stratebi
open business intelligence

2019

Usando Machine Learning para Data Quality

BIG DATA – BUSINESS INTELLIGENCE – MACHINE LEARNING

Índice

1. INTRODUCCIÓN 2

2. CARACTERÍSTICAS ENTRADA DE DATOS 3

3. CARACTERÍSTICAS MÓDULOS IMPLEMENTADOS 4

MÓDULO TELÉFONOS 4
MÓDULO EMAILS 5
MÓDULO DOCUMENTOS DE IDENTIDAD 5
MÓDULO NOMBRES Y APELLIDOS 6
MÓDULO DIRECCIONES 7
MÓDULO DE-DUPLICACIÓN 8

4. TECNOLOGÍAS USADAS Y ESCALABILIDAD 9

5. CONCLUSIONES 10

6. INFORMACIÓN SOBRE STRATEBI 11

7. REFERENCIAS STRATEBI 12

8. EJEMPLOS DE DESARROLLOS ANALYTICS 13

1. INTRODUCCIÓN

Las empresas cada vez necesitan almacenar y procesar más datos sobre sus clientes, proveedores, personal o pedidos.

Sin embargo, a mayor Volumen de datos, mayor es la probabilidad de que existan datos incorrectos, como las direcciones o teléfonos que afecten de forma negativa a nuestro negocio. Por ejemplo, un pedido con una dirección errónea será devuelto, reduciendo el beneficio de la empresa y la confianza del cliente.

Teniendo en cuenta esta problemática, Stratebi propone una solución: **ST Big Data Quality**.

Solución para el tratamiento de datos de personas con características Big Data (Volumen, Variedad, Velocidad), basada en la aplicación de procesos de limpieza, normalización y validación a un grupo de datos muy comunes en cualquier base de datos. Esta solución escalable y extensible, incluye módulos para la limpieza, normalización validación, de-duplicación y corrección de datos de personas y direcciones en los siguientes tipos de datos:

Teléfonos
Emails
Documentos de Identidad
Nombres y Apellidos
Direcciones

Nuestra solución de Data Quality para Big Data se ha implementado usando la tecnología Apache Spark, la cual permite la escalabilidad del procesamiento a cualquier volumen de datos sin reescribir ni una línea de código. De esta forma, ST Big Data Quality se puede ejecutar en una única máquina y, si el tamaño de los datos de entrada, diccionarios de nombres, callejeros o datos recuperados de las API lo requieren, puede ejecutarse en un clúster Spark como Databricks o Amazon EMR en la nube o en instalación Hadoop on-premise como Hortonworks o Cloudera.

Además, estos módulos permiten realizar la validación y corrección de datos en función de datos de referencias obtenidos mediante diccionarios y API's de nombres, callejeros (ej. Correos) o dominios frecuentes de emails. En los siguientes apartados, analizaremos las principales funcionalidades de cada uno de los módulos.

2. CARACTERÍSTICAS ENTRADA DE DATOS

La entrada de datos será tabular desde ficheros CSV, Parquet, Avro, tablas de bases de datos (conector JDBC). Pero hay que tener en cuenta que la entrada de datos puede ser muy diversa, dependiendo del proceso de negocio y fuentes de datos implicadas, incluyendo fuentes de datos proporcionadas por terceros a través de diccionarios o web APIs.

Es decir, se pueden encontrar documentos con diferente número de columnas, con la existencia o no de cabecera de datos, o diferentes nombres para aquellas columnas que muestran el mismo tipo de dato.

Para ello, se define un fichero de configuración con el que se consigue personalizar la estructura de datos en la entrada, permitiendo describir el contenido de los datos de origen y tener así un proceso flexible para diferentes tipos de entrada de datos

Como, por ejemplo; se puede definir si el documento posee cabecera, su delimitador, los nombres de las columnas de entrada, ya sea una o varias columnas para un mismo tipo de dato, etc. Y, de manera genérica, también se tendrá en cuenta la posibilidad de datos nulos como entrada. Esto no sería un problema para el tratamiento de los datos.

