Explora, Transforma, Logra

¡Desmitificando ETL en las vacantes de trabajo!

Agenda

1. Introducción a ETL

2. El Flujo de Datos en una Organización

3. Roles en el Mundo de los Datos

4. Ejemplo de ETL

Con qué puedo ayudar ?

Un poco sobre mi.

Hesus García: DevOps Engineer

Analista de datos 🚀 con 5 años de experiencia en consultoría IT y análisis de datos. Ayudo a interpretar datos y elevar su calidad para mejorar la toma de decisiones. ✅

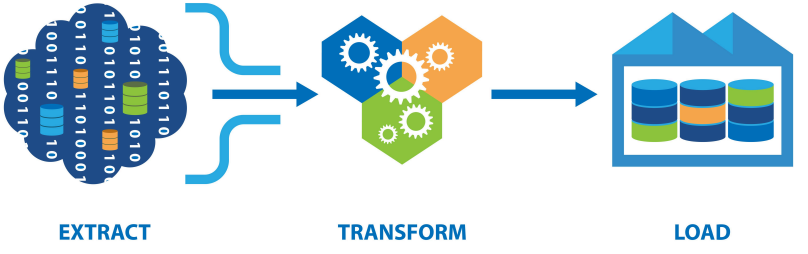
GitHub LinkedIn Mi Sitio Web

Introducción a ETL

¿Qué es ETL?

ETL es el acrónimo de Extract, Transform, Load (Extraer, Transformar, Cargar). Estos tres procesos forman una cadena de pasos utilizada para mover datos desde una o más fuentes a un destino, como puede ser un data warehouse.

- Extraer: Obtener datos de diversas fuentes.
- Transformar: Limpiar y organizar los datos según las necesidades.
- Cargar: Mover los datos transformados a su destino final.



La Analogía del Proceso de Ensamblaje en ETL

Imagina el ETL como una **línea de ensamblaje** en una fábrica, donde:

- **Extraer** equivale a seleccionar los recursos brutos necesarios.
- **Transformar** es como construir y modificar esos recursos para crear un producto final.
- **Cargar** representa entregar el producto terminado a su destino, listo para su uso o venta.

Así como un automóvil necesita ser ensamblado antes de llegar al consumidor, los datos requieren un proceso de ETL antes de ser analizados o utilizados en la toma de decisiones.

Ejemplo Práctico de ETL: Escenario Escolar

Consideremos una escuela que necesita consolidar y analizar datos de diferentes fuentes para mejorar la gestión de cursos y el seguimiento de los alumnos:

- Datos de inscripción de alumnos en diferentes cursos.
- Información sobre el rendimiento académico y asistencia de los estudiantes.

El objetivo es tener una visión integral que permita una mejor planificación educativa y seguimiento del progreso estudiantil.

Detalle del Proceso ETL: Escenario Escolar

En nuestro ejemplo de la escuela, el proceso ETL se desarrolla de la siguiente manera:

1. **Extraer:** Los datos se extraen de sistemas de gestión académica y plataformas de aprendizaje en línea, los cuales pueden estar en formatos diversos como bases de datos, hojas de cálculo o incluso informes en papel.

2. **Transformar:** Esta etapa convierte los datos extraídos en un formato estandarizado. Por ejemplo, unificar los formatos de fecha, consolidar listas de alumnos inscritos y rendimiento académico en una estructura tabular coherente, y limpiar los datos de alumnos duplicados o incorrectos.

3. **Cargar:** Los datos transformados se cargan en un sistema centralizado, como una base de datos escolar, donde pueden ser accedidos para generar informes de rendimiento por curso, analizar tendencias de inscripción, y monitorear la asistencia y el progreso de los estudiantes.

¡Sigan Adelante!

¡Lo están haciendo increíblemente bien! Estamos aprendiendo juntos y avanzando paso a paso en el mundo de ETL. Recuerden, cada experta alguna vez fue principiante.

