Herramientas ETL

Herramientas ETL (Extraer, Transformar, Cargar)

Actualizado feb 2024 · 12 min leer

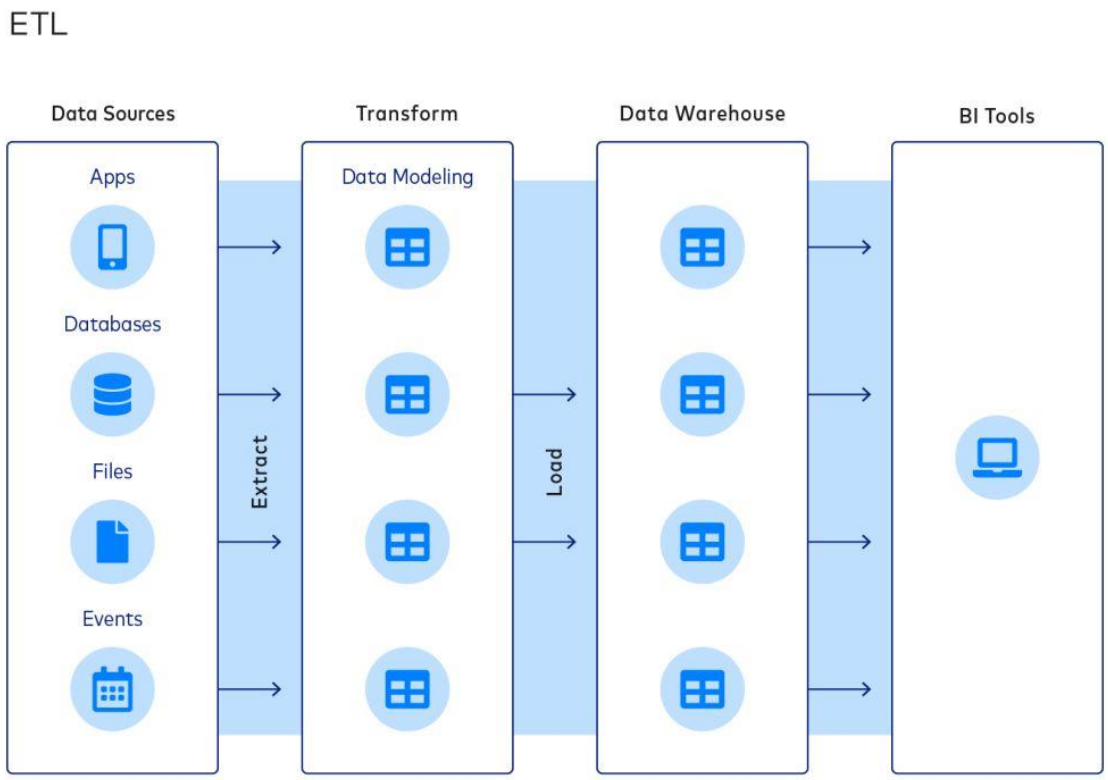
Contenido

Las herramientas ETL (Extraer, Transformar, Cargar) son una parte importante para resolver estos problemas. Hay muchas herramientas ETL diferentes entre las que elegir, lo que da a las empresas la posibilidad de seleccionar la mejor opción. Sin embargo, revisar todas las opciones disponibles puede llevar mucho tiempo.

¿Qué es ETL?

ETL es un enfoque común para integrar datos y organizar pilas de datos. Un proceso ETL típico consta de las siguientes etapas:

- **Extraer** datos de las fuentes
- **Transformar** datos en modelos de datos
- **Cargar** datos en almacenes de datos



El paradigma ETL es popular porque permite a las empresas reducir el tamaño de sus almacenes de datos, lo que puede ahorrar costes de cálculo, almacenamiento y ancho de banda.

Sin embargo, estos ahorros de costes son cada vez menos importantes a medida que desaparecen estas limitaciones. Como resultado, el ELT (Extraer, Cargar, Transformar) es cada vez más popular. En el proceso ELT, los datos se cargan en un destino tras la extracción, y la transformación es el paso final del proceso. A pesar de ello, muchas empresas siguen confiando en el ETL.

¿Qué son las herramientas ETL?