Un ejemplo sería el caso de los teléfonos. Nos podemos encontrar varias columnas con datos, como puede ser para teléfono fijo y teléfono móvil, ambos datos podrán ser procesados sin ningún inconveniente. Al ser configurable, no existe un número limitado de columnas que sea posible procesar.

Datos de entrada:

Tel_fijo	Tel_móvil
954873488	674993281
688 075 156	727 496 241

Fichero de configuración:

```
columnas_entrada_telefonos
 {
  columnas_telefono ="Tel_fijo",”Tel_móvil”
 }
```

CARACTERÍSTICAS MÓDULOS IMPLEMENTADOS

Para cada uno de los módulos tendremos que tener en cuenta algunas características propias para ser capaces de sacar el máximo partido a esta solución.

MÓDULO TELÉFONOS

El objetivo de este módulo es la limpieza y estandarización de los datos de los números de teléfonos correspondientes tanto a móviles, como fijos nacionales.

Se extraerán los números de teléfonos a los que se les realizará una limpieza previa, eliminando caracteres no deseados, para proceder a la clasificación en fijo o móvil según la estructura del dato informada. Esto se consigue, gracias a la aplicación de técnicas de limpieza, como expresiones regulares, para cumplir con el Plan de Numeración Nacional (PNN) para fijos y móviles. Será la parte responsable de cada una de las verificaciones que se realizan específicamente para este módulo.

Además, existirá la verificación de la provincia con respecto al prefijo nacional, en el caso de los números de teléfonos fijos. Esto va a llevar a obtener una columna de salida dónde se medirá la calidad del dato, ya sean de alta; si el teléfono cumple con el PNN y se corresponde con la provincia, media; si se cumple el PNN, pero no coincide la provincia o baja calidad; si no se cumple el PNN.

Aquí podemos ver un ejemplo de cómo sería la entrada y salida para el módulo de teléfonos:

Id	Teléfono	Provincia	Normalización_Teléfono	Tipo_Teléfono	Calidad_Teléfono
1	954786976	Madrid	954786976	Fijo	Media
2	954 810 851	Sevilla	954810851	Fijo	Alta
3	487 997887	Madrid	487997887	Desconocido	Baja
4	600 13 76 54	Navarra	600137654	Móvil	Alta
5	987	Granada	987	Desconocido	Baja

MÓDULO EMAILS

Para la estandarización de las direcciones de emails, una vez que se extraen los datos, se validan según una expresión regular que verifica el estándar actual de email RFC 5322. Previamente a validar las direcciones según determina la RFC se realizan tareas de limpieza para obtener una salida lo más clara posible.

Al igual que nos encontramos una columna de calidad de dato en los teléfonos, en los correos electrónicos también vamos a ser capaces de ver de un solo vistazo aquellos datos que cumplen los requisitos de la RFC 5322 con una calidad alta y, en su defecto, a los que no los cumplan se les asignará una calidad baja. Aquí podemos ver un ejemplo de ello:

Id	Email	Normalización_Email	Calidad_Email
1	ET@scelerisquedui.co.uk	et@scelerisquedui.co.uk	Alta
2	Nullam.suscripit@malesua	nullam.suscripit@malesua	Baja
3	sem@aliquameu.org	sem@aliquameu.org	Alta
4	Lore$$@StrateBI.com	lore$$@stratebi.com	Baja
5	hola@hotmail.com <hola@hotmail.com>	hola@hotmail.com	Alta

Dicho módulo se puede extender según las necesidades de la empresa. Agregando la validación de los emails frente a servicios de API web de pago, y poder cubrir también la corrección de los errores comunes en la escritura de los dominios más empleados, además de comprobar la existencia del dominio y que se encuentre activo.

MÓDULO DOCUMENTOS DE IDENTIDAD

Respecto a los documentos de identidad o NIF, se van a tratar respecto a la legislación vigente. Se realizará la validación con lo requerido por la Agencia Tributaria, por lo que se engloba cualquier tipo de NIF, desde todos los casos asociados a personas físicas, como a entidades jurídicas, este último conocido antiguamente como CIF.

En este módulo, también se quiere abordar la posibilidad de completar aún más la información pudiendo calcular el dígito de control de aquellos documentos que no los tuviera informado. Enriqueciendo así la calidad de nuestros datos.

