Data Governance- Azure Data Catalog

BIG DATA – BUSINESS INTELLIGENCE – MACHINE LEARNING

CONTENIDO

1. GOBIERNO DEL DATO

1.1 INTRODUCCIÓN

En el presente documento se pretende dar una pincelada al gobierno del dato, destacando los conceptos principales y resumiendo ciertos aspectos a tener en cuenta para llevar a cabo un proyecto de este tipo. Se propone la herramienta Azure Data Catalog como software de apoyo para el desarrollo.

Dentro de un concepto global Data Management, el concepto de Data Governance se encarga de la dirección y supervisión de los datos, identificando además las siguientes áreas que están interrelacionadas:

- Arquitectura de datos: Define el plan de gestión de los activos de datos alineándose con la estrategia organizativa para establecer los requisitos de datos estratégicos y los diseños para cumplir con estos requisitos.

- Modelización y diseño de datos: Es el proceso de descubrir, analizar, representar y comunicar los requisitos de datos en una forma precisa llamada modelo de datos.

- Almacenamiento y operaciones de datos: Incluye el diseño, la implementación y el mantenimiento de los datos almacenados para maximizar su valor. Las operaciones se realizan a lo largo del ciclo de vida de los datos, desde la planificación hasta la eliminación de los mismos.

- Seguridad de los datos: La seguridad de los datos garantiza que la privacidad y la confidencialidad de los datos se mantengan, que no se violen los datos y que se acceda a ellos de manera adecuada.

- Integración e interoperabilidad de datos: Incluye procesos relacionados con el movimiento y consolidación de datos dentro y entre almacenes de datos, aplicaciones y organizaciones.

- Gestión de documentos y contenidos: Incluye las actividades de planificación, implementación y control utilizadas para gestionar el ciclo de vida de los datos y la información que se encuentran en una serie de medios, especialmente los documentos necesarios para apoyar los requisitos de cumplimiento legal y reglamentario.

- Datos de referencia y maestros: Incluye la conciliación y el mantenimiento continuos de datos compartidos fundamentales para permitir el uso coherente en todos los sistemas de la versión más exacta, oportuna y pertinente de la verdad sobre las entidades comerciales esenciales.

- Data warehousing y BI: Incluye los procesos de planificación, ejecución y control para gestionar los datos de apoyo a la toma de decisiones y permitir a los analistas de datos obtener valor de los datos mediante el análisis y la presentación de informes.

- Metadatos: Incluye las actividades de planificación, ejecución y control para permitir el acceso a metadatos integrados y de alta calidad que incluyen definiciones, modelos, flujos de datos y otra información crítica para comprender los datos y el sistema a través del cual se crean, se mantienen y se accede a ellos.

- Calidad de los datos: Incluye la planificación y aplicación de técnicas de gestión de la calidad para medir, evaluar y mejorar la idoneidad de los datos para su uso dentro de una organización.

Ilustración 1. DAMA-DMBOK2 Framework

Existen diversos frameworks que nos pueden ayudar a la hora de abordar un proyecto de este tipo estableciendo criterios, equipo de trabajo, actividades a desarrollar, entregables y buenas prácticas para la correcta gestión del dato como son DAMA, DGI o DCAM

1.2. EQUIPO DE TRABAJO, ACTIVIDADES Y ENTREGABLES

En este punto mencionaremos tres aspectos importantes a destacar.

- Equipo de trabajo: En lugar de centrarnos en "títulos" de puestos concretos, vamos a destacar las habilidades que debemos reunir para cubrir todas las áreas a desarrollar:
    - Procesos de negocio: Comprender los requisitos y determinar el impacto en los datos
    - Modelado y arquitectura de datos: Establecer un roadmap y arquitectura
    - Habilidades técnicas de arquitectura: Diseño de dw, construcción y gestión
    - Manipulación de datos: Resolución de problemas estructurales, relación entre datos y datasources
    - Analítica de datos: Interpretación de datos, representación gráfica
    - Habilidades con el idioma: Definir glosarios comprensibles
    - Habilidades estratégicas de negocio: Comprender las necesidades de la organización
- Actividades:
    - Evaluación de la situación actual, identificación de necesidades e identificación de casos de uso
    - Definición de la estrategia y el framework operativo centrado en casos de uso
    - Gestión del cambio, creación del glosario de negocio (business glosary) y coordinación con otras áreas
    - Hacer operativo el modelo para toda la organización
- Entregables
    - Plan estratégico y roadmap
    - Análisis de procesos
    - Framework operativo y roles de trabajo
    - Glosario de negocio

- Cuadro de mando resumen
- Plan de gestión del cambio, comunicación y formación
- Buenas prácticas, guías

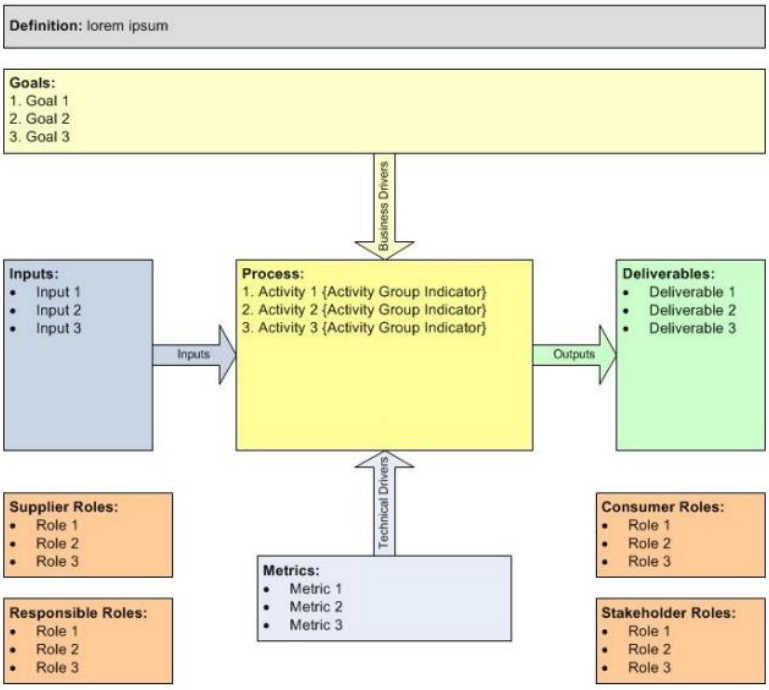


Ilustración 2. Template para documentación proceso de negocio

2. AZURE DATA CATALOG

En este documento nos centraremos en la herramienta Azure Data Catalog que nos permite definir un catálogo central y gobernado de datos de confianza.

Se trata de definir un catálogo de datos a nivel corporativo que cubra todas las fuentes de datos de la organización permitiendo una trazabilidad e identificación del dato entre las distintas fuentes. En otras palabras, Azure Data Catalog es un repositorio de metadatos que permite el registro, enriquecimiento, comprensión y consumo de los datos de una organización de manera transversal.

Dicho catálogo puede ser compartido y elaborado de manera colaborativa fácilmente. Puede descubrir, perfilar, organizar y documentar automáticamente sus metadatos y hace que sea fácil de buscar.

Tal y como se menciona en el punto anterior, el data governance es solo una pieza dentro del data management y por tanto, no hay que restar importancia a otros procesos como:

- Data Quality: limpieza, integración, profiling
- Definir los BPM de nuestro negocio
- Gestión de los workflows y responsabilidades entre los roles de la organización

2.1 PRINCIPALES CARACTERÍSTICAS

Podemos clasificar la funcionalidad que nos ofrece en tres bloques:

- **Publicación**: La publicación de datos se puede ejecutar manualmente. Se puede publicar datos mediante bases de datos, REST APIs y herramientas nativas a Azure. Para consultar las fuentes de datos soportadas, acceder a este enlace.
- **Consumo**: Las funcionalidades de búsqueda de datos en detalle es el punto fuerte de Azure Data Catalog. La implementación de metadatos permite profundizar y relacionar los datos entre sí.
- **Gestión**: Donde empieza el gobierno del dato en sí y el análisis transversal de repositorios de datos. En esta etapa se aplican las políticas de calidad de datos (integración), las definiciones de roles (quién puede visualizar cual conjunto de datos) y otras acciones. Importante resaltar que la implementación de seguridad aplicada en esta etapa es válida para la capa de metadatos creada dentro de Azure. En otras palabras, ninguna regla de seguridad o de negocio creada en Azure Data Catalog será aplicada en los entornos en que residen las fuentes de los datos. Por lo tanto, se puede decir que Azure Data Catalog es responsable por la seguridad de datos a partir del punto de publicación en la herramienta.

2.2 ARQUITECTURA Y DESCRIPCIÓN DEL SERVICIO

El servicio de Data Catalog es gestionado por Azure y por lo tanto los costes se restringen a la licencia. Es necesario crear una cuenta Azure y activar el *Directory Tentant*.

