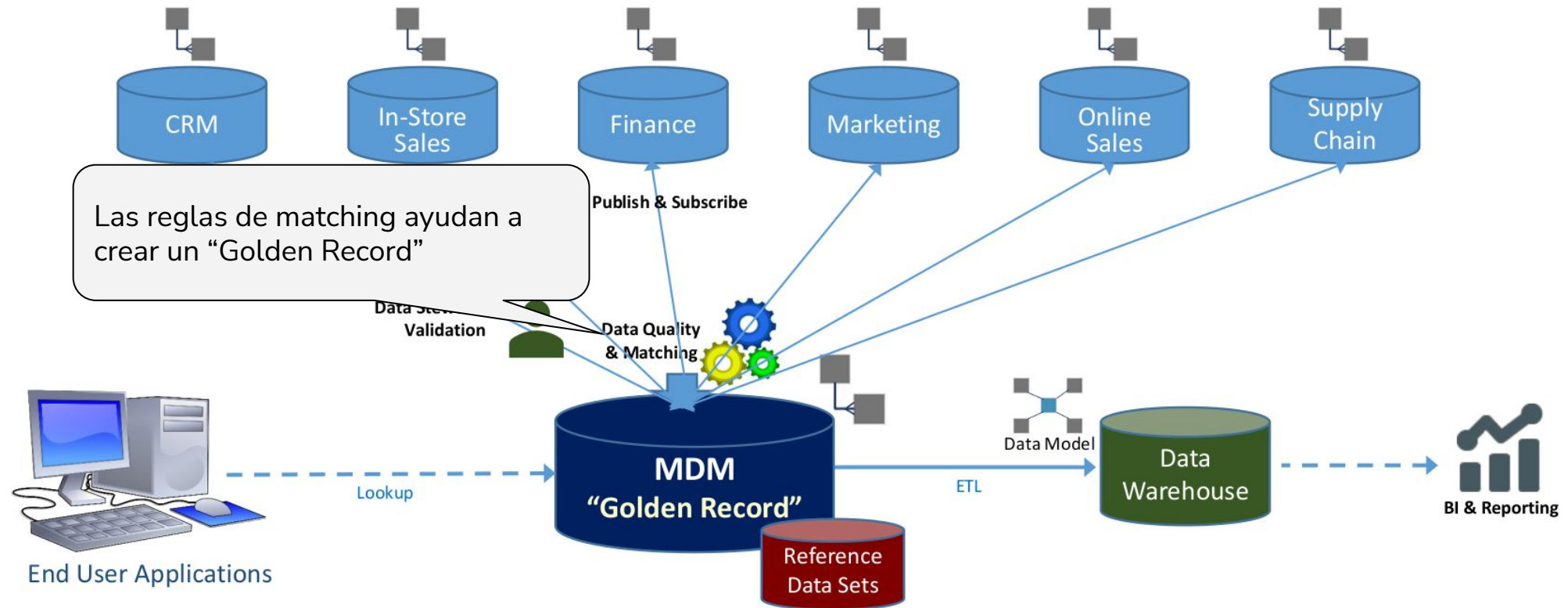
# Gestión de proyectos MDM

## Ejemplo: modelo MDM



# Gestión de proyectos MDM

## Ejemplo: matching



# Gestión de proyectos MDM

## Ejemplo: matching

Las combinaciones de campos candidatos para el matching, a menudo se alinean con las primary keys del modelo de datos lógicos.

### Customer

Customer ID
Date of Birth (AK1.1)
SSN (AK1.2,AK2.1)
First Name
Last Name (AK2.2)
Maiden Name
Middle Name
Name Prefix
Name Suffix
Phone Number
Email
Gender

### Matching on Primary Key

Ideally, if all systems use the same unique identifier, matching is easier.  
But this isn't often realistic in "real world" systems.

### Matching on Alternate Keys

- First, match on Date of Birth + SSN
- Then, match on SSN + Last Name
- Etc.

# Gestión de proyectos MDM

## Ejemplo: fuzzy matching

Se puede usar fuzzy matching, que es particularmente útil para hacer coincidir campos de cadenas como nombres y direcciones, donde los errores humanos o diferentes estándares de entrada de datos entre sistemas pueden causar ligeras variaciones en valores similares, p. Ej.

- "101 Main St" frente a "101 Main Street"
- "John Smith" frente a "J Smith"

Además, se pueden crear sinónimos para ayudar con la coincidencia, por ejemplo

- "St", "St.", "Street", etc. para direcciones
- "Tim", "Timothy" para nombres y apodos.