Id	NIF	Normalización_NIF	Tipo_NIF
1	69738859f	69738859F	NIF
2	69738859M	69738859M	ERROR
3	Z5387415	Z5387415S	NIF
4	P147 3721	P1473721G	CIF
5	X4638945y	X4638945Y	NIF
6	47653C	00047653C	NIF
7	37j3	37j3	ERROR

MÓDULO NOMBRES Y APELLIDOS

El objetivo de este módulo es la limpieza y estandarización de los datos de nombres y apellidos de las personas registradas en las fuentes de datos, así como la clasificación de los individuos por género (hombres o mujeres), incluso si éste no viniera informado, existirá un proceso final gracias al cual vamos a poder asignar un sexo u otro dependiendo de la naturaleza del nombre.

Como se ha comentado anteriormente, se va a hacer uso de fuentes de datos externas para contrastar la información recibida y así poder alimentar nuestro sistema de un gran número de datos ya clasificados. Nuestra solución se integra con diccionarios o fuentes de referencia para la validación y corrección de datos como las del INE o listas de personas que no quieren comunicaciones comerciales, como la lista Robinson.

Aprovechando la información ofrecida por la Lista Robinson, ST Big Data Quality ayuda a cumplir con la GDPR actual, al evitar comunicaciones a personas que están en las listas de exclusión como la Robinson, con la que se hace “matching” tras aplicar los módulos de calidad del dato.

Para este módulo se aplicarán técnicas estadísticas avanzadas basadas en el uso de distancias de edición, técnicas comúnmente usadas en los correctores de los procesadores de texto como MS Word, para completar o corregir nombres y apellidos que presenten errores tipográficos. También se podrá verificar el sexo de un individuo y hacer uso de ello para validar los diferentes nombres.

Para un mejor resultado, en toda la información se verificará la presencia de hipocorísticos en nombres o apellidos, que gracias a la lógica del programa se transformarán en nombres o apellidos completos a partir de una tabla de equivalencias.

MÓDULO DIRECCIONES

El objetivo de este proceso es la limpieza y estandarización de los datos de la dirección postal asociada a cada individuo teniendo en cuenta que se puede hacer uso de abreviaturas típicas para la definición de direcciones.

Se separará el contenido del resto del domicilio mediante 'tokenización' y nuestra solución separa mediante inferencia los elementos del domicilio 'Portal', 'Escalera', 'Piso' y 'Puerta' en función del orden en el que se indiquen y la cantidad de éstos.

Para su corrección y validación se apoyará en diccionarios para provincias y poblaciones. ST Big Data Quality se integra con callejeros como el de Correos y otros, frecuentemente usados, como el INE.

Cabe destacar que se utilizarán técnicas estadísticas avanzadas de Fuzzy Matching para asociar un nombre de población incompleto o erróneo a uno correcto.

```
val c4aux = dirdf
 .join(c123aux,  dirdf(colId) === c123aux(colId) , "left_anti")
 .join(callejero_ext,
  !callejero_ext(col_provincia_nueva+"2").contains(dirdf(col_provincia_nueva)) &&
    dirdf(col_poblacion_nueva) =!= callejero_ext(col_poblacion_nueva+"3") &&

    levenshtein(dirdf(col_direccion_nueva), callejero_ext(col_direccion_nueva+"2")) <= distanciaLevMax &&

    dirdf(col_cpostal_nueva) === callejero_ext(col_cpostal_nueva+"2")
  , "inner"
 )
 .withColumn(col_inter_distancia,levenshtein(col(col_direccion_nueva), col(col_direccion_nueva+"2")))

.withColumn(colResultado,udf(lDir).apply(concat_ws(",",col(col_direccion_nueva+"_completa"),col(col_portal_nuev
a),
 col(col_resto_nueva),col(col_poblacion_nueva),col(col_cpostal_nueva),col(col_provincia_nueva))))

c4aux.cache()
```