Ahora, pongamos a prueba lo que hemos aprendido con un par de preguntas:

¿Qué significa ETL?

A) Extraer, Transformar, Listar

B) Entrar, Tocar, Leer

C) Extraer, Transformar, Cargar

D) Enviar, Traducir, Lanzar

Pueden responder usando el chat de zoom. Recuerden seleccionar la opción que consideren correcta. ¡Estamos aquí para aprender y crecer!

¿En qué etapa del proceso ETL se limpian los datos?

A) Extraer

B) Transformar

C) Cargar

D) Ninguna de las anteriores

Pueden responder usando el chat de zoom. Recuerden seleccionar la opción que consideren correcta. ¡Estamos aquí para aprender y crecer!

Roles en el Mundo de los Datos

Descubre quién hace qué en el ecosistema de datos.

El Rol del Analista de Datos

Responsables de interpretar los datos para ofrecer insights y recomendaciones accionables.

Los analistas de datos son responsables de interpretar los datos para ofrecer insights y recomendaciones accionables.

- **Habilidades:** Dominio de SQL para la extracción y manipulación de datos, capacidad para realizar análisis estadísticos avanzados y experiencia en visualización de datos para comunicar resultados de manera efectiva.
- **Objetivo:** Ayudar a la organización a tomar decisiones informadas y estratégicas basadas en datos sólidos y análisis detallados.
- **Herramientas:** Uso experto de herramientas como Power BI y Looker Studio para análisis y visualización de datos, así como Google BigQuery para el procesamiento de grandes volúmenes de datos.

Los analistas de datos convierten los datos en historias que impulsan el cambio organizacional y facilitan la toma de decisiones informadas.

El Rol del Científico de Datos

Especialistas que utilizan métodos científicos para modelar y entender complejos conjuntos de datos.

- **Habilidades:** Programación (Python/R), machine learning, estadística avanzada.
- **Objetivo:** Crear modelos predictivos y algoritmos para extraer insights.
- **Herramientas:** JupyterNotebook, TensorFlow, SciKit Learn.

Los científicos de datos buscan patrones y conexiones ocultas para predecir futuros comportamientos.

El Rol del Ingeniero de Datos

Los ingenieros de datos son fundamentales en la construcción y mantenimiento de la infraestructura necesaria para el almacenamiento, procesamiento y análisis de grandes volúmenes de datos.

- **Habilidades clave:**
  - Sistemas de bases de datos
  - ETL (Extract, Transform, Load)
  - Arquitectura de datos
- **Objetivo principal:**
  - Garantizar que los datos sean accesibles y estén en un formato listo para ser analizado.
- **Herramientas comunes:**
  - Hadoop
  - Spark
  - Kafka

