



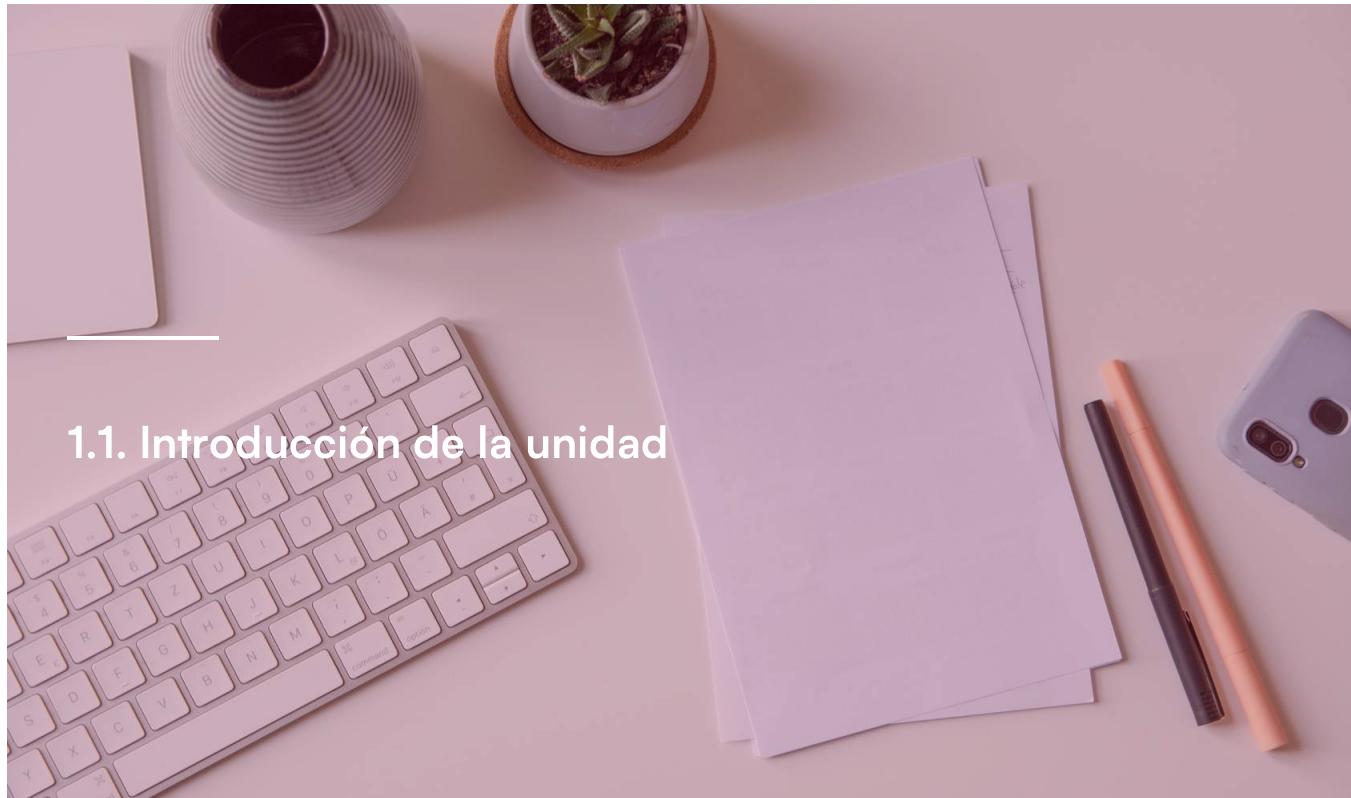
Análisis de datos masivos aplicados al negocio



- ≡ I. Introducción
- ≡ II. Objetivos
- ≡ III. Datos externos
- ≡ IV. Demo
- ≡ V. Resumen
- ≡ VI. Caso práctico con solución
- ≡ VII. Glosario

QUESTION BANKS

I. Introducción



1.1. Introducción de la unidad

En esta unidad se tratará uno de los temas clave para obtener un conjunto de datos completo para la empresa u organización. En este caso se trata del análisis de la información externa a la compañía, para poder aportar información adicional a la toma de decisiones.

El mundo de los **datos abiertos** proporcionados por Gobiernos y empresas gubernamentales es cada vez más grande. Junto con técnicas de **web scrapping**, **API abiertas**, **datos de terceros de pago o compartidos**, forma el principal conjunto de fuentes de datos que se puede integrar junto con los datos internos de la compañía.

La incorporación de datos externos o de terceros es una parte importante de los programas de análisis de datos, ya que las empresas buscan información estratégica desde fuera de sus empresas.

i Una estrategia de datos exitosa convierte los datos de una empresa en importantes conocimientos y ganancias para los departamentos de las organizaciones que los utilizan. Pero no debe limitarse a la información que proviene del interior de la empresa. Existe información meteorológica histórica, de preferencias de clientes y de tendencias de compra a la que las organizaciones tienen acceso. Esto es una gran cantidad de conjuntos de datos fuera de su ámbito natural, algunos son de pago y otros, de carácter gratuito o colaborativo. Con ello pueden agudizar los análisis e impulsar los resultados finales.

Las organizaciones más maduras, analíticamente hablando, utilizan fuentes de datos como **datos de clientes, datos de proveedores, reguladores e incluso competidores**. Las organizaciones analíticamente más innovadoras, o las empresas que incorporan la analítica en la mayoría de los aspectos de la toma de decisiones, tienen más probabilidades que las organizaciones menos maduras a la hora de utilizar ciertas fuentes de datos. Y es más probable que utilicen una variedad de tipos de datos, incluidos datos móviles, de redes sociales y, por últimos, públicos o de terceros.

Existen incluso organizaciones que comparten sus propios datos con clientes, proveedores, agencias gubernamentales e incluso competidores. Así, generan una mayor influencia en su ecosistema empresarial. Esta colaboración compartiendo datos se realiza en la búsqueda de mejoras de procesos, principalmente.

En palabras de Asif Muhammad Syed, “En todo tipo de áreas, la gente está utilizando datos de terceros para aumentar los datos que ya tienen [...]. En la mayoría de los casos, usted no se puede construir modelos predictivos de alta calidad con solo datos internos”.¹

¹Asif Muhammad Syed, vicepresidente de estrategia de datos de Hartford Steam Boiler Inspection and Insurance Co. (2020).

En este módulo se verá lo siguiente:

- *Open data* y API abiertas.
- Comprar datos.
- Compartir datos.
- Datos de redes sociales y *web scrapping*.

II. Objetivos



2.1. Objetivos de la unidad

1

Comprender los datos externos a la compañía.

2

Conocer las tipologías de datos externos.

3

Identificar datos de terceros útiles para la organización.

4

Extraer un ejemplo de datos de terceros.

III. Datos externos

Los datos externos a menudo siguen siendo un recurso sin explotar por la mayoría de las organizaciones, que se centran en los datos generados por ellas mismas.

Con la creciente cantidad de datos disponibles a través de la web, obtenidos de proveedores de datos especializados, los datos externos son cada vez más relevantes. Los datos externos complementan los datos internos mejorando y enriqueciendo la información de la que se dispone; para así realizar análisis avanzado y optimizar procesos como el del área comercial (por ejemplo, con datos de geolocalización, clima o tráfico). También permiten crear nuevos servicios de los que no se disponía mediante los datos internos y enriquecerlos.

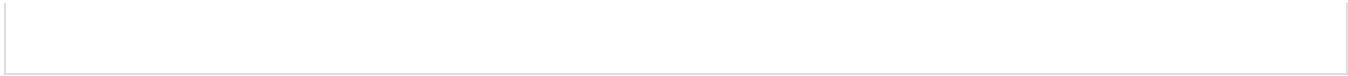
Pero ¿qué son los datos externos?

Los datos externos a menudo se asocian con el mundo **open data**, pero son más que esta increíble e inmensa fuente de datos. La siguiente definición ha sido desarrollada por el Competence Center Corporate Data Quality (CC CDQ): "Los datos externos se refieren a cualquier tipo de datos que hayan sido capturados, procesados y proporcionados desde fuera de la empresa".²

Así, se pueden clasificar los datos externos en cuatro categorías:

1	2	3	4
---	---	---	---

Datos abiertos.



Datos pagos.



Datos compartidos.



Datos de redes sociales.

²Krasikov, P.; Legner, C.; Harbich, M.; Eurich, M. *Open data use cases. Framework for the generation and documentation of open data use cases*; 2020.

	Open data	Datos de pago	Datos compartidos	Redes sociales
Origen	Gobiernos, Organizaciones no gubernamentales	Proveedores profesionales de datos. Intermediarios	Datos internos de las empresas, fuentes autorizadas	Contenido generado por usuario
Acceso	Plataformas de datos abiertos, meta plataformas, enlaces directos	Portales dedicados, software	Intercambio bilateral o intermediario	Conexión a puntos de acceso oficiales o red web de plataformas y redes sociales
Precio	Gratis	Costes por uso, suscripción, etc.	Los intermediarios pueden cobrar ciertas tarifas	Disponible de forma gratuita pero sujetos a copyright
Estructura	Semiestructurados, desestructurados	Estructurados	Estructurados, semiestructurados, desestructurados	Desestructurados

Tabla 1. Datos externos.

Fuente: elaboración propia.

3.1. *Open data* o datos abiertos

Los datos abiertos se pueden definir como “datos que están disponibles gratuitamente y que todos pueden usar y volver a publicar sin restricciones de derechos de autor o patentes”.³

Los datos abiertos se basan en ocho principios fundamentales:

³Braunschweig, K.; Eberius, J.; Thiele, M.; Lehner, W. *The state of open data. Limits of current open data platforms*. Technische Universität Dresden; 2012.

Principios 1

Completos

Todos los datos públicos se ponen a disposición, son datos que no están sujetos a limitaciones de privacidad, seguridad o privilegios válidos.

Con ellos se pretende que la población pueda tomar mejores decisiones para el bien general y/o reutilizarlos en interés propio o de terceros. Por ello se puede conocer y reutilizar el contenido de los datos que contienen los Gobiernos, ya que se ha pagado anteriormente por ellos mediante los impuestos.

Principios 2

Primarios

Los datos se recogen en la fuente al nivel de detalle mayor, no hay agregaciones o modificaciones. Para ello se deben cumplir tres propiedades:

- Nivel alto de detalle.
- Originales, sin tratamientos previos.
- Trazables para validar su origen.

Principios 3

Oportunos

Los datos se publicarán tan pronto como sea necesario para preservar su valor. Algunos datos solo tienen valor en un periodo de tiempo acotado. Por ello se deben publicar rápido y actualizar frecuentemente.

Principios 4

Accesibles

Para la mayor parte de usuarios posible y para diversos objetivos.

Principios 5

Procesables por máquinas

Se deben publicar los datos pensando en que se automatizará su procesamiento. Estos formatos deben estar adecuadamente documentados y aclarados.

Por ello se evita publicar en formatos no estructurados, como textos libres, archivos PDF, JPG o PNG. Hay que publicar al menos un formato legible y automatizable, como CSV, XML, JSON, RDF, etc.

Principios 6

No discriminatorios

Los datos están disponibles para cualquier persona, sin necesidad de registrarse en un portal o web, presentándolos mediante API públicas.

Principios 7

No propietarios

No se usan formatos propietarios para publicar los datos, evitando el pago por uso de aplicaciones específicas para explotar los datos. Se priorizan los siguientes formatos: CSV, XML, SVG, etc.

Principios 8

Libres de licencia

Los datos no están sujetos a ningún derecho de autor, patentes, marcas o regulación. Se permiten restricciones razonables de privacidad y seguridad.

CONTINUAR

En abril de 2020, había alrededor de 4000 fuentes de datos abiertas en todo el mundo. Entre ellas, por ejemplo, el Portal Europeo de Datos Abiertos.

La variedad de temas y temas cubiertos por los conjuntos de datos abiertos es la base para muchos escenarios de uso en el contexto empresarial. Por ejemplo, las estadísticas demográficas y económicas mejoran el marketing y los análisis de orientación al cliente.

Se pueden utilizar múltiples códigos y estándares (códigos HS, mercancía peligrosa, GTIN, códigos ISO de país, etc.) para enriquecer los datos de las empresas existentes.

Los datos de los registros corporativos oficiales pueden ayudar a mejorar la calidad de los datos de los socios comerciales al eliminar duplicados o agregar nuevas entradas.

¿Cómo se utilizan los datos abiertos en las empresas?

Aunque hay un número creciente de conjuntos de datos abiertos, las empresas rara vez los utilizan debido a la falta de transparencia, la calidad incierta y los diferentes formatos de las fuentes. Muchas empresas ni siquiera conocen los conjuntos de datos abiertos disponibles y su relevancia para los procesos o decisiones comerciales. Aquellos que lo hacen no están seguros de la calidad de los datos y de si cumplen con los estándares de su empresa. Además, la integración es difícil porque el acceso a datos abiertos, las licencias y las condiciones son muy diferentes.

Los datos de *open data* se suelen ofrecer por sus propietarios como ficheros descargables, como web services o API, e incluso como accesos a bases de datos.