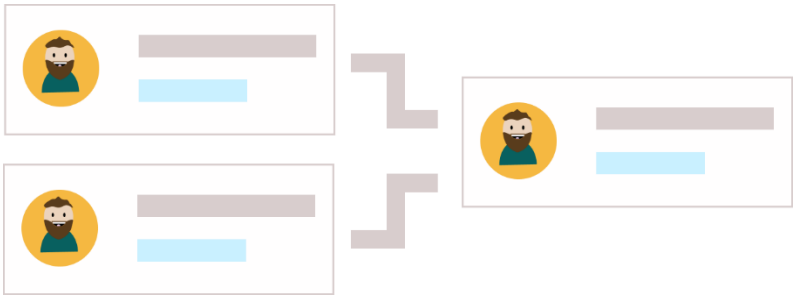
MÓDULO DE-DUPLICACIÓN

El objetivo de dicho módulo es determinar si dos registros de personas son duplicados y para ello se recurre a la aplicación de un algoritmo para la detección y eliminación de todos los posibles duplicados en los datos de entrada. Basándonos en los campos principales como pueden ser el DNI, Nombre y Apellidos. Aquellos que tienen más peso para determinar si se trata de la misma persona.

```
object DuplicatesRemoval {

 def main(args: Array[String]) {


  if (args.length > 0) {

   Logger.getLogger("org").setLevel(Level.ERROR)
   Logger.getLogger("akka").setLevel(Level.ERROR)

   /* SPARK SESSION */

   val spark = SparkSession
     .builder
     .master("local") // local
     .appName(nombre_proceso)
     .getOrCreate()
```

Se utilizará también el algoritmo de la distancia de Levenshtein entre cadenas de dos registros, como las de los campos DNI, Nombre y Apellido y, además, en la ponderación de los resultados en función del peso que puede tener cada uno de estos campos por separado en la determinación de la unicidad de un registro. Funcionando como algoritmos de clasificación de aprendizaje supervisado como el K-NN. Clasificará cada dato calculando la distancia del elemento nuevo con los existentes y los ordenará de menor a mayor distancia.

Además, cuando se detectan dos registros como duplicados, es posible usar, según la configuración, la información del registro, como, por ejemplo, el nombre entero más completo o la fecha que nos muestra el registro más reciente para decidir con cuál de todos los duplicados nos quedamos en la tabla procesada resultante.



3. TECNOLOGÍAS USADAS Y ESCALABILIDAD

Con el objetivo de abordar dicho problema de la manera más eficiente posible, se ha usado la tecnología Big Data Apache Spark. Ofreciendo de esta manera todos los beneficios de las tecnologías Big Data:

- **Escalabilidad:** El programa puede ejecutarse en 1 o n máquinas de un clúster sin hacer ningún cambio en el programa
- **Rapidez de procesamiento:** Gracias a la escalabilidad y a la arquitectura de Spark basada en el uso distribuido de la memoria RAM.
- **Tolerancia a fallos:** Incluso en procesos de Data Quality, que implican cantidades ingentes de datos, los procesos siempre terminarán.
- **Extensibilidad:** para añadir nuevas funcionalidades de calidad de datos.

Hemos optado por esta tecnología por la gran cantidad y variedad de fuentes de datos que es necesario procesar para lograr la calidad del dato, ya sean los datos de entrada como los diccionarios que se puedan utilizar para la corrección y validación de los datos.

Apache Spark es una tecnología Open Source con una comunidad muy activa. Es una herramienta donde los cálculos se ejecutan entre 10 y 100 veces más rápidos que otras plataformas. Además, en una misma herramienta nos permite combinar SQL in memory, streaming, Machine Learning y grafos.

Como se ha comentado, Spark se puede ejecutar en diversos entornos, ya sea en Spark standalone, Mesos, clúster en la nube en Databricks, EMR o HDInsight, distribuciones Hadoop on-premise, como Hortonworks o Cloudera.

4. CONCLUSIONES

Como hemos mostrado, **ST Big Data Quality** es la solución escalable y extensible de Stratebi para garantizar la calidad de datos en cualquier proceso de negocio.