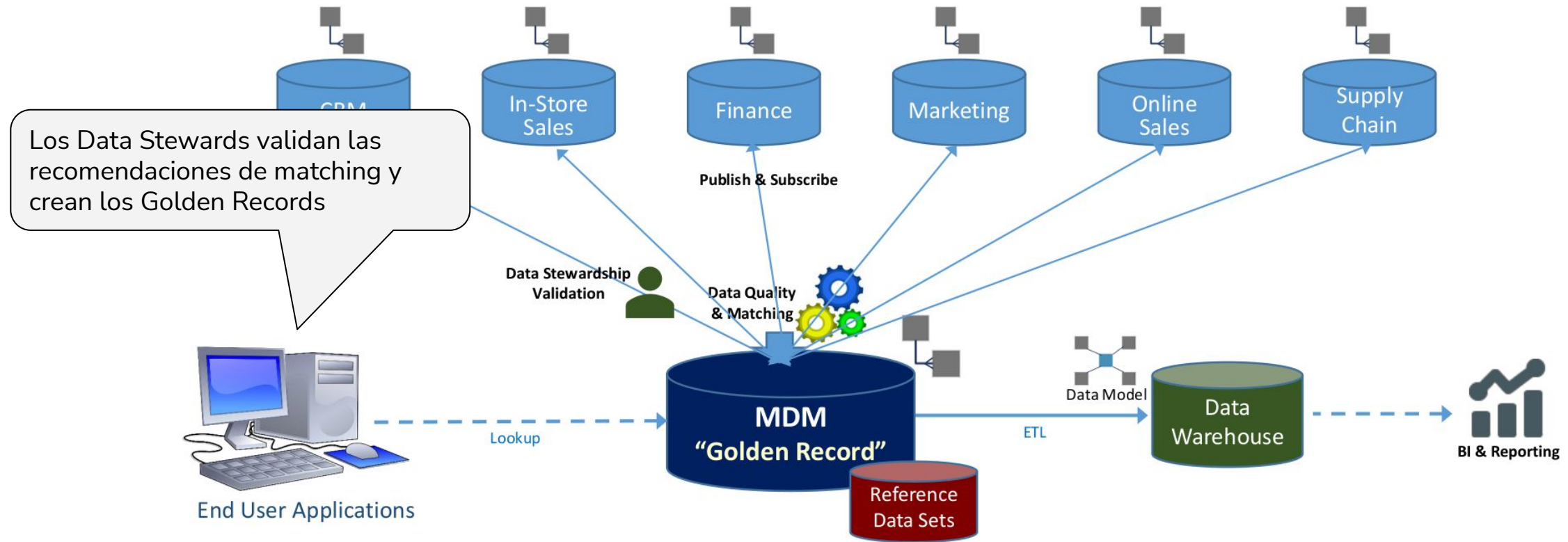
Cuando se utiliza fuzzy matching, se pueden definir umbrales de calidad de datos para la aprobación automática.

- Se crean puntuaciones de coincidencia para cada coincidencia aproximada, por ejemplo, 0.9 indicaría una coincidencia fuerte y 0.2 una débil.
- Utilizando estas puntuaciones como guía, se pueden definir umbrales para los cuales las coincidencias se pueden aprobar automáticamente, las que se pueden rechazar automáticamente y las que necesitan la revisión humana de un administrador de datos.