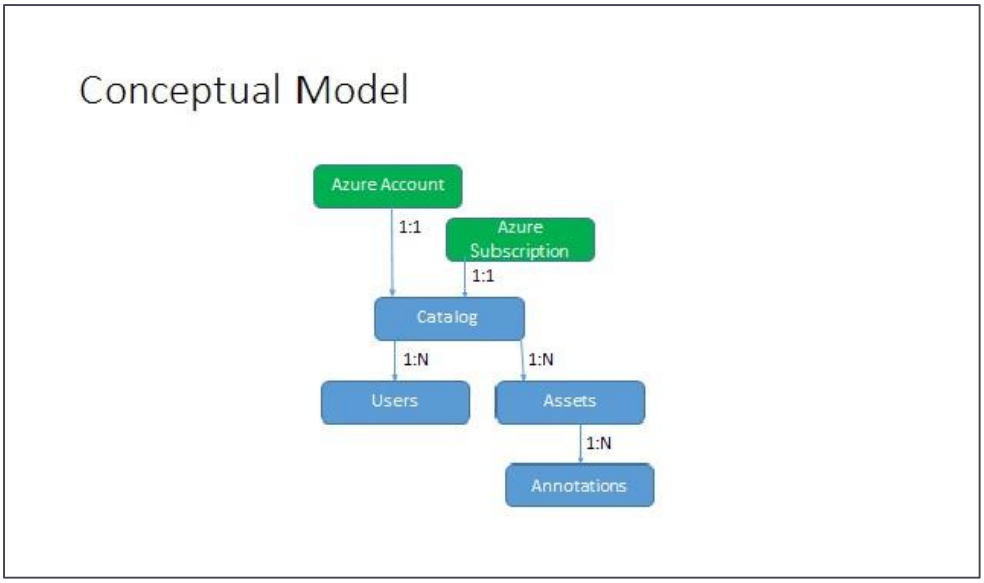


Ilustración 3. Modelo conceptual en Azure

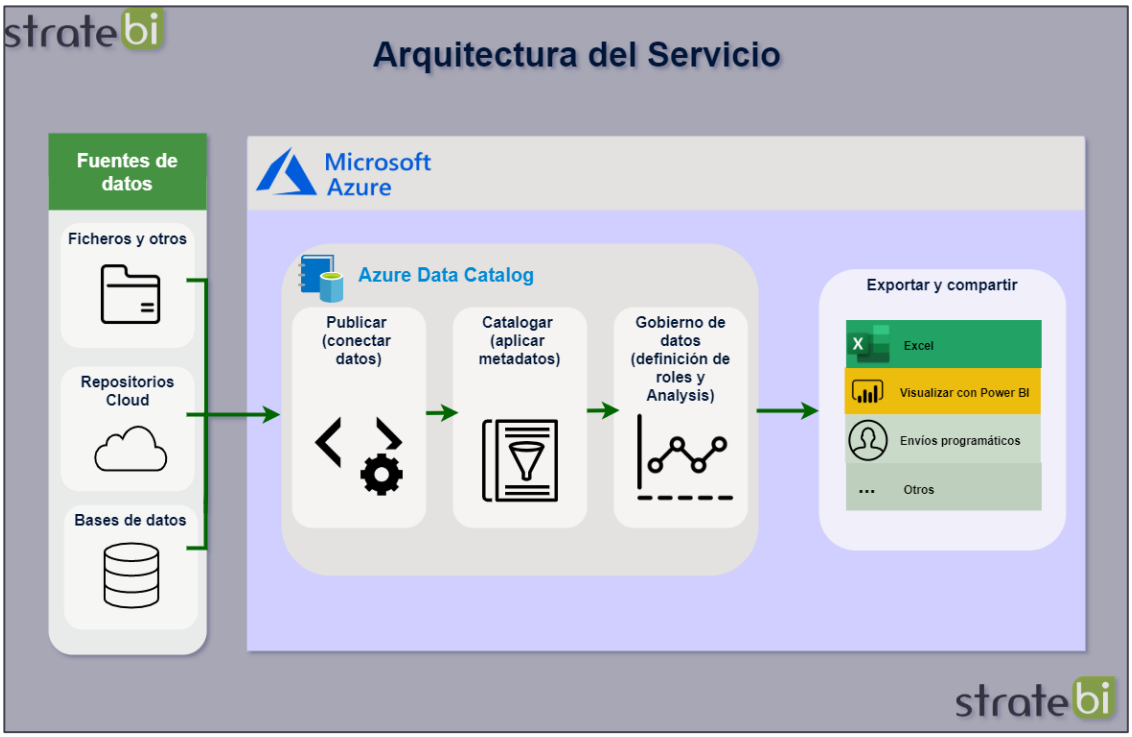


Ilustración 4. Arquitectura del servicio

Junto a Azure Data Factory y otros componentes de Azure, encaja perfectamente en una arquitectura de tipo Big Data siendo usado de manera transversal al flujo de datos de un proyecto, conteniendo las definiciones y el repositorio centralizado de los datos de todas nuestras fuentes:

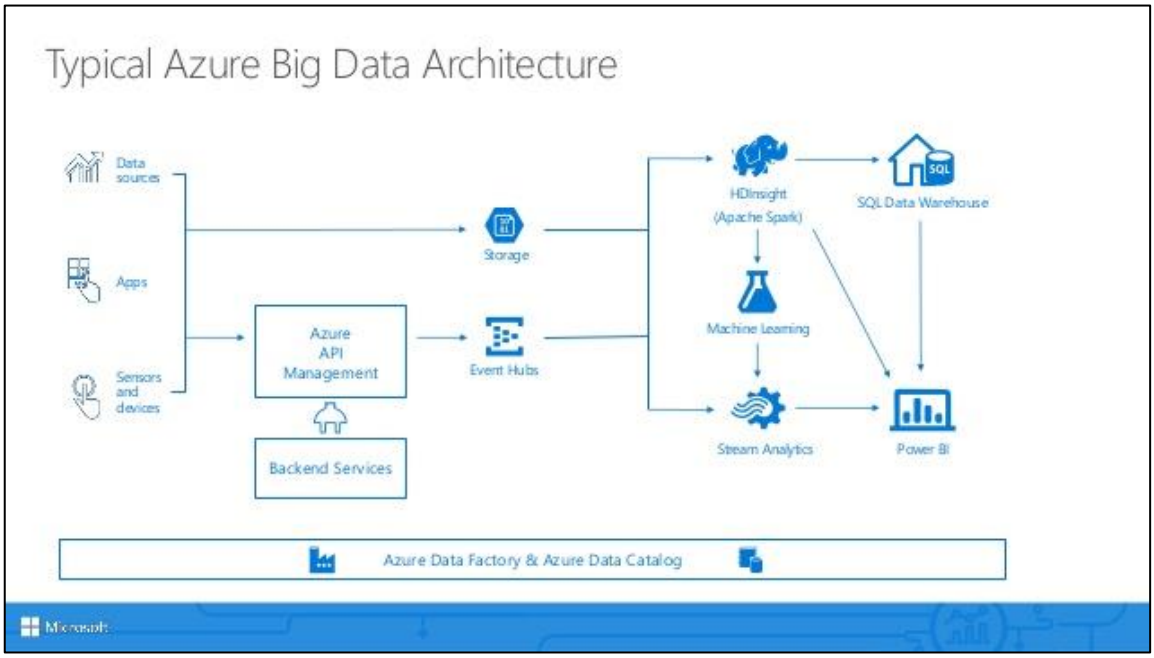


Ilustración 5. Arquitectura tipo big data

Existen dos modalidades de precio para Azure Data Catalog:

- ***Free***: La edición gratuita tiene por objetivo proporcionar una experiencia integral de la utilización del servicio.
    - **Ventajas**: Permite a cualquier usuario registrar, enriquecer, comprender, descubrir y consumir datos de fuentes registradas en el Catálogo de Datos.
    - **Desventajas**: Cualquier objeto registrado es visible para todos los usuarios autentificados. No incluye la funcionalidad glosario empresarial.
    - **Precio por usuario por mes**: gratuito.
        - **Número máximo de usuarios**: ilimitado
        - **Número máximo de Objetos Catalogados**: 5,000 unidades

- ***Standard:*** Además de todas las características de la edición gratuita, la edición estándar proporciona capacidades de gobernabilidad que permiten a los usuarios tomar la propiedad de los activos registrados para un mayor control.
    - **Ventajas**: incluye la funcionalidad glosario empresarial. Es clave disponer de un glosario empresarial.
    - **Desventajas**: Coste más elevado en comparación a la alternativa Free.
    - **Precio por usuario por mes**: € 0.85
        - **Número máximo de usuarios**: ilimitado
        - **Número máximo de Objetos Catalogados**: 100,000 unidades

Puedes simular el coste de acuerdo con la necesidad de tu organización puedes utilizar la calculadora de precios de Azure Data Catalog ofrecida por Azure.

2.3 GLOSARIO EMPRESARIAL

Esta funcionalidad es exclusiva para la versión Estándar. Mediante el uso del glosario empresarial en Azure Data Catalog y del etiquetado, se puede identificar, administrar y detectar recursos de datos de forma sencilla.

El glosario empresarial promueve que los miembros de la organización aprendan el vocabulario empresarial. El glosario también admite la captura de metadatos descriptivos, lo que simplifica el conocimiento y la detección de recursos. Más información sobre el glosario en este enlace.

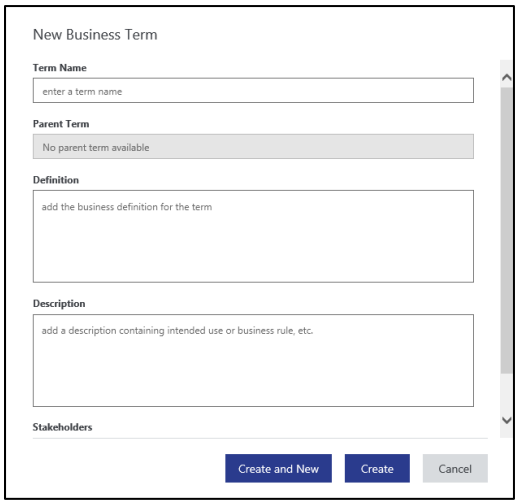


Ilustración 6. Definición de glosario

2.4 RECOLECCIÓN DE DATOS EN AZURE DATA CATALOG

Una vez que se accede a la página principal, podemos ver los *Assets* disponibles. Tenemos varios opciones en la primera página:

- Buscar por un objeto a través del buscador principal (*Search Data Catalog*).
- Publicar datos.
- Abrir un *Asset* directamente (*My Assets*).