Flujos de Datos y Pipelines

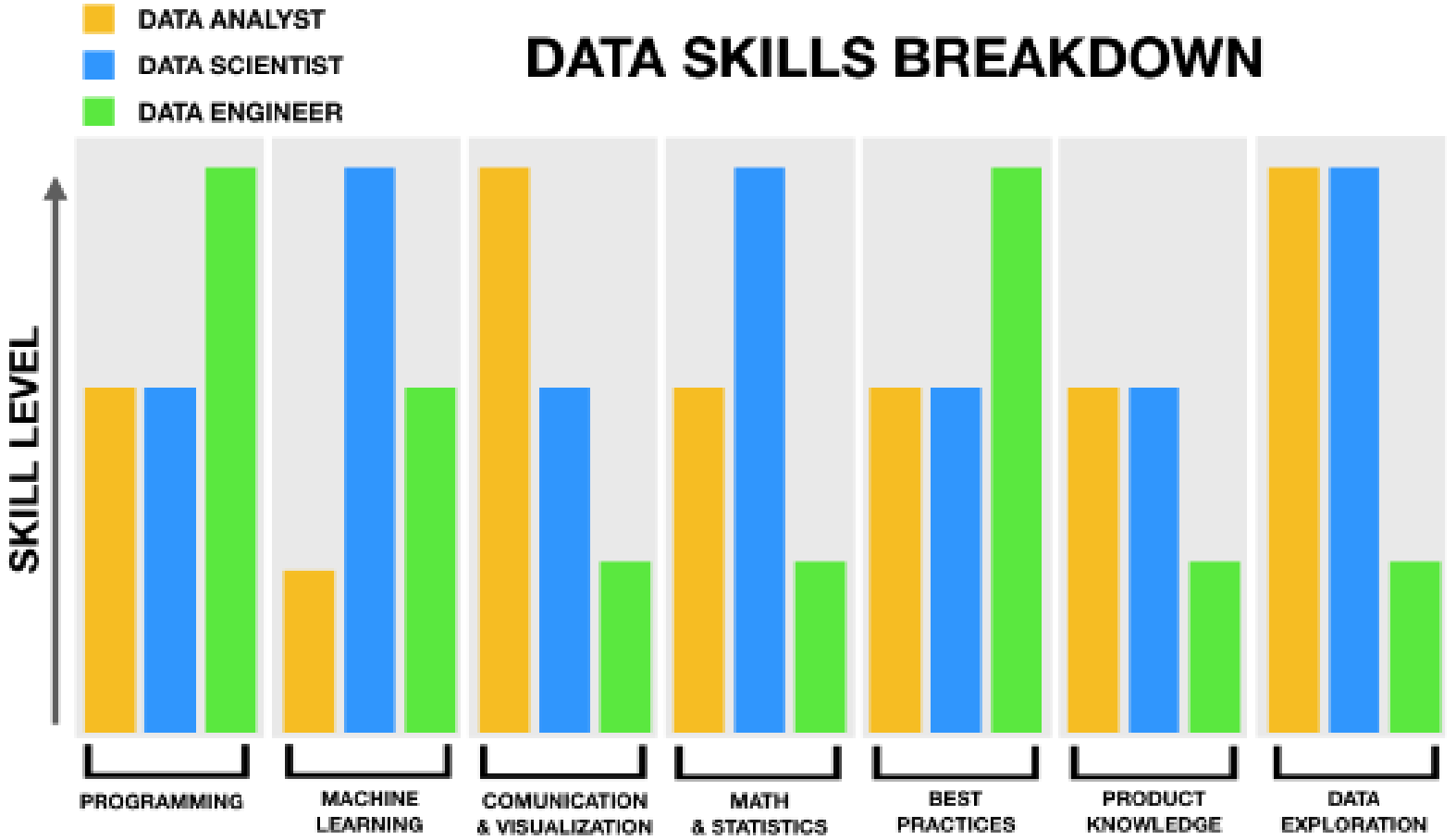
Los ingenieros de datos crean y mantienen los flujos de datos y pipelines, esenciales para asegurar que los datos sean accesibles, confiables y listos para el análisis.

Un pipeline en la ingeniería de datos es:

- Una serie de procesos automatizados.
- Mueve y transforma datos desde su origen hasta un destino final para análisis o almacenamiento.
- Fundamental para el manejo eficiente y ordenado de datos en ciencia de datos y la ingeniería de datos.

Esto permite un flujo constante de datos a través de diferentes etapas, facilitando el análisis y la toma de decisiones basadas en datos.

Diagrama de Skills en Data



Interacción del Analista de Datos con ETL

Los analistas de datos desempeñan roles variados en ETL, influenciados por:

- **Madurez organizacional:** En firmas bien establecidas, se centran más en análisis. En startups, pueden abarcar todo el proceso ETL.
- **Herramientas tecnológicas:** El stack tecnológico (ej., Google BigQuery, Airflow) determina su grado de implicación técnica.
- **Objetivos del proyecto:** La complejidad dicta si su enfoque es más hacia la transformación de datos o análisis puro.
- **Colaboración:** La interacción con equipos de ingeniería o ciencia de datos puede ampliar su rol en la calidad de datos.

