Introducción, análisis y utilización de la herramienta Hop

CONTENIDO

1. DESCRIPCIÓN

Hop es un programa en desarrollo que surge como una bifurcación de Pentaho Data Integration (PDI), de código abierto, y gracias a las múltiples contribuciones recibidas, este pretende ser más intuitivo, eficiente y modular a su antecesor.

Se desarrolla con la intención de facilitar todos los aspectos de la orquestación de datos y metadatos, manteniendo, y en muchos casos, mejorando la estabilidad ya conseguida por PDI a lo largo de los años. Además de incluir soporte nativo para Apache Beam, incorpora nuevas funcionalidades tales como la creación y diferenciación de entornos de ejecución, la integración total con GIT para un control total de las versiones o la creación de pruebas unitarias sobre los procesos desarrollados con la herramienta.

2. OBJETIVO

Su objetivo es presentar nuevos conceptos e ideas, a la vez que se mejoran las funciones de PDI menos destacadas suprimiendo aquellas funcionalidades más obsoletas.

Su versión preliminar introduce mejoras en la interfaz de usuario y consigue eliminar código obsoleto para mejorar los tiempos de ejecución y conectividad de sus propios elementos. Además, su configuración es mucho más uniforme.

3. OBTENCIÓN DE LA HERRAMIENTA Y TECNOLOGÍA

Durante la realización de este documento, se optó por utilizar la versión 0.3 liberada el 5 de junio de 2020. Una versión más actual de la herramienta puede ser descargada en el siguiente enlace:

www.project-hop.org/download/download

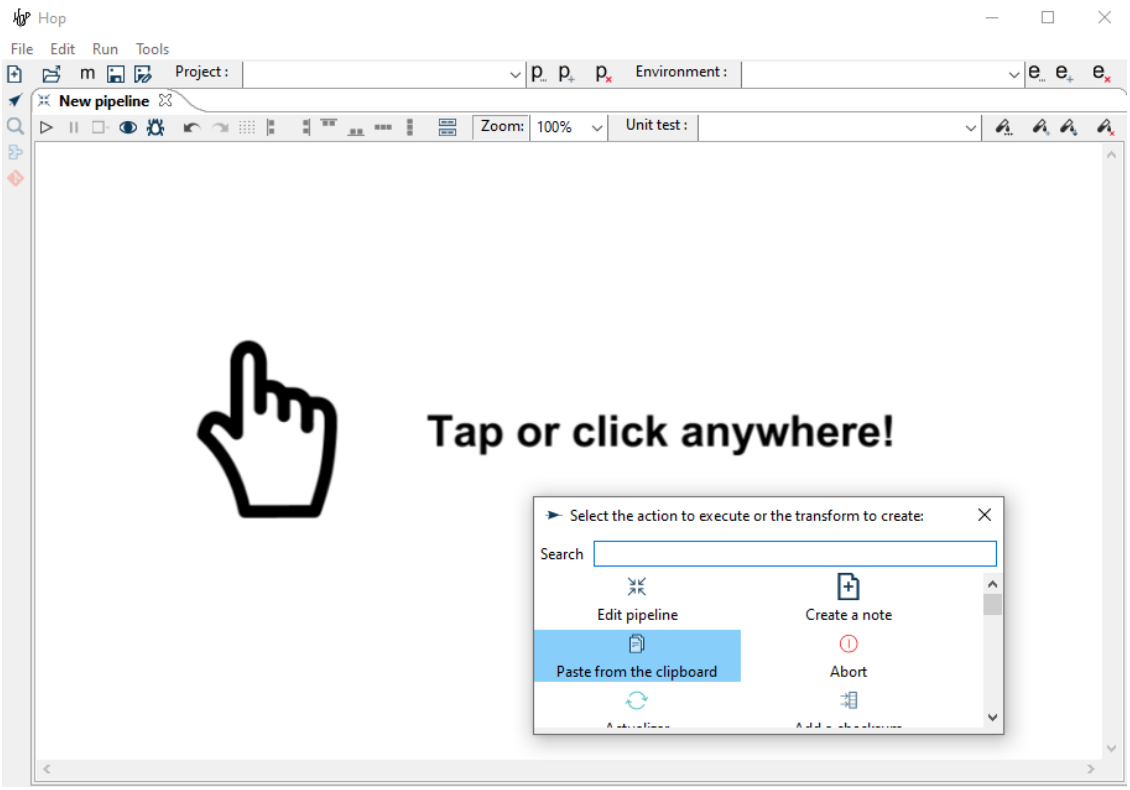
Una vez se descarga el fichero, se descomprime y se arranca la aplicación mediante el archivo **hop-gui**. En caso de tener un sistema operativo Windows, el archivo que se debe ejecutar para iniciar la herramienta es **hop-gui.bat**. Si el sistema operativo objetivo es Linux, se ejecutará el archivo **hop-gui.sh**.