Para este caso de uso, vamos a publicar nuevos datos utilizando la opción *Publish Data*:

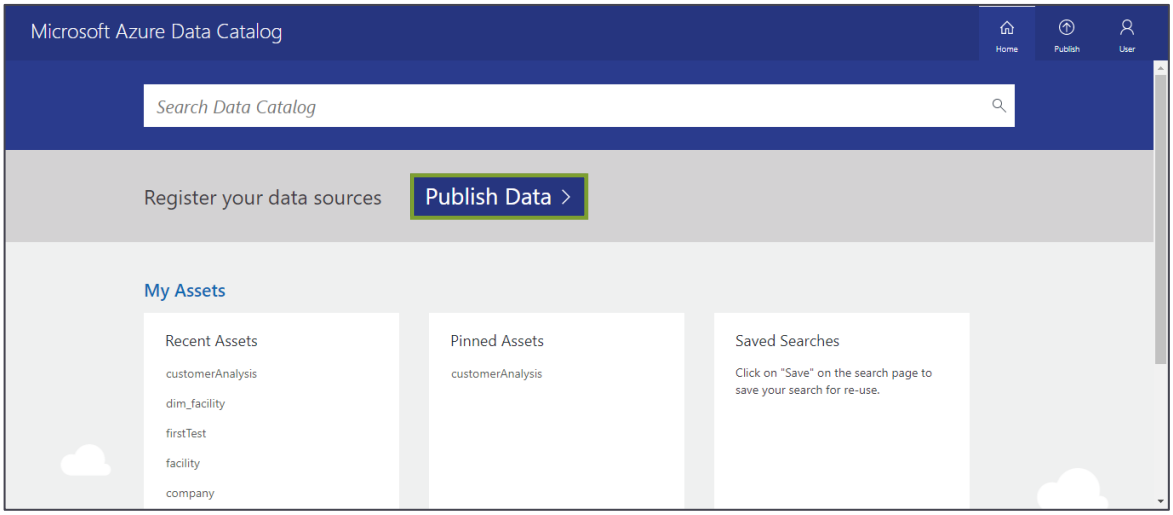


Ilustración 7. Gestión de Assets

Podemos definir una conexión manual o utilizar la aplicación de escritorio. Para este caso de uso, utilizaremos la aplicación escritorio:

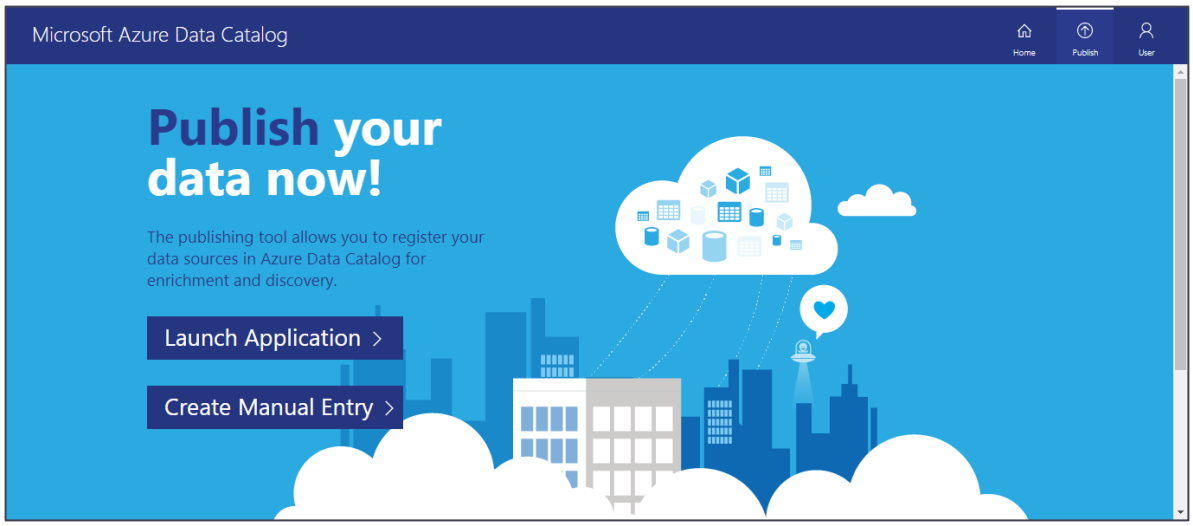


Ilustración 8. Ventana inicial Azure Data Catalog

Tenemos la posibilidad de elegir una de las distintas opciones de conexión de datos. Todas las posibles conexiones se encuentran en este enlace.

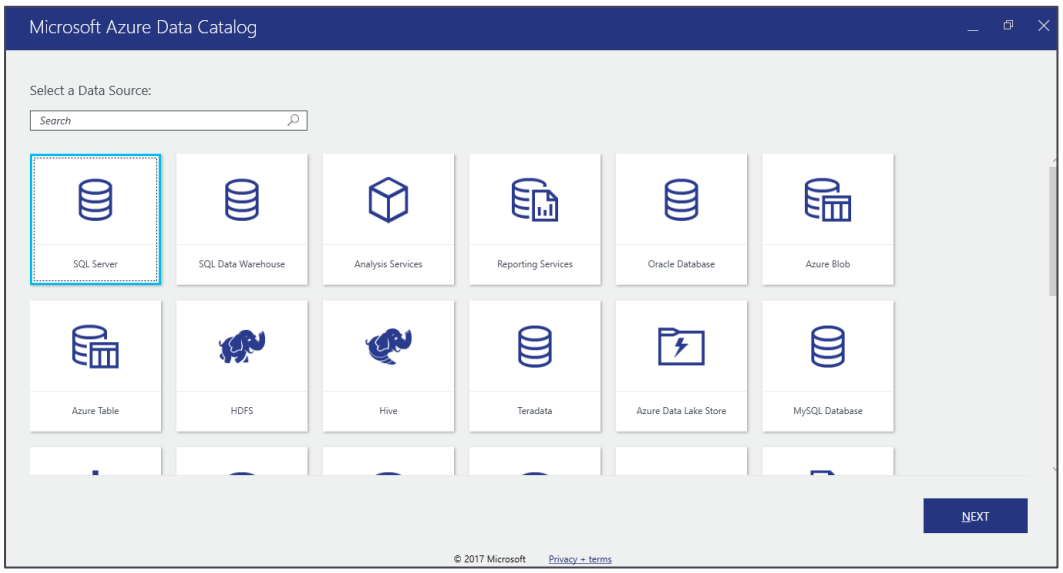


Ilustración 9. Conexiones a orígenes de datos

2.4.1 EXTRACCIÓN DE METADATOS DE UNA BASE DE DATOS

Una vez elegida el tipo de conexión, la creamos utilizando una base de datos existente:

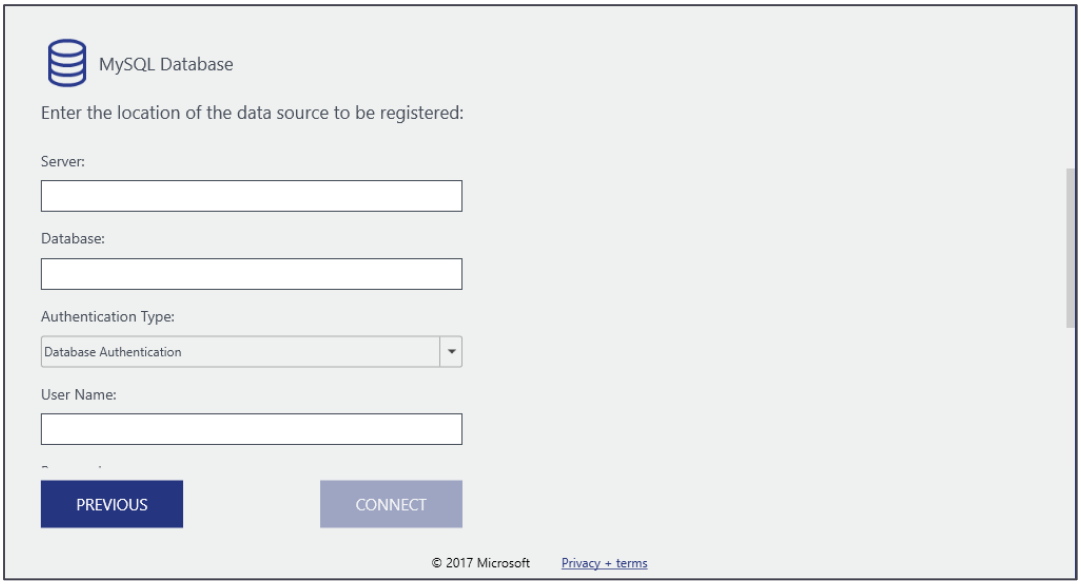


Ilustración 10. Definición de conexión a BBDD

Una vez creada la conexión, se debe elegir el tipo de objeto para importar a Azure Data Catalog. En la sección "*Available objects*", podemos ver los objetos disponibles en esta conexión, como su tipo, en la columna "*Type*":

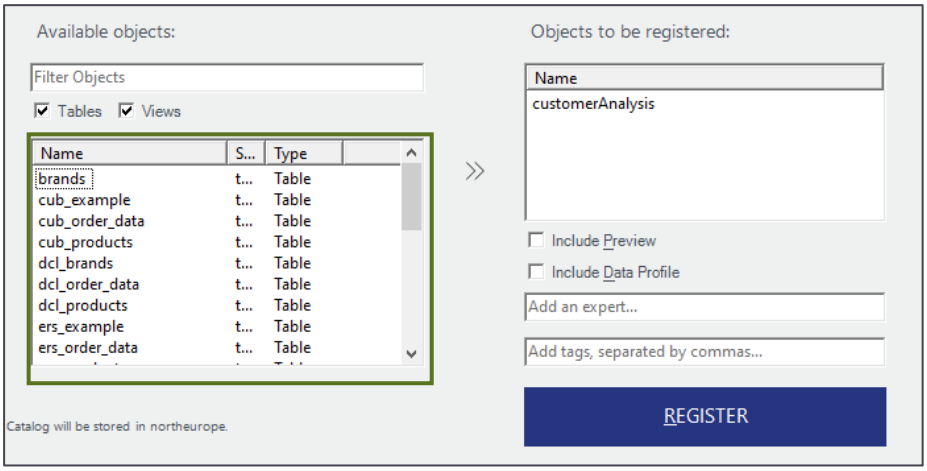


Ilustración 11. Registro de objetos de BBDD

Podemos ver ahora el *Asset* creado y disponible para uso en la plataforma web.

