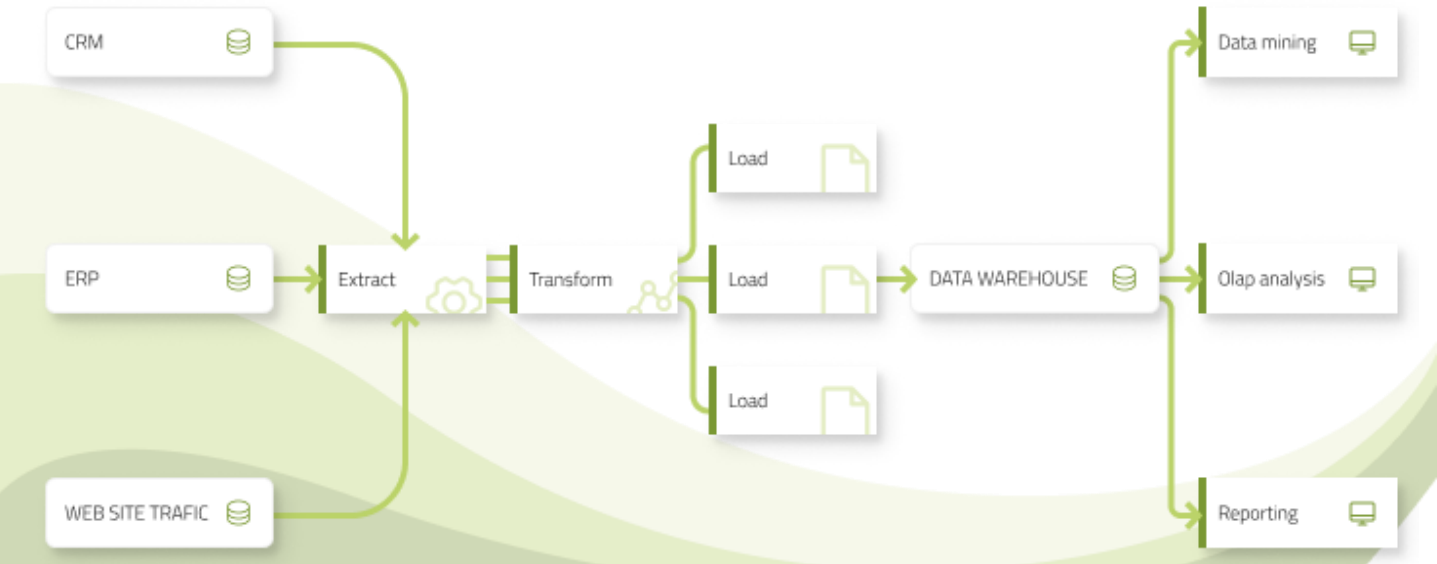
Top Open Source Data Integration Tools

Top 11 aplicaciones, con tutoriales y casos de uso



Emilio Arias

Estas son las mejores herramientas y tecnologías **Open source** que nos permiten **ingestar y transformar datos:**

ETL: Extract, **T**ransform, **L**oad

ELT: Extract, **L**oad, **T**ransform

CDC: Change **D**ata **C**apture

Captura de cambios en los datos en tiempo real a medida que ocurren en las fuentes de datos, lo que permite una integración continua y actualizada.

Federated ETL/ELT

Procesos distribuidos que extraen datos de múltiples fuentes heterogéneas sin consolidarlos en un repositorio central, lo que permite la integración de datos sin la necesidad de replicar grandes volúmenes de datos.

Data Virtualization

Permite acceder y manipular datos de múltiples fuentes como si estuvieran en una única fuente, sin necesidad de replicarlos físicamente.

Data Replication

Replica datos de una fuente a otra para propósitos de integración y respaldo, manteniendo los datos sincronizados entre diferentes sistemas.

Data Wrangling

Herramientas que permiten la limpieza, transformación y manipulación de datos de manera visual e interactiva, facilitando la preparación de datos para su integración.