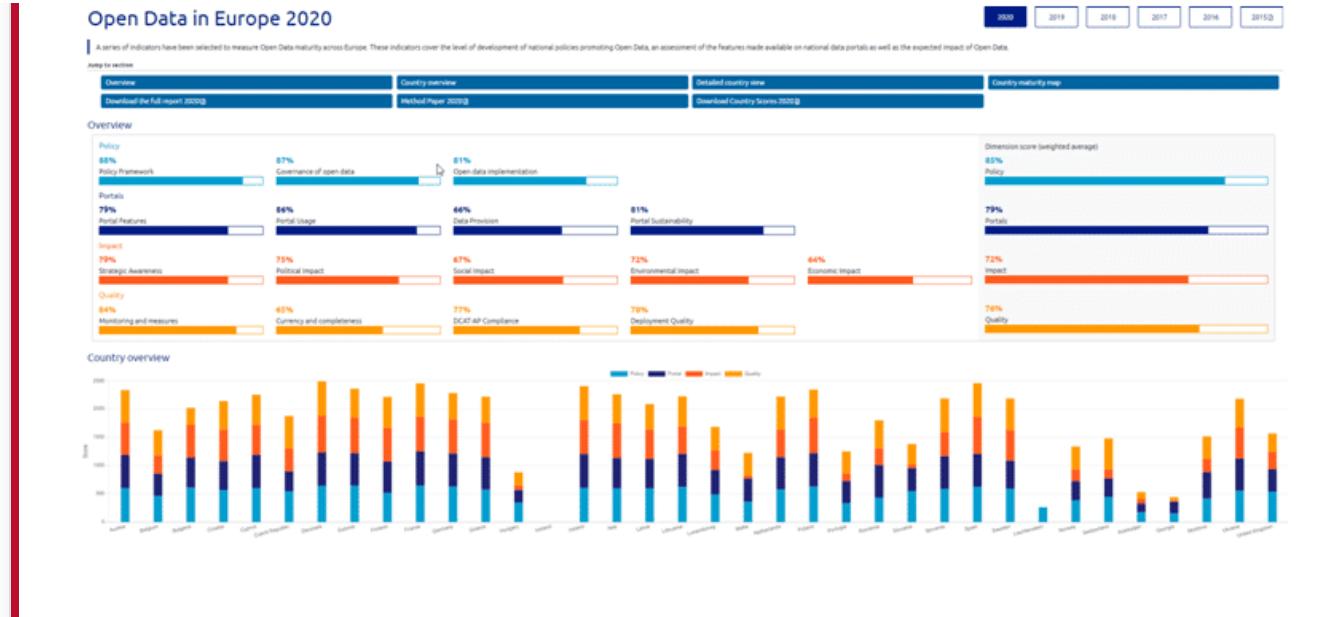
¿Cómo se encuentran los conjuntos de datos abiertos para empresas?

Para resolver estos problemas, el CC CDQ, junto con la Universidad de Lausana, lanzó un proyecto para proporcionar una “*app store* de datos abiertos” que ayude a las empresas a encontrar, integrar y utilizar datos abiertos. El proyecto de dos años está financiado por InnoSuisse, en estrecha colaboración con Nestlé, SBB y Swisscom. CDQ AG es el principal socio de implementación.

A continuación, se puede ver un informe de madurez respecto a los datos abiertos facilitado por el Portal Europeo de Datos Abiertos en el que evalúa, país por país, lo referente a cuatro conjuntos de métricas respecto a política, portales, impacto y calidad.

Figura 1. Informe de madurez de Open Data Europa.

Fuente: <https://data.europa.eu/en/dashboard/2020>



3.1.1. Clasificación según el grado y usabilidad de los datos abiertos

Tim Berners-Lee, miembro del World Wide Web Consortium (W3C), inventor de la World Wide Web, y, posteriormente, de los *linked data*, ha impulsado un esquema de desarrollo de cinco estrellas, utilizado de manera global, para medir en qué grado son abiertos y usables los datos que ofrece una institución.

1 estrella *

En este nivel los datos deben estar disponibles en la web, sea en el formato que sea, y con licencia abierta. Los formatos para cumplir este nivel suelen ser el PDF o formatos de imagen como JPG, PNG, etc.

Este nivel permite visualización, impresión y almacenamiento local. Permite también la ingestión de los datos en otro sistema, su modificación y la posibilidad de compartirlos. Pero, para estas tareas, se requiere de la creación de software para su extracción del documento o su copiado a mano.

2 estrellas **

Para conseguir las dos estrellas los datos estarán disponibles de forma estructurada, para que sean legibles por máquinas. Formatos tipo: XSL, DOC, MDB.

Al sistema destino estos formatos le permiten procesar directamente para realizar modificaciones, cálculos o visualizaciones, como, por ejemplo, gráficas, pero para ello es necesario el uso de software propietario.

3 estrellas ***

El nivel de tres estrellas es parecido al de dos, pero en este caso los formatos son abiertos, es decir, no propietarios. Entre ellos se encuentran XML, JSON o CSV.

Al ser formatos no propietarios, el consumidor podrá hacer todas las cosas que puede hacer con el nivel dos estrellas, pero sin la limitación que impone el uso de un software en concreto.

4 estrellas ****

El nivel de cuatro estrellas debe cumplir el nivel de tres estrellas y es necesario el uso de estándares abiertos de W3C (Consorcio World Wide Web). Para ello los datos deben ser identificados mediante una URI (identificador uniforme de recursos) y que así estén integrados en la web. Para este nivel una forma de representación de los datos es RDF.

Sin embargo, la estructura de los datos suele ser más difícil de entender.

Un editor invertirá en este nivel más tiempo y esfuerzo en el análisis de los datos, la preparación para su representación, la asignación de las URI y la búsqueda y creación de patrones para aplicarlos a la información. Pero tendrá un gran control sobre los datos para realizar optimizaciones a todos los niveles. Además, ofrece la posibilidad a otros editores de enlazar a sus datos de modo que sean promocionados al nivel cinco estrellas.

RDF (*resource description framework* o marco de descripción de recursos) permite la interoperabilidad entre aplicaciones que intercambian información comprensible por la página web, para proporcionar una infraestructura que soporte actividades de metadatos. RDF permite realizar consultas contra los datos utilizando varios lenguajes de consulta entre los que destaca SPARQL.

5 estrellas *****

Las cinco estrellas se consiguen logrando los niveles anteriores y vinculando, además, los datos con los que otras personas o instituciones publican, de modo que se proporcione un contexto para ellos. El formato para este nivel es el *linked* RDF.

CONTINUAR

3.1.2. API abiertas

Las API (application program interfaces) son interfaces con una serie de bibliotecas o paquetes de software preparados para que otro sistema o aplicación pueda llamarlos, utilizarlos y descargar información. Hoy día, son comúnmente utilizadas en cualquier aplicación.

Las API se usan en gran variedad de sectores para permitir la integración de conjuntos de datos o servicios en aplicaciones o webs. Se usan para conectar diversos sistemas software, aplicaciones y programas que mantienen la actividad operacional del día a día en funcionamiento.

En pocas palabras, las API son un medio de comunicación entre sistemas. Por ejemplo, se usan para sincronizar distintas aplicaciones que realizan funciones similares dentro de una organización multinacional. O permiten el intercambio de información entre sistemas o aplicaciones.

Las API pueden ser cerradas o abiertas.

Cerradas

Cerrada significa que la empresa propietaria solicita el pago por la divulgación de los datos que provee la API.

Abiertas

Sin embargo, muchas compañías, en la actualidad, ofrecen información de forma abierta. Esto significa que los desarrolladores pueden

acceder libremente para

Ejemplo

Imagínese, por ejemplo, que se pretende presentar información meteorológica en una aplicación móvil. En este caso se puede recurrir a una API para conectarse a un servicio web que ofrezca dicha información, en vez de programarlo todo desde cero, ahorrando mucho en tiempo y costes.

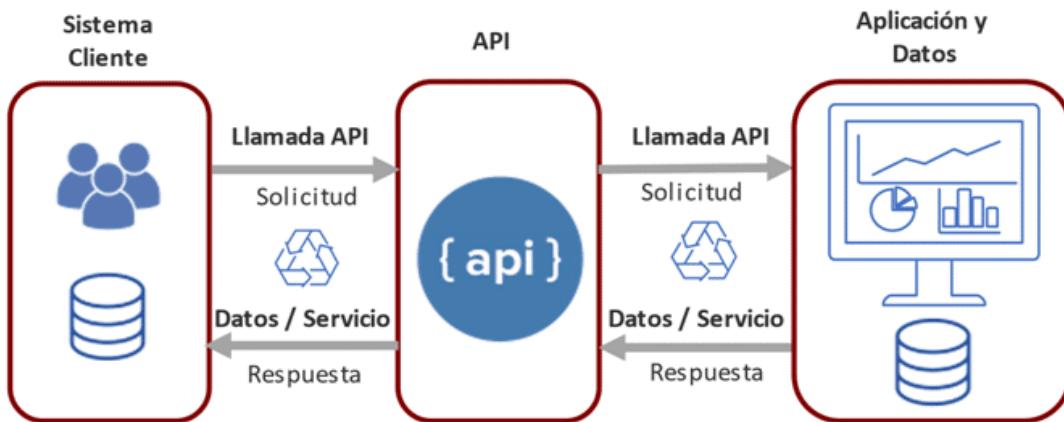


Figura 2. Estructura de API.

Fuente: elaboración propia.

- i** Tanto en el ejemplo como en el caso práctico se verá cómo acceder a API abiertas de la NASA para consultar en tiempo real los datos que proveen. Estos datos se importarán usando PDI y la máquina virtual a una base de datos.

3.2. Datos de pago

Los datos pagados son datos disponibles comercialmente, adquiridos directamente de proveedores de datos especializados (o intermediarios) y mercados de datos, y se ofrecen a un cierto coste.

Uno de los proveedores típicos de datos pagados es Dun & Bradstreet (D&B), con sus soluciones de datos maestros D&B, que ofrecen datos comerciales, análisis e información para las empresas. Otros proveedores populares son Nielsen, para datos de investigación de mercado, o Reuters, para datos financieros.

En ciertos sectores en los que participan ciertos intermediarios independientes de la organización, estos pueden vender sus datos a las empresas, siempre cumpliendo la legislación. Pero aportando datos más específicos de ventas, canales de distribución e incluso datos de la competencia.

Dado que los datos externos pagados se proporcionan a un coste, lo mejor para el proveedor es entregar los datos en alta calidad y proporcionar una descripción exhaustiva.

Por lo general, los datos pagados se proporcionan como información estructurada o incluso se entregan con el conocimiento extraído directamente al usuario final.

CONTINUAR

3.3. Datos compartidos

Este tipo de datos externos se refiere a los datos que se comparten entre empresas dentro de ecosistemas empresariales (por ejemplo, dentro de la comunidad de intercambio de datos de CDQ o plataformas industriales, como **Skywise** o **GDSN**).

Dentro de un entorno protegido, las empresas pueden compartir sus datos internos con sus socios comerciales y beneficiarse de los esfuerzos de la comunidad en términos de mantenimiento, integridad y actualización. Ejemplos de entornos de intercambio e intercambio incluyen:

CDQ Data Sharing Community —

En la que las empresas multinacionales líderes no solo enriquecen sus propios datos con fuentes públicas, sino que también validan sus datos con los registros de sus socios comerciales. Este concepto innovador va más allá del simple intercambio de datos, también ofrece la oportunidad de compartir el conocimiento común, como las reglas comerciales.

Red global de sincronización de datos (GDSN) —

Proporcionada por GS1 como grupo de datos global para las industrias minorista y de bienes de consumo.

Skywise —

Una plataforma de datos, iniciada por **Airbus y Palantir Technologies**, que conecta la cadena de valor de la aviación e incluye más de cien aerolíneas en todo el mundo, así como proveedores.

Algunos ejemplos de cómo Skywise podría ayudar a las aerolíneas a definir y mejorar sus modelos comerciales usando datos de Airbus incluyen:

- Mayor fiabilidad operativa de la flota mediante mantenimiento predictivo y preventivo.
- Eficiencia operativa mejorada para flotas heredadas.
- Análisis rápidos de la causa raíz de los problemas en servicio.
- Optimización del rendimiento de cada aeronave a través del análisis de datos de operaciones de vuelo.
- Seguimiento de la eficacia del mantenimiento a lo largo del tiempo.
- Flujos de trabajo de informes con un solo clic, que incluyen informes complejos a los organismos reguladores.



Figura 3. Ejemplos del beneficio compartir datos.

Fuente: [Skywise](#).

3.4. Datos de redes sociales

Los datos de redes sociales se refieren a los datos compartidos por los usuarios de las plataformas de redes sociales, incluidos los metadatos (por ejemplo, ubicación, hora, idioma, datos biográficos). En resumen, los datos de las redes sociales son la **información recopilada** de estas que muestra cómo los usuarios comparten, ven o interactúan con su contenido o perfiles.

En el apogeo de las redes sociales, los especialistas en marketing estaban obsesionados con métricas de vanidad, como el número de seguidores y los me gusta. Este tipo de datos pueden parecer impresionantes a simple vista, pero significan muy poco por sí mismos.