Ilustración 12. Assets recientes

Al seleccionar el *Asset*, podemos abrirlo y empezar a trabajar sobre este objeto. En la pestaña a la izquierda del objeto podemos realizar búsquedas por otros Assets y filtrar la búsqueda por los Tags creados. Se puede filtrar por tipo de objeto, tags y tipo de fuentes de datos:

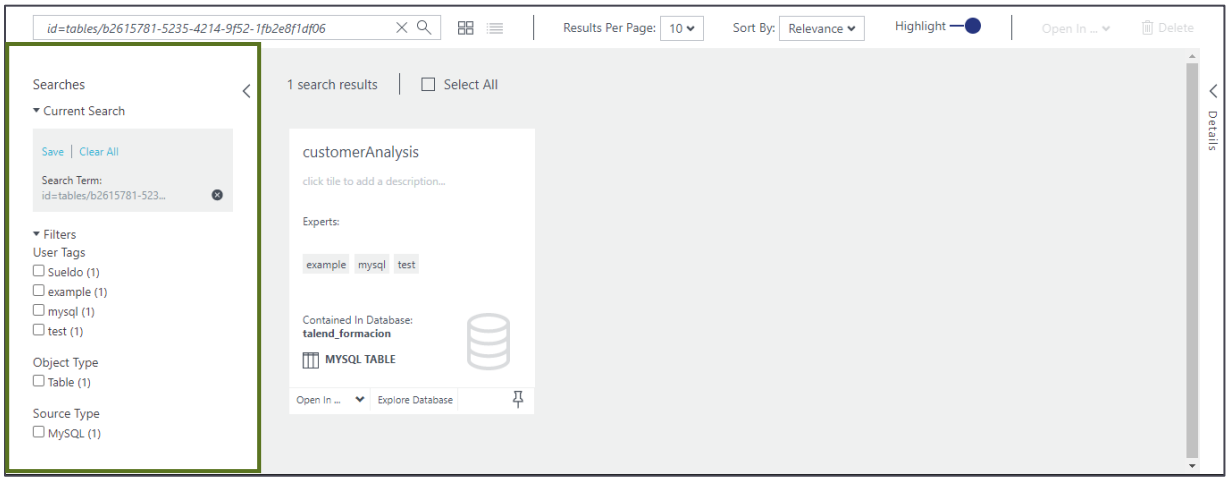


Ilustración 13. Búsqueda de Assets

La sección central de la página nos presenta el objeto en cuestión.

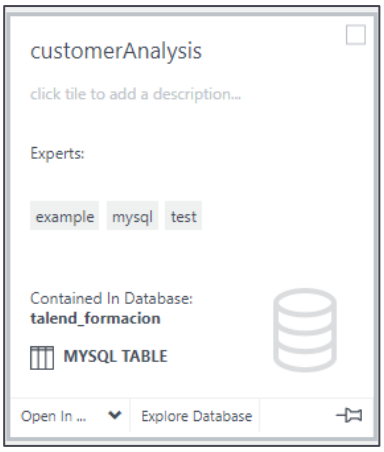


Ilustración 14. Selección de Asset

Para crear metadatos, simplemente pinchamos sobre la tarjeta y las opciones de edición se abrirán en la parte derecha:

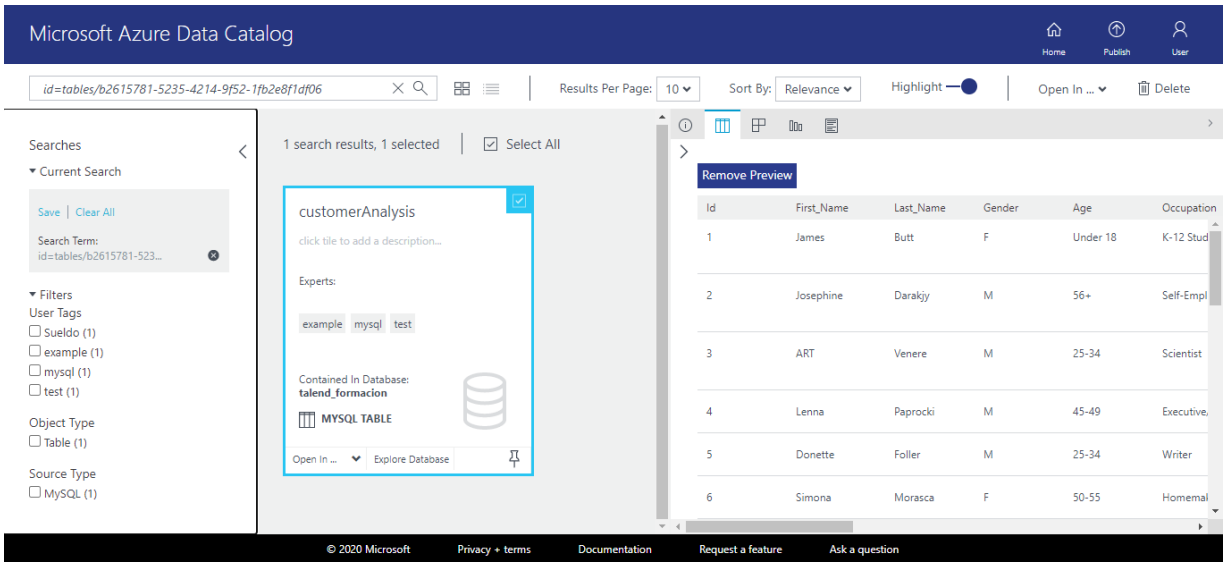
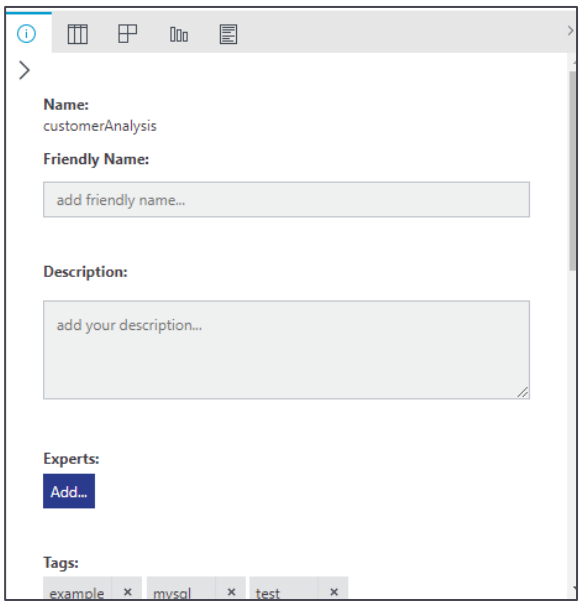


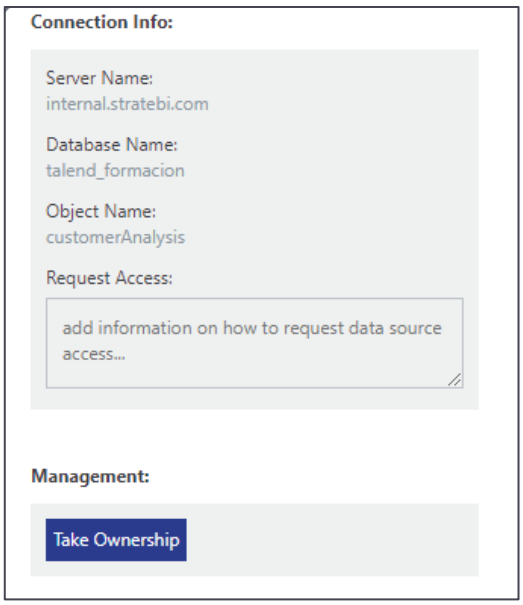
Ilustración 15. Visualización de datos y propiedades

2.5 PROPIEDADES DE LOS METADATOS

Las opciones de edición y creación de metadatos son las siguientes:

Propiedades: Podemos introducir un nombre de búsqueda para el objeto (*friendly name*), crear tags, visualizar datos de conexión, dar de alta a los *Managers* de los objetos, y ver los metadatos de creación y actualización:





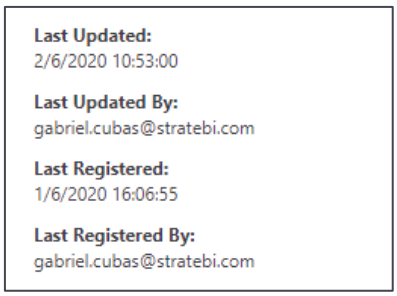


Ilustración 16. Propiedades del objeto

En esta pestaña, podemos dar de alta las cuentas responsables de este activo, estando disponibles en Active Directory:

Ilustración 17. Definición de responsables del objeto

También podemos habilitar nuestra cuenta actual como propietaria del objeto:

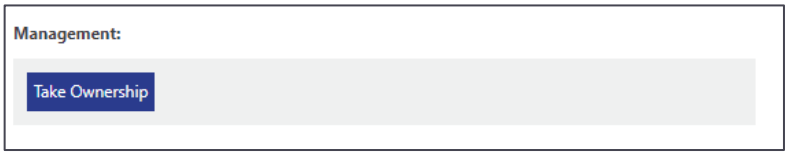
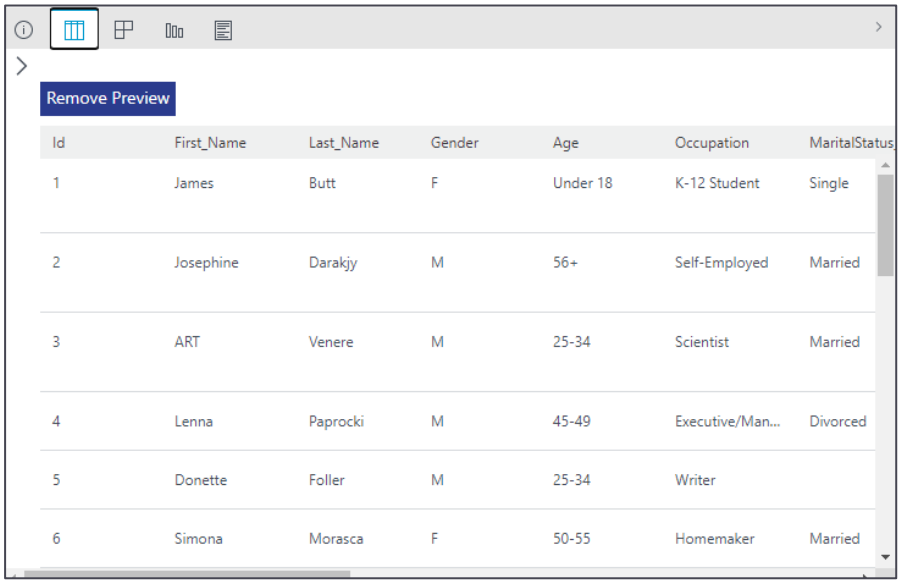


Ilustración 18. Tomar control de un objeto

View: Muestra datos para este objeto. En nuestro caso, se trata de una tabla de BBDD:



Id	First_Name	Last_Name	Gender	Age	Occupation	MaritalStatus
1	James	Butt	F	Under 18	K-12 Student	Single
2	Josephine	Darakjy	M	56+	Self-Employed	Married
3	ART	Venere	M	25-34	Scientist	Married
4	Lenna	Paprocki	M	45-49	Executive/Man...	Divorced
5	Donette	Foller	M	25-34	Writer	
6	Simona	Morasca	F	50-55	Homemaker	Married

Ilustración 19. Previsualización de datos

Columns: Permite realizar cambios sobre las columnas de la tabla y también insertar *Tags* y descripciones para cada campo:

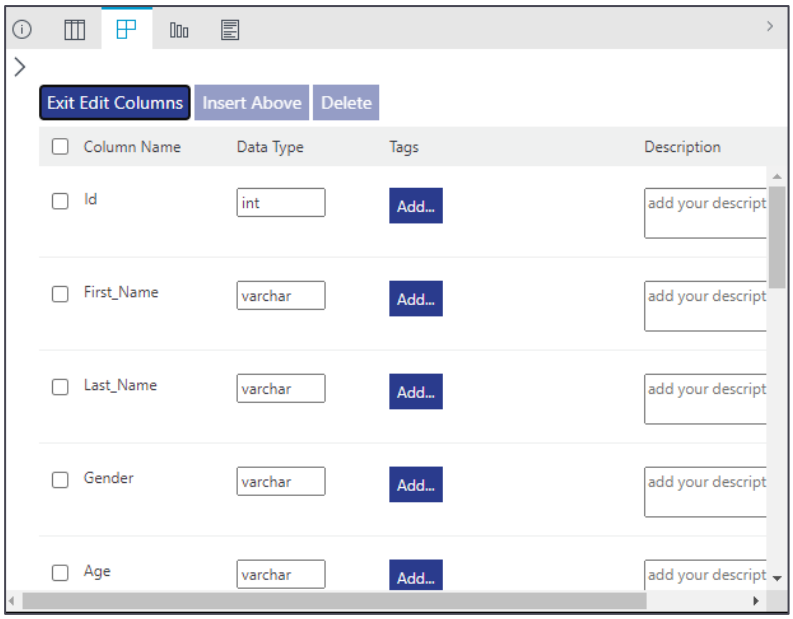


Ilustración 20. Definición y asignación de tags

Profile: realiza un resumen rápido sobre el objeto, a nivel de tabla y columna:

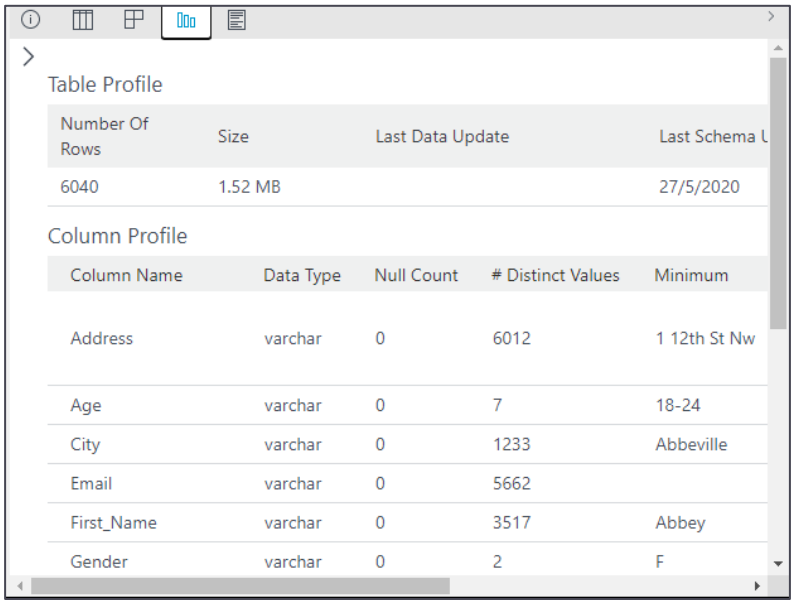


Table Profile

Number Of Rows	Size	Last Data Update	Last Schema U
6040	1.52 MB		27/5/2020

Column Profile

Column Name	Data Type	Null Count	# Distinct Values	Minimum
Address	varchar	0	6012	1 12th St Nw
Age	varchar	0	7	18-24
City	varchar	0	1233	Abbeville
Email	varchar	0	5662	
First_Name	varchar	0	3517	Abbey
Gender	varchar	0	2	F

Ilustración 21. Profiling de las columnas

Documentation: Permite realizar una documentación detallada sobre el objeto:

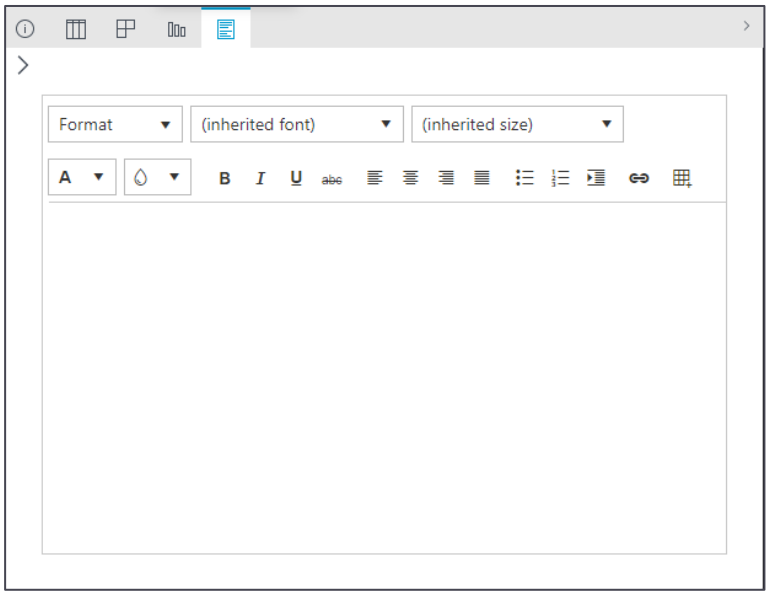


Ilustración 22. Documentación sobre el objeto

A partir del momento de creación de los metadatos, los objetos puedes ser buscados en el buscador principal utilizando los Tags, perfiles, y otras propiedades asignadas:

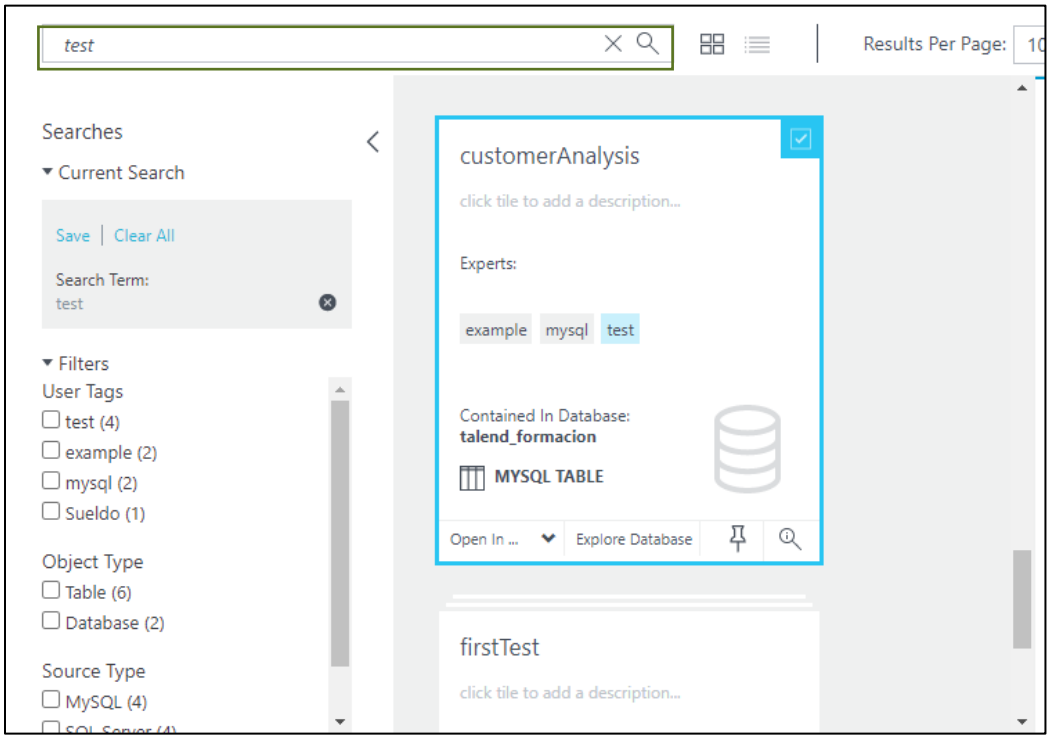


Ilustración 23. Búsqueda en metadatos

2.6 EXPORTACIÓN E INTEGRACIÓN CON OTRAS HERRAMIENTAS

Azure Data Catalog posibilita la exportación e integración directa con Excel, Power Query y Power BI. En la cabecera de la página principal, vemos la opción "Open in...":

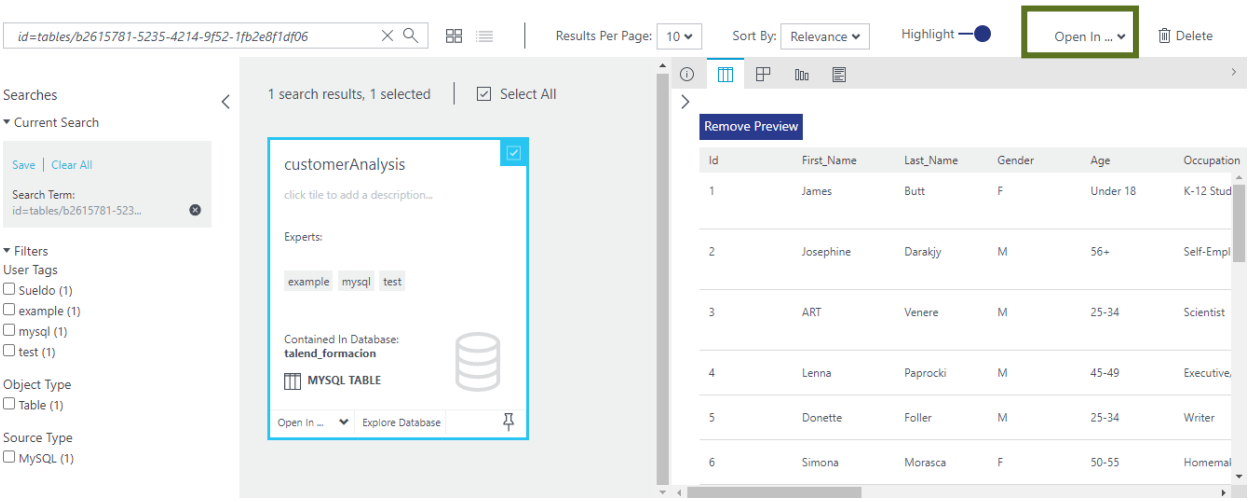
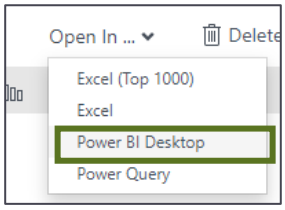


Ilustración 24. Abrir definición con otras herramientas

En el caso de Power BI, descargamos un fichero .pbix y podemos utilizar el repositorio de metadatos que hemos creado con Power BI Desktop:



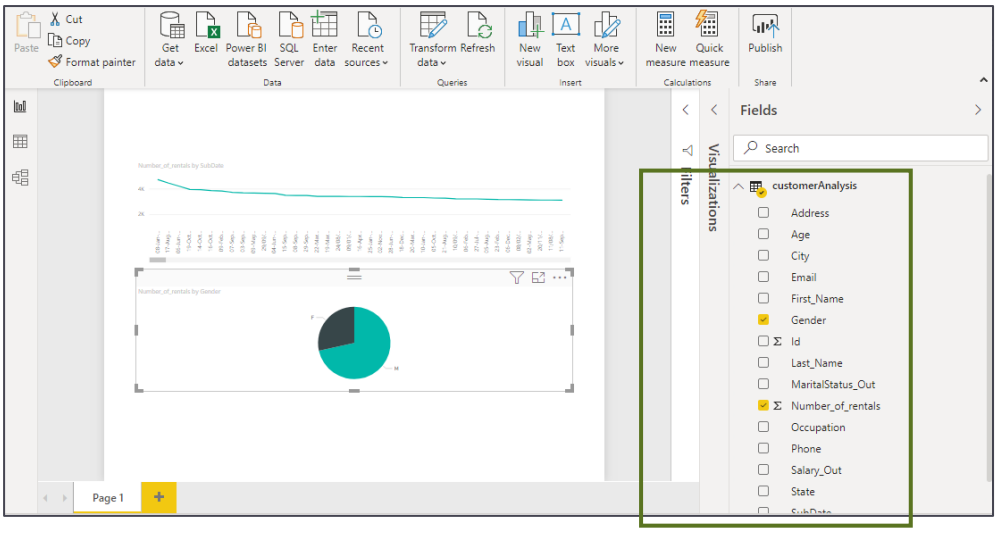


Ilustración 25. Edición en Power BI

2.7 PROBLEMAS COMUNES

2.7.1 ERROR CON LOGIN EN APLICACIÓN DE ESCRITORIO

Este error ocurre cuando no está configurado la versión TLS 1.2, necesario para hacer funcionar la aplicación escritorio de Azure Data Catalog. Los pasos a seguir para solventarlo son:

1. Configurar TLS 1.2 en el registro.
   1.1. Instalar TLS 1.2 en registry: https://www.youtube.com/watch?v=8q9qJzteBn8
2. Configurar .NET para que utilice TLS 1.2 por defecto:
   2.1. Abrir regitry editor (windows+R y escribir regedit)
   2.2. Abrir la ruta HKEY_LOCAL_MACHINE/SOFTWARE/Microsoft/.NETFramework/v4.0.30319
   2.3. En la ruta indicada, crear un nuevo **DWORD**, nombrarlo **schUseStrongCrypto** y asignarle el valor (value data) a 1 en hexadecimal
   2.4. En la ruta indicada, crear un nuevo **DWORD**, nombrarlo **SystemDefaultTlsVersions** y asignarle el valor (value data) a 1 en hexadecimal

