BBVA INNOVATION CENTER

BBVAOpen4U

Herramientas de **visualización de datos**

BBVA INNOVATION CENTER | BBVAOpen4U

01 La caja de herramientas del científico de datos

La ciencia del dato se erige en nuestros días como una profesión multidisciplinar. Esta pretende ser una guía básica de recursos en cada una de las facetas desempeñadas por estos profesionales.

La ciencia del dato se erige en nuestros días como una profesión multidisciplinar en la cual conocimientos técnicos de diversas áreas se solapan formando un perfil más propio del Renacimiento que del superespecializado siglo XXI.

Dada la escasez de formación estructurada en la materia, los científicos de datos se ven obligados a ir coleccionando conocimientos, habilidades y herramientas que les permitan desarrollar de forma óptima sus competencias.

Este artículo pretende ser una guía básica no exhaustiva de recursos en cada una de las facetas desempeñadas por estos profesionales.

HERRAMIENTAS Y LENGUAJES

- SQL
- Sqlite
- SQlite3
- RSQlite
- Toad
- Tora
- RapidMiner
- Knime
- Pentaho
- RODBC
- RJDBC
- pyODBC
- mxODBC
- SQLAlchemy
- pandas
- data.table
- XML
- Jsonlite
- json

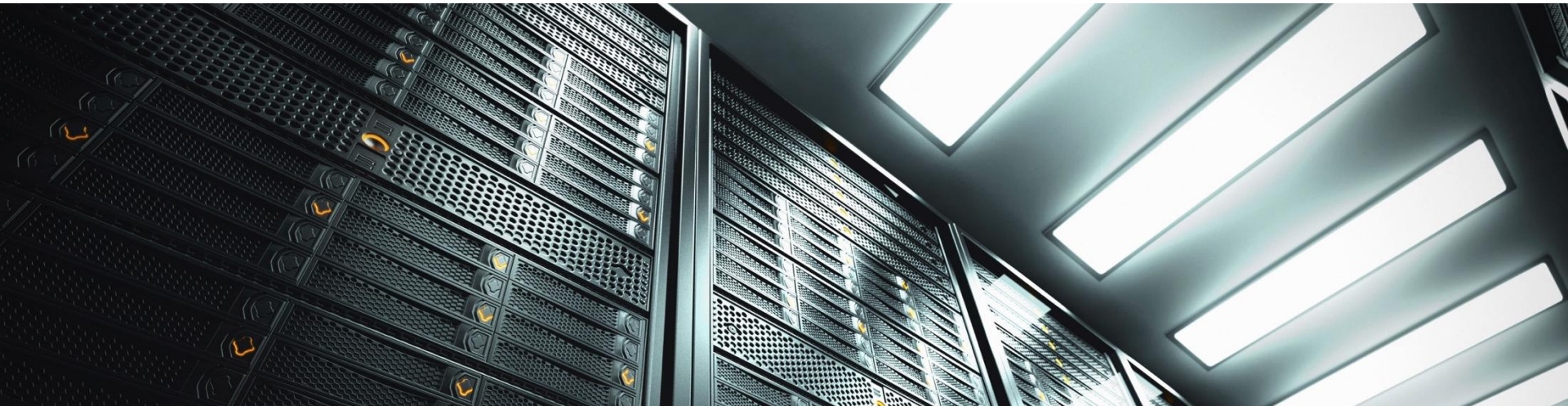
Gestión de datos

Parte del trabajo del científico de datos es la captura, depuración y almacenamiento de la información en un formato adecuado para su tratamiento y análisis.

El caso más frecuente será el acceso a una réplica de los datos para una captura puntual o periódica.

Será necesario conocer SQL para acceder a la información almacenada en bases de datos relacionales. Cada base de datos tiene una consola de comandos para ejecutar las sentencias SQL, aunque son mayoría los que prefieren un entorno gráfico con información sobre tablas, campos e índices. Entre las herramientas más valoradas están Toad, versión comercial para plataforma Microsoft y Tora, versión libre multiplataforma.

Una vez extraídos los datos podemos guardarlos en ficheros de texto plano que luego cargaremos en nuestro entorno de trabajo para *machine learning* o utilizar una herramienta como SQlite.



SQlite es una base de datos relacional ligera, sin dependencias externas y que no necesita la instalación en un servidor. Transportar una base de datos es tan fácil como copiar un solo fichero. En nuestro caso, cuando vayamos a procesar la información lo haremos sin necesidad de concurrencia ni de múltiples escrituras en los datos origen, lo cual se adapta perfectamente a las características de SQlite.