# Gestión de proyectos MDM

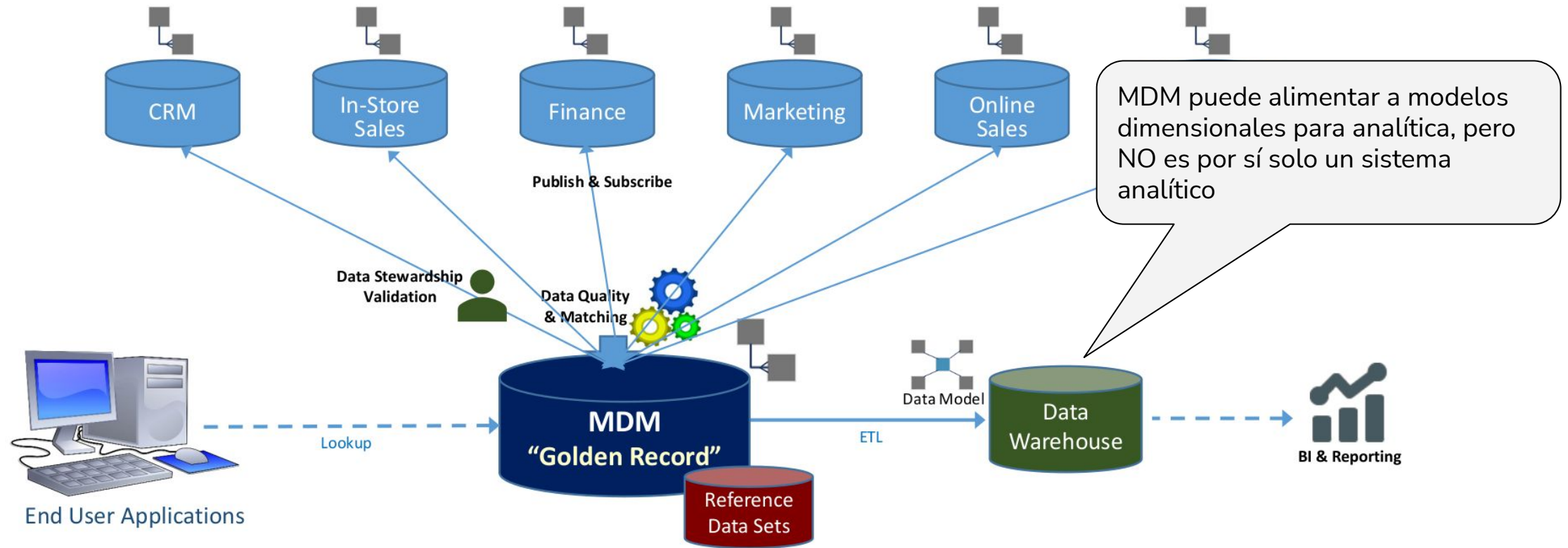
## Ejemplo: matching



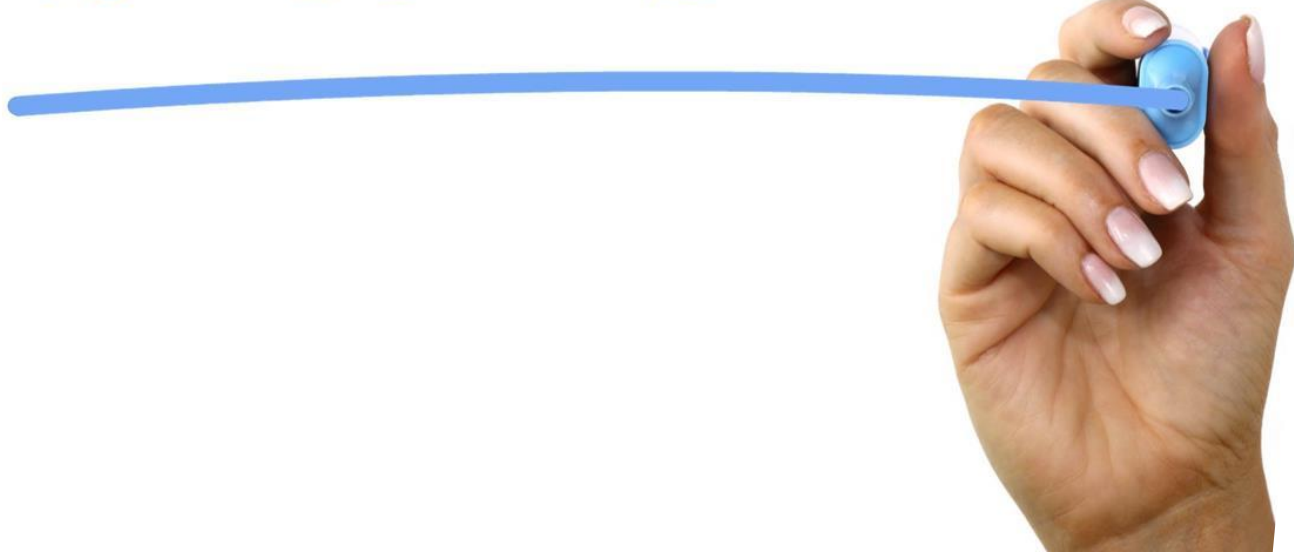


# Gestión de proyectos MDM

## Ejemplo: fuente para analítica



# INDEX



## Índice

1. Introducción a MDM
2. Arquitectura MDM
3. Valor de Master data
4. Gestión de proyectos MDM
- 5. Conclusiones**
6. Herramientas

# Conclusiones

## Barreras comunes que impiden el éxito de MDM

- Considerar MDM como puramente una iniciativa tecnológica.
- Asumir que los datos sucios son solo un problema de TI
- Administrar la vasta complejidad de múltiples dominios de datos sin las técnicas adecuadas, incluidos modelos de datos comunes, API de integración y funciones habilitadas para servicios web.
- Falta de enfoque en el gobierno de datos, priorización, personas y procesos.
- Subestimar el nivel de patrocinio ejecutivo requerido para el éxito
- Priorizar ineficazmente la financiación y la gestión de costes.

# Conclusiones

## Recomendaciones

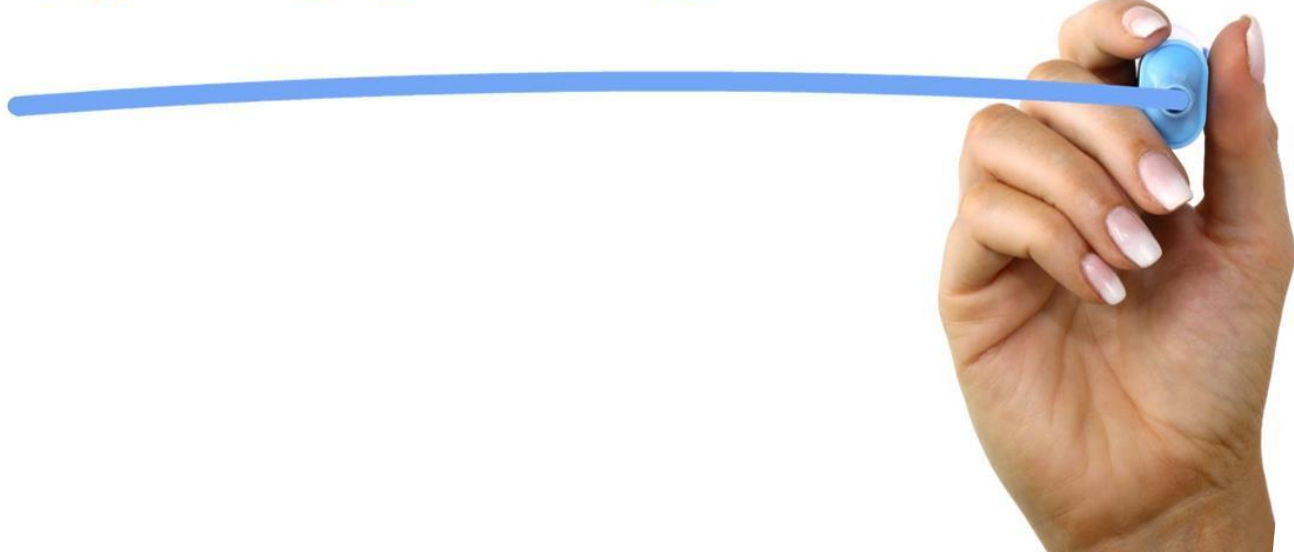
Considere las estrategias de calidad de datos que respaldan las demandas de la empresa:

- Priorice sus objetivos de calidad de datos enfocándose en elementos de datos que respalden sus procesos más críticos para el negocio.
- Comience con la calidad de datos basada en proyectos.
- Súbete a las iniciativas de gestión de datos entre empresas.
- Adopte la gobernanza de datos para permitirle pasar de DQ basado en proyectos a MDM de clase empresarial.

Los datos maestros se convertirán en el punto principal en la arquitectura SOA

- Las soluciones MDM independientes de la aplicación proporcionarán un contexto más rico para una SOA que los enfoques específicos de la aplicación

# INDEX



## Índice

1. Introducción a MDM
2. Arquitectura MDM
3. Valor de Master data
4. Gestión de proyectos MDM
5. Conclusiones
6. **Herramientas**

# Herramientas

## Informatica



Informatica ofrece una plataforma MDM modular que incluye sus herramientas MDM multidominio, resolución de identidad, cliente 360, relación 360, proveedor 360 y producto 360. Incorpora capacidades de integración de datos, calidad de datos, seguridad y gestión de procesos de negocios, y puede usarse para crear aplicaciones personalizadas que le permitan visualizar las relaciones entre sus grandes datos. También aprovecha las capacidades de aprendizaje automático.

Figure 1. Magic Quadrant for Master Data Management Solutions



Source: Gartner (January 2020)

# Herramientas

## TIBCO



Adquirió Orchestra Networks, una empresa especializada en software de Master Data Management, y como tal asegura soluciones diseñadas con departamentos de IT en mente.

Gran variedad de opciones de MDM: almacenamiento de datos semánticos, metamodelos, arquitectura distribuida D3, integraciones con múltiples sistemas de terceros, entre ellos otros grandes proveedores como Informatica, IBM, Microsoft, Oracle,... Ofrece apps nativas para iOS y Android y el software puede lanzarse desde diversas plataformas.

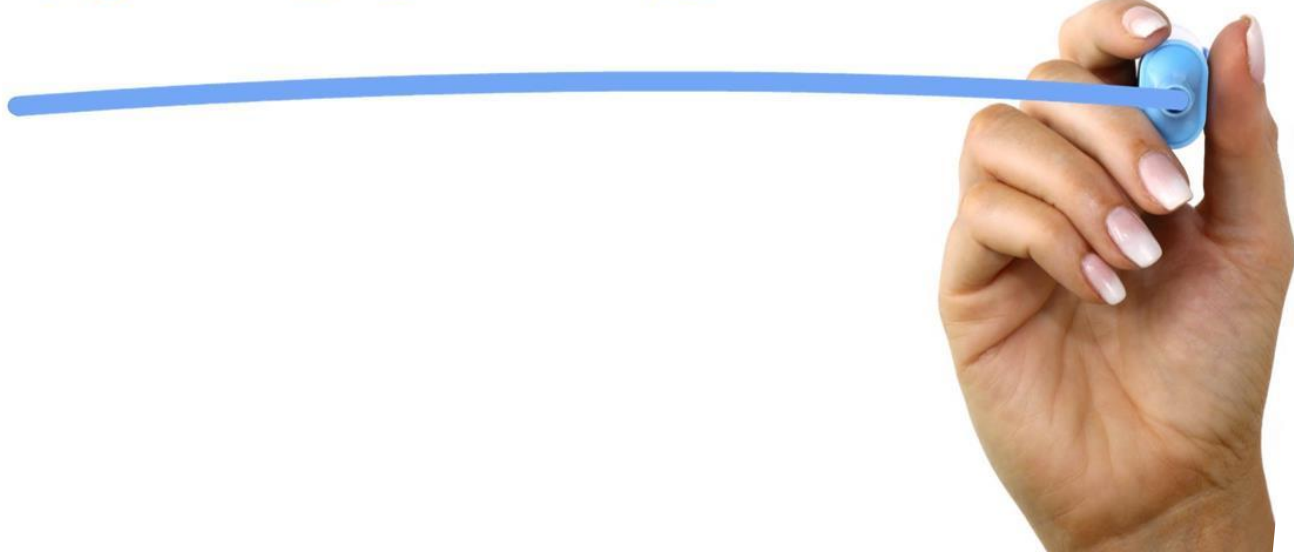
Figure 1. Magic Quadrant for Master Data Management Solutions



Source: Gartner (January 2020)



# INDEX



## Índice

1. Introducción a MDM
2. Arquitectura MDM
3. Valor de Master data
4. Gestión de proyectos MDM
5. Conclusiones
6. Herramientas
7. **Ejemplo práctico**