¿Qué significan miles de seguidores si no se traducen en ingresos? ¿Cuál es el compromiso de los seguidores con el contenido publicado?

Un enfoque distinto de los datos de las redes sociales permite realizar otro tipo de análisis. Por ejemplo, si una publicación alcanza una popularidad casi viral, sería entonces conveniente publicar contenidos similares. Otro ejemplo es el uso de los *hashtags* si se determina que alguno de ellos genera mayor número de clics.

La disponibilidad de estos datos varía significativamente de una plataforma de redes sociales (por ejemplo, Facebook, LinkedIn o Twitter) a otra, principalmente debido al formato no estructurado o semiestructurado.

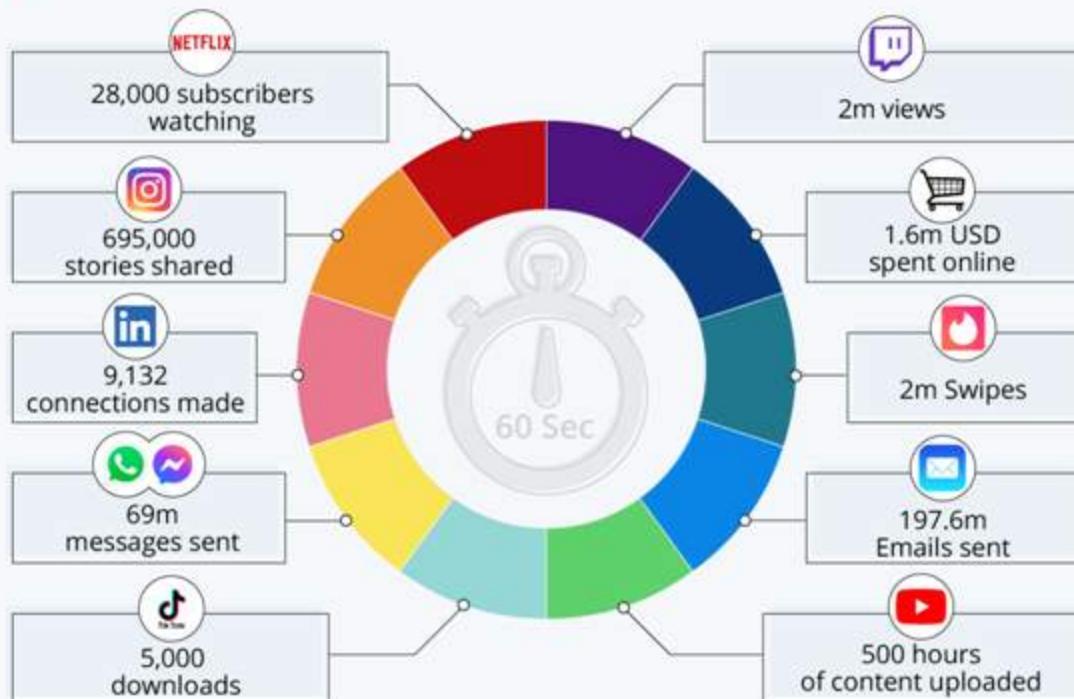
Desde una perspectiva empresarial, las redes sociales son una fuente de información de valor incalculable. Por ejemplo, la publicidad personalizada puede basarse en publicaciones anteriores en redes sociales y en las preferencias de los usuarios.

Los *hashtags*, que se utilizan ampliamente en múltiples plataformas de redes sociales, permiten analizar tendencias y son una fuente importante de análisis de mercado.

- i La información de contacto proporcionada por los usuarios de las plataformas de redes sociales puede considerarse como una fuente de información actualizada para validar los registros internos de la empresa.

A Minute on the Internet in 2021

Estimated amount of data created on the internet in one minute



Source: Lori Lewis via AllAccess



statista

Figura 4. Un minuto en internet.

Fuente: Lewis, L. vía Statista.

Uno de los mayores desafíos es la extracción de información de las plataformas de redes sociales. Por la naturaleza del contenido, que aparece muy rápido, es necesario un software de monitoreo o *web scrapping* para asegurar el flujo constante de la información. Aunque no existen tarifas directas aplicables a la extracción de datos de las plataformas de redes sociales, los esfuerzos de procesamiento pueden ser costosos.

WEB SCRAPPING

WEB CRAWLING

Los procesos de *web scrapping* se pueden aplicar a todo tipo de webs, por lo tanto, muchas empresas aplican técnicas de *web scrapping* a su propia competencia para obtener datos de política de precios en cada momento, por ejemplo.

WEB SCRAPPING

WEB CRAWLING

También se cuenta con técnicas de *web crawling*, rastreador web o araña web que rastrea un sitio web en su totalidad, si este está autorizado por los archivos “robots.txt”. Así, se puede conseguir el “sitemap.xml” de un sitio, y saber cuáles son las URL públicas de las que se compone, para futura obtención de datos.

Con el fichero “sitemap.xml” se puede saber toda la información de un sitio web. Existen librerías en Python y otros lenguajes que permiten mejorar estas gestiones.

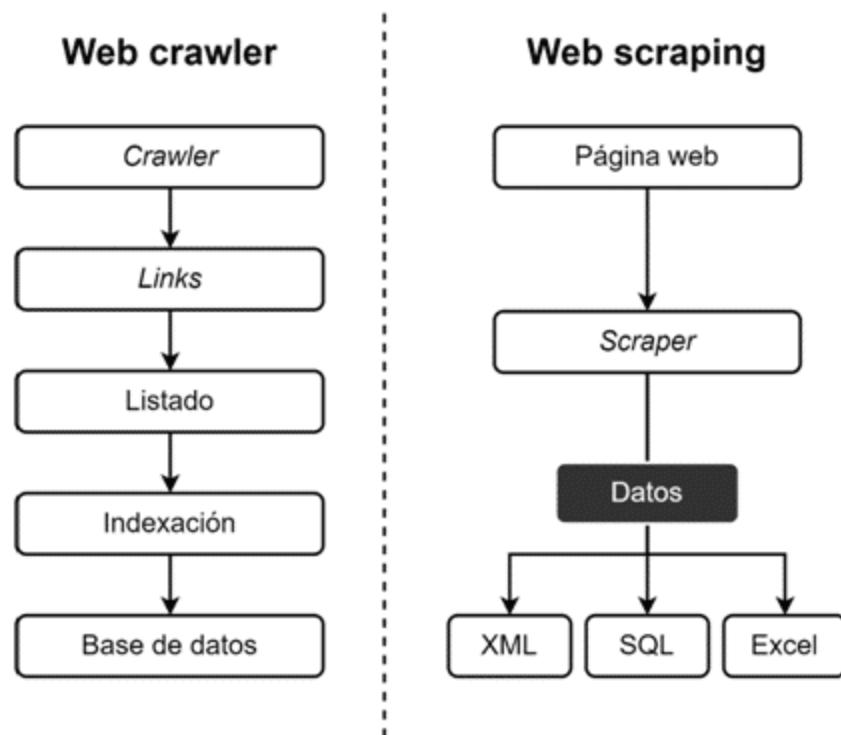


Figura 5. Web crawler y web scraping.

Fuente: elaboración propia

CONTINUAR

3.5. ¿Cuáles son los desafíos típicos del abastecimiento y la gestión de datos externos?

Según estudios recientes, en la mayoría de las empresas:

- No coordinan el suministro de datos externos. Y no tienen un proceso formal de abastecimiento de datos externos.

- No aprovechan el potencial de las fuentes de datos externas: no tienen la función de “cazadores de datos”, y se especializan en descubrir nuevas fuentes de datos alternativas.
- Es un desafío identificar, integrar y mantener fuentes de datos externas dada la heterogeneidad y la falta de transparencia (sobre la calidad de la semántica, los ciclos de actualización, etc.)

Normalmente, las empresas que eligen explotar información de terceros optan por el uso de fuentes de datos de pago, por su fiabilidad y por su estabilidad. Solo aquellas organizaciones más avanzadas analíticamente incorporan datos extraídos de redes sociales o *web scrapping* a sus fuentes de datos fiables. Por último, el mundo del *open data* genera ciertas incertidumbres en cuanto a fiabilidad de los datos y estabilidad del servicio.

3.6. Cómo utilizar datos externos: ejemplos de uso de datos externos en empresas

Los datos externos también pueden ser útiles en las siguientes situaciones:

Proporcionar conocimientos basados en datos

El análisis de datos se puede mejorar con datos externos en áreas operativas, como la gestión de relaciones con los clientes, recursos humanos, cadena de suministro y almacenamiento. Por ejemplo, las grandes ciudades pueden pronosticar el estado futuro del tráfico y la contaminación con la ayuda de datos externos mediante datos meteorológicos, datos de los servicios públicos de transportes, carreteras, etc.

Mejora de los procesos comerciales

Muchas empresas ya utilizan datos de geolocalización, meteorológicos y de tráfico para planificar y gestionar sus entregas; información adicional sobre eventos excepcionales, como desastres, puede ayudar

a evitar interrupciones en la cadena de suministro. En el mundo de hoy, donde gran parte de la cadena de valor de las empresas está subcontratada, estas empresas subcontratadas pueden vender o compartir sus datos, para que la organización pueda mejorar en procesos comerciales y operativos.

Mejora de las capacidades de gestión de datos

La obtención de datos externos reduce los esfuerzos de mantenimiento de datos. También se puede utilizar para enriquecer los datos internos y mejorar la calidad de los datos.

Habilitación de nuevos servicios

Los datos externos también se utilizan para innovar e introducir nuevos productos y servicios que se adapten a las necesidades de los consumidores.

IV. Demo

En esta unidad se han mostrado ejemplos de algunos casos de aplicación de datos externos. En la demo se realizará la integración de datos externos abiertos en el sistema de prácticas. Para ello se va a integrar una fuente de datos abierta proporcionada por la NASA.

El **portal de API de la NASA** permite que los datos de la NASA, incluidas imágenes, sean accesibles para los desarrolladores de aplicaciones. El catálogo es muy grande y crece constantemente: <https://api.nasa.gov/>

No es necesario autenticarse para explorar los datos de la NASA. Sin embargo, si se van a utilizar intensamente las API para una aplicación móvil, por ejemplo, es preciso registrarse para obtener una clave de desarrollador de la NASA.

Los servicios web establecen unos límites de cantidad de solicitudes para las API. Los límites de tarifas pueden variar según el servicio.

En este caso se usará la **DEMO_KEY**. Esta clave de API se puede usar para explorar las API antes de registrarse, pero tiene límites mucho más bajos, por lo que se recomienda registrarse para obtener una propia clave de API si se planea hacer un uso extensivo. Los límites de tarifa para DEMO_KEY son los siguientes:

- Límite por hora: 30 solicitudes por dirección IP.
- Límite diario: 50 solicitudes por dirección IP.

4.1. Data set

Para la demo se va a usar una API de la NASA que devuelve datos en formato JSON. En este caso se verá un conjunto de datos del área 'SSD/CNEOS: Dinámica del sistema solar y Centro de estudios de objetos cercanos a la Tierra'.



Figura 6. CNEOS Apophis.
Fuente: NASA / JPL-Caltech.

Este servicio proporciona una interfaz en formato JSON, relacionada con SSD y CNEOS.

Esta API consta de varios componentes. En la demo se usará la API “CAD Asteroide y cometa cercanos se acercan a los planetas en el pasado y el futuro”.

Esta API proporciona acceso a los datos actuales de aproximación cercana para todos los asteroides y cometas en **SBDB** (*small-body database*) de **JPL**. El resultado devuelve unos valores predeterminados según unos parámetros fijos para la consulta, en este caso:

- **NEO Earth** se acerca a menos de 0,05 au (distancia lunar).
- En los próximos 60 días.
- Ordenados por fecha.

Esta API invoca el servicio GET mediante [este enlace](#).

Esta API es parametrizable, como se puede ver en su [web oficial](#).

Ejemplo

Por ejemplo, si se quieren ver datos de aproximación cercana a la Tierra para NEO dentro de cinco distancias lunares a partir del 2020-Jan-01, ordenados por distancia, se usarán los siguientes parámetros:

- dist-max=5.
- date-min=2020-01-01.
- sort=dist.
- Diameter=1.

Generando esta URL, <https://ssd-api.jpl.nasa.gov/cad.api?dist-max=5&date-min=2020-01-01&diameter=1&sort=dist>, con el siguiente resultado:

Figura 7. SSD/CNEOS API

Fuente: NASA

El resultado en formato JSON se compone de los siguientes datos

Las solicitudes de consulta correctas dan como resultado una carga útil de datos en formato JSON. El contenido específico depende del modo de consulta. Si una búsqueda es demasiado restrictiva, es posible obtener un resultado de conteo cero.