Los lenguajes que utilizaremos para nuestros algoritmos tienen conectividad con SQlite (Python a través de SQlite3 y R a través de RSQlite) por lo que podemos optar por importar los datos antes de preprocesarlos o hacer parte en la base de datos, lo cual nos evitará más de un problema a partir de un volumen medio de registros.

Otra alternativa para la captura en lotes de los datos es la utilización de una herramienta que incluya el ciclo ETL completo (extracción, transformación y carga), entre las que destacan RapidMiner, Knime y Pentaho. Con ellas podremos definir el ciclo de captura y depuración de los datos de manera gráfica mediante conectores.

Cuando tengamos garantizado acceso al origen de datos durante el preproceso podemos optar por una conexión ODBC (RODBC y RJDBC en R y pyODBC, mxODBC y SQLAlchemy en Python) y beneficiarnos de realizar uniones (JOIN) y agrupaciones (GROUP BY) utilizando el motor de la base de datos e importando posteriormente los resultados.

Para el procesado externo a la base de datos pandas (librería de Python) y data.table (paquete de R) son primera elección. En el caso de R, data.table permite soslayar uno de los puntos débiles de éste, la gestión de la memoria, realizando operaciones vectoriales y agrupaciones por referencia, es decir, sin tener que duplicar temporalmente los objetos.

Un tercer escenario sería el acceso a información generada en tiempo real y que sea transmitida en formatos como XML o JSON. Serían proyectos denominados de incremental *learning* entre los que se encuentran los sistemas de recomendación, publicidad online y trading de alta frecuencia.

Utilizaremos herramientas como XML o jsonlite (paquetes para R) o xml y json (módulos de Python). Con ellos haremos una captura en *streaming*, calcularemos la predicción, la devolveremos en el mismo formato y actualizaremos nuestro modelo una vez el sistema de origen nos facilite, más adelante, el resultado observado en la realidad.

Análisis de datos

Si bien las áreas de ***business intelligence***, ***data warehousing*** y ***machine learning*** **son objetos de la ciencia de datos,** es esta última la más diferencial en el sentido que necesita de un número mayor de utilidades específicas.

En cuanto a lenguajes de programación, imprescindibles en nuestra caja de herramientas son R y Python, los más utilizados para el aprendizaje automático.

Para Python destacamos la suite scikit-learn que cubre casi todas las técnicas, salvo quizás las redes neuronales. Para estas tenemos varias alternativas interesantes, como Caffe y Pylearn2. Pylearn2 utiliza como base Theano, una interesante librería de Python que permite definiciones simbólicas y uso transparente de los procesadores GPU.

Si necesitamos modificar algún paquete de R requeriremos C++ y disponer de utilidades que nos permitan volver a generarlos: Rtools o devtools facilitan todos los procesos relacionados con el desarrollo.

Entre los paquetes para R más utilizados destacan:

- *Gradient boosting*: gbm y xgboost.
- Ensamblado de árboles de regresión y clasificación: randomForest y randomForestSRC.
- Máquinas de soporte de vectores: e1071, LiblineaR y kernlab.
- Regresión con regularización (Ridge, Lasso y ElasticNet): glmnet.
- Modelos generalizados aditivos: gam.
- *Clustering*: cluster.

Utilidades que nos harán la vida más fácil en R:

- Data.table: Lectura rápida de ficheros texto, creación, modificación y borrado de columnas por referencia, unión de tablas por una clave común o agrupación y resumen de datos.
- Foreach: Ejecución de procesos en paralelo contra un *backend* previamente definido con alguna utilidad como doMC o doParallel.
- Bigmemory: Manejar grandes matrices y compartirlas entre varias sesiones o ejecuciones.
- Caret: Comparación modelos, control de particiones de datos (*splitting*, *bootstrapping*, *subsampling*) y ajuste de parámetros (*grid search*).
- Matrix: Manejo de matrices dispersas y transformación de variables categóricas a binarias (*onehot encoding*) mediante la función sparse.model.matrix.

Una mención especial requieren los entornos distribuidos. Si hemos trabajado con datos procedentes de una entidad o empresa de cierto tamaño probablemente tengamos experiencia con el denominado ecosistema Hadoop. Hadoop es en su origen un sistema distribuido de ficheros (HDFS) dotado de unos algoritmos (MapReduce) que permiten realizar procesamiento de la información en paralelo.

Algunas de las herramientas de aprendizaje automático que conviven con Hadoop:

- Vowpal Wabbit: Métodos para *online learning* basado en gradiente descendente.
- Mahout: Suite de algoritmos entre los que destacan los sistemas de recomendación, *clustering*, regresión logística, *random forest*.
- h2o: Quizás la herramienta en fase de mayor crecimiento, con un gran número de algoritmos paralelizables. Puede ejecutarse desde un entorno gráfico propio o bien desde R o Python.