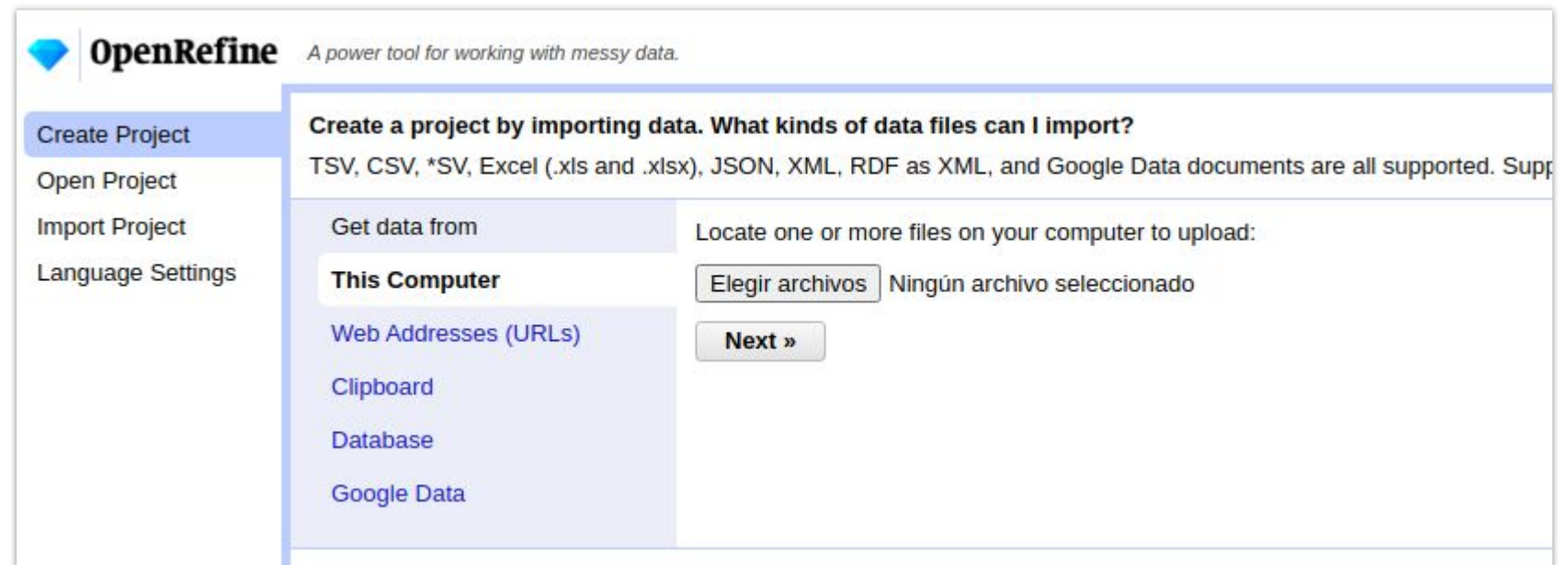
# Ejemplo práctico

## Fuzzy Matching

OpenRefine: <https://openrefine.org/download.html>

Librería de matching: <http://okfnlabs.org/reconcile-csv/>

Datasets: <https://tinyurl.com/yxg5tl9k>



# Ejemplo práctico

## Arrancar Open Refine

Arrancamos la aplicación:

**Linux:** en ssh, ejecutar `./refine`

<https://github.com/OpenRefine/OpenRefine/wiki/Installation-Instructions#linux>

**Windows:** double-click en openrefine.exe o refine.bat si falla el primero

<https://github.com/OpenRefine/OpenRefine/wiki/Installation-Instructions#windows>

Acceder a <http://127.0.0.1:3333/>

# Ejemplo práctico

## Arrancar el servicio de matching

Seguir los pasos de <http://okfnlabs.org/reconcile-csv/>

1. Descargar la librería jar.
2. Descargar los dos csv del caso práctico: <https://tinyurl.com/yxg5tl9k>
3. Poner tanto el jar como el fichero match.csv en la misma carpeta
4. Ejecutar el comando (desde ssh para linux, desde cmd para windows)