APIs de Integración de datos

Existen muchas APIs diseñadas para facilitar la integración de datos entre sistemas y aplicaciones. Estas APIs pueden abarcar desde servicios web simples hasta interfaces más complejas.

Orchestration Tools

Herramienta de transformación de datos de código abierto que se utilizan para orquestar y ejecutar transformaciones de datos en el Data Lake.

01 Debezium

Debezium es una **plataforma de código abierto** que facilita la **captura de cambios en bases de datos relacionales** y la **emisión de eventos de cambio en tiempo real**. Utiliza el **log de transacciones de la base de datos** (como el **binlog de MySQL** o el **WAL de PostgreSQL**) para capturar los cambios realizados en los datos. Estos cambios se convierten en eventos que pueden ser consumidos por otras aplicaciones o sistemas en tiempo real.

Debezium proporciona conectores para varias bases de datos, incluyendo **MySQL, PostgreSQL, MongoDB, SQL Server, Oracle**, y otros más. Estos conectores son responsables de **monitorear los cambios en las bases de datos** y **emitir eventos de cambio correspondientes**.

La plataforma Debezium es **muy útil en arquitecturas de sistemas distribuidos** y en aplicaciones que requieren una **replicación de datos en tiempo real**, como **sistemas de procesamiento de eventos, microservicios, materialización de vistas**, y más. Permite **mantener sincronizados distintos sistemas y aplicaciones** de forma eficiente y en tiempo real.

https://debezium.io/

https://todobi.com/change-data-capture-cdc-comparativa-amazon-vs-debezium/

02 Pentaho Data Integration

Pentaho Data Integration (PDI), anteriormente conocido como **Kettle**, es una **herramienta de código abierto para la integración de datos y la transformación de datos** (ETL: Extract, Transform, Load). Es parte de la **suite Pentaho**, que incluye también **herramientas para el análisis de datos, generación de informes y visualización de datos**.

Pentaho Data Integration permite a los usuarios **extraer datos de diversas fuentes**, **transformarlos de acuerdo con las necesidades del negocio** y **cargarlos en un almacén de datos, data lakes, bases de datos relacionales u otros destinos**. Proporciona una **interfaz gráfica intuitiva** para **diseñar y ejecutar flujos de trabajo de transformación de datos sin necesidad de escribir código**, aunque también soporta la **escritura de scripts para operaciones más avanzadas**.

https://www.hitachivantara.com/es-latam/products/pentaho-platform/data-integration-analytics/pentaho-community-edition.html

(disponemos también del código ejecutable de las últimas versiones preservado en nuestros servidores. Contacta con el equipo de Stratebi)

1. https://todobi.com/curso-gratuito-herramientas-analytics-open-source/
2. https://todobi.com/integracion-de-sharepoint-con-pentaho-data-integration/
3. https://todobi.com/sap-connection-tools-for-process-automation-microsoft-pentaho-talend/
4. https://todobi.com/como-integrar-pentaho-mondrian-con-vertica/
5. https://todobi.com/como-hacer-deep-learning-con-pentaho/

03 Talend Open Studio

Talend Open Studio es una **herramienta de integración de datos de código abierto** que permite a los usuarios realizar una **amplia variedad de tareas relacionadas con la gestión de datos**. Es una **suite de software** que ofrece funcionalidades para el **diseño, desarrollo, prueba y despliegue de procesos de integración de datos**.