Interesará también al científico de datos estar al corriente de las **nuevas tendencias de cambio generacional de Hadoop hacia Spark**.

Spark tiene varias ventajas sobre Hadoop para el procesamiento de la información y la ejecución de algoritmos. La principal de ellas la velocidad, dado que es hasta cien veces mayor debido a que, a diferencia de Hadoop, Spark utiliza la gestión 'en memoria' y sólo escribe a disco cuando es necesario.

Spark puede ejecutarse de forma independiente o puede convivir como un componente más de Hadoop, de forma que la migración puede planificarse de manera no traumática. Puede por ejemplo utilizar HBase como base de datos, aunque Cassandra se está imponiendo como solución de almacenamiento por su redundancia y escalabilidad.

Como muestra de los aires de cambio, Mahout desde el pasado año trabaja para integrarse con Spark, distanciándose de MapReduce y Hadoop, y H2O.ai ha lanzado Sparkling Water que es la versión de su suite h2o sobre Spark.

Visualización

Para terminar una breve referencia a la presentación de los resultados.

Las herramientas más utilizadas en R son sin duda lattice y ggplot2 y en Python Matplotlib, pero si necesitamos presentaciones profesionales integradas en entornos web la mejor opción sin duda es D3.js.

Entre los entornos integrados de *business intelligence*, con un enfoque claro a la presentación, destacar Tableau, el más conocido, y como alternativas para la exploración gráfica de datos, Birst y Necto.

BBVA INNOVATION CENTER | BBVAOpen4U

02 Cinco herramientas

de visualización de datos que no debes perder de vista

Te presentamos algunas de las mejores herramientas de visualización de datos que puedes usar en tu negocio para sacar el mayor provecho a la gran cantidad de información que se crea cada día en el mundo digital.

ÍNDICE DE HERRAMIENTAS DE VISUALIZACIÓN

Hoy en día, el universo digital está alcanzado nuevos umbrales. La cantidad de datos generada, tanto por usuarios particulares como por las empresas, está aumentando a un ritmo vertiginoso. De hecho, según un estudio de IDC y EMC, el universo de datos digitales está doblando su tamaño cada dos años y, en 2020, se habrán generado 44 *zettabytes* de información o, lo que es lo mismo, 44 trillones de gigabytes de datos estructurados y desesctructurados.

El hecho de crear y acceder a una página web, participar en un blog, aumentar nuestro número de seguidores, escribir comentarios, mandar un tuit o simplemente, navegar por internet, produce toda una serie de datos que, si se saben aprovechar correctamente, pueden ofrecer un gran valor para las empresas.

El gran reto, no obstante, es dar sentido a todos esos datos. Es decir, ser capaces de captar, relacionar, analizar y extraer su verdadero valor, de forma que la información se pueda presentar de manera atractiva, clara, concisa y comprensible. El objetivo es facilitar la toma de decisiones dentro de tu negocio. Explorar y analizar visualmente los datos de clientes puede llevarte, además, a descubrir nuevas vías para llegar hasta ellos, segmentarlos mejor, personalizar ofertas de productos o servicios y crear ideas innovadoras, entre otras muchas posibilidades, que pueden mantener el *engagement* entre tu marca y tus usuarios a lo largo del tiempo.