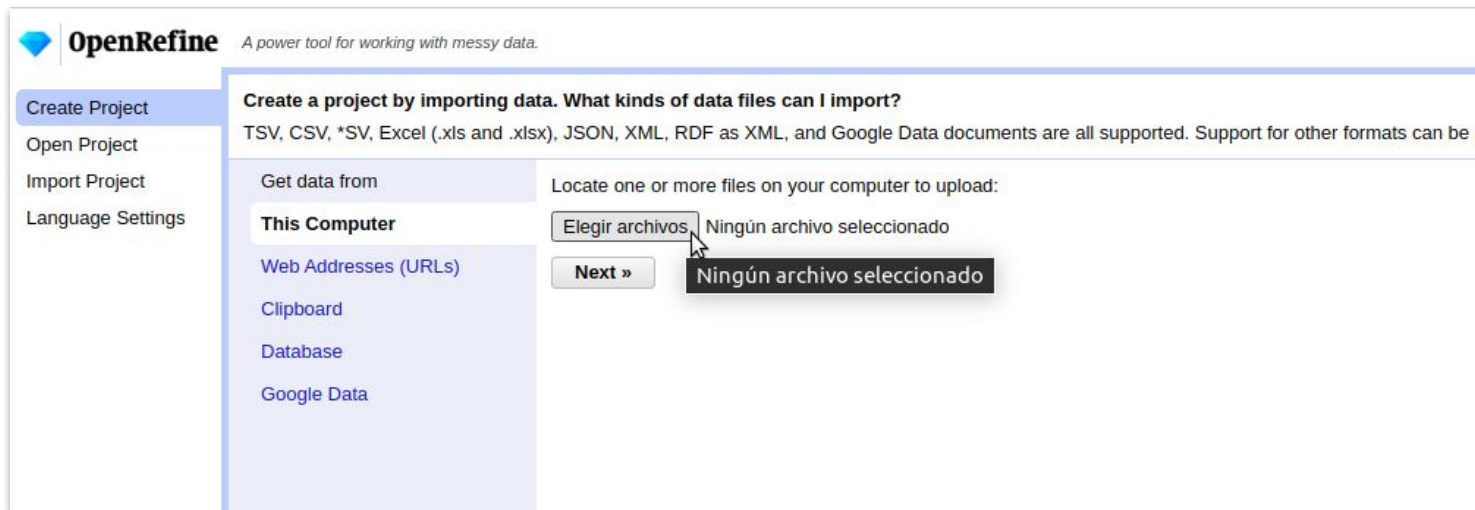
```
java -Xmx2g -jar reconcile-csv-0.1.2.jar match.csv match_name id_match
```

```
alfonsofernandez@alfonsofernandez:~/Descargas$ java -Xmx2g -jar reconcile-csv-0.1.2.jar match.csv match_name id_match
Starting CSV Reconciliation service
Point refine to http://localhost:8000 as reconciliation service
2021-01-07 09:12:38.027:INFO:oejs.Server:jetty-7.x.y-SNAPSHOT
2021-01-07 09:12:38.047:INFO:oejs.AbstractConnector:Started SelectChannelConnector@0.0.0.0:8000
```

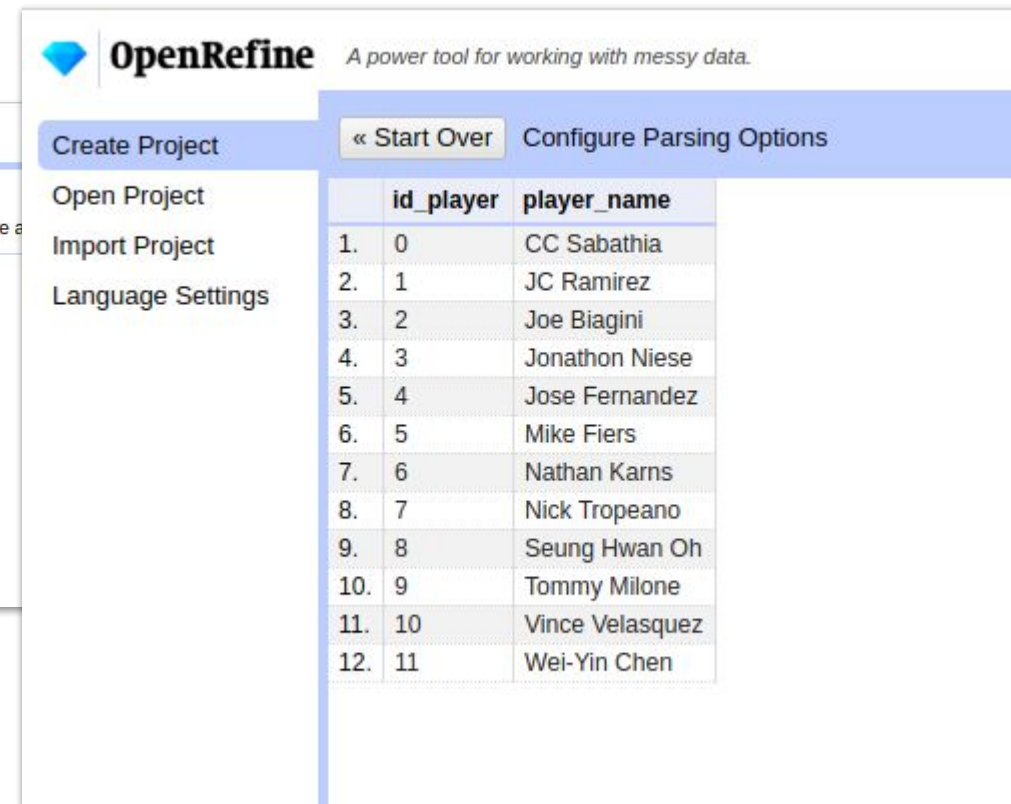
# Ejemplo práctico

## Importar fichero

Ir a open refine, e importar el fichero players.csv



The screenshot shows the 'Create Project' step in OpenRefine. The left sidebar has 'Create Project' selected. The main area asks 'What kinds of data files can I import?' and lists supported formats: TSV, CSV, \*SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents. Under 'Get data from', 'This Computer' is selected. A text box says 'Locate one or more files on your computer to upload:'. Below it is a button 'Elegir archivos' (Choose files) and a 'Next »' button. A tooltip over the 'Elegir archivos' button says 'Ningún archivo seleccionado' (No file selected).