Pudiendo abordar, con los diferentes módulos del proceso, la corrección y validación de:

- Teléfonos
- Documentos de Identidad
- Emails
- Nombres y Apellidos
- Dirección
- De-duplicaciones

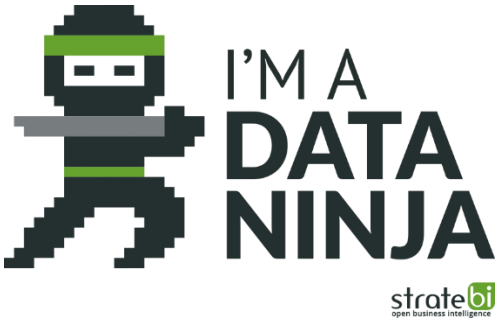
Sin duda, para tomar buenas decisiones de negocios, es fundamental tener una buena calidad del dato, sino sólo se consiguen decisiones pobres. Por lo que es fundamental garantizar dicha calidad para el análisis y toma de decisiones.

Con ST Big Data Quality obtenemos una mayor calidad del dato gracias a la integración con diccionarios de nombres y direcciones, las aplicaciones de expresiones regulares para que se respete normativas específicas aplicables a algunos de los módulos o técnicas de estadística avanzada o de Machine Learning para evitar elementos duplicados entre nuestros datos.

Todo ello gracias a las tecnologías Big Data que hoy día nos ofrece un rápido procesamiento, además de escalabilidad tanto tecnológica como a nivel funcional según las necesidades de cada cliente. Se ha convertido en una de las soluciones más utilizadas en el mundo para conocer cómo mejorar tu negocio.

En definitiva, con herramientas como ST Big Data Quality se pueden abordar un grupo de datos para conseguir información fiable y aumentar la rentabilidad de aquellas empresas que lo empleen.

5. INFORMACIÓN SOBRE STRATEBI



Stratebi es una empresa española, con sede en Madrid y oficinas en Barcelona, Alicante y Sevilla, creada por un grupo de profesionales 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 a través de Stratebi.

Stratebi es la única empresa española que ha estado presente en los 11 Pentaho Developers celebrados en Europa.

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.**

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.

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

Stratebi ha sido elegida como **Caso Éxito del Observatorio de Fuentes Abiertas de Cenatic**.

Observatorio Nacional del Software de Fuentes Abiertas | cenatic

6. REFERENCIAS STRATEBI

SECTOR PRIVADO

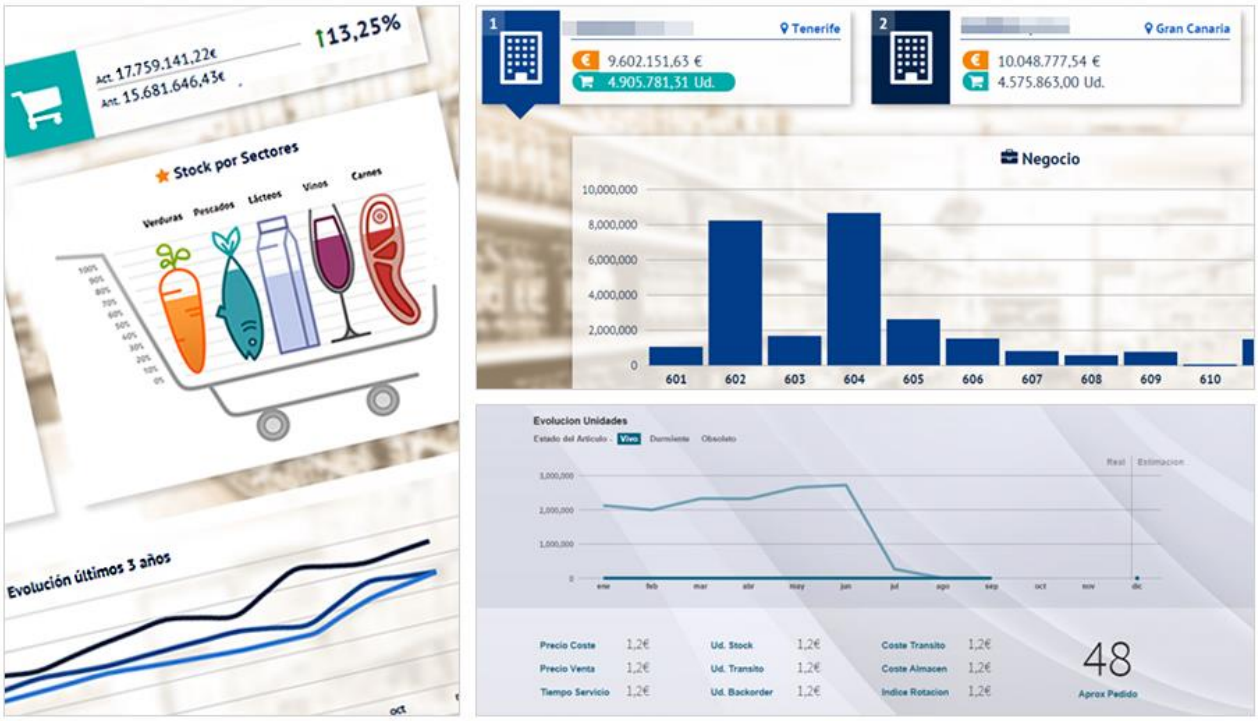
GLOBALIA
ITALFARMACO
pepecar.com
20:20mobile
EROSKI
amper
Unilever
educaweb (*)
EQUIFAX
BBVA
Blanco
PROSEGUR
FEMUR
Telefónica
FERMAX
BOSCH
AMNESTY INTERNATIONAL