1

Cada registro CAD se empaqueta como una matriz de campos (correspondientes a los enumerados) en el siguiente orden:

- *des*: designación principal del asteroide o cometa (por ejemplo, 443, 2000 SG344).
- *orbit_id*: ID de órbita.
- *jd*: tiempo de aproximación cercana (JD Ephemeris Time, TDB).
- *cd*: hora de aproximación cercana (fecha/hora del calendario formateado, TDB).
- *dist*: distancia de aproximación nominal (au).
- *dist_min*: distancia de aproximación mínima (3-sigma) (au).
- *dist_max*: distancia de aproximación máxima (3-sigma) (au).
- *v_rel*: velocidad relativa al cuerpo de aproximación en aproximación cercana (km/s).
- *v_inf*: velocidad relativa a un cuerpo sin masa (km/s).
- *t_sigma_f*: incertidumbre 3-sigma en el tiempo de aproximación cercana (formateado en días, horas y minutos; los días no se incluyen si son cero; el ejemplo "13:02" es 13 horas, 2 minutos; el ejemplo "2_09: 08" es 2 días, 9 horas, 8 minutos).
- *body*: nombre del cuerpo de aproximación cercana (por ejemplo, la Tierra).
 - Solo salida si los parámetros de consulta del cuerpo están configurados en TODOS.
- *h*: magnitud absoluta H (mag).
- *diámetro*: diámetro del cuerpo (km).
 - Opcional: solo salida si se solicita con el parámetro de consulta de diámetro.
 - Nulo si no se conoce.
- *diámetro_sigma*: incertidumbre 1-sigma en el diámetro del cuerpo (km).
 - Opcional: solo salida si se solicita con el parámetro de consulta de diámetro.
 - Nulo si no se conoce.

CONTINUAR

4.2. Carga en PDI

En este módulo se usará PDI para extraer la información desde las API a un fichero o tabla de base de datos.

Para ello se utilizará PDI y se creará una transformación que realizará una llamada a la API de datos abiertos de la NASA seleccionada y transformará el fichero JSON en datos estructurados.

En **primer lugar**, se realizará la **llamada al servicio Rest**, que devuelve los datos en formato JSON. Para ello se usarán dos objetos:

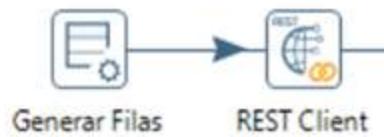


Figura 8. Llamada Rest API.

Fuente: elaboración propia.

Generador de filas

Actúa como *trigger* para el siguiente paso, si no, no se ejecuta. En este solo se especifica la URL que proporciona los datos.

Figura 9. Generador de fila.

Fuente: elaboración propia.

Generar Filas

Nombre paso	Generar Filas
Límite	1
Never stop generating rows	<input type="checkbox"/>
Interval in ms (delay)	5000
Current row time field name	now
Previous row time field name	FiveSecondsAgo

Campos :

#	Nombre	Tipo	Formato	Longitud	Precision	Moneda	Decimal	Grupo	Valor
1	url	String							https://ssd-api.jpl.nasa.gov/cad.api?dist-max=

Buttons: Help, Vale, Previsualizar, Cancelar

Cliente rest

Este objeto realiza la llamada a la URL y descarga los datos.

Figura 10. Cliente rest API.

Fuente: elaboración propia.

REST Client

Step name: REST Client

General Authentication SSL Headers Parameters Matrix Parameters

Settings

URL	https://ssd-api.jpl.nasa.gov/cad.api?dist-max=5LD&date-min=2020-01-
Accept URL from field?	<input checked="" type="checkbox"/>
URL field name	url
HTTP method	GET
Get Method from field	<input type="checkbox"/>
Method field name	
Body field	
Application type	JSON

Output fields

Result field name	result
HTTP status code field name	
Response time (milliseconds) field name	
Response header field name	

Buttons: Help, Vale, Cancelar

A partir de aquí se realiza el **procesado de los datos en formato JSON**. Para ello se necesita un conversor de datos en formato JSON. Y luego, dada la estructura en la que se encuentra ese fichero JSON, se realizan un reemplazo de caracteres y una separación de campos.

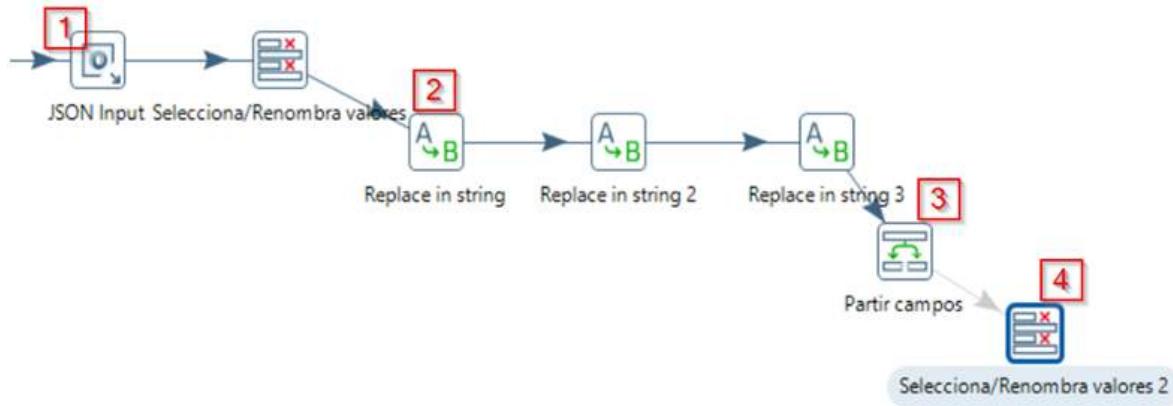


Figura 11. Segunda parte.

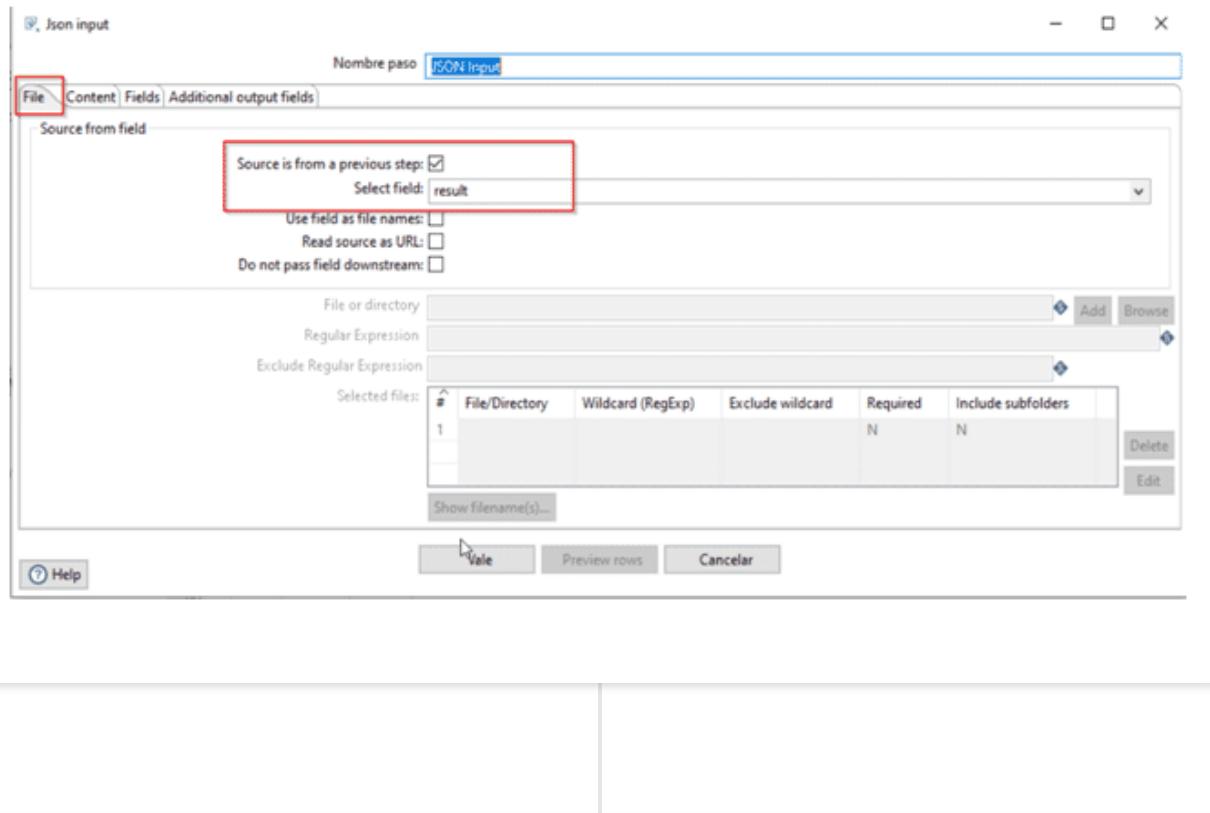
Fuente: elaboración propia.

1) Json input: parsea el formato JSON

Primero se configura la pestaña de fichero, indicando la entrada.

Figura 12. File JSON.

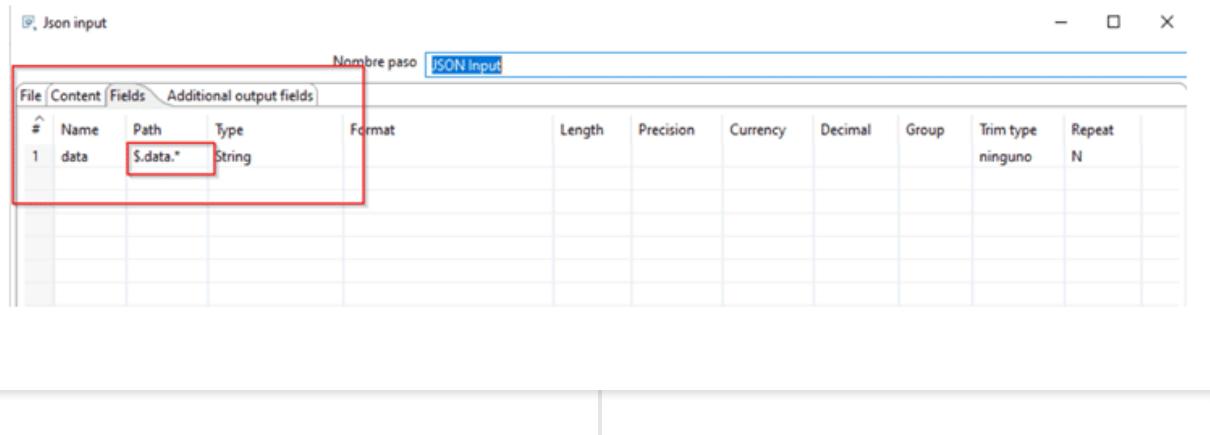
Fuente: elaboración propia.



Después se configura la pestaña de campos, donde se especifica la ruta de los datos dentro del fichero JSON. Esta sección depende de la composición del fichero JSON.

Figura 13. Fields JSON.

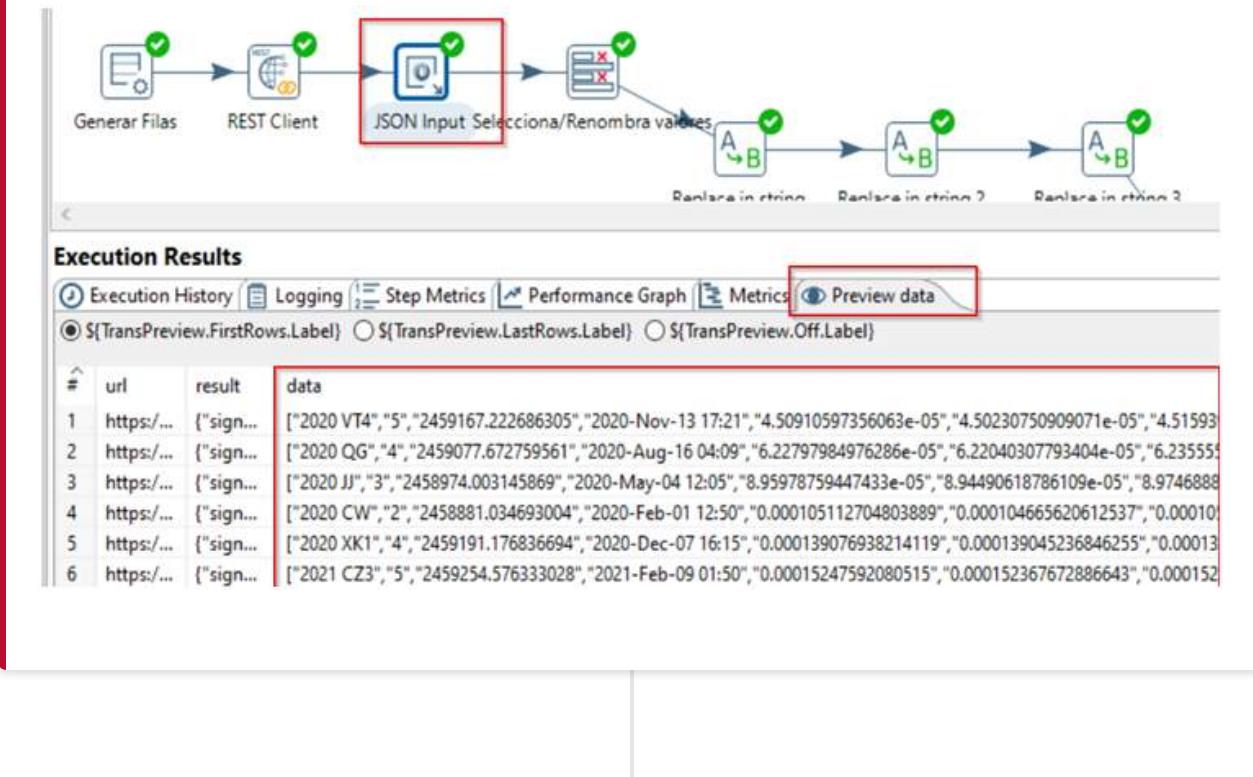
Fuente: elaboración propia.



Si se ejecuta el paquete de carga hasta este punto se verá que ya se dispone de los datos, pero en **una única columna** con corchetes y los campos separados por comas:

Figura 14. Preview.

Fuente: elaboración propia.



2) Se reemplazan caracteres del resultado que sobran, como [,] , “ ”

Figura 15. Replace.

Fuente: elaboración propia.

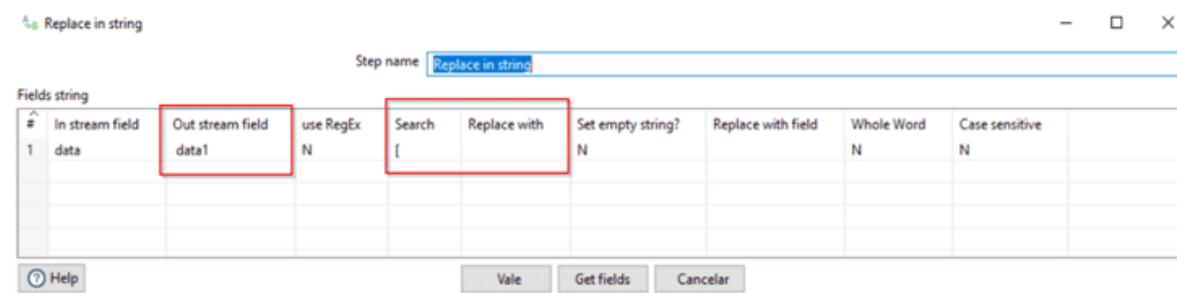
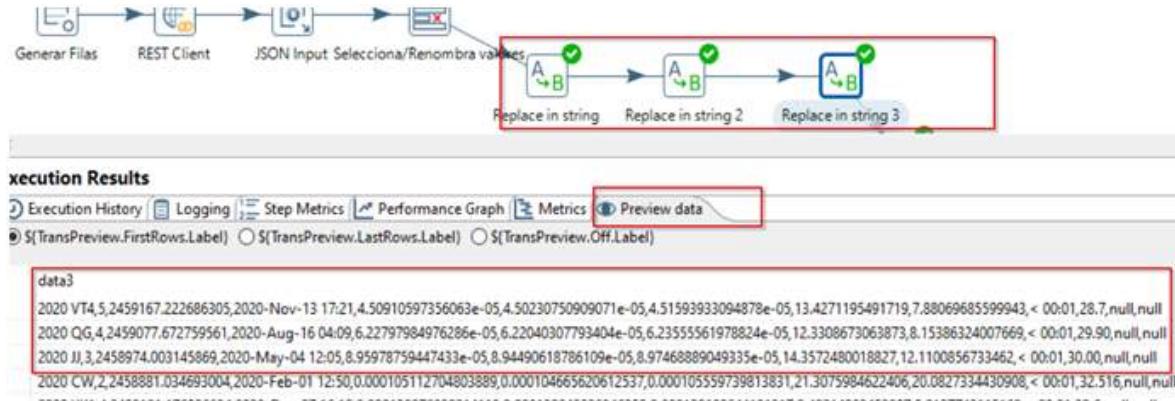


Figura 16. Resultado de *replace*.

Fuente: elaboración propia.



3) Se separan los campos por un separador dado

Figura 17. Separador por columnas.

Fuente: elaboración propia.

Partir campo

Nombre paso: Partir campo

Campo a partir: data3

Separador: ,

Enclosure:

Campos

#	Nuevo campo	ID	Eliminar ID?	Tipo	Longitud	Precisión	Formato	Grupo	Decimal	Moneda	NuloSi	Default	Trim type
1	Desc	N		String							ninguno	ninguno	ninguno
2	Orbit_id	N		String							ninguno	ninguno	ninguno
3	Jd	N		String							ninguno	ninguno	ninguno
4	Cd	N		String							ninguno	ninguno	ninguno
5	Dist	N		String							ninguno	ninguno	ninguno
6	dist_min	N		String							ninguno	ninguno	ninguno
7	dist_max	N		String							ninguno	ninguno	ninguno
8	v_rel	N		String							ninguno	ninguno	ninguno
9	v_inf	N		String							ninguno	ninguno	ninguno
10	t_sigma_f	N		String							ninguno	ninguno	ninguno
11	h	N		String							ninguno	ninguno	ninguno
12	diameter	N		String							ninguno	ninguno	ninguno
13	diameter_sigma	N		String							ninguno	ninguno	ninguno

4) Se seleccionan solo los campos que se quieren de salida

Primero, se selecciona la opción para traer todos los campos, y luego se filtran solo aquellos que se quieren mostrar.

Figura 18. Seleccionar columnas.

Fuente: elaboración propia.

Selección/Renombrar valores

Nombre paso: Selección/Renombrar valores

Selección & Modifica | Eliminar | Meta-information

Campos:

#	Nombre campo	Renombrar	Longitud	Precisión
1	Desc			
2	Orbit_id	2		
3	Jd			
4	Cd			
5	Dist			
6	dist_min			
7	dist_max			
8	v_rel			
9	v_inf			
10	t_sigma_f			
11	h			
12	diameter			
13	diameter_sigma			

1 Dibujar campos a seleccionar

Edit Mapping

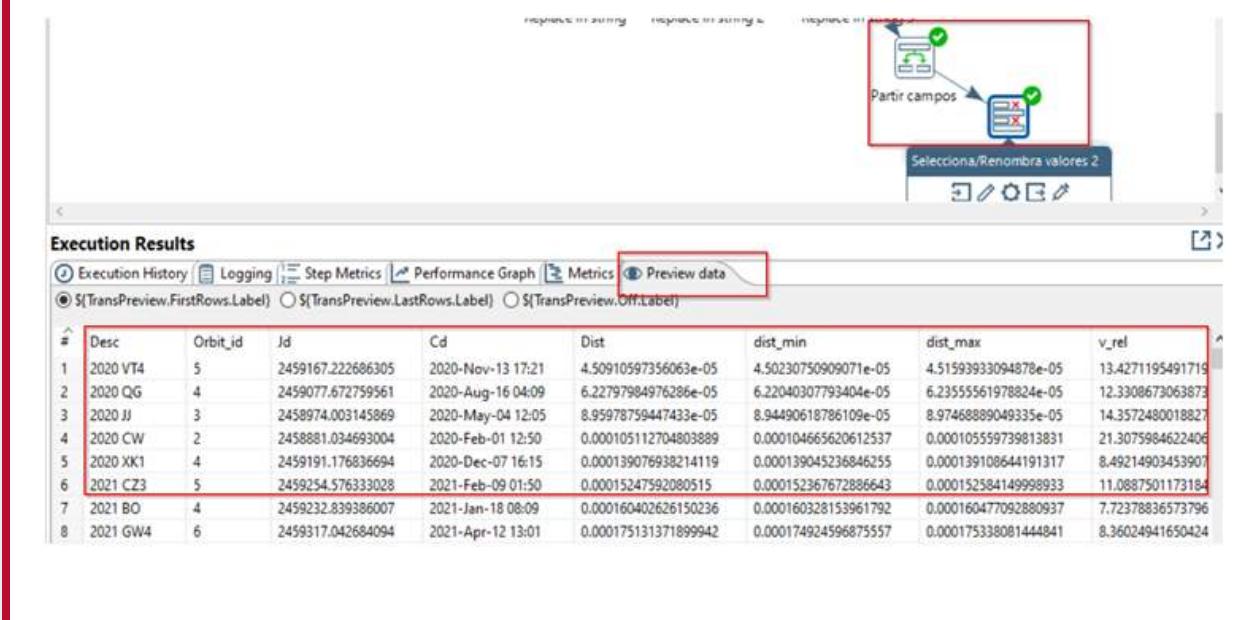
Include unspecified fields, ordered by name

Vale Cancelar

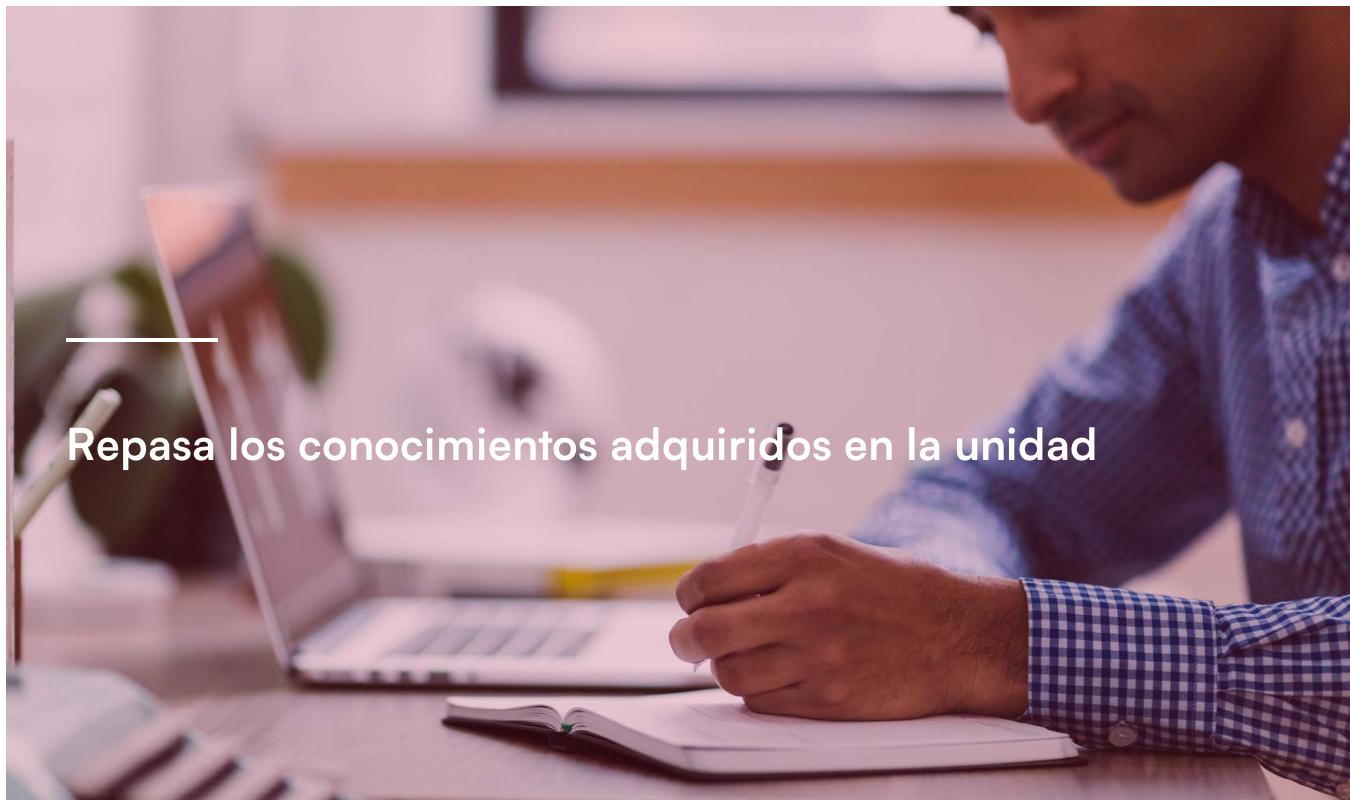
Help

Figura 19. Resultado final.

Fuente: elaboración propia.



V. Resumen



Repasa los conocimientos adquiridos en la unidad

En esta unidad se han mostrado ejemplos del uso de datos externos a la compañía. Estos datos externos a la compañía vienen de diversas fuentes. Desde **datos abiertos** de Gobiernos y empresas gubernamentales, los cuales son gratuitos pueden ser explotados por cualquier persona, pasando por **datos de pago**, que se venden entre compañías colaboradoras o empresas especializadas para enriquecer los datos de las organizaciones, y **datos compartidos** entre empresas para mejorar los

procesos operativos y mejorar el rendimiento de empresas colaboradoras entre sí, hasta **datos de redes sociales** y captados mediante *web scraping*.

Los datos abiertos están creciendo cada vez más y ya son un requisito indispensable para cualquier Gobierno local o nacional.

El **open data** se engloba dentro del marco de las *smart cities*. Mediante los datos abiertos se puede acceder, compartir y utilizar de forma gratuita datos que permiten conectarse e interactuar mejor con las ciudades. Las aplicaciones incluyen **horarios de autobuses en tiempo real, información sobre viviendas sociales, iniciativas de cuidado, grupos de juego y contratos públicos, etc.**

Los datos de pago son muy comunes hoy en día entre empresas que se subcontratan servicios, por ejemplo. Y también se solicitan datos especializados de marketing a empresas profesionales de estos servicios, como **Nielsen**.

Los datos compartidos son muy comunes en empresas con contratos a largo plazo y de gran coste. También en empresas cuya criticidad es mayor, como, por ejemplo, en el mundo de la aviación. **Airbus**, en este caso, comparte datos con todas las empresas que operan con aviones de su compañía; así, se avanza en mantenimiento predictivo y mejoras constantes para que sus aviones sean más seguros y pierdan las mínimas horas de servicio.

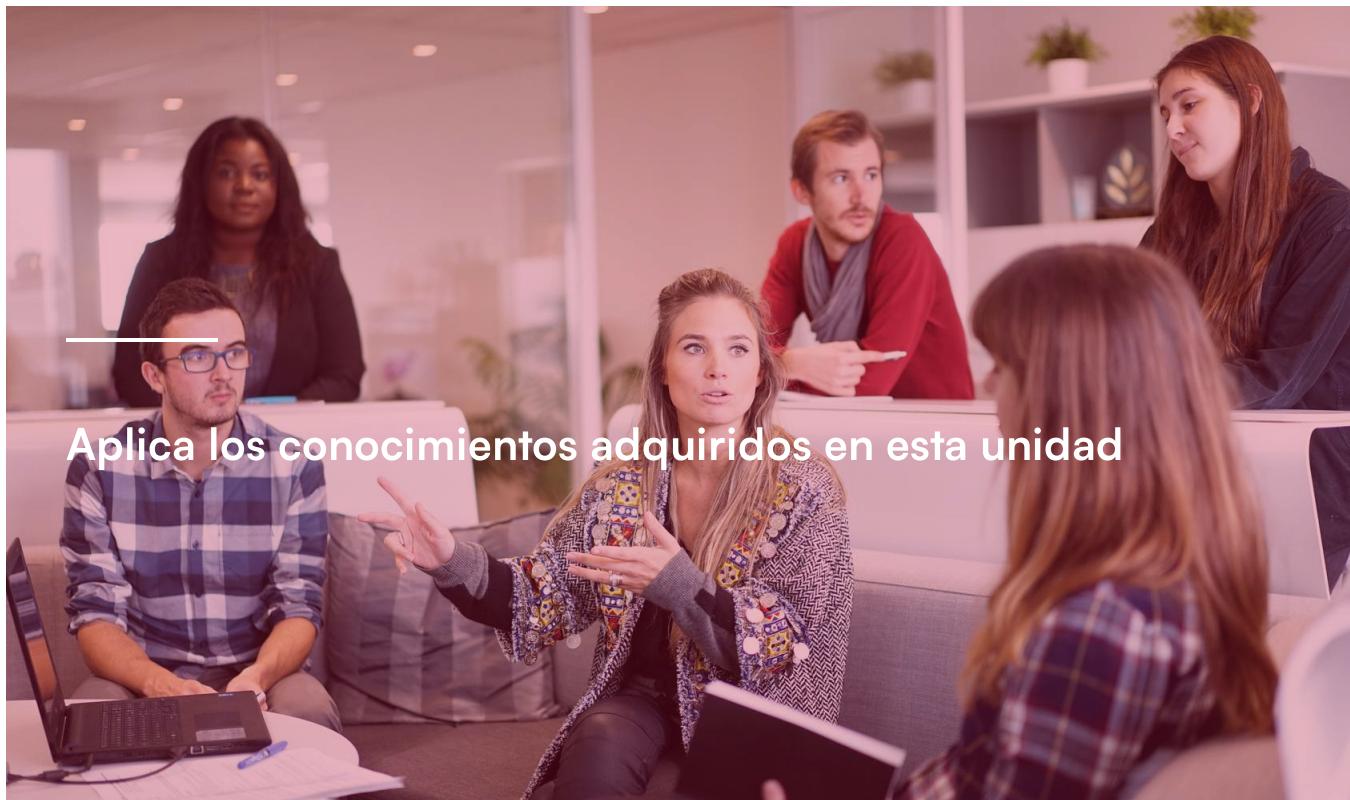
Por último, los datos de **redes sociales** y webs mediante técnicas de **web scraping** permiten al mundo del marketing cuantificar campañas y reacciones recibidas. La integración de estos datos varía según la red social que se esté analizando. Y las técnicas de *web scraping* funcionan siempre y cuando no varíe la estructura de la web que se está estudiando.

En conclusión, muchas empresas no usan datos externos ya que no se coordina el suministro de datos externos. Y no tienen un proceso formal de abastecimiento de datos externos. Supone un desafío identificar, integrar y mantener fuentes de datos externas dadas la heterogeneidad y la falta de transparencia sobre estos.

En el caso práctico y en la demo se ve cómo se puede integrar una fuente de datos abiertos usando PDI. En este caso se usa una fuente de datos provista por la NASA, la cual, mediante *web services*, proporciona acceso a los datos actuales de aproximación cercana para todos los asteroides y cometas.

Usando PDI, se automatizarán estos datos para transformarlos a un formato relacional que permita analizar la información de forma más sencilla.

VI. Caso práctico con solución



DATOS

Haciendo uso de Pentaho Data Integrator, se va a automatizar la integración de un conjunto de datos.

Como anteriormente en la demo, se van a usar datos abiertos de la Nasa: [NASA Open APIs](#).

En este caso se quiere automatizar los datos de la API de datos de *fireballs*. Esta proporciona un método para solicitar registros específicos del conjunto de datos disponible. Cada consulta exitosa devolverá

contenido que representa uno o más registros de datos de bola de fuego. [Fireball Data API \(nasa.gov\)](#).

Para ello se realizará una petición de los últimos 50 sucesos y se cargará de forma automática y actualizando los datos con PDI.

Nota: si la API estuviera fuera de servicio temporalmente, se deja un fichero JSON en la carpeta de recursos de la máquina virtual.

SE PIDE

Realizar un análisis del conjunto de datos y realizar el proceso de carga con PDI.

[VER SOLUCIÓN](#)

SOLUCIÓN

1. Data set

La API de datos de bola de fuego proporciona un método para solicitar registros específicos del conjunto de datos disponible. Cada consulta exitosa devolverá contenido que representa uno o más registros de datos de bola de fuego. Consultese la página de CNEOS sobre bolas de fuego para obtener detalles sobre este conjunto de datos.

Esta API consta de los siguientes parámetros:

- *date-min*: excluir datos anteriores a esta fecha AAAA-MM-DD o fecha/hora AAAA-MM-DDThh:mm:ss.
- *date-max*: excluir datos posteriores a esta fecha AAAA-MM-DD o fecha/hora AAAA-MM-DDThh:mm:ss.
- *energy-min*: excluir los datos con energía radiada total menor que este valor positivo en julios × 1010 (por ejemplo, 0,3 = 0,3 × 1010 julios).
- *energy-max*: excluir datos con energía total radiada mayor que esto (ver energía-min).
- *impact-e-min*: excluir datos con energía de impacto estimada menor que este valor positivo en kilotonnes (kt) (por ejemplo, 0,08 kt).
- *impact-e-max*: excluir datos con energía total radiada mayor que esto (ver impacto-e-min).
- *vel-min*: excluir datos con velocidad en el pico de brillo menor que este valor positivo en km/s (p. ej., 18,5).
- *vel-max*: excluir datos con una velocidad en el pico de brillo mayor que este valor positivo en km/s (por ejemplo, 20).
- *alt-min*: excluir datos de objetos con una altitud menor que esta (por ejemplo, 22 significa objetos más pequeños que esto).
- *alt-max*: excluir datos de objetos con una altitud mayor que esta (por ejemplo, 17,75 significa objetos más grandes que esto).
- *req-loc*: ubicación (latitud y longitud) requerida; cuando se establece en verdadero, excluye los datos sin una ubicación.
- *req-alt*: altitud requerida; cuando se establece en verdadero, excluye los datos sin una altitud.

- *req-vel*: velocidad requerida; cuando se establece en verdadero, excluye los datos sin una velocidad.
- *req-vel-comp*: componentes de velocidad requeridos; cuando se establece en verdadero, excluye los datos sin componentes de velocidad.
- *vel-comp*: incluir componentes de velocidad.
- *sort*: ordenar datos en el campo especificado: "fecha", "energía", "impacto-e", "vel" o "alt" (el orden de clasificación predeterminado es ascendente: anteponer "-" para descender).
- *limit*: limitar los datos a los primeros N resultados (donde N es el número especificado y debe ser un valor entero mayor que cero).

Generando la siguiente URL, para limitar los últimos 50 eventos con el siguiente resultado: <https://ssd-api.jpl.nasa.gov/fireball.api?limit=50>

```
{
  "signature": {"source": "NASA/JPL Fireball Data API", "version": "1.0"}, 
  "count": 0
}
```

Los datos se devuelven en formato JSON como un solo objeto (tabla hash). El número de registros contenidos en el objeto se indica en la tecla "contar". En los casos en los que las restricciones especificadas por el usuario son demasiado estrictas (sin resultados coincidentes), solo se devuelven las claves de "recuento" y "firma", como en el siguiente ejemplo.

```
{"count": 0, "signature": {"version": "1.0", "source": "NASA / JPL Fireball Data API"}}
```

Cada registro se proporciona como un elemento del objeto "datos" y cada registro es una matriz de campos. Los nombres de cada campo contenido en cada registro se proporcionan en la matriz de

objetos “campos”.

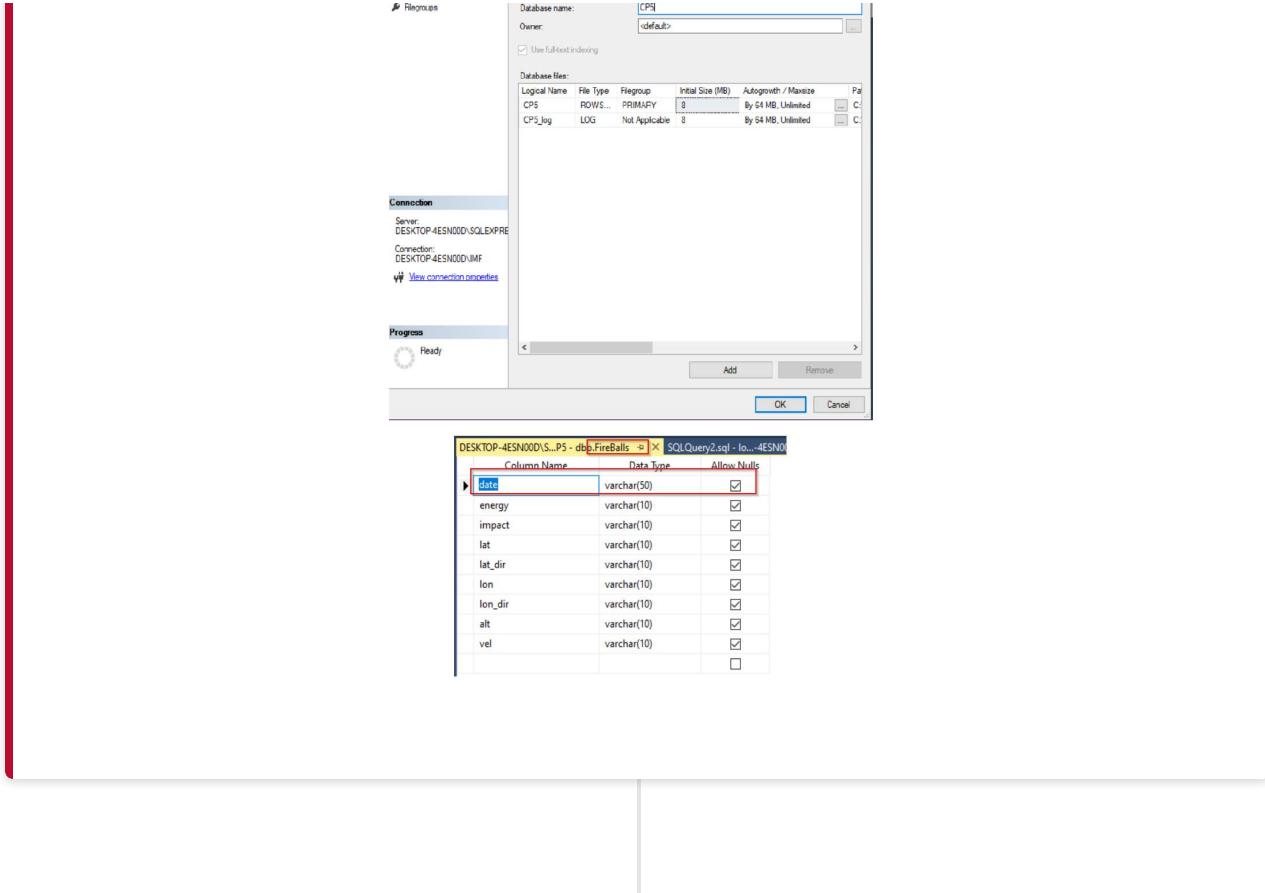
Los campos se definen de la siguiente manera:

- *date*: fecha/hora de brillo máximo (GMT).
- *lat*: latitud en el brillo máximo (grados).
- *lon*: longitud en el brillo máximo (grados).
- *lat-dir*: dirección de latitud (“N” o “S”).
- *lon-dir*: dirección de latitud (“E” o “W”).
- *alt*: altitud sobre el geoide en el brillo máximo (km).
- *vel*: velocidad con brillo máximo (km/s).
- *energy*: energía total radiada aproximada (1010 julios).
- *impact-e*: energía de impacto total aproximada (kt).
- *vx*: velocidad estimada previa a la entrada (componente X centrada en la Tierra, km/s).
- *vy*: velocidad estimada previa a la entrada (componente Y centrada en la Tierra, km/s).
- *vz*: velocidad est. previa a la entrada (componente Z centrada en la Tierra, km/s).