Por dónde empezar

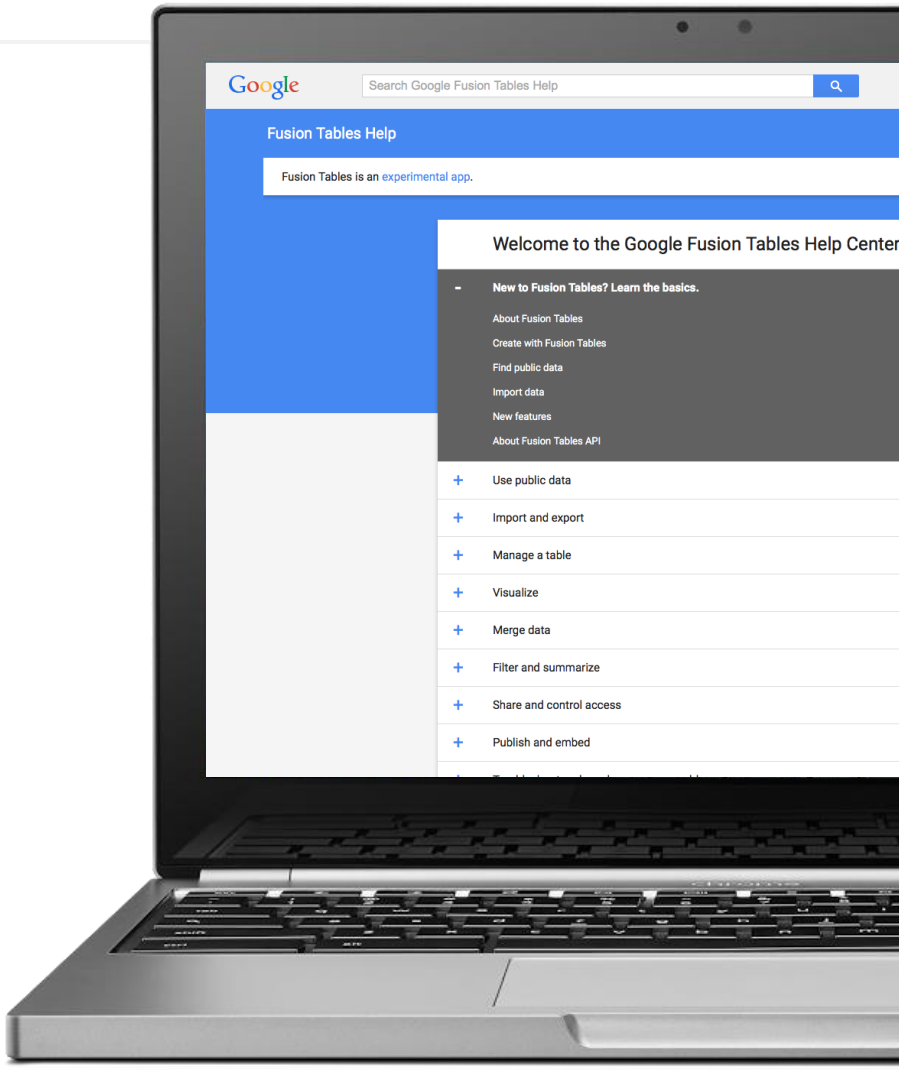
Puede que el primer paso dentro de la visualización de datos resulte intimidante. Por fortuna, al igual que el crecimiento de datos avanza, también lo hacen las herramientas que nos ayudan a sacar su valor. Te presentamos 5 herramientas recomendables para iniciarse en este mundo.

Google Fusion Tables

Es una excelente herramienta para principiantes o aquellas personas que no saben de programación. Además, para los usuarios más avanzados, existe una API que permite producir gráficas o mapas a partir de información.

Una de las ventajas de esta aplicación es la diversidad de representaciones de datos que hay a disposición del usuario. Además, ofrece la posibilidad de crear gráficos o mapas de manera relativamente rápida, incluyendo funciones GIS para analizar datos por geografías.

Esta herramienta es muy utilizada por The Guardian para producir mapas de una manera rápida y detallada.