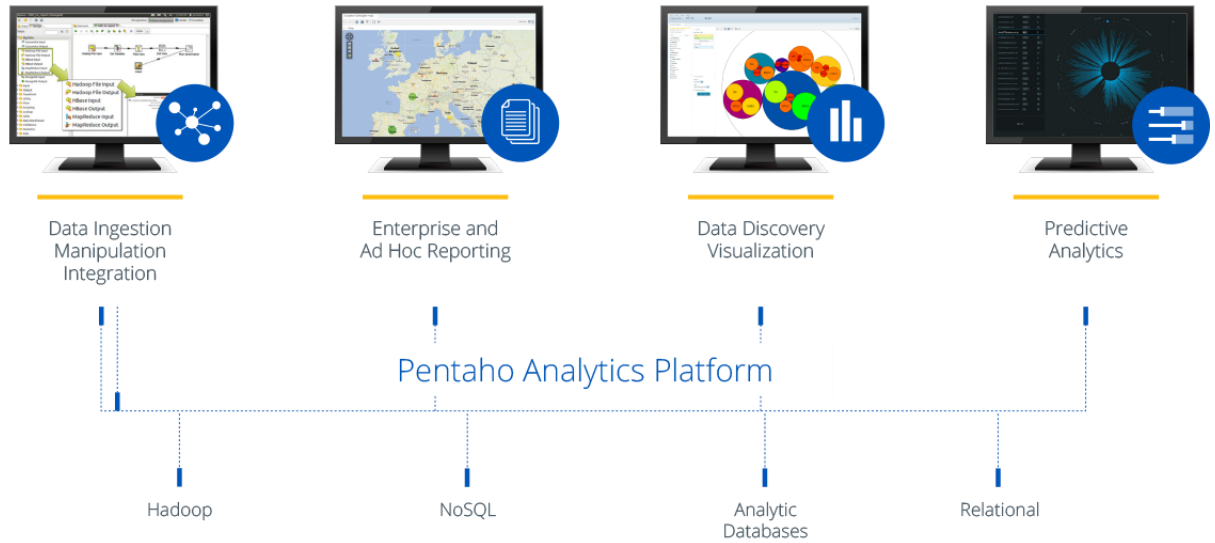
SECTOR PÚBLICO

Gobierno de Canarias
Consejería de Educación, Universidades y Sostenibilidad
UAM UNIVERSIDAD AUTONOMA DE MADRID
Caixa d'Enginyers
Diputació Tarragona
Ayuntamiento de Logroño
CSIC
Universitat de Lleida
IDI INSTITUT DE DIAGNÒSTIC PER LA IMATGE
Ayuntamiento de Arrecife
Gobierno de La Rioja
Dipsalut
Organisme Autònom de Salut Pública
Diputació de Girona
Universidad Zaragoza
AJUNTAMENT DE REUS
SaludMadrid
EOI Escuela de organización industrial