2. Carga en PDI

Primero se crean la base de datos en SQL Server y la tabla.

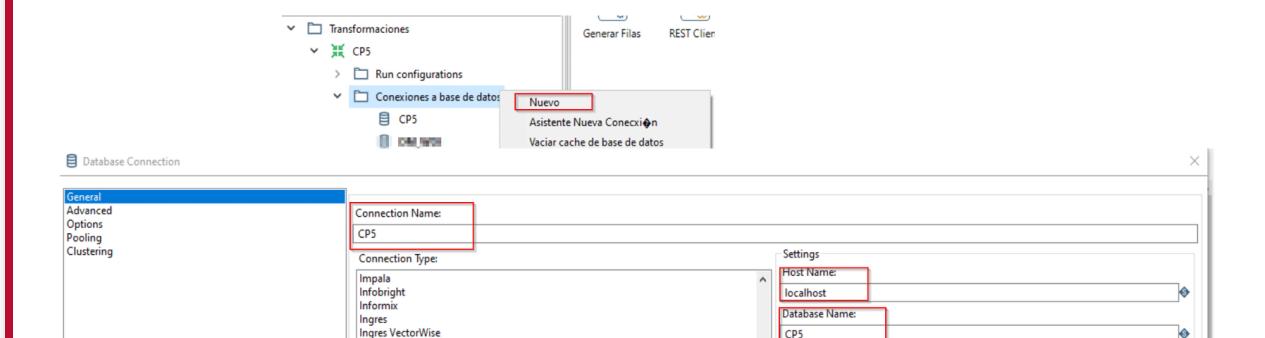


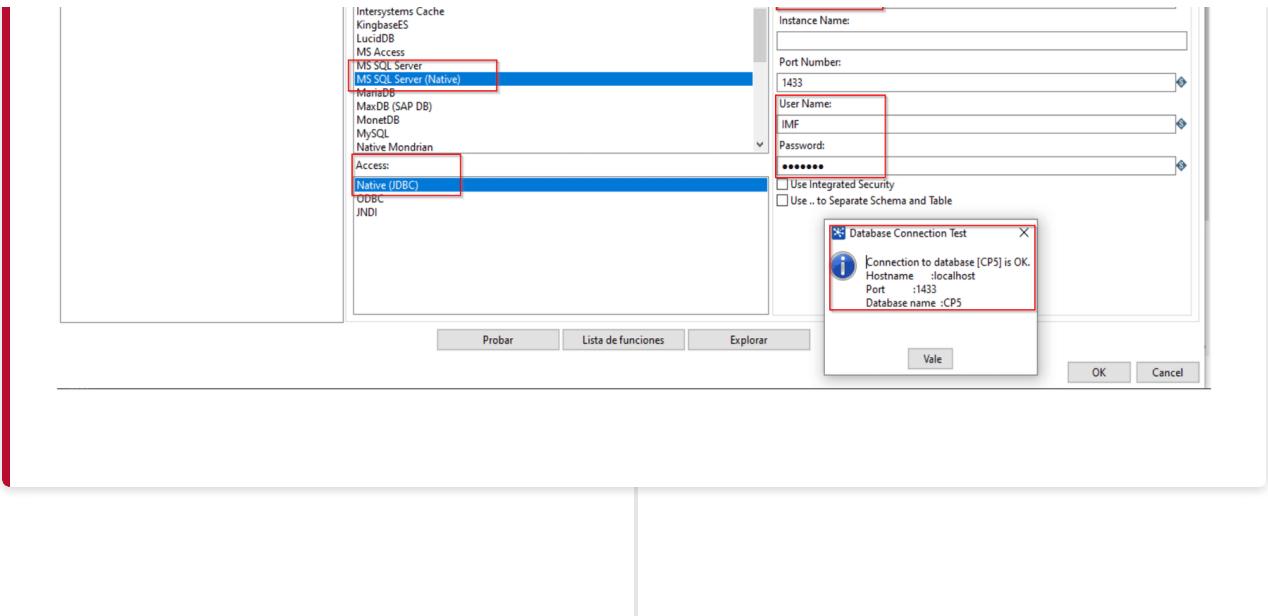


Como se ha visto anteriormente en este módulo, se usará PDI para extraer la información desde las API a un fichero o tabla de base de datos.

Para ello se usará PDI y se creará una transformación que realizará una llamada a la API de datos abiertos de la NASA seleccionada y transformará el fichero JSON en datos estructurados.

Lo **segundo** será generar la conexión a la base de datos que se va a usar para cargar datos.





En **tercer lugar**, se realizará la llamada al servicio **Rest**, que devuelve los datos en formato JSON. Para ello se usarán dos objetos:



Generador de filas. Actúa como trigger para el siguiente paso, si no, no se ejecuta. En este solo se especifica la URL que proporciona los datos.



Interval in ms (delay)	5000
Current row time field name	now
Previous row time field name	FiveSecondsAgo
Campos :	
Nombre	Tipo
url	String
Formato	
Longitud	
Precision	
Moneda	
Decimal	
Grupo	
Valor	<code>https://ssd-api.jpl.nasa.gov/fireball.api?limit=50</code>
Set N	

Cliente rest. Este objeto realiza la llamada a la URL y descarga los datos.

REST Client

Step name **REST Client**

General Authentication SSL Headers Parameters Matrix Parameters

Settings

URL `https://ssd-api.jpl.nasa.gov/cad.api?dist-max=5LD&date-min=2020-01`

Accept URL from field?

URL field name `url`

HTTP method `GET`

Get Method from field

Method field name

Body field

Application type `JSON`

Output fields

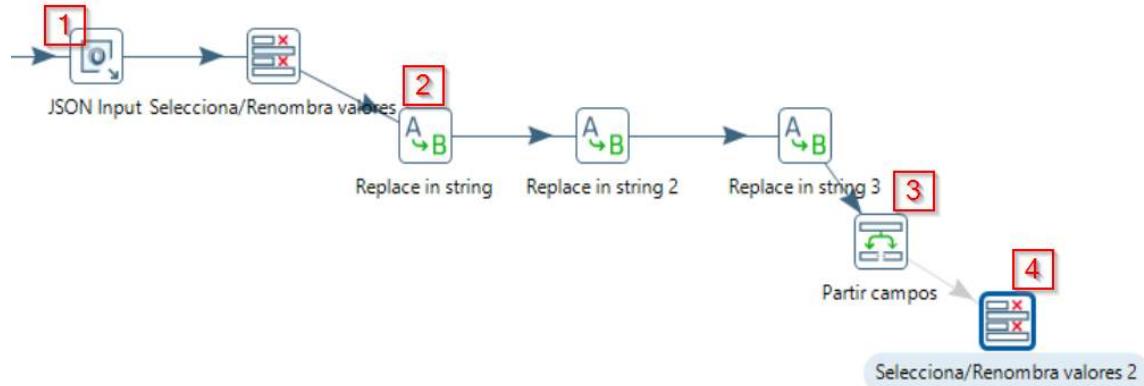
Result field name `result`

HTTP status code field name

Response time (milliseconds) field name

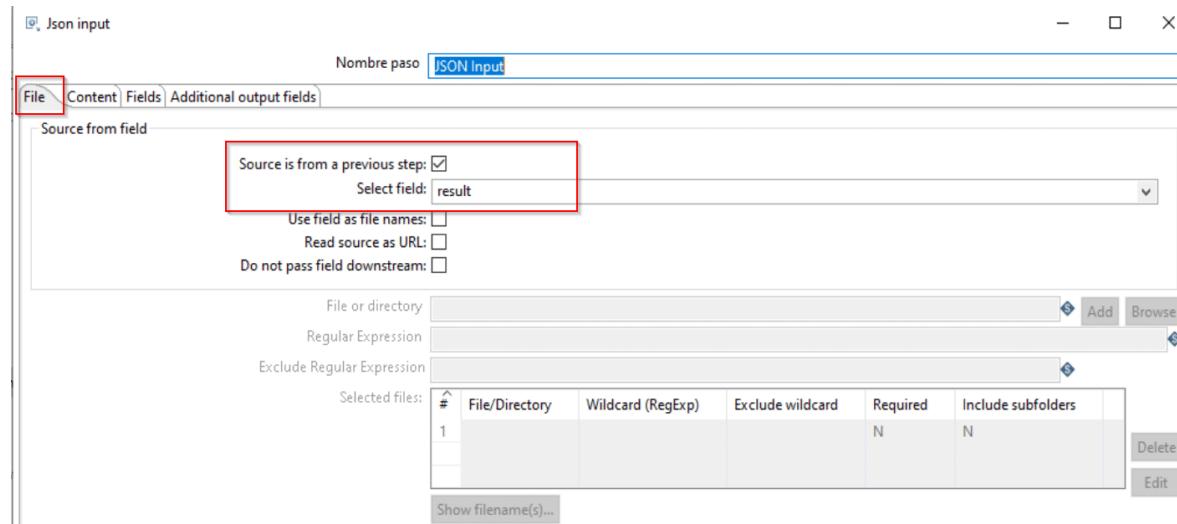
Response header field name

A partir de aquí se realiza el **procesado de los datos en formato JSON**. Para ello se necesita un conversor de datos en formato JSON. Y luego, dada la estructura en la que se encuentra ese fichero JSON, se realizan un reemplazo de caracteres y una separación de campos.



1) Json Input parsea el formato JSON:

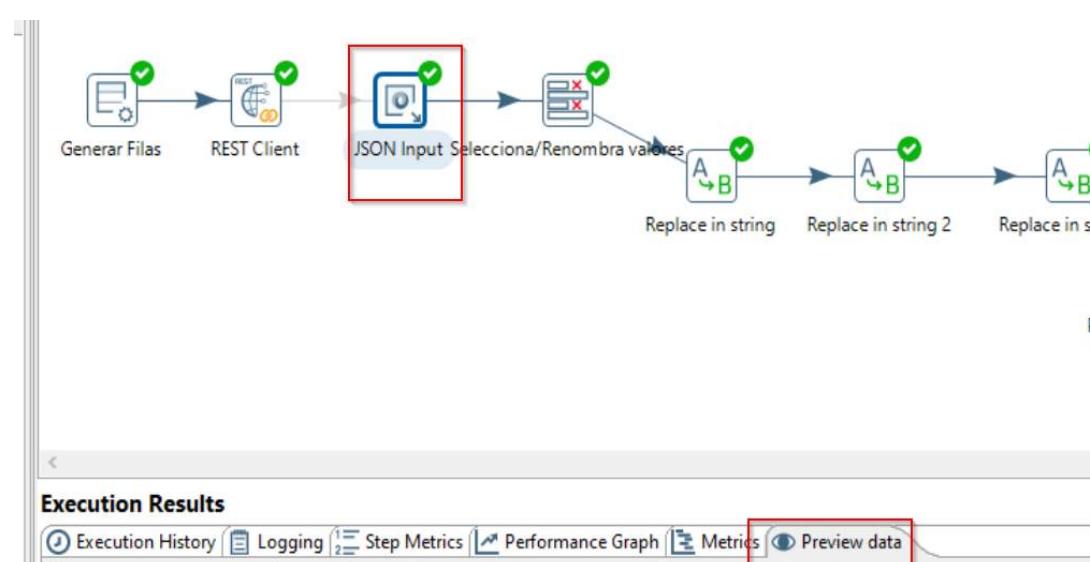
Primero se configura la pestaña de fichero, indicando la entrada.



Después se configura la pestaña de campos, donde se especifica la ruta de los datos dentro del fichero JSON. Esta sección depende de la composición del fichero JSON.

#	Name	Path	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type	Repeat
1	data	\$.data.*	String							ninguno	N

Si se ejecutamos el paquete de carga hasta este punto, se verá que ya se dispone de los datos, pero **en una única columna** con corchetes y los campos separados por comas:



① \${TransPreview.FirstRows.Label} ○ \${TransPreview.LastRows.Label} ○ \${TransPreview.Off.Label}		
#	url	data
1	https://ssd-api.jpl.nasa.gov/fireball.api?limit=50	["2021-07-30 08:06:34", "14.6", "0.42", "7.8", "S", "90.1", "E", "63.0", null]
2	https://ssd-api.jpl.nasa.gov/fireball.api?limit=50	["2021-07-29 13:19:57", "3.7", "0.13", "42.4", "N", "98.4", "E", "26.4", "14.7"]
3	https://ssd-api.jpl.nasa.gov/fireball.api?limit=50	["2021-07-07 13:41:14", "3.3", "0.11", null, null, null, null, null, null]
4	https://ssd-api.jpl.nasa.gov/fireball.api?limit=50	["2021-07-05 03:46:24", "74", "1.8", "44.3", "N", "164.2", "W", "43.4", "15.7"]

2) Se reemplazan caracteres del resultado que sobran como [,".

Replace in string

Step name: Replace in string

Fields string		use RegEx	Search	Replace with	Set empty string?	Replace with field	Whole Word	Case sensitive
#	In stream field	Out stream field	N	[N		N	N
1	data	data1						

Vale Get fields Cancelar

Execution Results

Execution History Logging Step Metrics Performance Graph Metrics Preview data

② \${TransPreview.FirstRows.Label} ○ \${TransPreview.LastRows.Label} ○ \${TransPreview.Off.Label}

```
data3
2020 VT4,5,2459167,222686305,2020-Nov-13 17:21,4.50910597356063e-05,4.50230750909071e-05,4.51593933094878e-05,13.4271195491719,7.88069685599943,< 00:01,28.7,null,null
2020 QG,4,2459077,672759561,2020-Aug-16 04:09,6.22797984976286e-05,6.22040307793404e-05,6.23555561978824e-05,12.3308673063873,8.15386324007669,< 00:01,29.90,null,null
2020 JJ,3,2458974,003145869,2020-May-04 12:05,8.95978759447433e-05,8.94490618786109e-05,8.97468889049335e-05,14.3572480018827,12.1100856733462,< 00:01,30.00,null,null
2020 CW,2,2458881,034693004,2020-Feb-01 12:50,0.000105112704803889,0.000104665620612537,0.000105559739813831,21.3075984622406,20.0827334430908,< 00:01,32.516,null,null
```

3) Se separan los campos por un separador dado.

The screenshot shows the 'Partir campo' (Split field) step configuration window. The 'Nombre paso' (Step name) is set to 'Partir campo'. The 'Campo a partir' (Field to split) is set to 'data3'. The 'Separador' (Separator) is set to ',' (comma). The 'Enclosure' (Enclosure) is set to an empty string. Below the configuration, there is a table titled 'Campos' (Fields) containing 12 rows of data. The first row is highlighted with a red border. The columns are: #, Nuevo campo (New field), ID, Eliminar ID? (Delete ID?), Tipo (Type), Longitud (Length), Precisión (Precision), Formato (Format), Grupo (Group), Decimal (Decimal), Moneda (Currency), NuloSi (Null value), Default, Trim type (Trim type), and ninguno (none). The data rows are:

#	Nuevo campo	ID	Eliminar ID?	Tipo	Longitud	Precisión	Formato	Grupo	Decimal	Moneda	NuloSi	Default	Trim type
1	Date	N		String									ninguno
2	Lat	N		String									ninguno
3	Lon	N		String									ninguno
4	Lat_dir	N		String									ninguno
5	Lon_dir	N		String									ninguno
6	Alt	N		String									ninguno
7	Vel	N		String									ninguno
8	energy	N		String									ninguno
9	impact	N		String									ninguno
10	vx	N		String									ninguno
11	vy	N		String									ninguno
12	vz	N		String									ninguno

4) Se seleccionan solo los campos que se quieren de salida. Primero se selecciona la opción para traer todos los campos y luego se filtran solo aquellos que se quieren mostrar.

The screenshot shows the 'Selecciona/Renombra valores' (Select/Rename values) step configuration window. The 'Nombre paso' (Step name) is set to 'Selecciona/Renombra valores 2'. There are three tabs at the top: 'Selecciona & Modifica' (Select & Modify), 'Eliminar' (Delete), and 'Meta-information'. The 'Campos:' (Fields:) table has a red border around its first column. The columns are: #, Nombre campo (Field name), Renombrar a (Rename to), Longitud (Length), and Precisión (Precision). The data rows are:

#	Nombre campo	Renombrar a	Longitud	Precisión
1	Date			
2	Lat			
3	Lon			
4	Lat_dir			
5	Lon_dir			
6	Alt			
7	Vel			
8	Energy			

On the right side of the window, there are two buttons: 'Obtener campos a seleccionar' (Get selected fields) and 'Edit Mapping'.

```
9 Impact
10 vx
11 vy
12 vz
```

Partir campos

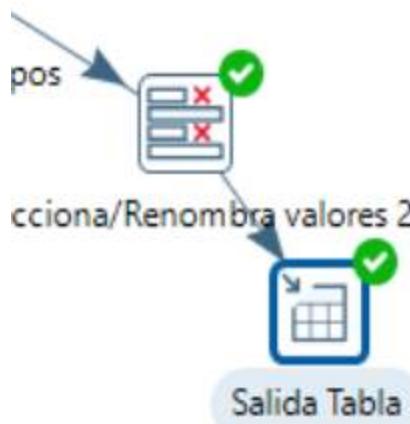
Selección/Renombrar valores 2

Execution Results

(1) Execution History (2) Logging (3) Step Metrics (4) Performance Graph (5) Metrics (6) Preview data
(7) \${TransPreview.FirstRows.Label} (8) \${TransPreview.LastRows.Label} (9) \${TransPreview.OffLabel}

#	Desc	Orbit_id	Jd	Cd	Dist	dist_min	dist_max	v_rel
1	2020 VT4	5	2459167.222686305	2020-Nov-13 17:21	4.50910597356063e-05	4.50230750909071e-05	4.51593933094878e-05	13.4271195491719
2	2020 QG	4	2459077.572759561	2020-Aug-16 04:09	6.2279784976286e-05	6.22040307793404e-05	6.2355561978824e-05	12.3308673063873
3	2020 JJ	3	2458974.003145869	2020-May-04 12:05	8.95978759447432e-05	8.94490618786109e-05	8.97468889049325e-05	14.3572490018827
4	2020 CW	2	2458881.034693004	2020-Feb-01 12:50	0.000105112704803889	0.000104665620612537	0.00010559739813831	21.3073984622405
5	2020 XK1	4	2459191.176836694	2020-Dec-07 16:15	0.000139076938214119	0.000139045236846255	0.000139108644191317	8.49214903453907
6	2021 CZ3	5	2459254.576333028	2021-Feb-09 01:50	0.00015247592080515	0.000152367672886643	0.00015284149998933	11.0887501173184
7	2021 BO	4	2459232.839386007	2021-Jan-18 08:09	0.000160402626150236	0.000160328153961792	0.0001604770923880937	7.72378836573796
8	2021 GW4	6	2459317.042684094	2021-Apr-12 13:01	0.000175131371899942	0.000174924596875557	0.000175338081444841	8.36024941650424

Por ultimo, hay que definir la salida a la tabla de SQL Server:



Reasignando el *mapping* del paso anterior (Seleccionar/renombrar valores 2):

Selecciona/Renombra valores

Nombre paso Selecciona/Renombra valores 2

Selecciona & Modifica | Eliminar | Meta-information

Campos:

#	Nombre campo	Renombrar a	Largo	Precisión
1	Date	date		
2	Energy	energy		
3	Impact	impact		
4	Lat	lat		
5	Lat_dir	lat_dir		
6	Lon	lon		
7	Lon_dir	lon_dir		
8	Alt	alt		
9	Vel	vel		

Obtener campos a seleccionar

Edit Mapping

Enter Mapping

Source fields: Target fields:

Date (JSON input)	
data1 (Replace in string)	
data2 (Replace in string 2)	

Add Delete

Mappings:

Date	(Partir campos) --> date
Energy	(Partir campos) --> energy
Impact	(Partir campos) --> impact
Lat	(Partir campos) --> lat
Lat_dir	(Partir campos) --> lat_dir
Lon	(Partir campos) --> lon
Lon_dir	(Partir campos) --> lon_dir
Alt	(Partir campos) --> alt
Vel	(Partir campos) --> vel

Include unspecified fields, ordered by name

Auto target selection? Auto source selection?
Hide assigned source fields? Hide assigned target fields?

Vale Guess Cancelar

Ejecutando la tarea se debe cargar en la tabla final de SQL Server:

Execution Results

([Execution History](#)) ([Logging](#)) ([Step Metrics](#)) ([Performance Graph](#)) ([Metrics](#)) ([Preview data](#))

([\\$\(TransPreview.FirstRows.Label\)](#)) ([\\$\(TransPreview.LastRows.Label\)](#)) ([\\$\(TransPreview.Off.Label\)](#))

#	date	energy	impact	lat	lat_dir	lon	lon_dir	alt	vel
1	2021-07-30 08:06:34	14.6	0.42	7.8	S	90.1	E	63.0	null
2	2021-07-29 13:19:57	3.7	0.13	42.4	N	98.4	E	26.4	14.7
3	2021-07-07 13:41:14	3.3	0.11	null	null	null	null	null	null
4	2021-07-05 03:46:24	74	1.8	44.3	N	164.2	W	43.4	15.7
5	2021-06-09 05:43:59	2.3	0.082	17.9	S	55.3	W	null	null
6	2021-05-16 15:51:09	3.8	0.13	52.1	S	171.2	W	37.0	null
7	2021-05-06 05:54:27	2.1	0.076	34.7	S	141.0	E	31.0	26.6
...

SQLQuery2.sql - loc_4ESN00D\IMF (60)* SQLQuery1.sql - loc_4ESN00D\IMF (54) + X

```
***** Script for SelectTopNRows command from SSMS *****
SELECT TOP (1000) [date]
    ,[energy]
    ,[impact]
    ,[lat]
    ,[lat_dir]
    ,[lon]
    ,[lon_dir]
    ,[alt]
    ,[vel]
    FROM [CP5].[dbo].[FireBalls]
```

100 %

Results Messages

	date	energy	impact	lat	lat_dir	lon	lon_dir	alt	vel
1	2021-07-30 08:06:34	14.6	0.42	7.8	S	90.1	E	63.0	null
2	2021-07-29 13:19:57	3.7	0.13	42.4	N	98.4	E	26.4	14.7
3	2021-07-07 13:41:14	3.3	0.11	null	null	null	null	null	null
4	2021-07-05 03:46:24	74	1.8	44.3	N	164.2	W	43.4	15.7
5	2021-06-09 05:43:59	2.3	0.082	17.9	S	55.3	W	null	null
6	2021-05-16 15:51:09	3.8	0.13	52.1	S	171.2	W	37.0	null
7	2021-05-06 05:54:27	2.1	0.076	34.7	S	141.0	E	31.0	26.6
8	2021-05-02 14:12:49	2.5	0.089	12.3	N	43.4	W	null	-
9	2021-04-13 02:16:47	2.1	0.076	26.8	N	79.1	W	44.4	14.1

VII. Glosario



El glosario contiene términos destacados para la comprensión de la unidad

API

Un conjunto de definiciones y protocolos que se utiliza para desarrollar e integrar el *software* de las aplicaciones. API significa interfaz de programación de aplicaciones. Las API permiten que sus productos y servicios se comuniquen con otros, sin necesidad de saber cómo están implementados.

Competence Center Corporate Data Quality (CC CDQ)

Consorcio de investigación de gestión de datos y una comunidad de expertos que se ocupa de los desafíos derivados de la digitalización y las estrategias basadas en datos. El objetivo principal del CC CDQ es transferir conceptos innovadores y resultados de investigación científica en el dominio de la gestión de datos a la práctica empresarial diaria para ayudar a las empresas a gestionar los datos como un activo.

Jet Propulsion Laboratory (JPL)

Centro dedicado a la construcción y operación de naves espaciales no tripuladas para la agencia espacial estadounidense NASA.

Nielsen Corporation

Anteriormente conocida como ACNielsen, es una firma de investigación de mercados global.

Una de las creaciones más conocidas de Nielsen son sus calificaciones, un sistema de medición de audiencia que cuantifica las audiencias de televisión, radio y periódicos en sus respectivos mercados de medios. En 1950, adquirieron la empresa C. E. Hooper y comenzaron a conectar dispositivos de grabación a una muestra estadística de alrededor de 1200 televisores de consumo en los Estados Unidos. Estos dispositivos utilizaban películas fotográficas en cartuchos enviados por correo para registrar los canales vistos por el consumidor y, así, determinar el tamaño de la audiencia. Más tarde, Nielsen desarrolló métodos electrónicos de recopilación y transmisión de datos.

RDF

Marco de descripción de recursos (del inglés, *resource description framework*); una familia de especificaciones de la World Wide Web Consortium (W3C) originalmente diseñada como un modelo de datos para metadatos. Ha llegado a ser usado como un método general para la descripción conceptual o modelado de la información que se implementa en los recursos web, utilizando una variedad de notaciones de sintaxis y formatos de serialización de datos.

Skywise —

Plataforma de datos, iniciada por Airbus y Palantir Technologies, que conecta la cadena de valor de la aviación e incluye más de cien aerolíneas en todo el mundo, así como proveedores.

SSD/CNEOS —

Solar System Dynamics y Center for Near-Earth Object Studies.

Web services o servicio web —

Tecnología que utiliza un conjunto de protocolos y estándares que sirven para intercambiar datos entre aplicaciones. El servicio web se utiliza para REST, SOAP y XML-RPC para la comunicación, mientras que la API se utiliza para cualquier estilo de comunicación. El servicio web solo admite el protocolo HTTP, mientras que la API admite el protocolo HTTP/HTTPS. El servicio web admite XML, mientras que la API admite XML y JSON. Todos los servicios web son API, pero no todas las API son servicios web.