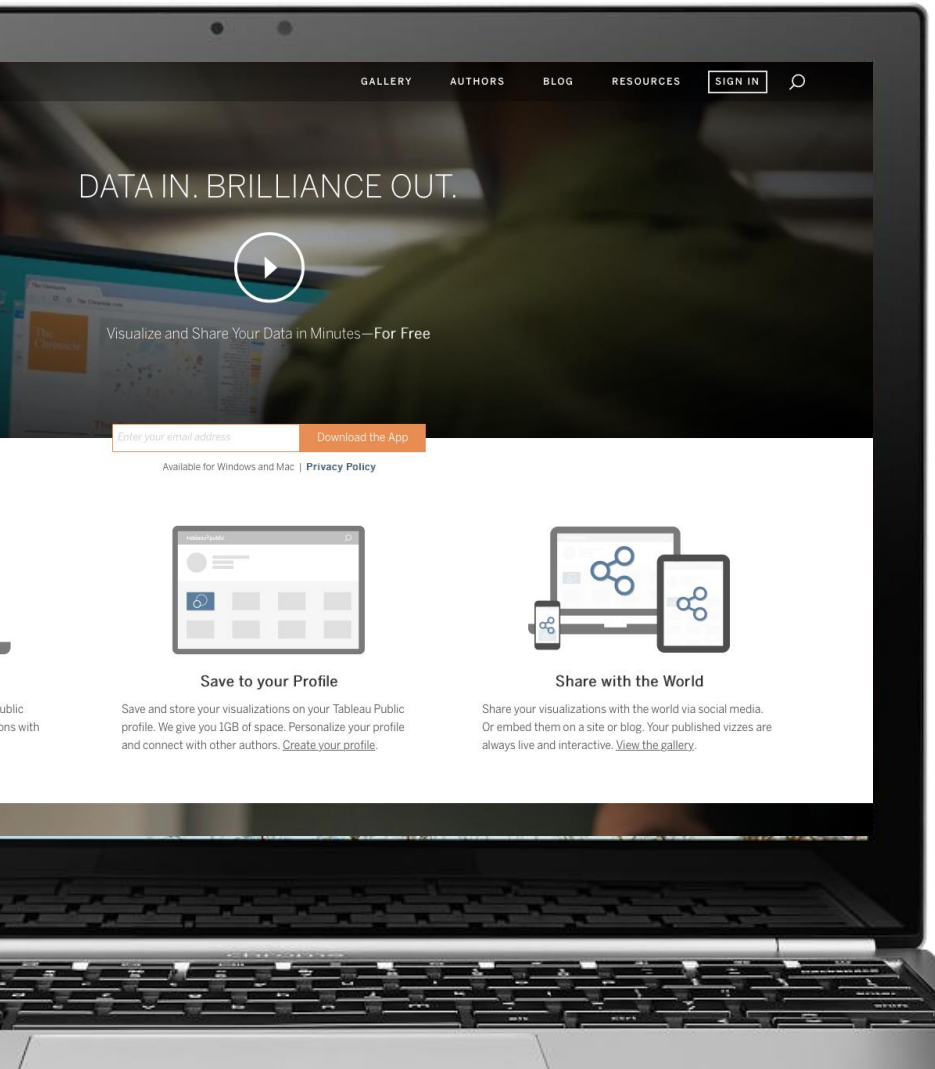
CartoDB

Se trata de un servicio *open source* dirigido a cualquier usuario, independientemente del nivel técnico que tenga, con una interfaz muy amigable. Permite crear una gran variedad de mapas interactivos, lo que permite elegir entre el catálogo que el mismo servicio ofrece, incluir mapas de **Google Maps**, o agregar a la lista tus propios mapas personalizados.

Lo más interesante es que es posible tener acceso a los datos de **Twitter** para ver cómo los usuarios reaccionan ante una marca, una determinada campaña de marketing o un evento. Un buen ejemplo de ello lo podemos ver en el mapa de seguimiento de tuits que se creó el pasado año con motivo del lanzamiento del último álbum de **Beyonce** en el que, claramente, se puede observar los lugares donde más impacto tuvo. Toda una fuente de información visual para los profesionales del marketing y de negocio.

También hay que destacar su activo grupo de desarrolladores que aporta gran cantidad de documentación y ejemplos. Además, el carácter abierto de su **API** hace que continuamente se estén desarrollando nuevas integraciones y aumentando las capacidades de la herramienta con nuevas librerías.

Tableau Public



Con **Tableau Public** puedes crear mapas interactivos, gráficos de barras, tartas, etc. de forma sencilla. Una de sus ventajas es que, al igual que con **Google Fusion Tables**, es posible importar tablas de **Excel** para facilitar tu trabajo. En cuestión de minutos, puedes crear un gráfico interactivo, embeberlo en tu página web y compartirlo. Por ejemplo, el medio de comunicación **Global Post** creó una serie de gráficos sobre cuáles son los mejores países para hacer negocios en África.

Recientemente, lanzaron su versión 8.2. En ella también podemos encontrar la nueva herramienta OpenStreetMap que permite generar mapas muy detallados a partir de datos locales, como cafeterías o tiendas. **Tableau Public** es una herramienta gratuita, aunque existe también una versión de pago.

iCharts

Con esta herramienta puedes iniciarte en el mundo de la visualización de datos. iCharts cuenta con una versión gratuita (Basic) y dos de pago (Platinum y Enterprise). Con esta herramienta puedes crear visualizaciones en pocos pasos exportando documentos de **Excel** y **Google Drive**, o añadir datos manualmente.