7. EJEMPLOS DE DESARROLLOS ANALYTICS

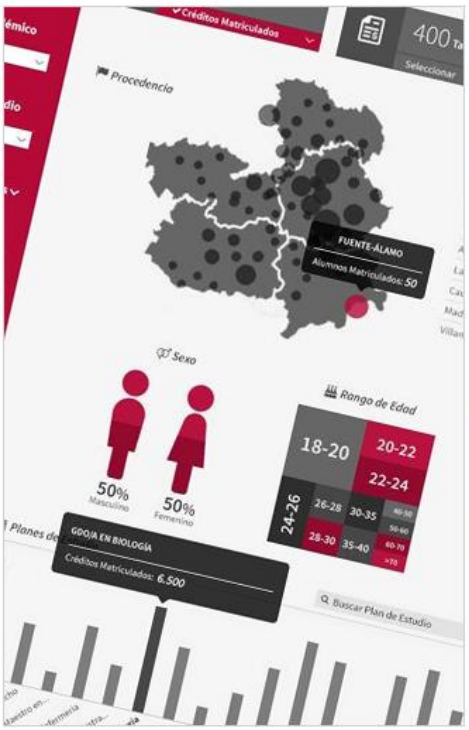
A continuación, se presentan **ejemplos de algunos screenshots** de cuadros de mando diseñados por Stratebi, con el fin de dar a conocer lo que se puede llegar a obtener, así como Demos Online en la web de Stratebi:





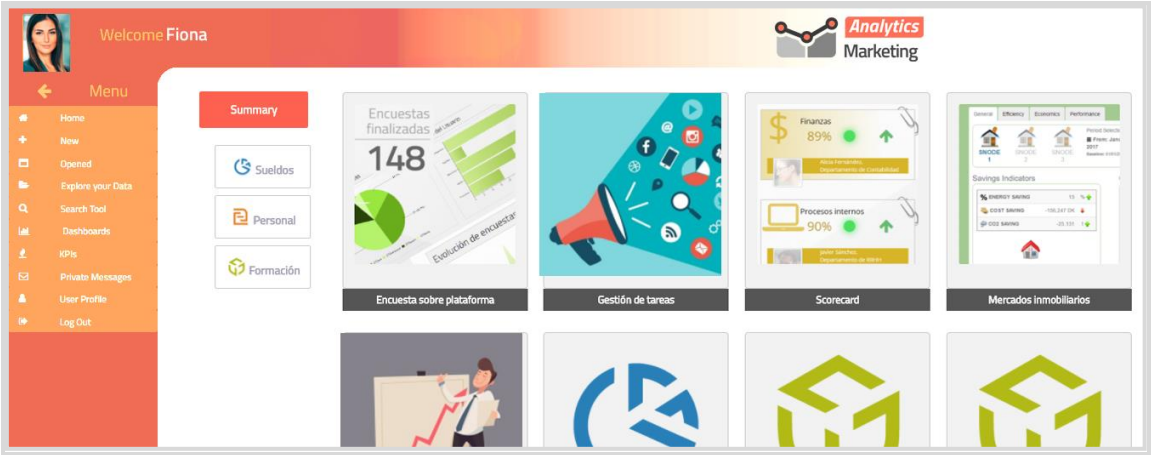
facebook DASHBOARD COMENTARIOS
Sexo
65%
35%
Sentimiento
28%
60%
12%
Localización
Tipo de recurso
Evolución Comentarios por días

smart MADRID
Último Actualización
18:24
TOP 5 PUNTOS DE MEDIDA
PROXIMIDAD DE PUNTOS DE MEDIDA
EVOLUCIÓN INTENSIDAD DE TRÁFICO


Créditos Matriculados
Procedencia
FUENTE-ÁLAMO
Alumnos Matriculados: 50
Sexo
50% Masculino
50% Femenino
Rango de Edad
18-20
20-22
22-24
24-26
26-28
28-30
30-35
35-40
GDO/A 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
TERRITORIAL: Madrid - Capital


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
Finanzas
89%
Procesos internos
90%
Encuesta sobre plataforma
Gestión de tareas
Scorecard
Mercados inmobiliarios

stratebi
open business intelligence

- Big Data
- ETL y Data Quality
- Business Intelligence
- Machine Learning
- Soluciones Verticales