Un modelo ETL

Un Vistazo desde la Trinchera del Análisis de Datos

Contexto

Una empresa de educación en línea busca optimizar sus servicios y ofertas educativas analizando el comportamiento de los usuarios, el rendimiento de los cursos, y las tendencias de inscripción. Utiliza un proceso ETL en Google BigQuery, SQL de BigQuery, Google Sheets, y Google Analytics, para alimentar dashboards en Looker Studio y Power BI, con replicación en otros sistemas mediante Servicios Comunes de Windows.

La compañía gestiona datos de diversas fuentes:

- **Google Analytics:** Datos del comportamiento de usuarios en el sitio y plataforma.
- **Google Sheets:** Registro de campañas promocionales y inscripciones manuales.
- **Moodle:** Como LMS, detalla actividades y progreso de los estudiantes.

Fase de Extracción

En esta fase, se recopilan datos de múltiples fuentes para prepararlos para su análisis:

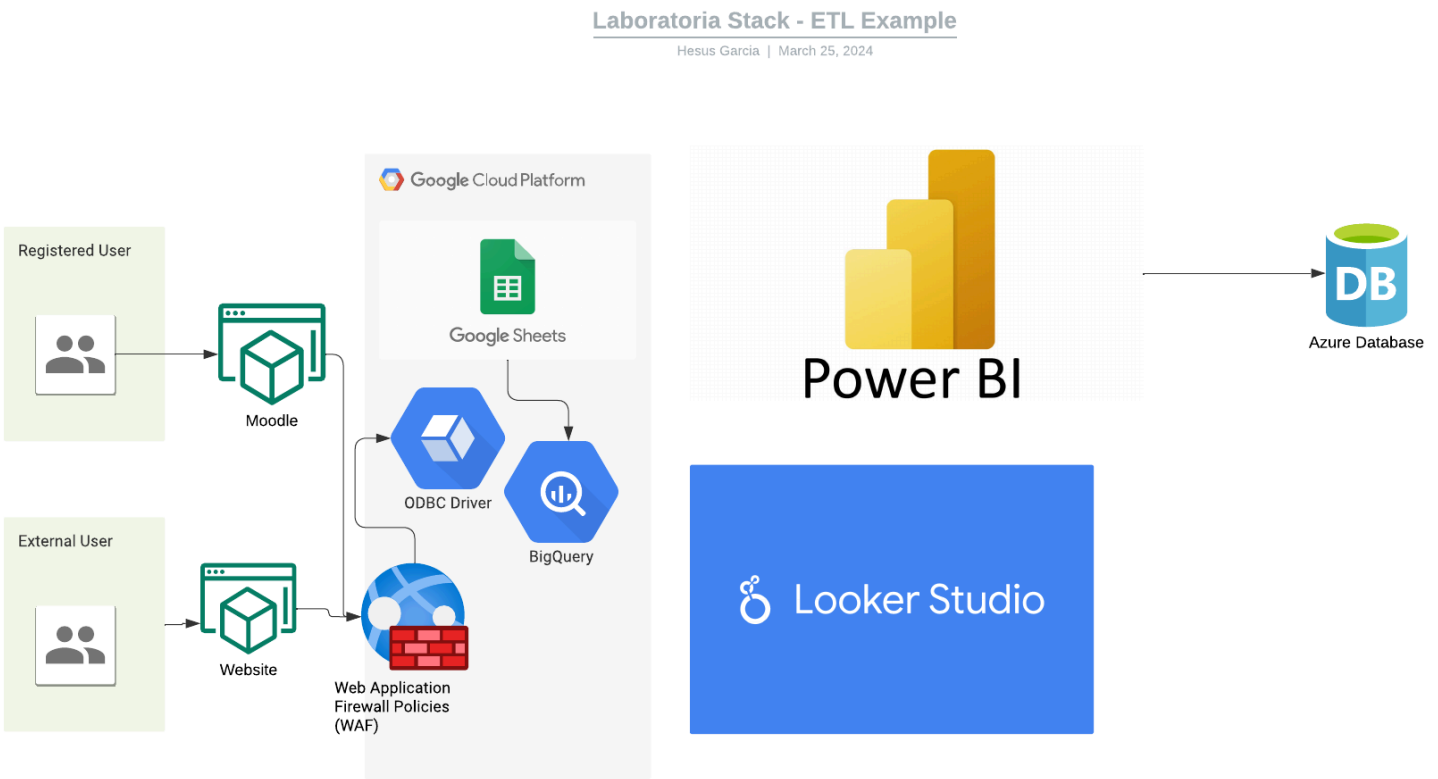
- **De Google Analytics:** Se utilizan las API de Google Analytics para extraer datos sobre las interacciones de los usuarios en la plataforma, como sesiones, duración de visitas, páginas vistas, y conversiones.
- **De Google Sheets:** Se aprovecha la integración de Google BigQuery con Google Sheets para extraer datos de forma directa, capturando información sobre inscripciones en eventos especiales y campañas promocionales.
- **De Moodle:** Se extraen datos completos sobre la participación de los estudiantes, incluyendo detalles de actividades, resultados de exámenes, y progreso en los cursos, lo cual es crucial para entender el rendimiento de los cursos y la eficacia de los materiales didácticos.

Fase de Transformación

En esta etapa crítica, los datos brutos se transforman en información valiosa mediante SQL de BigQuery:

- **Limpieza de Datos:** Se normalizan formatos, se corrigen errores, y se eliminan duplicados para asegurar la calidad y la coherencia de los datos.
- **Enriquecimiento de Datos:** Se combinan los datos de diferentes fuentes, como Google Analytics y Moodle, para obtener una visión holística del comportamiento del usuario, desde la visita inicial hasta la inscripción en cursos y su progreso académico.
- **Aggregación:** Se calculan métricas clave como las tasas de inscripción por curso, el promedio de duración de las sesiones, la tasa de finalización de cursos, y la efectividad de las campañas promocionales, proporcionando insights críticos para la toma de decisiones.

Diagrama de Arquitectura



Fase de Carga

Alimentación de Dashboards

- **Looker Studio:** Utiliza una conexión directa con BigQuery para visualizar las métricas clave, facilitando al equipo de producto y educativo la comprensión del rendimiento de los cursos y el engagement de los estudiantes.
- **Power BI:** Similarmente, Power BI se conecta a BigQuery para crear un dashboard dirigido al equipo de marketing y ventas, centrado en las métricas de rendimiento de las campañas y las tasas de conversión. Este dashboard se replica o se carga en otro sistema para ampliar la accesibilidad empresarial, utilizando los Servicios Comunes de Windows para su integración.

Fase de Carga

Integración Financiera

- Automatizar la replicación del dashboard de Power BI a una base de datos específica del departamento de finanzas. Esto permite un análisis financiero más detallado del rendimiento de los cursos y la efectividad de las campañas de marketing, facilitando la toma de decisiones basada en datos para presupuestos y asignaciones de recursos.

Análisis Avanzado

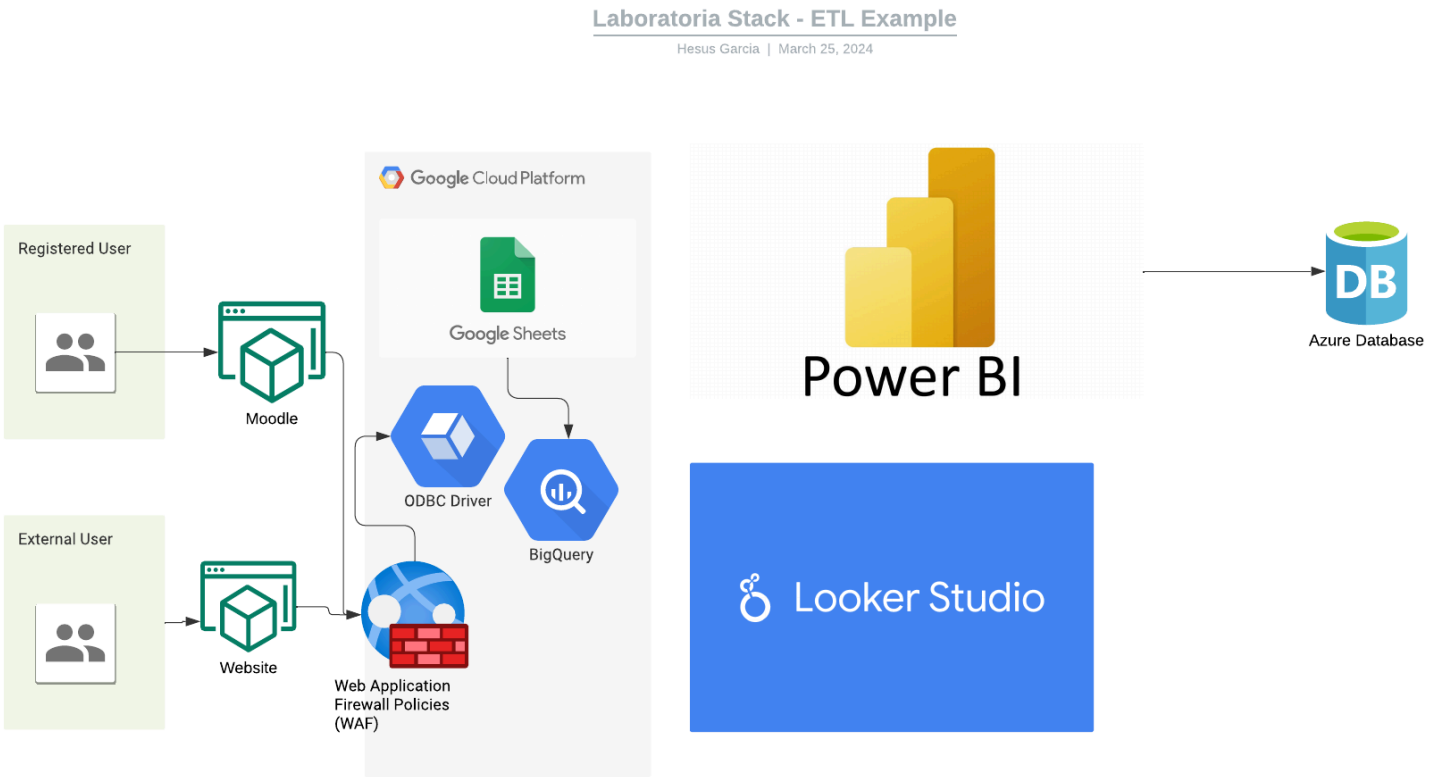
- Crear réplicas de las bases de datos en entornos dedicados a la investigación y el desarrollo de modelos predictivos. Esto puede incluir la creación de bases de datos para el análisis de tendencias a largo plazo, predicciones de inscripciones, y la optimización de estrategias de marketing y contenido educativo.
- Estas bases de datos pueden ser diseñadas para soportar análisis complejos, como machine learning y minería de datos, ofreciendo insights profundos y acciones recomendadas para mejorar la oferta educativa y la experiencia del usuario.

Herramientas Comunes

Es importante destacar que ETL (Extract, Transform, Load) es más un concepto que una herramienta específica

- Low-Code
  - Google Cloud DataFlow (AWS Glue , Azure Data Factory)
  - Pentaho
  - DataStage
  - Herramientas de integración como SSIS, ODI
- Prográmaticas
  - Python
  - R
  - Airflow
  - dbt
  - Herramientas de integración como SSIS, ODI (que soporten scripting)

Diagrama de Arquitectura



Preguntas