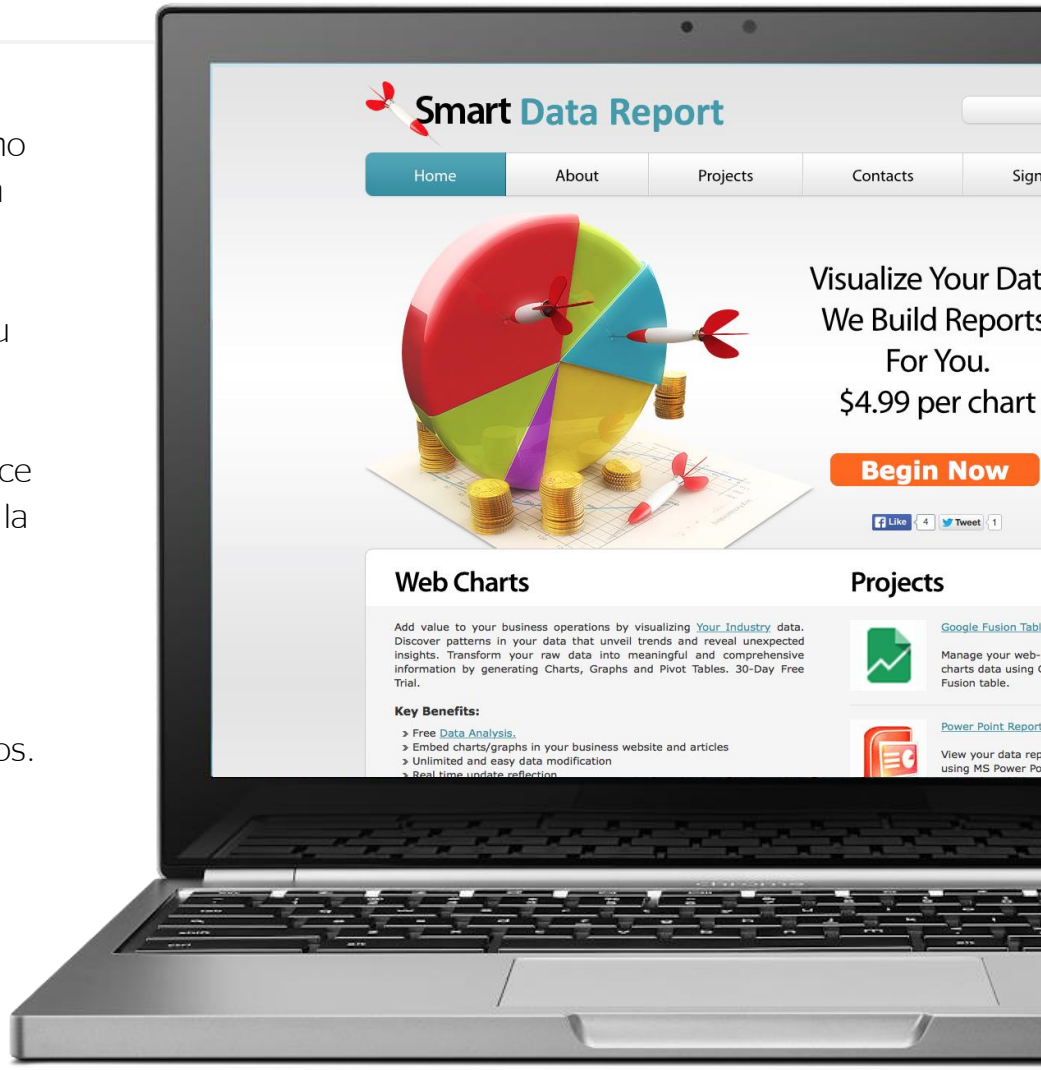
A través de esta herramientas también es posible compartir tus gráficos con tus colaboradores de forma privada, además de poder editar y actualizar dichos gráficos con nuevos datos a través de su servicio de *cloud computing*. Incluso, puedes compartirlos con tus clientes a través de mensajes de correo electrónico, boletines de noticias o redes sociales.

Entre las empresas que utilizan este servicio nos encontramos con la prestigiosa consultora IDC, que utiliza **iChart** para ofrecer un aspecto visual a los datos más relevantes de sus informes.

Smart Data Report

Finalmente, queremos destacar la solución **Smart Data Report** que, aunque no sea tan potente como las anteriores, tiene la ventaja de ser una solución de visualización de datos asequible para emprendedores y pequeñas empresas cuyos trabajadores no disponen de mucho tiempo en su día a día.