Como su nombre indica, las herramientas ETL son un conjunto de herramientas de software que se utilizan para **extraer**, **transformar** y **cargar** datos de una o varias fuentes en un sistema o base de datos de destino. Las herramientas ETL están diseñadas para automatizar y simplificar el proceso de extraer datos de diversas fuentes, transformarlos en un formato coherente y limpio y cargarlos en el sistema de destino de forma oportuna y eficiente. En la siguiente sección, veremos las consideraciones clave que los equipos de datos deben aplicar al considerar una herramienta ETL.

Consideraciones clave sobre las herramientas ETL

He aquí tres consideraciones clave para las herramientas ETL de una empresa.

1. **El alcance de la integración de datos**. Las herramientas ETL pueden conectarse a diversas fuentes y destinos de datos. Los equipos de datos deben optar por herramientas ETL que ofrezcan una amplia gama de integraciones. Por ejemplo, los equipos que quieran trasladar datos de Google Sheets a Amazon Redshift deben seleccionar herramientas ETL que admitan este tipo de conectores.
2. **Nivel de personalización**. Las empresas deben elegir sus herramientas ETL en función de sus requisitos de personalización y de los conocimientos técnicos de su equipo informático. A una empresa emergente pueden bastarle los conectores y transformaciones integrados en la mayoría de las herramientas ETL; una gran empresa con recopilación de datos a medida probablemente necesitará flexibilidad para elaborar transformaciones a medida con la ayuda de un sólido equipo de ingenieros.
3. **Estructura de costes.** Al elegir una herramienta ETL, las organizaciones deben tener en cuenta no solo el coste de la propia herramienta, sino también los costes de la infraestructura y los recursos humanos necesarios para mantener la solución a largo plazo. En algunos casos, una herramienta ETL con un coste inicial más elevado, pero con menores requisitos de tiempo de inactividad y mantenimiento, puede ser más rentable a largo plazo. Por el contrario, existen herramientas ETL gratuitas y de código abierto que pueden tener elevados costes de mantenimiento.

Otras consideraciones son:

- El nivel de automatización proporcionado
- El nivel de seguridad y cumplimiento
- El rendimiento y la fiabilidad de la herramienta

Las 19 mejores herramientas ETL a tener en cuenta

1. Informatica PowerCenter

Informatica PowerCenter es una de las mejores herramientas ETL del mercado. Dispone de una amplia gama de conectores para lagos y almacenes de datos en la nube, como AWS, Azure, Google Cloud y SalesForce. Sus herramientas con poca o ninguna programación están diseñadas para ahorrar tiempo y simplificar los flujos de trabajo.

Informatica PowerCenter incluye varios servicios que permiten a los usuarios diseñar, implementar y supervisar pipelines de datos. Por ejemplo, Repository Manager ayuda en la gestión de usuarios, Designer permite a los usuarios especificar el flujo de datos de la fuente al destino y Workflow Manager define la secuencia de tareas.

2. Apache Airflow

Apache Airflow es una plataforma de código abierto para crear, programar y supervisar flujos de trabajo mediante programación. La plataforma cuenta con una interfaz de usuario web y una interfaz de línea de comandos para gestionar y activar flujos de trabajo.

Los flujos de trabajo se definen mediante grafos acíclicos dirigidos (DAG), que permiten una visualización y gestión claras de las tareas y dependencias. Airflow también se integra con otras herramientas de uso común en ingeniería y ciencia de datos, como Apache Spark y Pandas.

Las empresas que utilizan Airflow pueden beneficiarse de su capacidad para escalar y gestionar flujos de trabajo complejos, así como de su activa comunidad de código abierto y su amplia documentación. Puedes aprender sobre Airflow en el siguiente curso de DataCamp.

3. IBM Infosphere Datastage

Infosphere Datastage es una herramienta ETL ofrecida por IBM como parte de su ecosistema Infosphere Information Server. Con su marco gráfico, los usuarios pueden diseñar pipelines de datos que extraigan datos de varias fuentes, realicen transformaciones complejas y entreguen los datos a las aplicaciones de destino.

IBM Infosphere es conocido por su velocidad, gracias a funciones como el equilibrio de carga y la paralelización. También admite metadatos, detección automática de fallos y una amplia gama de servicios de datos, desde el almacenamiento de datos a las aplicaciones de IA.

Igual que otras herramientas ETL empresariales, Infosphere Datastage ofrece una serie de conectores para integrar distintas fuentes de datos. También se integra perfectamente con otros componentes de IBM Infosphere Information Server, lo que permite a los usuarios desarrollar, probar, implementar y supervisar trabajos ETL.