Hop utiliza dos conceptos, los ***flujos de trabajo*** o ***workflows*** que no son más que una secuencia de operaciones realizadas secuencialmente de manera predeterminada (orquestación de tareas), y las ***tuberías*** o ***pipeline*** que realizan el trabajo de transformación de datos: leer, limpiar, formatear o escribir los datos en destino, entre otras cosas.

En cuanto a los *workflows*, como se ha mencionado, son los encargados de orquestar todas las acciones que se desean realizar en el proceso (otros *workflows* y pipelines además de otro tipo de acciones), y se componen de un punto de partida y uno o más puntos finales.

Los *workflows* trabajan mediante acciones y los *pipelines* con transformaciones, ambas conectadas por saltos (enlaces visuales). Estos enlaces son utilizados por los *workflows* para organizar la secuencia de ejecución de las acciones. Mientras que en los *pipelines* sirven para pasar datos de una transformación a otra.

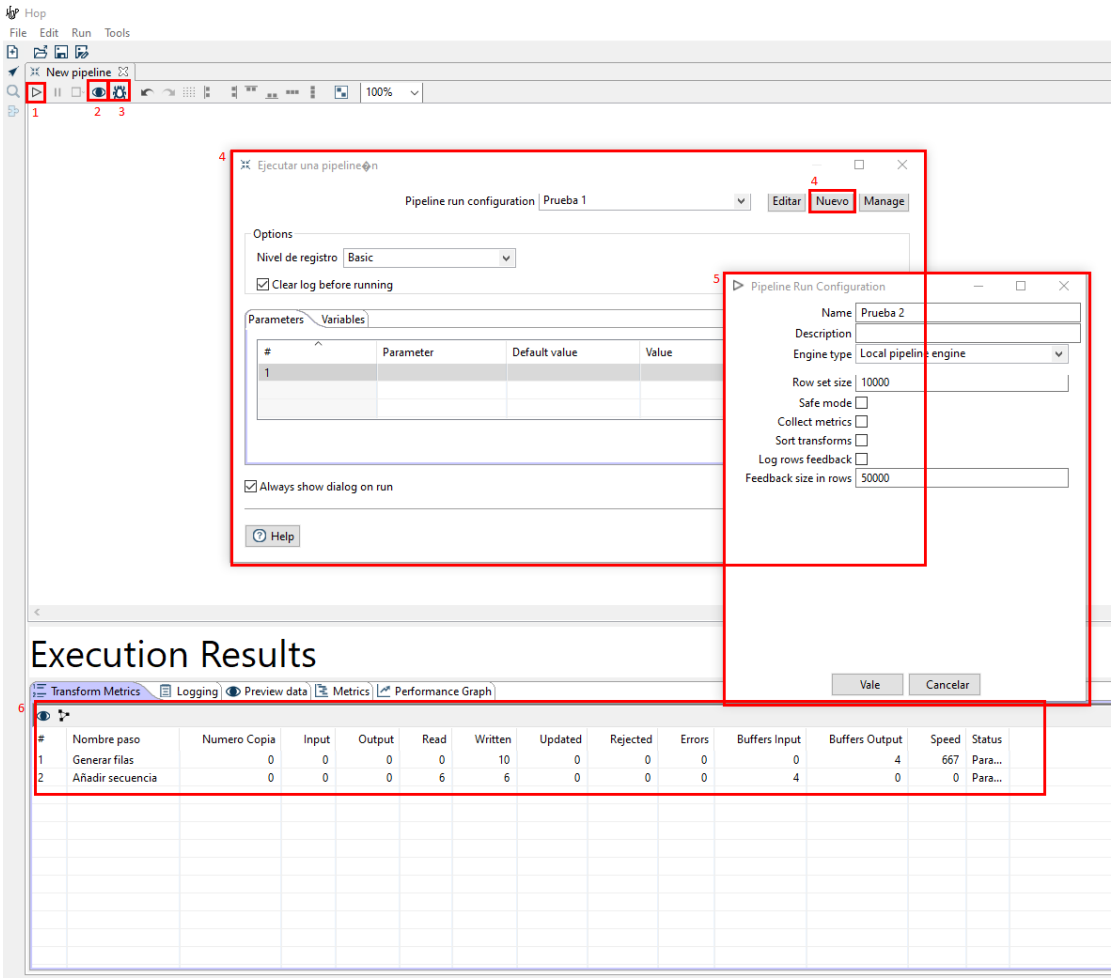
El primer paso para poder analizar su interfaz es crear un *pipeline*. Para ello, se debe seleccionar *pipeline* desde el menú de la aplicación, *File -> New -> Pipeline*. Aparecerá un lienzo, donde se podrá escoger cualquier tipo de transformación para construir la lógica del *pipeline* haciendo clic con el botón izquierdo del ratón sobre cualquier parte del mismo. La interfaz se visualiza como sigue.