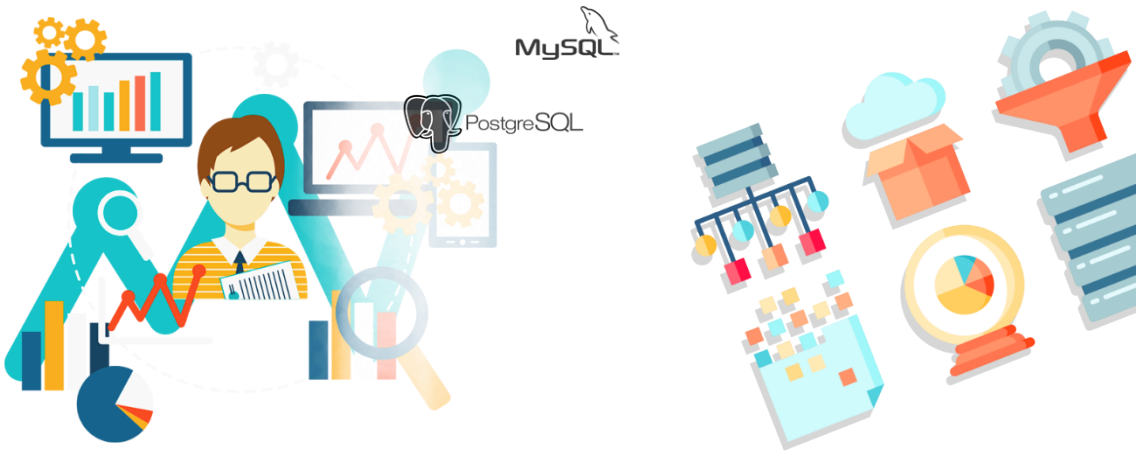
2.7.2 ERROR DE CONEXIÓN A BBDD

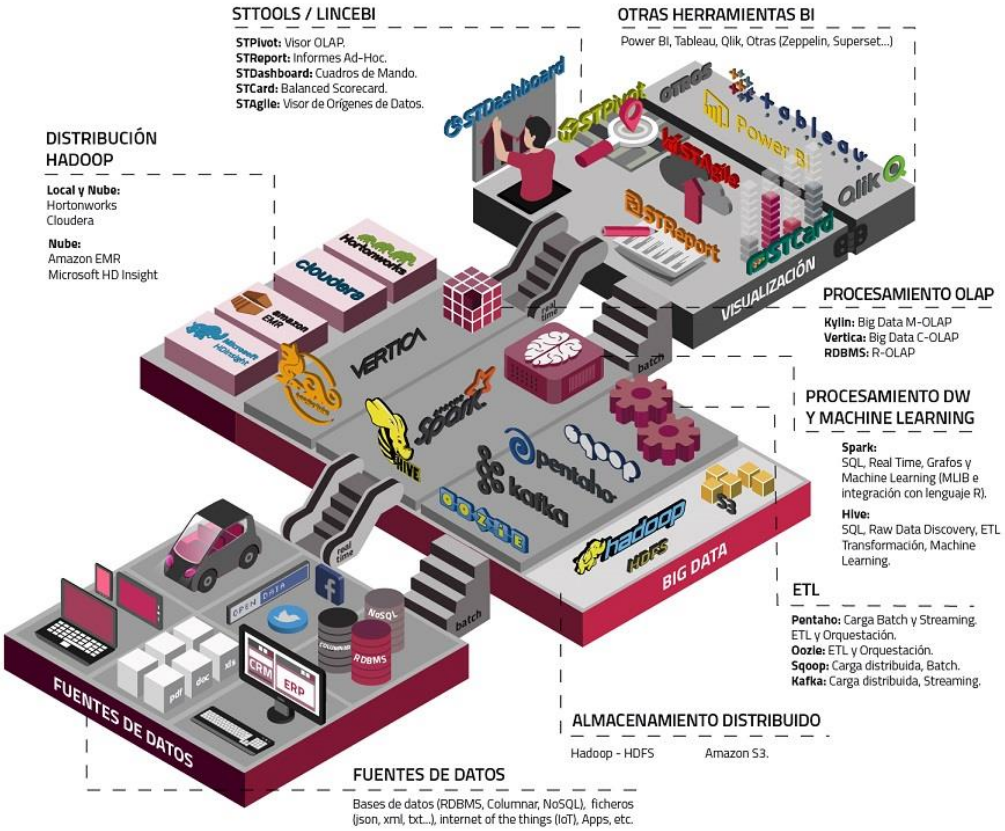
Ese error ocurre cuando .NET no identifica los conectores de base de datos. En el caso de MySQL o Postgre, para solucionarlo hay que descargar e instalar los siguientes ficheros:

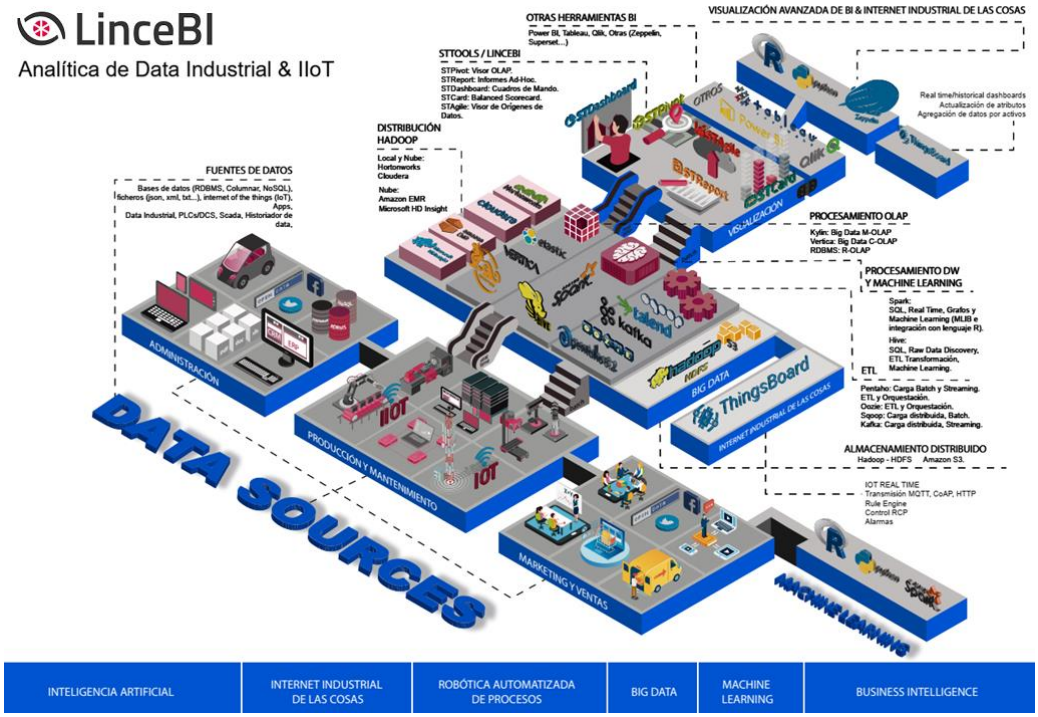
- MySQL: https://dev.mysql.com/downloads/connector/net/
- PostgreSQL: https://ftp.postgresql.org/pub/odbc/versions/msi/psqlodbc_12_02_0000-x64.zip

TECNOLOGÍAS

Trabajamos con las principales tecnologías y somos Partners Certificados de Vertica, Talend, Microsoft, Snowflake, Kylligence, Pentaho, etc.




STTOOLS / LINCEBI
STPivot: Visor OLAP.
STReport: Informes Ad-Hoc.
STDashboard: Cuadros de Mando.
STCard: Balanced Scorecard.
STAgile: Visor de Orígenes de Datos.
OTRAS HERRAMIENTAS BI
Power BI, Tableau, Qlik, Otras (Zeppelin, Superset...)
DISTRIBUCIÓN HADOOP
Local y Nube:
Hortonworks
Cloudera
Nube:
Amazon EMR
Microsoft HD Insight
VISUALIZACIÓN
PROCESAMIENTO OLAP
Kylin: Big Data M-OLAP
Vertica: Big Data C-OLAP
RDBMS: R-OLAP
PROCESAMIENTO DW Y MACHINE LEARNING
Spark:
SQL, Real Time, Grafos y Machine Learning (MLIB e integración con lenguaje R).
Hive:
SQL, Raw Data Discovery, ETL Transformación, Machine Learning.
BIG DATA
ETL
Pentaho: Carga Batch y Streaming. ETL y Orquestación.
Oozie: ETL y Orquestación.
Sqoop: Carga distribuida, Batch.
Kafka: Carga distribuida, Streaming.
FUENTES DE DATOS
ALMACENAMIENTO DISTRIBUIDO
Hadoop - HDFS
Amazon S3.
FUENTES DE DATOS
Bases de datos (RDBMS, Columnar, NoSQL), ficheros (json, xml, txt...), internet of the things (IoT), Apps, etc.


LinceBI
Analítica de Data Industrial & IIoT
OTRAS HERRAMIENTAS BI
VISUALIZACIÓN AVANZADA DE BI & INTERNET INDUSTRIAL DE LAS COSAS
STTOOLS / LINCEBI
DISTRIBUCIÓN HADOOP
FUENTES DE DATOS
PROCESAMIENTO OLAP
PROCESAMIENTO DW Y MACHINE LEARNING
ETL
ALMACENAMIENTO DISTRIBUIDO
ADMINISTRACIÓN
PRODUCCIÓN Y MANTENIMIENTO
MARKETING Y VENTAS
MACHINE LEARNING
DATA SOURCES
INTELIGENCIA ARTIFICIAL
INTERNET INDUSTRIAL DE LAS COSAS
ROBÓTICA AUTOMATIZADA DE PROCESOS
BIG DATA
MACHINE LEARNING
BUSINESS INTELLIGENCE

INFORMACIÓN SOBRE STRATEBI

Stratebi es una empresa española, con sede en Madrid y oficinas en Barcelona, Alicante y Sevilla, con amplia experiencia en sistemas de información, soluciones tecnológicas y procesos relacionados con soluciones de Open Source y de inteligencia de Negocio.

Esta experiencia, adquirida durante la participación en proyectos estratégicos en compañías de reconocido prestigio a nivel internacional, se ha puesto a disposición de nuestros clientes.

Somos **Partners Certificados en Microsoft PowerBI** con una dilatada experiencia

Stratebi es la única empresa española que ha estado presente todos los Pentaho Developers celebrados en Europa habiendo organizado el de España.

En Stratebi nos planteamos como **objetivo** dotar a las compañías e instituciones, de herramientas escalables y adaptadas a sus necesidades, que conformen una estrategia Business Intelligence capaz de rentabilizar la información disponible. Para ello, nos basamos en el desarrollo de soluciones de Inteligencia de Negocio, mediante tecnología Open Source.

Stratebi **son** profesores y responsables de proyectos **del Master en Business Intelligence de la Universidad UOC, UCAM, EOI…**

Los profesionales de Stratebi son los creadores y autores del primer weblog en español sobre el mundo del Business Intelligence, Data Warehouse, CRM, Dashboards, Scorecard y Open Source. Todobi.com

Stratebi es partner de las principales soluciones Analytics: Microsoft Power BI, Talend, Pentaho, Vertica, Snowflake, Kyligence, Cloudera…

Todo Bi, se ha convertido en una referencia para el conocimiento y divulgación del Business Intelligence en español.

OTROS

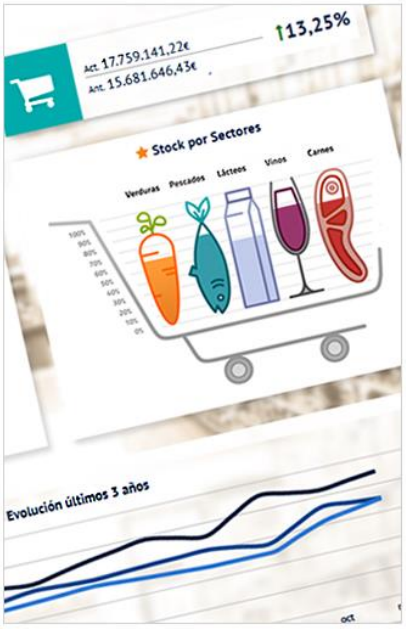
Trabajamos en los principales sectores y con algunas de las compañías y organizaciones más importantes de España.