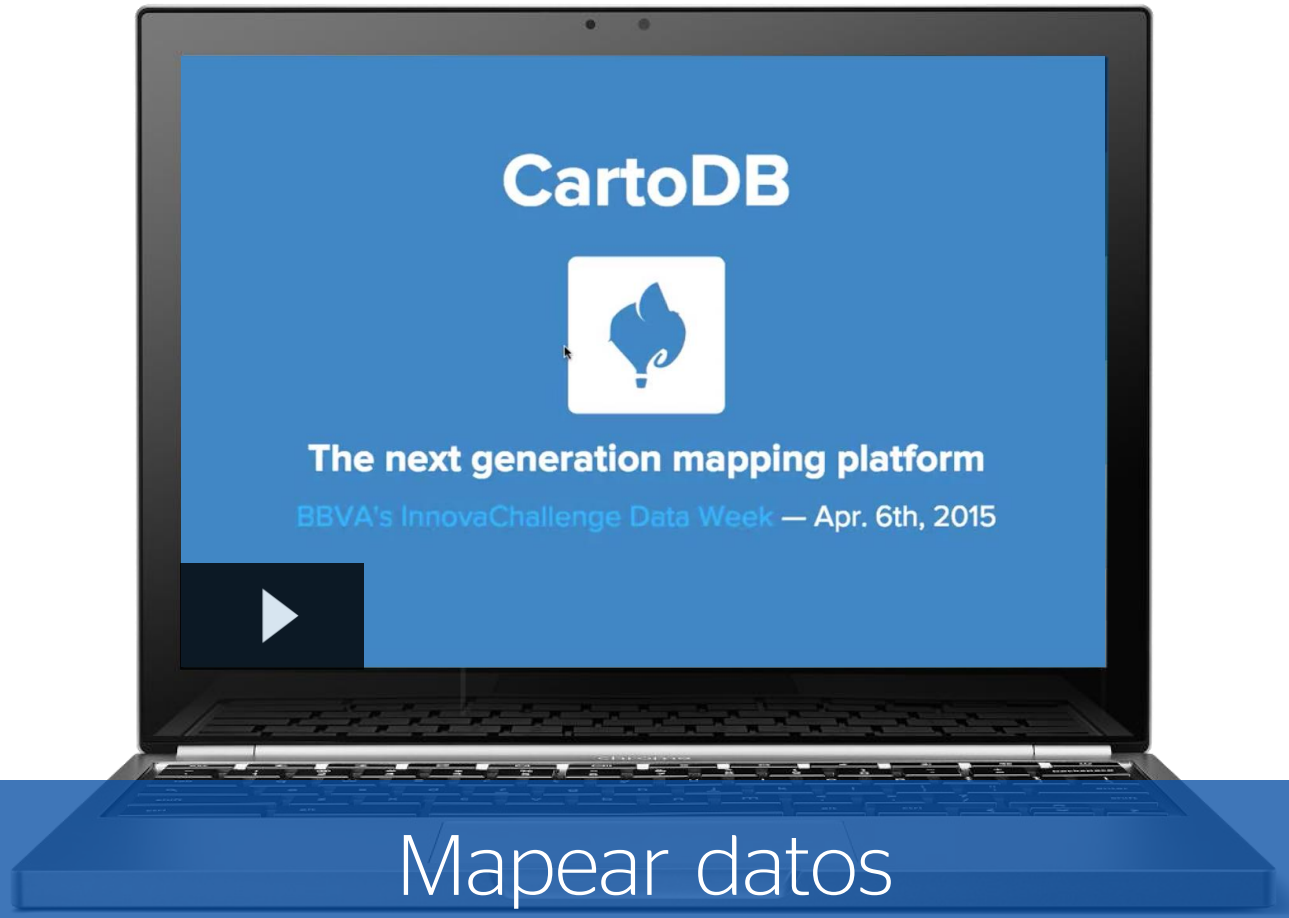
Esto se debe a que entre los servicios que te ofrece esta web se incluye el análisis de datos gratuito y la generación de informes que pueden enviarse por email, sin que sea necesario que lo haga uno mismo. Una vez el servicio ha preparado tu informe, se generan un código **HTML** para poder embeberlo en tu web corporativa o en tus artículos.



BBVA INNOVATION CENTER | BBVAOpen4U

03 Saca provecho a los datos con estos cuatro tutoriales

Mapear datos, visualizarlos en apps geoespaciales y aplicar el aprendizaje automático. Ponemos en práctica nuestros conocimientos con la ayuda de estos vídeos tutoriales.