En cuanto a la ejecución de un *pipeline*, este puede ser de tres modos:

- Directamente (RUN).
- Mediante una previsualización para conocer el resultado de la ejecución.
- Modo de depuración.

Cuando se selecciona la opción de ejecución directa (1), aparecerá una ventana emergente donde se podrá ajustar los parámetros de lanzamiento. Una vez terminada la ejecución, Hop muestra el resultado de la misma por pantalla en la parte inferior de la ventana.

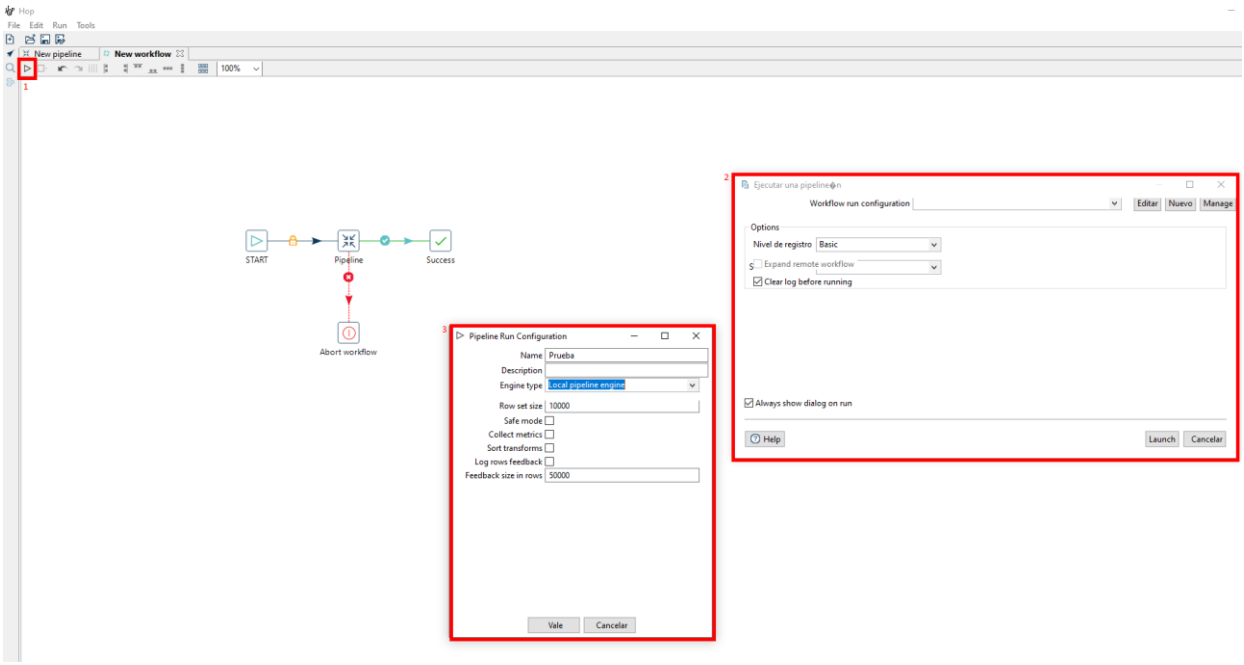


En cuanto a los *workflows*, estos tienen un comportamiento similar al de los *pipelines*.

A continuación, se crea un *workflow* sencillo, añadiendo al lienzo una serie de acciones. Este está compuesto por cuatro acciones:

- Un inicio, desde donde el *workflow* inicia la ejecución (***START***).
- Un *pipeline* general que puede ser el del ejemplo anterior.
- Un éxito, acción con la que el *workflow* finaliza en caso de que el resto de las acciones se ejecuten correctamente ***(Success)***.
- Un abortar, para detener la ejecución del workflow en caso de falla **(*Abort workflow*).**

La ejecución del *workflow* es exactamente igual a la de los *pipelines*. Ambos contienen los mismos modos de ejecución, así como las ventanas emergentes necesarias para configurar sus lanzamientos.


Hop
File Edit Run Tools
New pipeline
New workflow
100%
1
START
Pipeline
Success
Abort workflow
2
Ejecutar una pipeline
Workflow run configuration
Editar
Nuevo
Manage
Options
Nivel de registro
Basic
Expand remote workflow
Clear log before running
Always show dialog on run
Help
Launch
Cancelar
3
Pipeline Run Configuration
Name
Prueba
Description
Engine type
Local pipeline engine
Row set size
10000
Safe mode
Collect metrics
Sort transforms
Log rows feedback
Feedback size in rows
50000
Vale
Cancelar

4. EJEMPLO CARGA DE ALMACENES DE DATOS CON HOP

A continuación, se mostrará un ejemplo práctico de un proceso de migración de datos con Hop para la carga de datos en un Data Warehouse (DW).