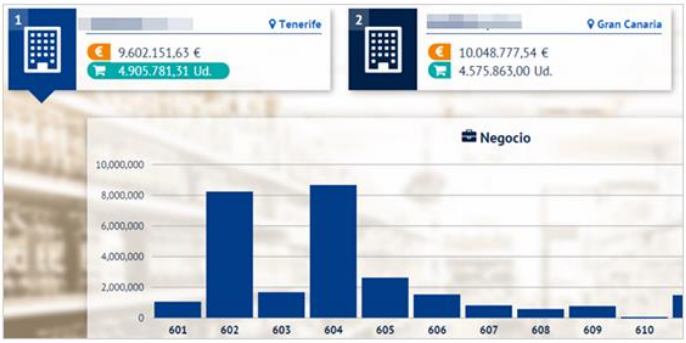
SECTOR PRIVADO

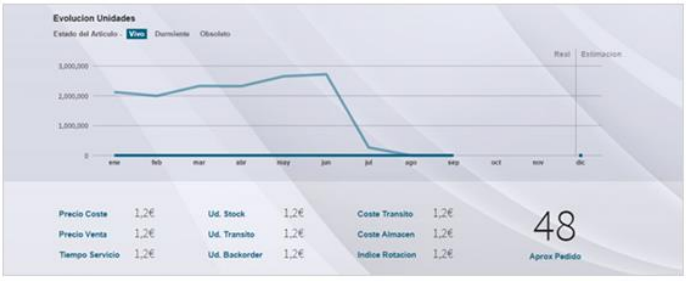
SECTOR PÚBLICO

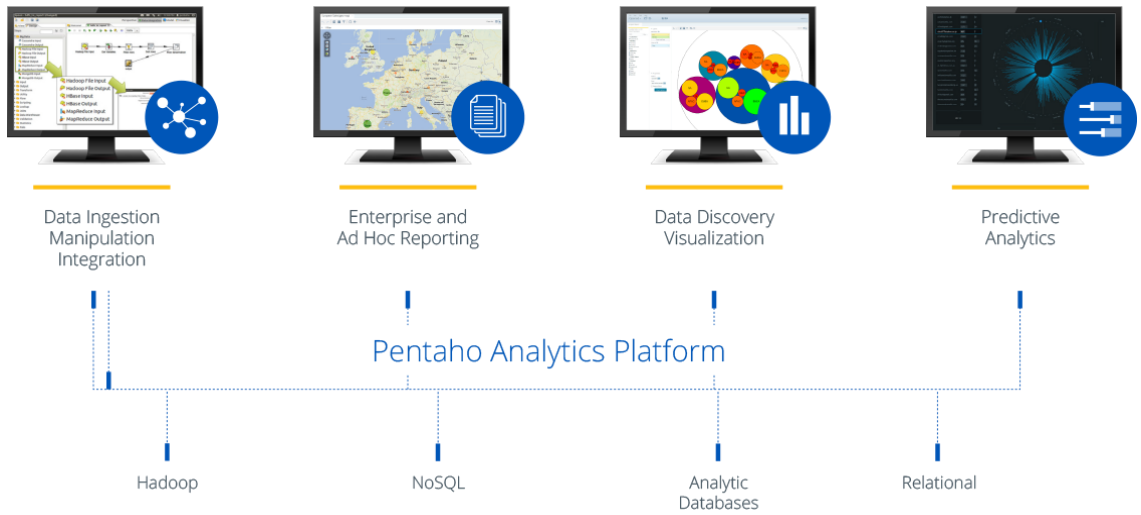
EJEMPLOS DE DESARROLLOS ANALYTICS

A continuación, se presentan **ejemplos de algunos screenshots** de cuadros de mando diseñados por Stratebi, con el fin de dar a conocer lo que se puede llegar a obtener, así como Demos Online en la web de Stratebi:



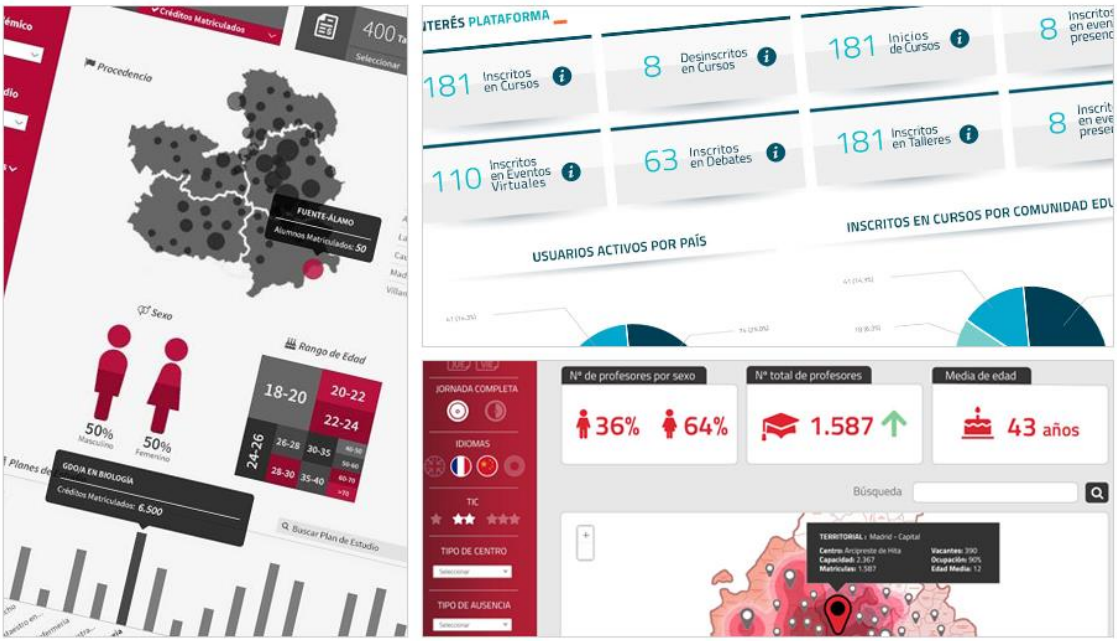


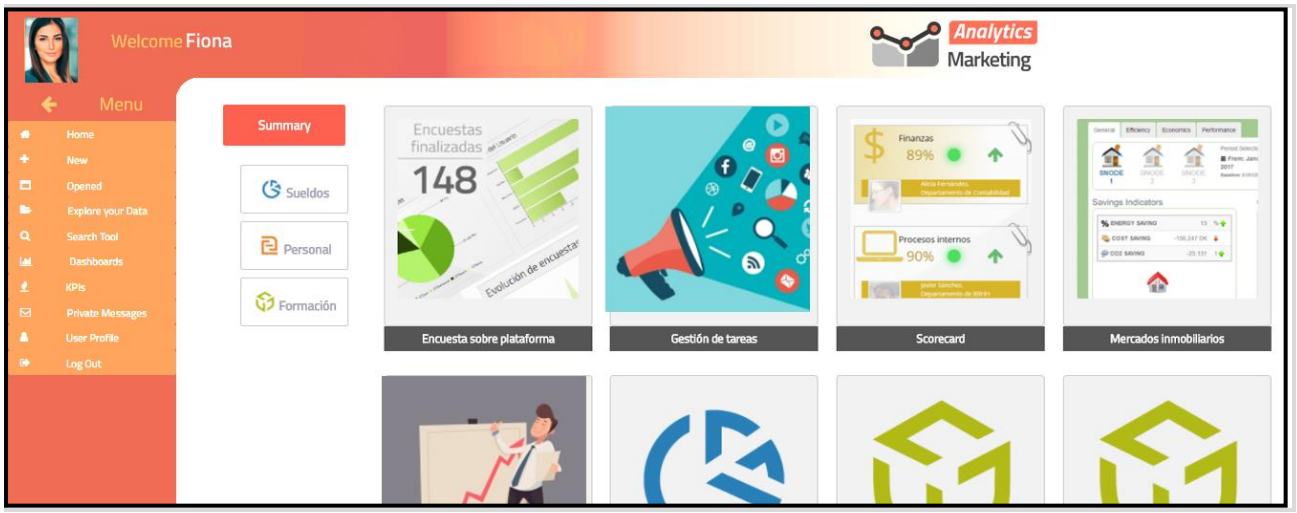





facebook DASHBOARD COMENTARIOS
Sexo
Sentimiento
Localización
Tipo de recurso
Evolución Comentarios por días


smart MADRID
Última Actualización
18:24


Procedencia
Sexo
Rango de Edad
50% Masculino
50% Femenino
18-20
20-22
22-24
24-26
26-28
28-30
30-35
35-40
GRADO EN BIOLOGÍA
Créditos Matriculados: 6.500
Buscar Plan de Estudio
181 Inscritos en Cursos
8 Desinscritos en Cursos
181 Inicios de Cursos
110 Inscritos en Eventos Virtuales
63 Inscritos en Debates
181 Inscritos en Talleres
USUARIOS ACTIVOS POR PAÍS
INSCRITOS EN CURSOS POR COMUNIDAD EDU
JORNADA COMPLETA
IDIOMAS
TIC
TIPO DE CENTRO
TIPO DE AUSENCIA
Nº de profesores por sexo
36%
64%
Nº total de profesores
1.587
Media de edad
43 años
Búsqueda


Welcome Fiona
Analytics
Marketing
Menu
Home
New
Opened
Explore your Data
Search Tool
Dashboards
KPIs
Private Messages
User Profile
Log Out
Summary
Sueldos
Personal
Formación
Encuestas finalizadas
148
Evolución de encuestas
Encuesta sobre plataforma
Gestión de tareas
Finanzas
89%
Procesos internos
90%
Scorecard
Savings Indicators
Mercados inmobiliarios