Talend Open Studio destaca por su **interfaz gráfica intuitiva** que permite a los usuarios **diseñar y ejecutar procesos de integración de datos sin necesidad de escribir código**. Además, ofrece una **amplia variedad de componentes predefinidos y reutilizables** que **simplifican el desarrollo** y **aceleran la implementación de soluciones de integración de datos**.

https://github.com/Talend

(disponemos también del código ejecutable de las últimas versiones preservado en nuestros servidores. Contacta con el equipo de Stratebi)

https://www.youtube.com/watch?v=Y96X1wWsdRo&t=166s

https://todobi.com/sap-connection-tools-for-process-automation-microsoft-pentaho-talend/

https://todobi.com/como-aprender-y-usar-etls-y-data-governance-con-talend/

https://todobi.com/creacion-y-usos-de-apis-con-talend-2/

https://todobi.com/descarga-el-paper-con-tips-para-talend/

04 Apache Airflow

Apache Airflow es una **plataforma de código abierto** utilizada para **programar, monitorear y administrar flujos de trabajo (workflows) de datos**. Fue desarrollado originalmente por **Airbnb** y posteriormente donado a la **Apache Software Foundation**, donde ahora es un **proyecto de alto nivel**.
Airflow permite a los usuarios definir **flujos de trabajo como DAGs (Directed Acyclic Graphs - Grafos Acíclicos Dirigidos)** en **Python**.

Cada **DAG** representa un **conjunto de tareas y las relaciones entre ellas**. Las tareas pueden ser de **diversos tipos**, como **ejecutar scripts de Python**, **invocar comandos de shell**, **ejecutar operaciones SQL**, **enviar correos electrónicos**, entre otras.

https://airflow.apache.org/

https://todobi.com/definiendo-un-data-lake-open-source/

https://todobi.com/disponible-videotutorial-del-workshop-business-intelligenceopen-source/

https://todobi.com/que-es-apache-airflow/

05 Apache Kafka

Apache Kafka es una **plataforma de streaming distribuida de código abierto** utilizada para la **transmisión de datos a gran escala en tiempo real**. Fue desarrollada originalmente por **LinkedIn** y posteriormente se convirtió en un **proyecto de Apache Software Foundation**.

Kafka se basa en un **modelo de publicación-suscripción** y está diseñado para **manejar flujos de datos en tiempo real** de manera **eficiente y confiable**. Es **altamente escalable y tolerante a fallos**, lo que lo hace adecuado para **entornos empresariales** que necesitan **procesar grandes volúmenes de datos** de manera **eficiente y sin perder información**.

https://kafka.apache.org/

https://todobi.com/how-a-modern-data-architecture-works/

https://todobi.com/caso-de-uso-de-apache-kafka-en-tiempo/

06 Apache Beam

Apache Beam es un **modelo de programación unificado** y una **API de procesamiento de datos distribuidos de código abierto**, desarrollado por **Google** y posteriormente trasladado al proyecto de **Apache Software Foundation**. Proporciona una forma unificada de **definir y ejecutar pipelines de procesamiento de datos** que pueden ejecutarse en múltiples motores de ejecución, como **Apache Flink, Apache Spark, Google Cloud Dataflow**, entre otros.

El objetivo principal de Apache Beam es **permitir el desarrollo de pipelines de procesamiento de datos de manera portable**, es decir, que puedan ejecutarse en distintos motores de procesamiento sin necesidad de realizar cambios en el código fuente. Para lograr esto, **Apache Beam define un modelo de programación unificado** que separa la **lógica del pipeline** de los **detalles de ejecución del motor subyacente**.

https://beam.apache.org/

https://todobi.com/apache-beam-introduccion/

https://todobi.com/dagster/

07 Apache Spark

Apache Spark es un **potente motor de procesamiento de datos distribuido y de código abierto** diseñado para realizar **análisis avanzados de datos de manera eficiente y escalable**. Fue desarrollado inicialmente en la **Universidad de California, Berkeley**, y ahora es un **proyecto de alto nivel de la Apache Software Foundation**.

Spark ofrece una **amplia gama de capacidades de procesamiento de datos**, incluyendo **procesamiento de datos en lotes (batch processing)**, **procesamiento de datos en tiempo real (stream processing)** y **procesamiento de datos de máquina a máquina (machine-to-machine processing)**.