The screenshot shows the 'Open Project' step in OpenRefine. The left sidebar has 'Open Project' selected. The main area shows a table of data with columns 'id\_player' and 'player\_name'. The table contains 12 rows of data. Above the table are buttons '« Start Over' and 'Configure Parsing Options'.

	id_player	player_name
1.	0	CC Sabathia
2.	1	JC Ramirez
3.	2	Joe Biagini
4.	3	Jonathon Niese
5.	4	Jose Fernandez
6.	5	Mike Fiers
7.	6	Nathan Karns
8.	7	Nick Tropeano
9.	8	Seung Hwan Oh
10.	9	Tommy Milone
11.	10	Vince Velasquez
12.	11	Wei-Yin Chen

# Ejemplo práctico

## Crear el proyecto

Elegir un nombre para el proyecto, y darle a “Create Project”



**OpenRefine** A power tool for working with messy data.

« Start Over    Configure Parsing Options

Project name:     Tags:     **Create Project »**

	id_player	player_name
1.	0	CC Sabathia
2.	1	JC Ramirez
3.	2	Joe Blagini
4.	3	Jonathon Niese
5.	4	Jose Fernandez
6.	5	Mike Fiers
7.	6	Nathan Karns
8.	7	Nick Tropeano
9.	8	Seung Hwan Oh
10.	9	Tommy Milone
11.	10	Vince Velasquez
12.	11	Wei-Yin Chen

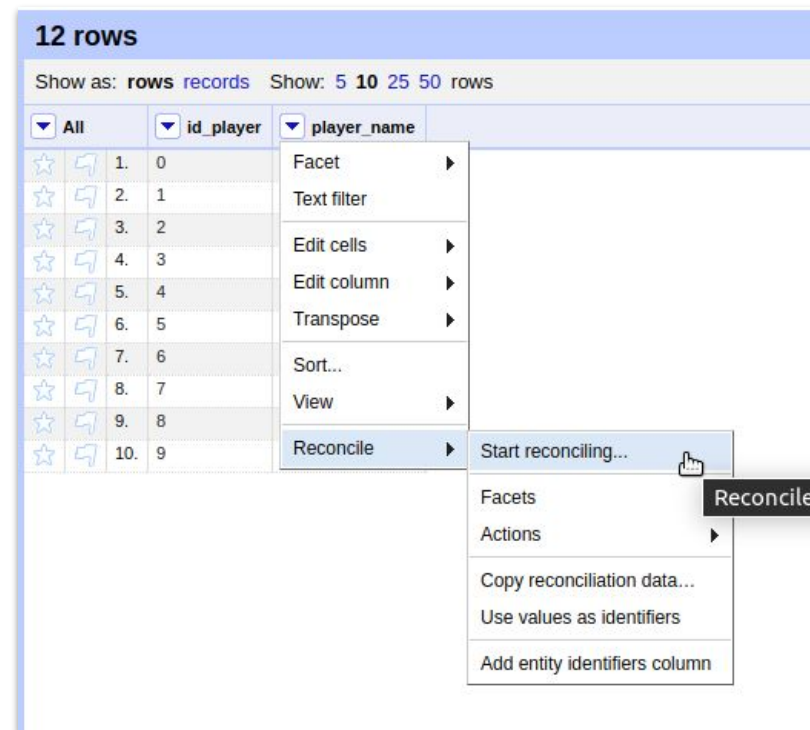
# Ejemplo práctico

## Matching de ambas columnas

Tenemos en el servicio que hemos arrancado en Java el fichero match.csv listo, y en el proyecto de open refine el fichero players.csv.

Ahora vamos a indicar desde open refine que nos haga una reconciliación de la columna player\_name sobre el servicio.

Como muestra la imagen, ir a la columna player\_name, ver las opciones, y elegir Reconcile → Start reconciling...





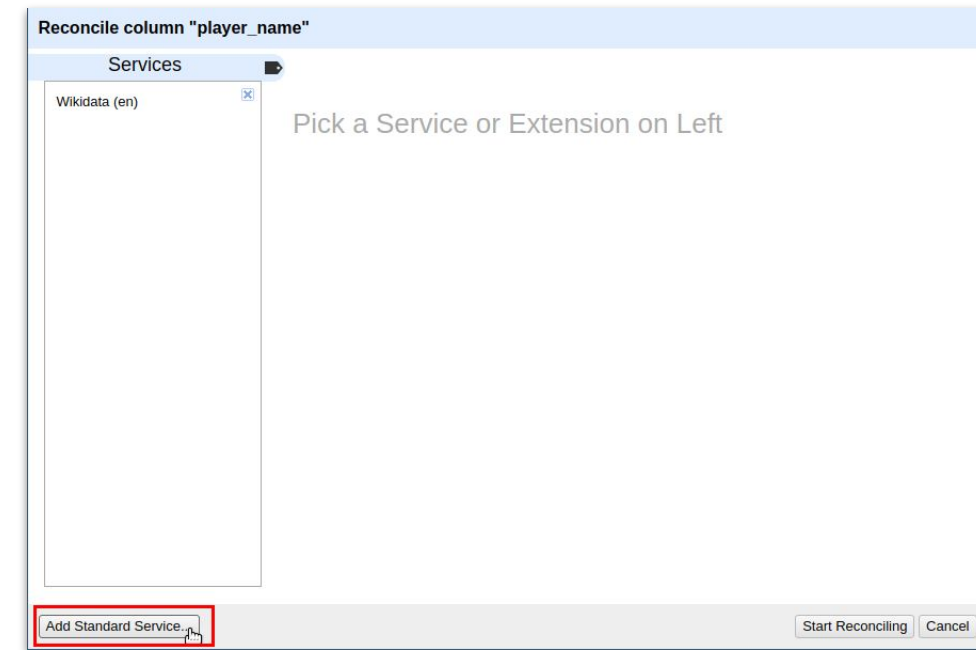
# Ejemplo práctico

## Matching de ambas columnas

Tenemos que añadir el servicio java que tenemos levantado.