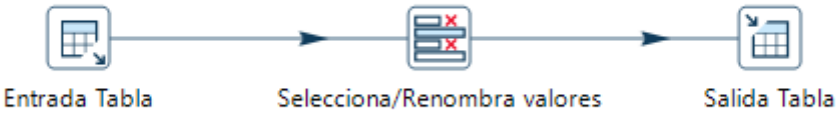
El Data Warehouse objetivo está compuesto por la tabla de hechos **Sales** y las tablas de dimensiones **Status**, **Customer**, **Product**, y **Time**.

Se asume que los datos provienen de archivos o bases de datos externas.

A modo de ejemplo se explicará únicamente el proceso de carga para la dimensión Cliente. Para el resto de las dimensiones, puesto que los procesos son similares, se mostrará solamente su visualización final y no su explicación.

Lo primero es crear un *pipeline* para cargar los datos de la dimensión. Para este caso, se realiza una extracción de los datos correspondientes a los clientes (customers) desde la base de datos. Luego se selecciona y se renombran algunos de estos campos. Por último, los datos extraídos se insertan en la tabla de la dimensión correspondiente.

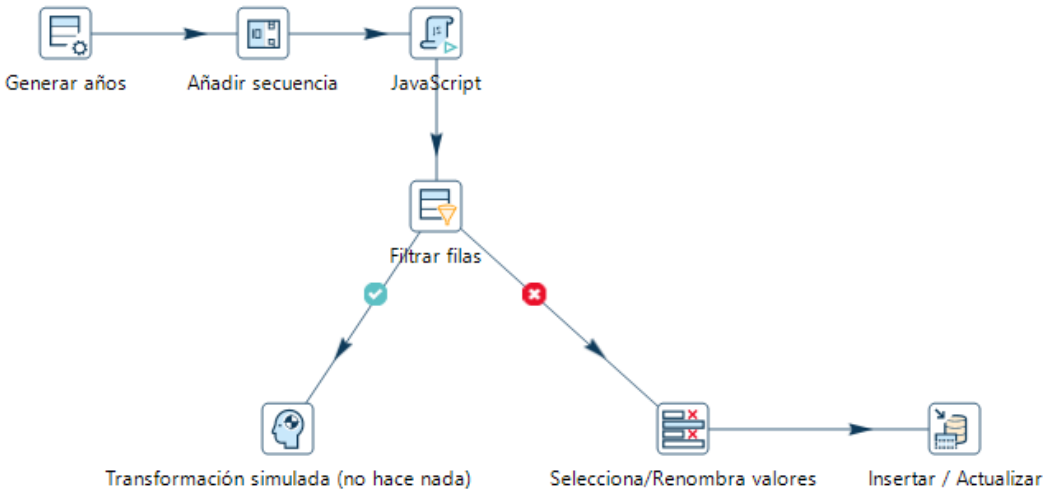
Dimensión cliente (t_dim_customer.hpl)



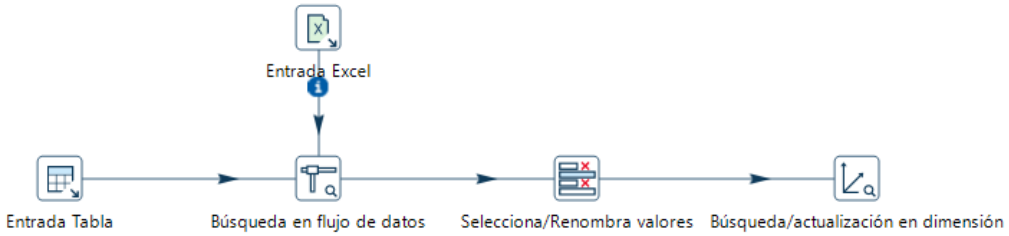
Después se cargan los datos correspondientes a las demás dimensiones de manera similar. Para cada una de ellas se crea un *pipeline* distinto. En cada caso se realizan los pasos necesarios para conseguir la máxima uniformidad de los datos. A continuación, se muestran los demás *pipelines* necesarios para la creación del DW objetivo.

Dimensión estado (t_dim_status.hpl)

Dimensión tiempo (t_dim_time.hpl)



Dimensión producto (t_dim_product.hpl)

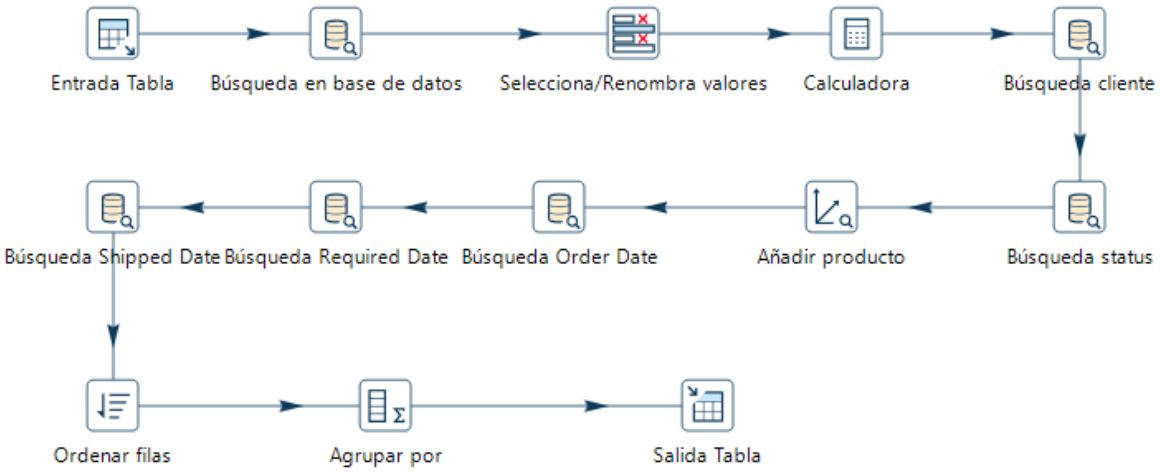


El próximo paso es la creación de un *workflow* que se encarga de iniciar y ejecutar los *pipelines* correspondientes a la carga de las dimensiones. Su composición consiste en un elemento **inicio (Start)** que arranca el proceso de inserción de los datos en las dimensiones, a través de los *pipelines* **cargar cliente (Load Customer)**, **cargar estado (Load Status)**, **cargar producto (Load Product)** y **cargar tiempo (Load Time)** y un elemento final de éxito **(Succes)** que indica si el proceso completo se ha ejecutado correctamente.

Carga de dimensiones (LoadDimensions.hpl)

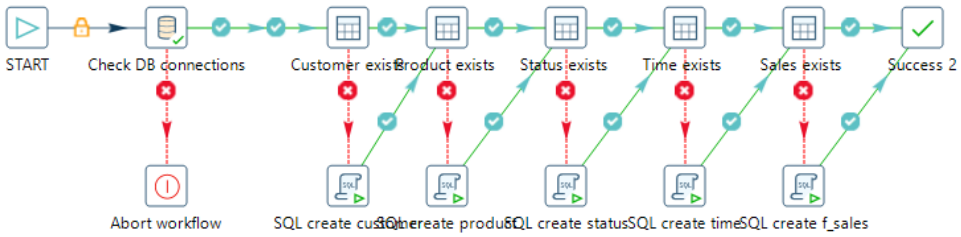
El siguiente paso, es crear un nuevo *pipeline* que carga los datos correspondientes a la tabla de hechos del Data Warehouse. Además, en este se aprecia la necesidad de formatear de los datos mediante diferentes transformaciones.

Tabla de hechos: ventas (t_h_sales.hpl)



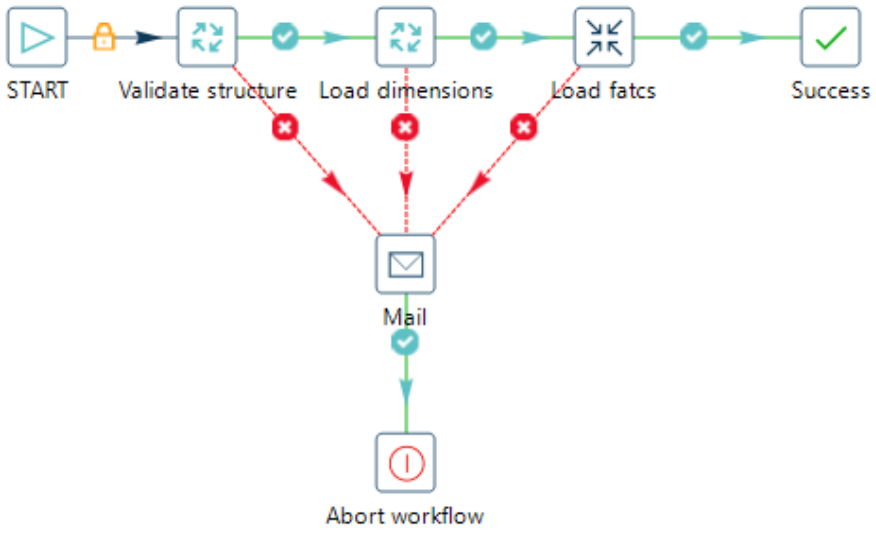
Otro paso interesante para la carga de datos de un Data Warehouse, es la de validar la conexión hacia la base de datos y sus tablas, en caso de que estas no existan, el proceso ha de crear todos los elementos necesarios para la inserción y la posterior recuperación de los datos. A continuación, su visualización.

Validar base de datos (ValidateDatabase.hwf)



Finalmente se crea un *workflow* que ejecuta el resto de los *pipelines* y *workflows* organizadamente con el objetivo de verificar, crear (en caso de ser necesario) y cargar datos en el DW correspondiente.

Cargar Data warehouse (Load_DW.hwf)



5. COMPARATIVA CON PDI

De momento, al comparar Hop con su predecesor notaremos muchas similitudes. Sin embargo, existen diferencias que valen la pena mencionar, y que se resumen en la siguiente tabla comparativas

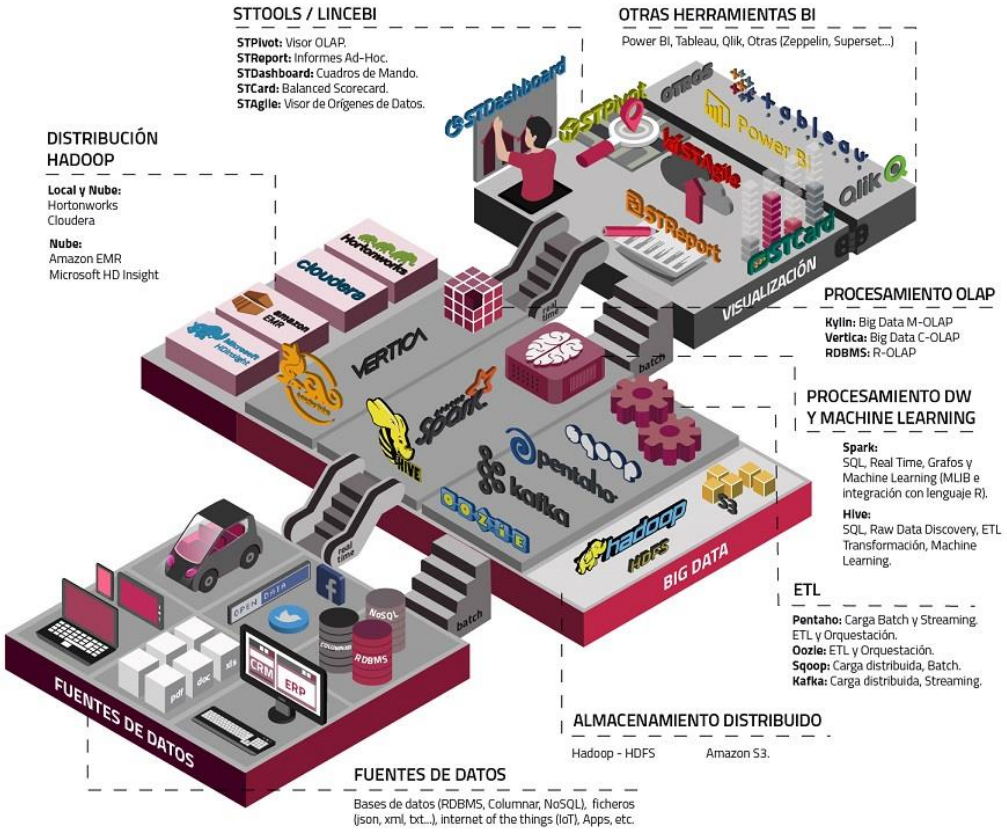
	PDI	Hop
Terminología	Trabajos Transformaciones Pasos Entrada de trabajo	Flujo de trabajo Tuberías Transformación Acción
Archivos	kjb (Kettle Job) ktr (Kettle transformation)	hwf (Hop Workflow) hpl (Hop PipeLine)
Utilidad de ejecución de procesos con interfaz	Spoon	Hop-gui
Utilidad de ejecución de procesos	kitchen para trabajos y pan para transformaciones	Hop-run para ambas
Utilidad para encriptar contraseñas	encr	hop-encrypt
Servidor web para la ejecución remota de procesos	carte	hop-server
Contenedores de datos	-	Soportado
Tiempo de arranque de la herramienta (aproximado)	El tiempo de inicio ronda el minuto	El tiempo de inicio es de unos pocos segundos

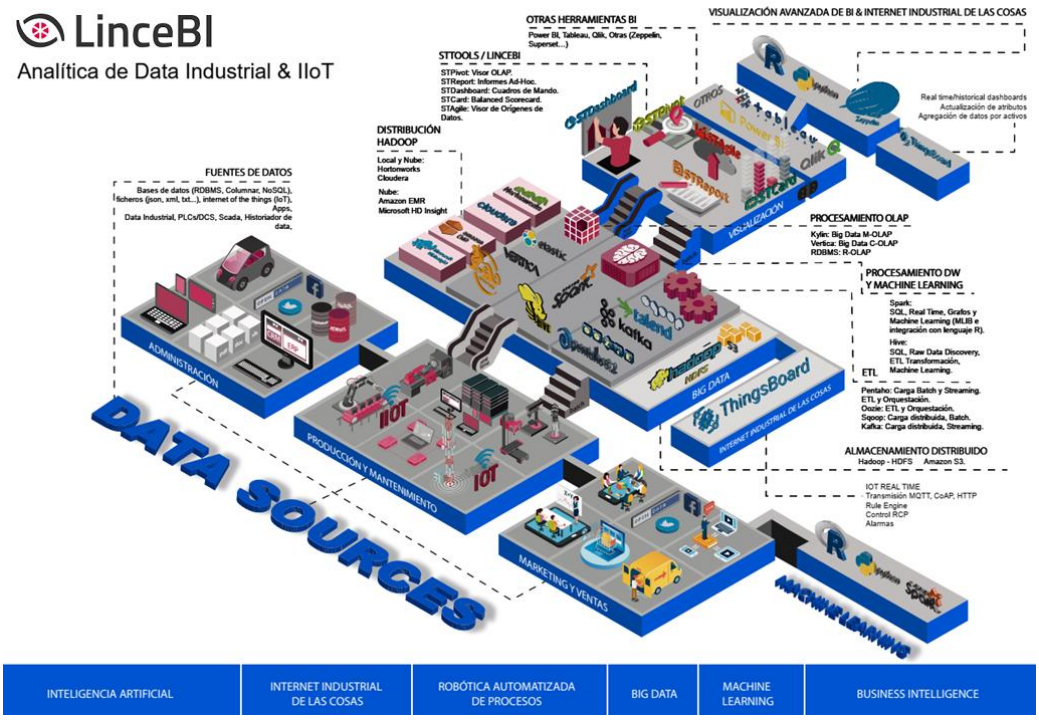
6. CONCLUSIÓN

Hop es una herramienta con mucho potencial, y aunque en apariencias es muy parecido a su predecesor, no cabe duda que sobrepasará a este en funcionalidades, extensibilidad y velocidad. La sola incorporación de Apache Beam extiende la posibilidad de la misma para ejecutar procesos en nuevos motores de ejecución. Su similitud con PDI hace que la curva de aprendizaje requerida sea casi inexistente, por lo que cualquier usuario con conocimientos en PDI podrá crear procesos en Hop sin inconvenientes.

7. TECNOLOGÍAS

Recientemente, hemos sido nombrados Partners Certificados de Vertica, Talend, Microsoft, Snowflake, Kylligence, Pentaho, etc.


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


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

8. 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.

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

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

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

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

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

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

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

9. OTROS

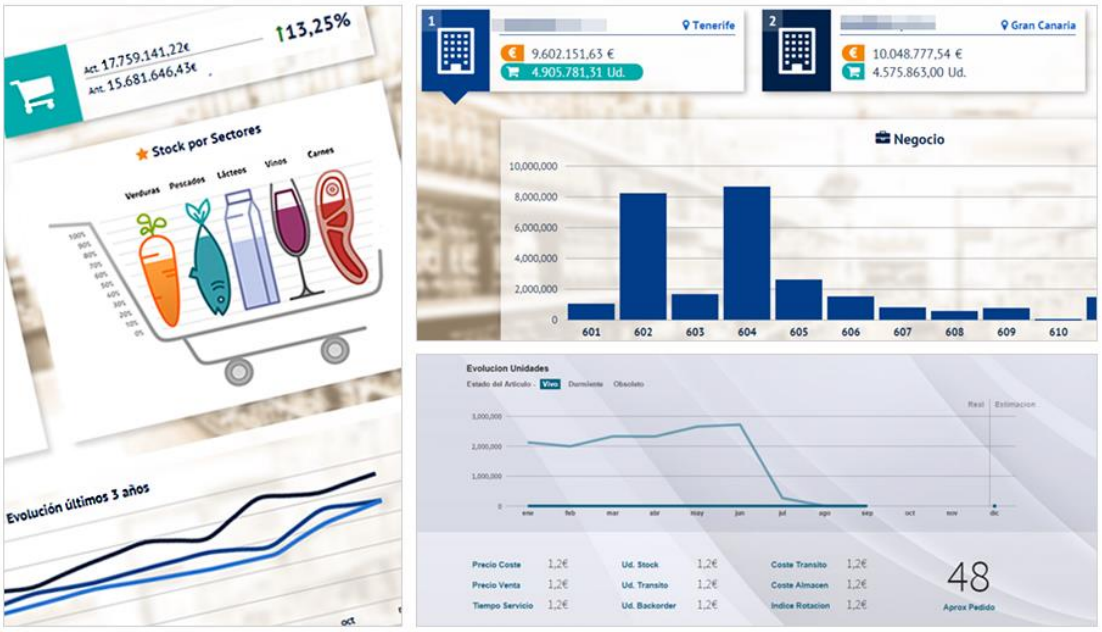
Trabajamos en los principales sectores y con algunas de las compañías y organizaciones más importantes de España.

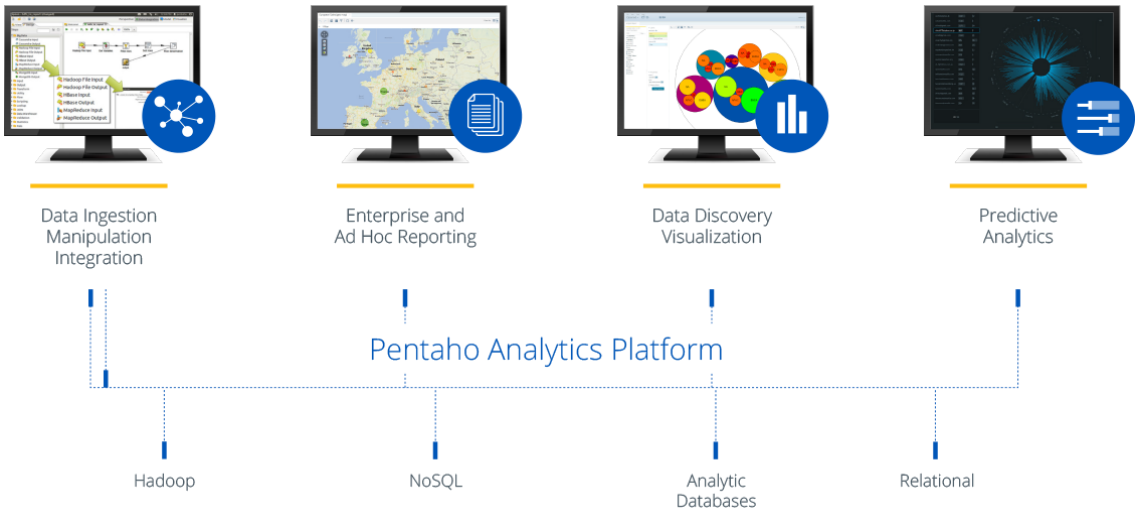
SECTOR PRIVADO

SECTOR PÚBLICO

10. EJEMPLOS DE DESARROLLOS ANALYTICS

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

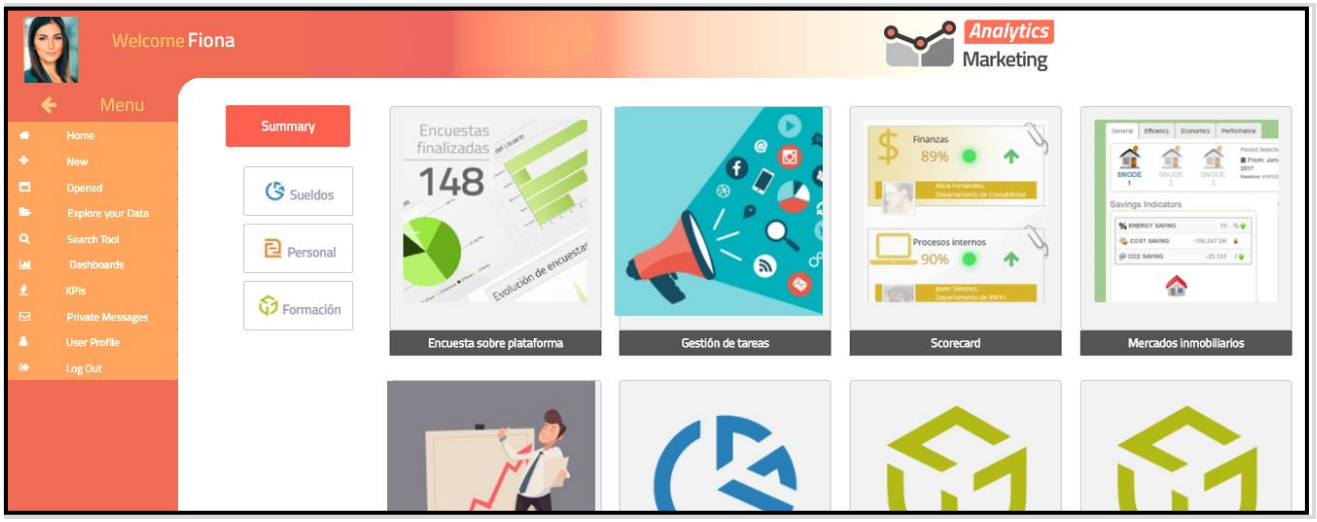





facebook DASHBOARD COMENTARIOS
Sexo
Sentimiento
Localización
Tipo de recurso
Evolución Comentarios por días


smart MADRID
Última Actualización
18:24


Créditos Matriculados
Procedencia
FUENTE-ÁLAMO
Alumnos Matriculados: 50
Sexo
50%
50%
Rango de Edad
18-20
20-22
22-24
24-26
26-28
28-30
30-35
35-40
GRADO EN BIOLOGÍA
Créditos Matriculados: 6.500
Buscar Plan de Estudio
INTERÉS PLATAFORMA
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
Encuesta sobre plataforma
Gestión de tareas
Finanzas
89%
Procesos internos
90%
Scorecard
Savings Indicators
Mercados inmobiliarios