- https://spark.apache.org/
- https://todobi.com/machine-learning-demo-tutorial/
- https://todobi.com/diccionario-de-arquitecturas-de-datos/
- https://todobi.com/c/
- https://todobi.com/25-tecnologias-que-necesitas-en-una-moderna-arquitectura-de-datos/

08 Apache HOP

Apache Hop, también conocido como **"Hop Orchestration"** y la evolución de **"Pentaho Data Integration" (PDI)**, es una **plataforma de integración de datos de código abierto** y un **proyecto de nivel superior de Apache Software Foundation**. Se originó como parte de la **suite Pentaho**, pero se ha convertido en un **proyecto independiente bajo el paraguas de Apache**.

Apache Hop ofrece capacidades de **extracción, transformación y carga (ETL) de datos**, similar a las que se encuentran en otras herramientas como **Talend** y **Apache NiFi**. Permite a los usuarios **crear flujos de trabajo para mover y transformar datos entre distintas fuentes y destinos**, lo que es fundamental para la **integración de datos en entornos empresariales y de Big Data**.

https://hop.apache.org/

El equipo de Stratebi realiza cursos completos y prácticos de Apache Hop. Pregúntales.

https://todobi.com/conociendo-hop-etl-open-source/

https://todobi.com/deberia-migrar-de-pentaho-data-integration-a-apache-hop/

https://todobi.com/apache-hop-instalacion-y-construccion-de-pipelines-con-dapr/

09 Airbyte

Airbyte es una **plataforma de código abierto para la integración de datos** que permite a las empresas **mover datos de una variedad de fuentes a un almacén de datos de manera eficiente y confiable**.

Fue creado con el objetivo de **facilitar el proceso de integración de datos** y **eliminar las complejidades asociadas con la creación y mantenimiento de canalizaciones de datos**.

https://airbyte.com/

https://todobi.com/comparando-airbyte-con-azure-data-factory/

10 dbt

dbt (abreviatura de **"Data Build Tool"**) es una herramienta de **transformación de datos** de código abierto y orientada a **SQL**. Su objetivo principal es permitir a los **analistas** y **científicos de datos** definir y ejecutar transformaciones de datos de manera **eficiente** y **reproducible**.

A diferencia de las herramientas tradicionales de **ETL** (Extract, Transform, Load), **dbt** se enfoca principalmente en la **capa de transformación de datos** y se integra bien con **almacenes de datos modernos** como **BigQuery**, **Snowflake**, **Redshift** y otros.

dbt se centra en la definición de **modelos de datos** y **transformaciones** utilizando SQL y sigue un enfoque de **"transformación en lugar de carga"**, lo que significa que los datos se transforman directamente en el almacén de datos en lugar de pasar por un proceso de **transformación intermedio**.

- https://www.getdbt.com/
- https://todobi.com/dbt-en-una-moderna-arquitectura-de-datos/
- https://todobi.com/prefect-data-orchestration-platform/

11 Windmill

Windmill Dev es una herramienta de **desarrollo** y **prueba** de código abierto diseñada para facilitar la **creación**, **ejecución** y **depuración** de pruebas de **interfaz de usuario (UI)** y **end-to-end (E2E)** para **aplicaciones web**. Es una herramienta relativamente **nueva** que se está desarrollando **activamente** y está diseñada para ser **fácil de usar** y **flexible** para los **desarrolladores** y **probadores de software**.

Convierte los **scripts** en **interfaces de usuario autogenerados**, **APIs** y **cron jobs**. Compone como **flujos de trabajo** o **data pipelines** y crea fácilmente **interfaces de usuario** complejos y con **muchos datos**.

https://www.windmill.dev/

Se trata de una de las herramientas de orquestación más recientes y que desde esta cuenta más nos gusta. Es espectacular. La estoy empezando a usar en proyectos con unos resultados increíbles.

El equipo de Stratebi prepara workshops y casos de uso.