Hacer click en Add Standard Service.

Añadir la URL: <http://localhost:8000/reconcile>



# Ejemplo práctico

## Matching de ambas columnas

En la parte derecha, puedes elegir incluir más columnas para que se muestren en el dataset resultado del matching. En este caso vamos a meter el id\_player.

En la parte inferior se puede elegir si se quiere hacer matching automático para valores de coincidencia altos. Lo marcamos y damos a Start reconciling.

The screenshot shows a dialog box titled "Reconcile column 'player\_name'". In the top right corner, there is a link "» Access Service API". The dialog is divided into two main sections. The left section, titled "Reconcile each cell to an entity of one of these types:", contains a list with one item: "CSV-recon /csv-recon", which is selected with a radio button. Below this list are three radio button options: "Reconcile against type:" (with an empty text field), "Reconcile against no particular type", and "Auto-match candidates with high confidence" (which is checked). Below these is a label "Maximum number of candidates to return" followed by an empty text field. The right section, titled "Also use relevant details from other columns:", contains a table with the following structure:

Column	Include?	As Property
id_player	<input checked="" type="checkbox"/>	<input type="text"/>

At the bottom of the dialog, there are three buttons: "Add Standard Service...", "Start Reconciling", and "Cancel".

# Ejemplo práctico

## Matching de ambas columnas

Como resultado vemos el dataset con la columna que queríamos machear (player\_name) y la que añadimos (player\_id). Por cada fila nos muestra el resultado de aproximación entre ambos datasets.

Por ejemplo: CC Sabathia:

C.C. Sabathia(0.762)

Jonathan Niese(0.174)

Nate Karns(0.105)

Joseph Biagini(0.087)

The screenshot shows the OpenRefine web interface. The main table displays 12 rows of data with columns for 'id\_player' and 'player\_name'. The 'player\_name' column is faceted by 'judgment' and 'best candidate's score'. A matching dialog is open on the right, showing a list of suggestions for the cell 'C.C. Sabathia' in the 'player\_name' column. The suggestions include 'C.C. Sabathia (0.762)', 'Jonathan Niese (0.174)', 'Nate Karns (0.105)', 'Joseph Biagini (0.087)', and 'J.C. Ramirez (0)'. The dialog also shows the 'match\_name' as 'C.C. Sabathia' and the 'id\_match' as '101'.

id_player	player_name
0	CC Sabathia
1	JC Ramirez
2	Joe Biagini
3	Jonathan Niese

# Ejemplo práctico

## Confirmar matching

Con esta recomendación, el usuario puede confirmar manualmente los resultados con dos opciones:

- Match this cell: confirma y realiza el cambio de esa celda
- Match all identical cells: para todas las filas idénticas, realiza el cambio.

Los cambios se puede deshacer (en la pestaña undo/redo). O volver a revisar el matching en cada celda (Choose new match)

OpenRefine players.csv Permalink

Facet / Filter Undo / Redo 2 / 2

Refresh Reset All Remove All

**player\_name: judgment** change

1 choices Sort by: name count

none 12

Facet by choice counts

**player\_name: best candidate's score** change reset

0.31 — 0.88

12 rows

Show as: rows records Show: 5 10 25 50 rows

			id_player	player_name
☆	1.	0		C.C. Sabathia Choose new match
☆	2.	1		JC Ramirez <input checked="" type="checkbox"/> J.C. Ramirez (0.7) <input checked="" type="checkbox"/> Tom Milone (0.111) <input checked="" type="checkbox"/> Michael Fiers (0.095) <input checked="" type="checkbox"/> Vincent Velasquez (0.08) <input checked="" type="checkbox"/> C.C. Sabathia (0) <input checked="" type="checkbox"/> Create new item Search for match
☆	3.	2		Joe Biagini <input checked="" type="checkbox"/> Joseph Biagini (0.696) <input checked="" type="checkbox"/> Jonathan Niese (0.174) <input checked="" type="checkbox"/> Nate Karns (0.105) <input checked="" type="checkbox"/> C.C. Sabathia (0.095)

# Ejemplo práctico

## ¿Qué hemos hecho?

Para dos datasets de personas diferentes (dos fuentes de datos), hemos aplicado un **algoritmo de fuzzy matching** para saber si alguno de los datos se “parecen” a los del otro. El sistema da una recomendación, y es el usuario el que debe confirmar o no.

Estos métodos son los que se usan en las **herramientas de MDM** para intentar generar la fuente única comparando datasets de distintas fuentes. En algunos casos, la recomendación es acertada y hay procesos de **machine learning** que ayudan a que el proceso se más preciso en cada iteración.

Estos sistemas **supervisados** ayudan a deduplicar registros y generar un único golden record.



**¡Muchas gracias!**



# Píldora DAMA LATAM



## Chapter 10: Reference and Master Data



**Keyla Dolores Méndez**

- Data Engineer en Farmacias Peruanas
- Microsoft Certified MCT
- Speaker
- Blogger