4. Oracle Data Integrator

Oracle Data Integrator es una herramienta ETL que ayuda a los usuarios a construir, implementar y gestionar almacenes de datos complejos. Viene con conectores listos para usar con muchas bases de datos, como Hadoop, EREP, CRM, XML, JSON, LDAP, JDBC y ODBC.

ODI incluye Data Integrator Studio, que proporciona a los usuarios empresariales y a los desarrolladores acceso a varios artefactos a través de una interfaz gráfica de usuario. Estos artefactos ofrecen todos los elementos de la integración de datos, desde el movimiento de datos hasta la sincronización, la calidad y la gestión.

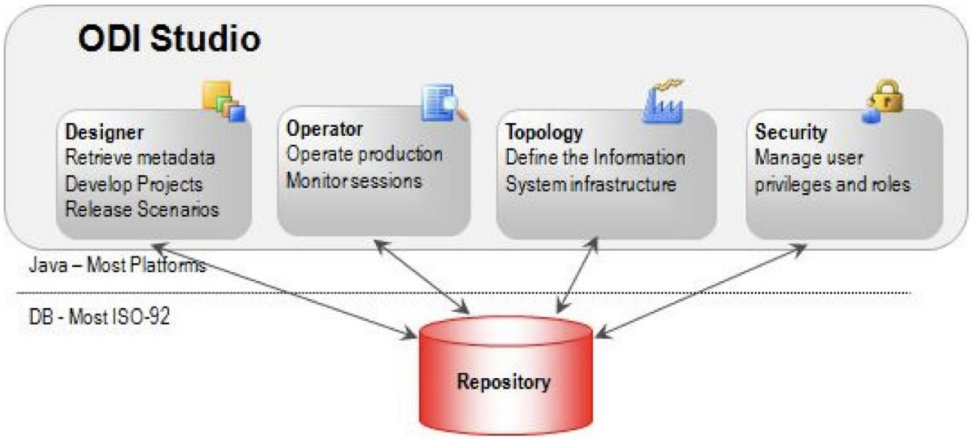


Figure 1 - ODI Studio Navigators connect to the repository

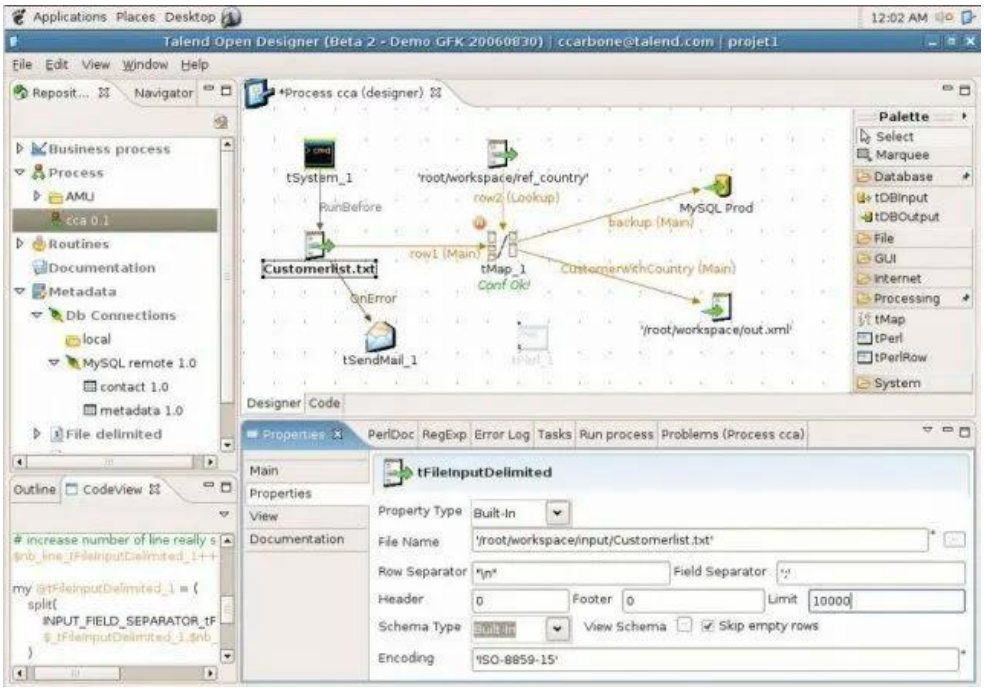
5. Microsoft SQL Server Integration Services (SSIS)

SSIS es una plataforma de nivel empresarial para la integración y transformación de datos. Viene con conectores para extraer datos de fuentes como archivos XML, archivos planos y bases de datos relacionales. Los profesionales pueden utilizar la interfaz gráfica de usuario del diseñador SSIS para construir transformaciones y flujos de datos.

La plataforma incluye una biblioteca de transformaciones integradas que reducen al mínimo la cantidad de programación necesaria para el desarrollo. SSIS también ofrece documentación completa para crear flujos de trabajo personalizados. Sin embargo, la pronunciada curva de aprendizaje y la complejidad de la plataforma pueden disuadir a los principiantes de crear rápidamente pipelines ETL.

6. Talend Open Studio (TOS)

Talend Open Studio es un popular software de integración de datos de código abierto que cuenta con una GUI fácil de usar. Los usuarios pueden arrastrar y colocar componentes, configurarlos y conectarlos para crear pipelines de datos. Entre bastidores, Open Studio convierte la representación gráfica en código Java y Perl.



Como herramienta de código abierto, TOS es una opción asequible con una amplia variedad de conectores de datos, incluidos conectores RDBMS y SaaS. La plataforma también se beneficia de una activa comunidad de código abierto que contribuye regularmente a la documentación y proporciona asistencia.

7. Pentaho Data Integration (PDI)

Pentaho Data Integration (PDI) es una herramienta ETL ofrecida por Hitachi. Captura datos de diversas fuentes, los limpia y los almacena en un formato uniforme y coherente.

Anteriormente conocido como Kettle, PDI cuenta con varias interfaces gráficas de usuario para definir pipelines de datos. Los usuarios pueden diseñar trabajos y transformaciones de datos utilizando el cliente PDI, Spoon, y luego ejecutarlos utilizando Kitchen. Por ejemplo, el cliente PDI puede utilizarse para ETL en tiempo real con Pentaho Reporting.

8. Hadoop

Hadoop es un marco de código abierto para procesar y almacenar big data en clústeres de servidores informáticos. Se considera la base del big data y permite almacenar y procesar grandes cantidades de datos.

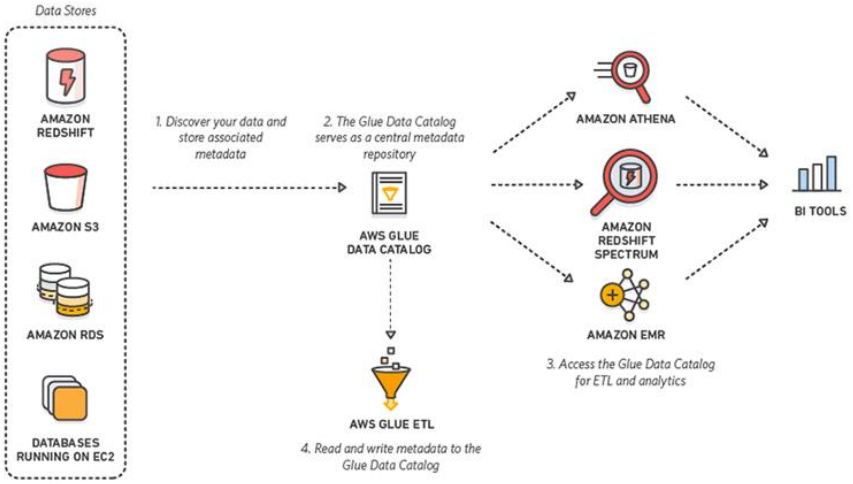
El marco Hadoop consta de varios módulos, como el sistema de archivos distribuido de Hadoop (HDFS) para almacenar datos, MapReduce para leer y transformar datos y YARN para la gestión de recursos. Hive se utiliza habitualmente para convertir consultas SQL en operaciones MapReduce.

Las empresas que piensen en usar Hadoop deben ser conscientes de sus costes. Una parte importante del coste de implementación de Hadoop procede de la potencia informática necesaria para el procesamiento y de la experiencia necesaria para mantener el ETL de Hadoop, más que de las herramientas o el almacenamiento en sí.

9. AWS Glue

AWS Glue es una herramienta ETL sin servidor ofrecida por Amazon. Descubre, prepara, integra y transforma datos de varias fuentes para casos de uso analíticos. Sin necesidad de configurar o gestionar la infraestructura, AWS Glue promete reducir el elevado coste de la integración de datos.

Mejor aún: al interactuar con AWS Glue, los profesionales pueden elegir entre una GUI de arrastrar y colocar, Jupyter Notebook o código Python/Scala. AWS Glue también ofrece compatibilidad con diversas cargas de trabajo y procesamientos de datos que satisfacen diferentes necesidades empresariales, como ETL, ELT, lotes y streaming.

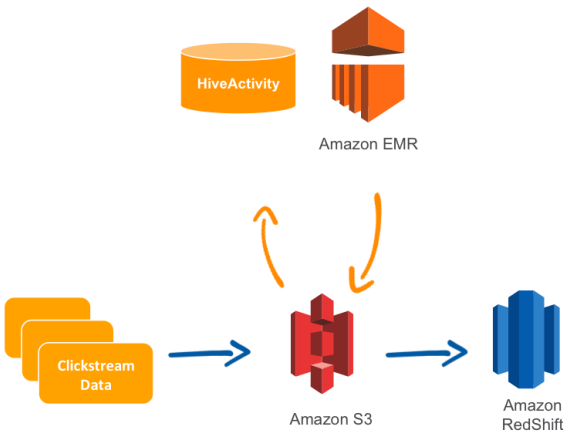


10. AWS Data Pipeline

AWS Data Pipeline es un servicio ETL gestionado que permite el movimiento de datos entre los servicios de AWS o los recursos locales. Los usuarios pueden especificar los datos que hay que mover, los trabajos o consultas de transformación y un calendario para realizar las transformaciones.

Data Pipeline es conocido por su fiabilidad, flexibilidad y escalabilidad, así como por su tolerancia a fallos y configurabilidad. La plataforma también cuenta con una consola de arrastrar y colocar para facilitar su uso. Además, es relativamente barata.

Un caso de uso común para AWS Data Pipeline es replicar datos desde el Servicio de Bases de Datos Relacionales (RDS) y cargarlos en Amazon Redshift.

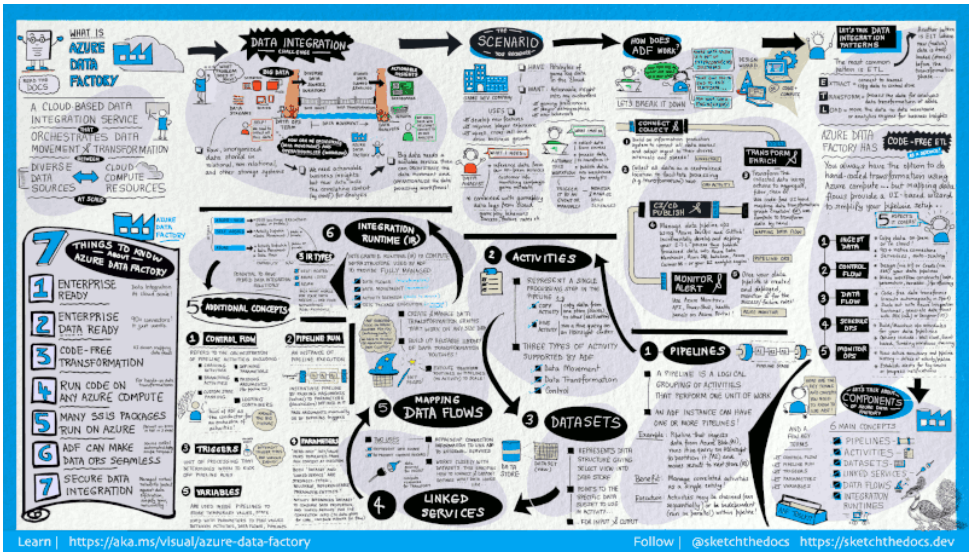


11. Azure Data Factory

Azure Data Factory es un servicio ETL en la nube ofrecido por Microsoft que se utiliza para crear flujos de trabajo que mueven y transforman datos a escala.

Comprende una serie de sistemas interconectados. Juntos, estos sistemas permiten a los ingenieros no solo introducir y transformar los datos, sino también diseñar, programar y supervisar los pipelines.

El punto fuerte de Data Factory reside en el gran número de conectores disponibles, desde MySQL a AWS, MongoDB, Salesforce y SAP. También es alabado por su flexibilidad: los usuarios pueden interactuar con una interfaz gráfica de usuario sin programación o con una interfaz de línea de comandos.



12. Google Cloud Dataflow

Dataflow es el servicio ETL sin servidor que ofrece Google Cloud. Permite el procesamiento de datos tanto en streaming como por lotes y no requiere que las empresas posean un servidor o clúster. Los usuarios solo pagan por los recursos consumidos, que se escalan automáticamente en función de los requisitos y la carga de trabajo.

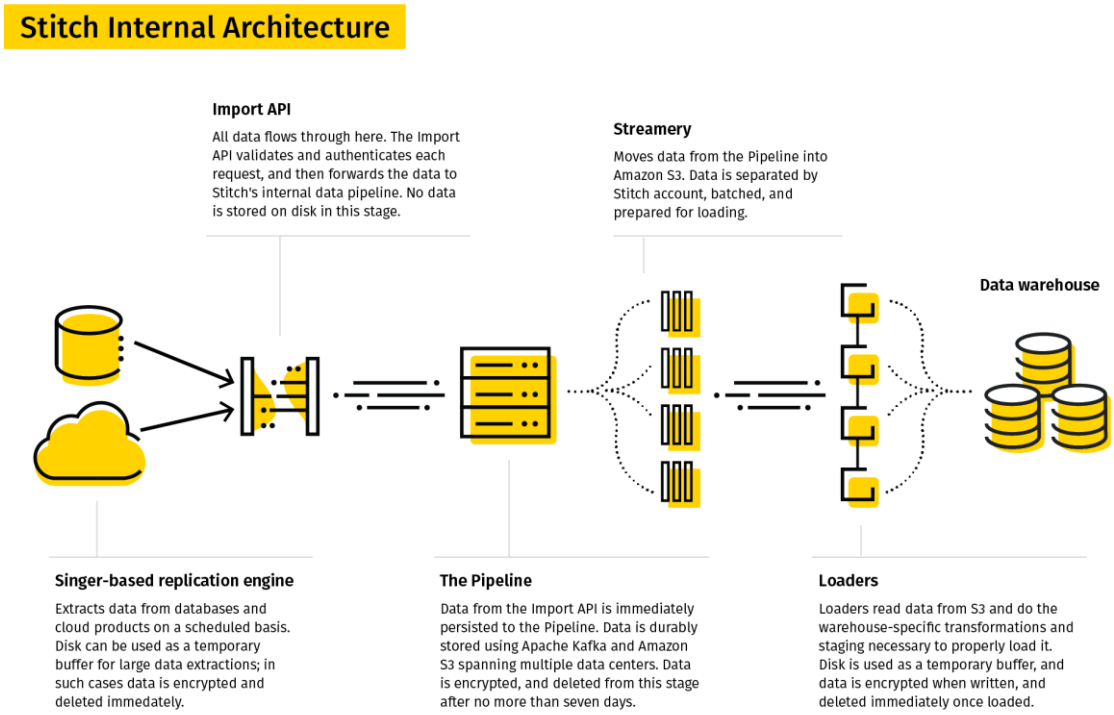
Google Dataflow ejecuta pipelines Apache Beam dentro del ecosistema Google Cloud Platform. Apache ofrece SDK de Java, Python y Go para representar y transferir conjuntos de datos, tanto por lotes como en streaming. Esto permite a los usuarios elegir el SDK adecuado para definir sus pipelines de datos.

13. Stitch

Stitch se describe a sí misma como una herramienta ETL sencilla y extensible creada para equipos de datos.

El proceso de replicación de Stitch extrae datos de diversas fuentes de datos, los transforma en un formato bruto útil y los carga en el destino. Sus conectores de datos incluyen bases de datos y aplicaciones SaaS. Los destinos pueden incluir lagos de datos, almacenes de datos y plataformas de almacenamiento.

Dada su sencillez, Stitch solo admite transformaciones sencillas, y no transformaciones definidas por el usuario.



14. SAP BusinessObjects Data Services

SAP BusinessObjects Data Services es una herramienta ETL empresarial que permite a los usuarios extraer datos de varios sistemas, transformarlos y cargarlos en almacenes de datos.

Data Services Designer proporciona una interfaz gráfica de usuario para definir pipelines de datos y especificar transformaciones de datos. Las reglas y los metadatos se almacenan en un repositorio, y un servidor de trabajos ejecuta el trabajo por lotes o en tiempo real.

Sin embargo, SAP Data Services puede ser caro, ya que el coste de la herramienta, el servidor, el hardware y el equipo de ingeniería puede aumentar rápidamente.

SAP Data Services es una buena opción para las empresas que utilizan SAP como sistema de planificación de recursos empresariales (ERP), ya que se integra perfectamente con SAP Data Services.

15. Hevo

Hevo es una plataforma de integración de datos para ETL y ELT que viene con más de 150 conectores para extraer datos de varias fuentes. Se trata de una herramienta de poca programación, lo que facilita a los usuarios el diseño de pipelines de datos sin necesidad de una amplia experiencia en programación.

Hevo ofrece una serie de funciones y ventajas, como la integración de datos en tiempo real, la detección automática de esquemas y la capacidad de manejar grandes volúmenes de datos. La plataforma también cuenta con una interfaz fácil de usar y asistencia al cliente 24 horas al día, 7 días a la semana.

16. Qlik Compose

Qlik Compose es una solución de almacenamiento de datos que diseña almacenes de datos y genera código ETL automáticamente. Esta herramienta automatiza el desarrollo y mantenimiento ETL, tediosos y propensos a errores. Esto acorta el plazo de ejecución de los proyectos de almacenamiento de datos.

Para ello, Qlik Compose ejecuta el código autogenerado, que carga los datos de las fuentes y los traslada a sus almacenes de datos. Estos flujos de trabajo pueden diseñarse y programarse utilizando Workflow Designer and Scheduler.

Qlik Compose también incluye la capacidad de validar los datos y garantizar su calidad. Los profesionales que necesiten datos en tiempo real también pueden integrar Compose con Qlik Replicate.

17. Integrate.io

Integrate.io, antes conocido como Xplenty, se gana un merecido puesto en nuestra lista de las mejores herramientas ETL. Su intuitiva interfaz abre la puerta a una gestión exhaustiva de los datos, incluso para los miembros del equipo con menos conocimientos técnicos. Como plataforma en la nube, Integrate.io elimina la necesidad de instalaciones voluminosas de hardware o software y proporciona una solución altamente escalable que evoluciona con las necesidades de tu empresa.

Su capacidad para conectar con una amplia variedad de fuentes de datos, desde bases de datos hasta sistemas CRM, lo convierte en una opción versátil para diversos requisitos de integración de datos. Al dar prioridad a la seguridad de los datos, ofrece funciones como el cifrado a nivel de campo y cumple normas clave como el RGPD e HIPAA. Con potentes funciones de transformación de datos, los usuarios pueden limpiar, formatear y enriquecer fácilmente sus datos como parte del proceso ETL.

18. Airbyte

Airbyte es una plataforma ELT líder de código abierto. Airbyte ofrece el mayor catálogo de conectores de datos (350 y aumentando) y 40 000 ingenieros de datos lo utilizan desde junio de 2023.

Airbyte se integra con dbt para su transformación de datos y con Airflow/Prefect/Dagster para la orquestación. Tiene una interfaz de usuario fácil de usar y dispone de API y Terraform Provider.

Airbyte se diferencia por su código abierto: se tardan 20 minutos en crear un nuevo conector con su creador de conectores sin programación, y puedes editar cualquier conector comercial, siempre que tengas acceso a su código. Además de su versión de código abierto, Airbyte ofrece una versión alojada en la nube (Airbyte Cloud) y una versión de pago autoalojada (Airbyte Enterprise) para cuando quieras poner en producción tus pipelines.

19. Astera Centerprise

Astera Centerprise es una herramienta ETL/ELT de nivel empresarial y 100 % libre de programación. Como parte de Astera Data Stack, Centerprise cuenta con una interfaz intuitiva con una curva de aprendizaje corta y permite a los usuarios de todos los niveles técnicos construir pipelines de datos en cuestión de minutos.

La herramienta automatizada de integración de datos ofrece una serie de funciones, como conectividad inmediata a varias fuentes y destinos de datos, extracción de datos con IA, asignación automática con IA, transformaciones avanzadas integradas y funciones de calidad de datos. Los usuarios pueden extraer datos estructurados y no estructurados, transformarlos y cargarlos fácilmente en el destino que elijan mediante flujos de datos. Estos flujos de datos pueden automatizarse para que se ejecuten con intervalos, condiciones o filedrops específicos utilizando el programador de trabajos integrado.