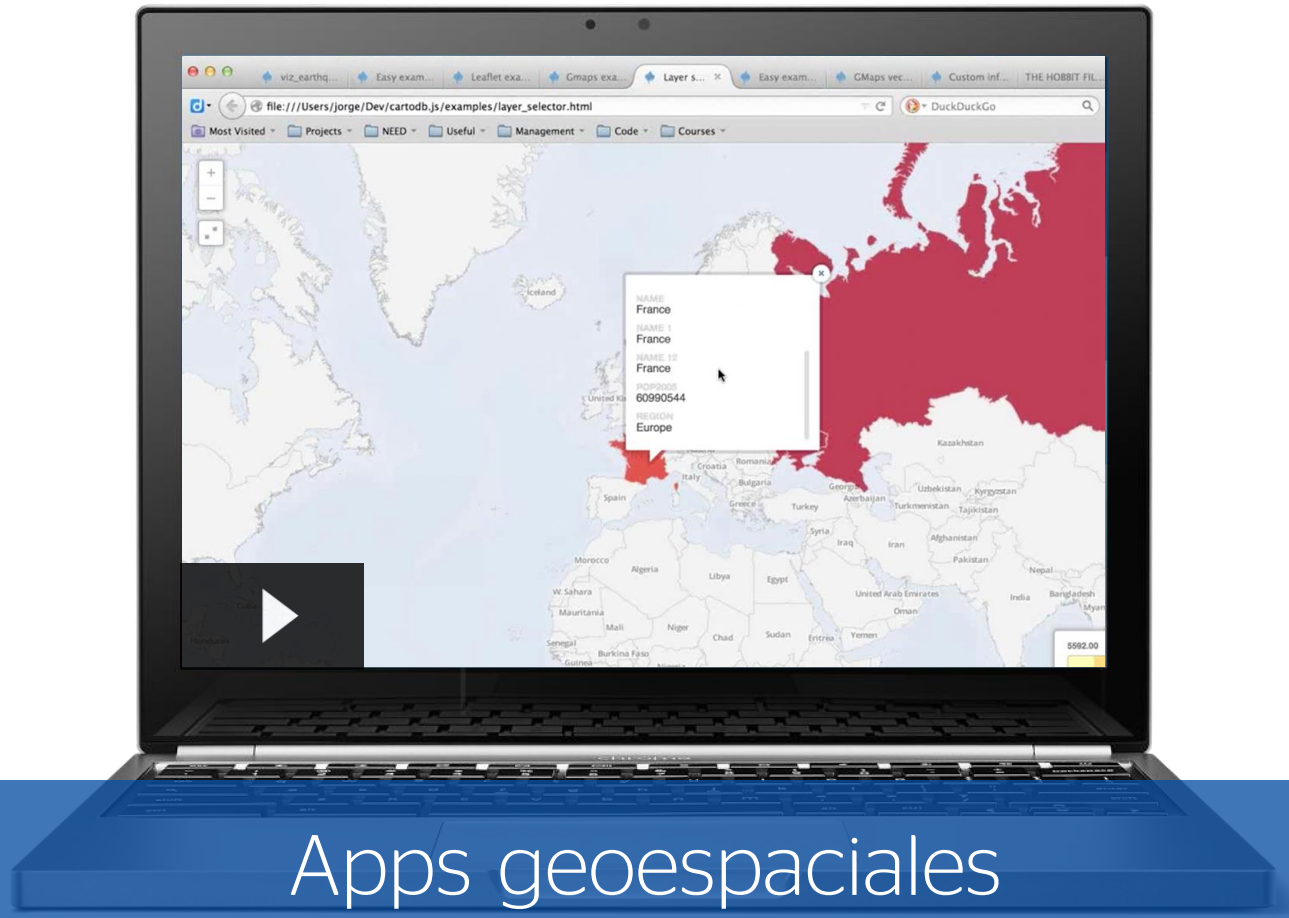


Mapear datos

CartoDB nos explica cómo convertir los datos de localización en conocimiento para tu negocio. En este tutorial podrás aprender a analizar, visualizar y construir aplicaciones de datos a través de su herramienta.

Aprendizaje automático

Ahora que se acerca el verano, Andrés González, responsable de soluciones de Big Data y Data Prediction en Clever Task, nos enseña cómo hacer predicciones de los datos en un caso muy concreto: el sector turístico.



Apps geoespaciales

Y si lo que quieres es aprender a crear aplicaciones con datos geoespaciales no te puedes perder este tutorial, también de CartoDB, en el que se explica cómo puedes sacar provecho de una API, en este caso la que abrió BBVA para el concurso InnovaChallenge, y así poder crear apps y visualizaciones.

Buenos ejemplos de visualización

Por último para cerrar esta recopilación, Alberto Cairo, profesor de visualización de datos en la Universidad de Miami, nos enseña las buenas prácticas en visualización de datos. Es bueno aprender de nuestros errores y de los aciertos de otros.

compartir

TE PUEDE INTERESAR

[Innovation Edge Big Data: generar valor de negocio con los datos]

[Emerging Tech: la visualización de datos más allá del ruido]

[Infografía: las claves de Big Data según DJ Patil]

[Infografía Big Data: el presente y el futuro de los datos]

[Caso de éxito de visualización de datos: Illustreets y CartoDB]

Regístrate
para estar al día de las últimas tendencias

BBVA INNOVATION CENTER

BBVAOpen4U

www.bbvaopen4u.com

conversa con nosotros en: