Tema 1. Computación en la nube en la era del big data y la inteligencia artificial

Servicios *big data* e IA de AWS

AWS ofrece una amplia gama de servicios especializados en *big data* e IA. Estos servicios permiten a las organizaciones recopilar, almacenar, procesar y analizar grandes volúmenes de datos, así como construir y desplegar modelos avanzados de IA para una variedad de aplicaciones. A continuación, profundizaremos en ellos:

- ***Big data.*** En la Figura 8 se muestra la amplia gama de servicios de *big data* e IA ofrecidos por esta plataforma.

Figura 8. Catálogo AWS *big data*. Fuente: Kyle Escosia, 2021.

- **AWS Glue.** Es un servicio de ETL (*extract, transform, load*) completamente administrado que les permite a las organizaciones descubrir, preparar y cargar datos de manera eficiente para el análisis. Proporciona capacidades de catalogación de datos, generación de código ETL automático y ejecución de trabajos de transformación de datos a escala. Las características clave son las siguientes:

  - **Catalogación de datos.** AWS Glue ofrece un catálogo de datos centralizado que les permite a los usuarios descubrir y entender fácilmente los datos disponibles en la organización.

  - **Generación de código ETL automático.** AWS Glue puede generar automáticamente código ETL (*extract, transform, load*) para transformar y preparar datos para el análisis. Utiliza un motor de generación de código basado en la lógica de transformación definida por el usuario y los metadatos del catálogo de datos.

  - **Ejecución de trabajos de transformación de datos.** AWS Glue les permite a los usuarios ejecutar trabajos de transformación de datos en un entorno completamente administrado y escalable. Utiliza Apache Spark como motor de ejecución para procesar grandes volúmenes de datos de manera rápida y eficiente.

  - **Integración con otros servicios de AWS.** AWS Glue se integra estrechamente con otros servicios de AWS, incluidos S3, Redshift, Athena y más. Esto les permite a los usuarios cargar datos desde diversas fuentes, realizar transformaciones complejas y cargar los datos transformados en almacenes de datos y sistemas de análisis.

- **AWS Kinesis.** Es una plataforma de *streaming* de datos completamente administrada que les permite a las organizaciones recopilar, procesar y analizar datos en tiempo real. Proporciona capacidades para la ingesta de grandes volúmenes de datos en tiempo real, procesar datos en tiempo real y almacenar datos durante períodos prolongados. Las características clave son las siguientes:

  - **Ingesta de datos en tiempo real y procesamiento de datos en tiempo real.** AWS Kinesis les permite a los usuarios la ingesta de grandes volúmenes de datos en

tiempo real desde una variedad de fuentes, incluidos los dispositivos IoT, las aplicaciones móviles y los sistemas de registro. Proporciona capacidades para procesar flujos de datos en tiempo real y almacenar datos en *streams* duraderos.

- Además, AWS Kinesis se integra con **otros servicios de AWS,** como S3, Redshift, DynamoDB y más. Esto les permite a los usuarios procesar datos en tiempo real y almacenar datos en *streams* duraderos para su análisis y uso futuro.

- **Amazon S3 (Simple Storage Service).** Es un servicio de almacenamiento en la nube altamente escalable y duradero que les permite a las organizaciones almacenar grandes volúmenes de datos de manera segura y accesible desde cualquier lugar. Es ideal para almacenar datos estructurados, semiestructurados y no estructurados, incluidos archivos de registros, imágenes, vídeos y más. Las características clave son la durabilidad, la escalabilidad, el cifrado de datos y la integración con otros servicios de AWS, como Athena, Glue y Redshift.

- **Amazon EMR (Elastic MapReduce).** Es un servicio de procesamiento de datos distribuido basado en Apache Hadoop y Apache Spark que permite ejecutar y escalar fácilmente clústeres de procesamiento de datos en la nube. Las características clave son la escalabilidad automática, el soporte para una variedad de motores de procesamiento de datos (Hadoop, Spark, Presto, Hive) y la integración con otros servicios de AWS, como S3, DynamoDB y Redshift.

- **Amazon Redshift.** Es un servicio de almacenamiento de datos en la nube diseñado para la creación y análisis de *data warehouses* a escala petabyte. Permite que las organizaciones analicen grandes volúmenes de datos de manera rápida y eficiente utilizando SQL estándar. Las características clave son el alto rendimiento, la escalabilidad masiva, la integración con herramientas de análisis de datos, como Tableau, Looker y Power BI.

Tema 1. Computación en la nube en la era del big data y la inteligencia artificial

En el área de la inteligencia artificial, la plataforma AWS ofrece los siguientes servicios:

- **Amazon SageMaker.** Es un servicio completamente gestionado que les permite a los científicos de datos y desarrolladores construir, entrenar y desplegar modelos de aprendizaje automático de manera rápida y sencilla.
  - **Características clave.** Los servicios de entrenamiento y despliegue automático de modelos, la integración con marcos de aprendizaje automático, como TensorFlow y PyTorch, y la capacidad de escalar y gestionar flujos de trabajo de aprendizaje automático de extremo a extremo.
- **Amazon Rekognition.** Es un servicio de visión artificial que permite analizar imágenes y vídeos para detectar y reconocer objetos, rostros, texto y más.
  - **Características clave.** Detección y reconocimiento de objetos y rostros en imágenes y vídeos, análisis de contenido visual en tiempo real, integración con otros servicios de AWS.
- **Amazon Comprehend.** Es un servicio de procesamiento de lenguaje natural (NLP) que permite analizar y extraer información de texto no estructurado.
  - **Características clave.** Análisis de sentimientos, detección de entidades y frases clave, clasificación de documentos, integración con otros servicios de AWS.
- **Amazon BedRock.** Es un servicio totalmente administrado que ofrece una selección de modelos fundacionales (FM) de alto rendimiento de las principales empresas de IA, como AI21 Labs, Anthropic, Cohere, Meta, Mistral AI, Stability AI y Amazon, a través de una sola API, junto con un amplio conjunto de funciones necesarias para crear aplicaciones de IA generativa con seguridad, privacidad e IA responsables.
  - **Características clave.** Con Amazon Bedrock, puede experimentar y evaluar con facilidad los mejores FM para su caso de uso, personalizarlos de forma privada con sus datos mediante técnicas como el ajuste y la generación aumentada de

recuperación (RAG), y crear agentes que ejecuten tareas utilizando los sistemas y orígenes de datos de su empresa. Dado que Amazon Bedrock no tiene servidores, no se tiene que administrar ninguna infraestructura y puede integrar e implementar de forma segura capacidades de IA generativa en sus aplicaciones mediante los servicios de AWS que ya conoce.

Estos son solo algunos ejemplos de los servicios de AWS en *big data* e IA. AWS ofrece una amplia gama de herramientas y servicios que permiten que las organizaciones aprovechen la potencia de la nube para gestionar datos a gran escala y construir soluciones avanzadas de inteligencia artificial para una variedad de aplicaciones.

Tema 1. Computación en la nube en la era del big data y la inteligencia artificial

1.4. Referencias bibliográficas

Amazon Web Services. (2017, noviembre 29). *AWS re:Invent 2017: #EarthonAWS: How NASA Is Using AWS (STG205)* [Vídeo]. YouTube. https://youtu.be/Sh7FB-tkYXM

Amazon Web Service. (2018a, noviembre 28). *AWS re:Invent 2018: Airbnb's Journey from Self-Managed Redis to ElastiCache for Redis (DAT319)* [Vídeo]. YouTube. https://youtu.be/eyd_8efUCwM

Amazon Web Service. (2018b, agosto 24). *Capital One Reimagines Banking Using AWS* [Vídeo]. YouTube. https://youtu.be/qU0HuWtzDC4

Amazon Web Services. (2021, abril 21). *Accelerating Enterprise-Wide AI/ML Innovation: GE Healthcare and Autodesk* [Vídeo]. YouTube. https://youtu.be/0vWIAOYkjRk

Amazon Web Services. (2024, abril 12). *AWS re:Invent 2023 - Pinterest extends existing data lake with generative AI | AWS Events* [Vídeo]. YouTube. https://youtu.be/p6B7QX2osRU

Amazon Web Service. (s. f.-a). *2024 Gartner Report. Magic Quadrant for Cloud AI Developer Services*. https://pages.awscloud.com/GLOBAL-multi-DL-gartner-mq-cips-2020-learn.html?pg=WIAWS

Amazon Web Services. (s. f.-b). *AWS Documentation.* https://docs.aws.amazon.com/es_es/

Amazon Web Services. (2015, febrero 27). *Netflix Delivers Billions of Hours of Content Globally by Running on AWS* [Vídeo]. YouTube. https://youtu.be/lQGHsBOZJBw

Tema 1. Computación en la nube en la era del big data y la inteligencia artificial

Andersson, J. C. (2023). *Learning Microsoft Azure: Cloud Computing and Development Fundamentals.* O'Reilly Media, Inc.

Bhatia, J. y Chaudhary, K. (2023). *The Definitive Guide to Google Vertex AI: Accelerate your machine learning journey with Google Cloud Vertex AI and MLOps best practices*. Packt Publishing.

California Privacy Protection Agency. (2024, Agosto). *California Consumer Privacy Act of 2018*. https://cppa.ca.gov/regulations/pdf/cppa_act.pdf

Eagar, G. (2023). *Data Engineering with AWS - Second Edition*. Packt Publishing.

Google Cloud. (2018, septiembre 10). *PayPal partners with Google Cloud* [Vídeo]. YouTube. https://youtu.be/9jJ6xLOSS3c

Google Cloud. (2020, julio 14). *Spotify uses Google Cloud to unlock infinite capacity and faster innovation* [Vídeo]. YouTube. https://youtu.be/55xgR_o4PGs

Google Cloud. (s. f.). *Documentación de Google Cloud.* https://cloud.google.com/docs?hl=es-419

Inderpreet Singh. (2024, marzo 14). *Beyond the Hype: Choosing the Right AI Platform for You — Vertex AI vs. AWS SageMaker*. Medium. https://medium.com/@inderpreet.khanduja/beyond-the-hype-choosing-the-right-ai-platform-for-you-vertex-ai-vs-aws-sagemaker-871beb47d909

Johnson Controls. (2023, junio 27). *Johnson Controls wins 2023 Microsoft Global ISV Partner of the Year for OpenBlue Building Solutions* [Vídeo]. YouTube. https://youtu.be/8qHLAZUIMXk

Kozlovski, S. (2018, abril 27). *A Thorough Introduction to Distributed Systems. Free Code Camp*. FreeCodeCamp. https://www.freecodecamp.org/news/a-thorough-introduction-to-distributed-systems-3b91562c9b3c/

Krzyzanowski, P. (2018). *Distributed Systems. 27. Engineering Distributed Systems*. Rutgers University. https://people.cs.rutgers.edu/~pxk/417/notes/content/27-engineering-ds-slides.pdf

Kyle Escosia. (2021, mayo 29). *Introduction to the AWS Big Data Portfolio* [Entrada de un foro]. AWS Community ASEAN. https://dev.to/awscommunity-asean/introduction-to-the-aws-big-data-portfolio-2539

Lee, R. (2018). *Big data, Cloud computing, Data Science & Engineering.* Springer.

Mahmood, Z., Puttini, R. y Erl, T. (2013). *Cloud computing: Concepts, Technology & Architecture.* Pearson.

Microsoft Learn. (2023, marzo 13). *What is a data mesh?* https://learn.microsoft.com/en-us/azure/cloud-adoption-framework/scenarios/cloud-scale-analytics/architectures/what-is-data-mesh

Microsoft Learn. (s. f.). *Documentación de Azure.* https://learn.microsoft.com/es-es/azure/?product=popular

Mr Anand M. (2024). *Microsoft Azure AI Fundamentals* [libro autopublicado].

Paganini, C. (2019, octubre 23). Primer: Distributed Systems and Cloud Native Computing. *The New Stack*. https://thenewstack.io/primer-distributed-systems-and-cloud-native-computing/

Rautenstrauch, R. (2023, septiembre 30). *Integración de OpenAI en el universo Azure.* Consultor365. https://www.consultor365.com/ia/integracion-openai-universo-azure/

Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos). *Diario Oficial de la Unión Europea*, de 27 de abril de 2016. https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=celex%3A32016R0679

Roland Berger. (2022, junio 1). *Microsoft Intelligent Manufacturing Award 2021 – BMW* [Vídeo]. YouTube. https://youtu.be/bACcdP7kz5A

Sanjay, M., Kumar Tyagi, A., Pluri, V. y Garg, L. (2022). *Artificial Intelligence for Cloud and Edge Computing.* Springer.

Srinivasan, S. (2019, junio 27). *Data and Analytics on Google Cloud Platform*. Medium. https://medium.com/@srivatsan88/data-and-analytics-on-google-cloud-platform-13bc92a4596f

Subramanian, S. y Nathu, S. (2021). *AWS Certified Machine Learning: Specialty.* Editorial Sybes.

Universidad Internacional de Valencia. (2017, febrero 22). *Sistemas distribuidos, características y clasificación*. https://www.universidadviu.com/es/actualidad/nuestros-expertos/sistemas-distribuidos-caracteristicas-y-clasificacion

Tema 2. Ingesta de datos para big data e inteligencia artificial en cloud

2.1. Introducción y objetivos

Considerando la importancia del *cloud computing* en el desarrollo e implementación de soluciones de IA y *big data*, los **objetivos** que persigue este tema son los siguientes:

- Entender el ciclo de vida del dato en el *big data* y la IA.
- Identificar las etapas que componen este ciclo de vida.
- Profundizar en la etapa de ingesta de datos, explicando principales tipologías, paradigmas y desafío.
- Conocer el principal servicio de AWS para ingesta de datos *batch*: AWS Glue.
- Aprender los principales componentes de AWS Glue: Data Catalog, Glue Studio, Glue Notebooks y Glue Pipelines.
- Aprender el funcionamiento de AWS Glue Data Catalog para inferir esquemas y crear un catálogo de metadatos.
- Entender el funcionamiento de AWS Glue Studio para desarrollar procesos ETL.
- Conocer el funcionamiento de AWS Glue Notebooks para desarrollar procesos de ETL en la ingesta de datos.

Tema 2. Ingesta de datos para big data e inteligencia artificial en cloud

2.2. Ingesta de datos batch

Ciclo de vida del dato en el big data y la IA

El **ciclo de vida de los datos** en *big data* e IA es un proceso que abarca desde la recolección inicial de los datos hasta la implementación y la mejora continua de modelos de IA. Estas etapas suelen estar interconectadas y pueden variar según el contexto y los requisitos específicos del proyecto, pero, a continuación, hay una descripción general de las etapas comunes:

- **Adquisición de datos.** En esta etapa, se recopilan los datos de diversas fuentes, que pueden incluir bases de datos empresariales, archivos de registros, datos de sensores, redes sociales, transacciones en línea, entre otros. Es crucial asegurarse de que los datos sean relevantes, precisos, completos y estén disponibles en el formato adecuado para su procesamiento posterior.

- **Almacenamiento de datos.** Los datos recopilados se almacenan en un sistema de almacenamiento adecuado, que puede ser un sistema de archivos distribuido, como Hadoop Distributed File System (HDFS), un almacén de datos en la nube, como Amazon S3 o Google Cloud Storage, o una base de datos distribuida, como Apache Cassandra o MongoDB. Es importante diseñar un esquema de almacenamiento eficiente que permita el acceso rápido a los datos y pueda escalar según sea necesario.

- **Procesamiento de datos.** En esta etapa, los datos almacenados se procesan para su análisis y extracción de información relevante. Esto puede implicar la limpieza de datos para eliminar valores atípicos y datos incompletos, la transformación de datos para adaptarlos a un formato adecuado para el análisis y la agregación de datos para obtener resúmenes y estadísticas útiles. Se utilizan herramientas como Apache Spark, Apache Flink o Hadoop MapReduce para procesar grandes volúmenes de datos de manera eficiente.

- **Análisis de datos.** Una vez procesados, los datos se analizan para extraer información significativa y obtener *insights* que puedan ser útiles para la toma de decisiones. Esto puede incluir análisis descriptivos para comprender patrones y tendencias en los datos, análisis predictivos para hacer predicciones sobre eventos futuros, y análisis prescriptivos para recomendar acciones basadas en los datos. Se utilizan técnicas como minería de datos, aprendizaje automático y análisis estadístico para realizar este tipo de análisis.

- **Desarrollo de modelos de IA.** En esta etapa, se desarrollan y entrenan modelos de inteligencia artificial utilizando técnicas de aprendizaje automático y procesamiento de lenguaje natural. Esto puede implicar la selección y la preparación de características, la elección de algoritmos de aprendizaje automático adecuados, el entrenamiento de modelos con datos de entrenamiento y la evaluación del rendimiento del modelo utilizando datos de prueba. Se pueden utilizar herramientas, como TensorFlow, scikit-learn y PyTorch, para desarrollar y entrenar modelos de IA.

- **Implementación y despliegue.** Una vez que se han desarrollado y entrenado los modelos, se implementan y despliegan en entornos de producción para su uso en aplicaciones y sistemas en tiempo real. Esto puede implicar la integración de modelos en aplicaciones existentes, la exposición de modelos, como servicios web API o contenedores Docker, y la monitorización del rendimiento del modelo en producción para garantizar su eficacia y fiabilidad.

- **Monitorización y optimización.** Finalmente, se monitorean continuamente los modelos desplegados en producción para detectar cualquier degradación en el rendimiento y tomar medidas correctivas según sea necesario. Esto puede implicar la recopilación y el análisis de datos de entrada y salida del modelo, la evaluación del rendimiento del modelo en tiempo real y la reentrenamiento periódico del modelo con nuevos datos para mejorar su precisión y generalización.

Tema 2. Ingesta de datos para big data e inteligencia artificial en cloud

El ciclo de vida de los datos en *big data* e IA es un proceso iterativo y continuo, donde cada etapa alimenta a la siguiente y donde el aprendizaje y la mejora son fundamentales para el éxito a largo plazo.

La Figura 1 muestra los componentes básicos, los datos y el ciclo de vida de los datos asociado (el círculo interior), y las herramientas/infraestructuras que permiten el proceso de investigación para poder aplicar IA (el círculo exterior).

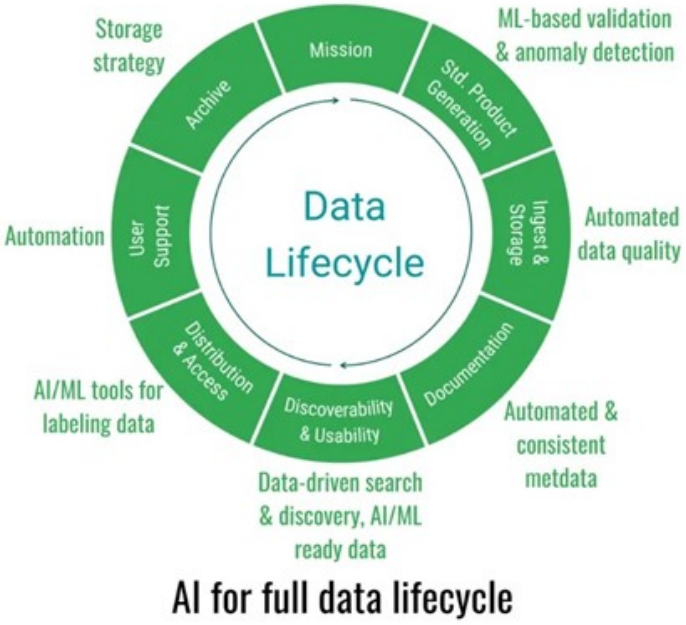


Figura 1. Ciclo de vida de los datos. Fuente: IMPACT Unofficial, 2022.

Tema 2. Ingesta de datos para big data e inteligencia artificial en cloud

Ingesta de datos

La ingesta de datos es una etapa crítica en el proceso de IA que se refiere a la recopilación y adquisición de datos de diversas fuentes para su posterior análisis y procesamiento. A continuación, se muestra una descripción detallada de la ingesta de datos para IA:

- **Recopilación de datos.** La recopilación de datos implica la obtención de datos de diversas fuentes, como bases de datos empresariales, sistemas de archivos, dispositivos IoT, registros en línea, redes sociales y más. Estos datos pueden ser estructurados, semiestructurados o no estructurados, y pueden provenir de diferentes formatos y ubicaciones.

- **Extracción de datos.** Una vez recopilados, los datos deben ser extraídos de sus fuentes originales y transformados en un formato adecuado para su procesamiento posterior. Esto puede implicar la limpieza de datos para eliminar valores atípicos y datos incompletos, la normalización de datos para asegurar la coherencia y la consistencia, y la conversión de datos a un formato compatible con las herramientas de análisis y modelado de IA.

- **Preprocesamiento de datos.** El preprocesamiento de datos implica la preparación de los datos para su uso en modelos de IA. Esto puede implicar la selección de características relevantes, la codificación de variables categóricas, la normalización de datos numéricos y la división de los datos en conjuntos de entrenamiento, validación y prueba. El preprocesamiento es crucial para garantizar la calidad y la eficacia de los modelos de IA.

- **Integración de datos.** En muchos casos, los datos provienen de múltiples fuentes y deben integrarse para su análisis y modelado. Esto puede implicar la combinación de datos de diferentes bases de datos, sistemas de archivos o API, y la unificación de esquemas y formatos de datos. La integración de datos es importante para obtener una visión completa y coherente de la información que se utilizará en los modelos de IA.

- **Almacenamiento de datos.** Una vez que los datos han sido recopilados, extraídos, preprocesados e integrados deben ser almacenados en un sistema de almacenamiento adecuado para su acceso y uso posterior. Esto puede implicar el uso de sistemas de almacenamiento en la nube, bases de datos distribuidas, sistemas de archivos distribuidos o almacenes de datos especializados. El almacenamiento de datos debe ser escalable, seguro y accesible para garantizar un procesamiento eficiente y eficaz de los datos.

- **Gestión de metadatos.** La gestión de metadatos es importante para mantener un registro de los datos y proporcionar información sobre su origen, significado y calidad. Esto puede incluir la creación de catálogos de datos, la documentación de esquemas de datos, la trazabilidad de los datos y la gestión de versiones de los datos. La gestión de metadatos facilita la búsqueda, la comprensión y la utilización de los datos en proyectos de IA.

- **Automatización y orquestación.** La automatización y la orquestación son importantes para gestionar el flujo de datos de manera eficiente y escalable. Esto puede implicar el uso de herramientas de automatización para programar y ejecutar tareas de ingesta de datos de manera automatizada, y el uso de sistemas de orquestación para coordinar y gestionar flujos de datos complejos entre diferentes sistemas y servicios.

En resumen, la **ingesta de datos** es una etapa fundamental en el proceso de IA que implica la recopilación, extracción, preprocesamiento, integración, almacenamiento, gestión de metadatos y automatización de datos para su uso en modelos de IA. Una ingesta de datos eficiente y bien gestionada es crucial para el éxito de los proyectos de IA, como podemos ver en la Figura 2.

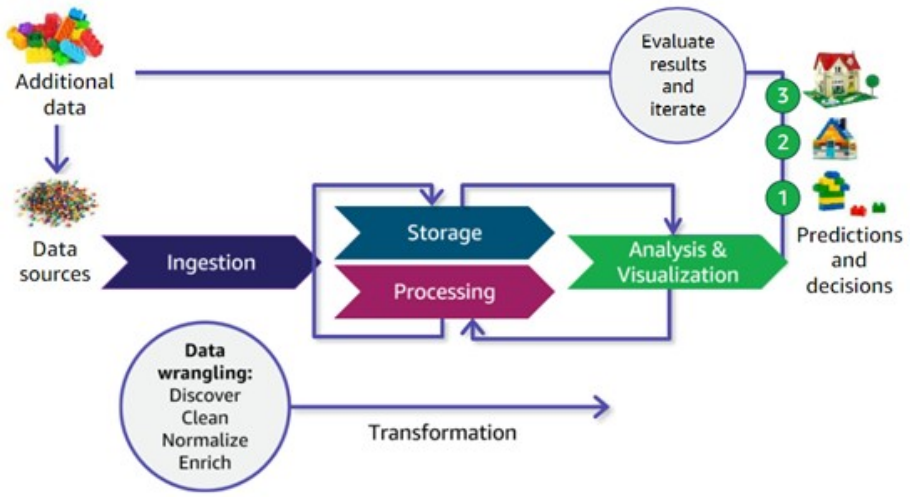


Figura 2. Ingesta de datos. Fuente: Aitor Medrano, s. f.

Tipologías de ingesta de datos

La ingesta de datos en el contexto de la informática y la IA se refiere al proceso de **recopilación y adquisición de datos** desde diversas fuentes para su posterior procesamiento y análisis. Hay varias tipologías de ingesta de datos, que se clasifican según la naturaleza de los datos y cómo se obtienen. A continuación, se mencionan algunas de las tipologías más comunes:

- **Ingesta *batch*.** Implica la recopilación y procesamiento de grandes volúmenes de datos en bloques o lotes periódicos. Los datos se recopilan durante un período de tiempo y se procesan en lotes, generalmente en intervalos programados. Esto es útil para casos de uso donde el análisis no necesita ser en tiempo real y se puede tolerar cierto retraso en la disponibilidad de los resultados.

- **Ingesta *streaming*.** Implica la recopilación y procesamiento de datos en tiempo real a medida que se generan. Los datos se transmiten continuamente desde las fuentes de origen a los sistemas de destino, donde se procesan y analizan en tiempo real. Esto es útil para casos de uso donde se requiere una respuesta rápida a los datos entrantes y se necesita un análisis en tiempo real.

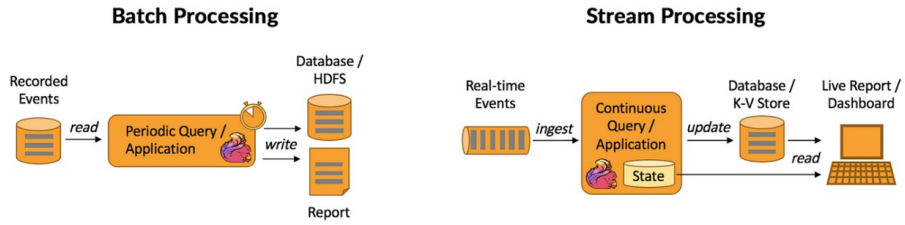


Figura 3. Ingesta *batch* y *streaming*. Fuente: Murat Sivri, 2022.

- **Ingesta incremental.** Implica la recopilación de solo los datos nuevos o actualizados desde la última ejecución. En lugar de procesar todos los datos cada vez, solo se procesan los datos que han cambiado desde la última ingesta. Esto es útil para reducir el tiempo y los recursos necesarios para procesar datos, especialmente en entornos con grandes volúmenes de datos que cambian con frecuencia.

- **Ingesta de datos estructurados.** Implica la recopilación de datos que siguen un formato predefinido y organizado, como las bases de datos relacionales, los archivos CSV o JSON con una estructura definida. Estos datos suelen ser fáciles de procesar y analizar debido a su organización y coherencia.

- **Ingesta de datos semiestructurados.** Implica la recopilación de datos que tienen una estructura parcialmente definida, como documentos HTML, archivos XML o registros de servidor. Estos datos pueden contener elementos que no siguen una estructura estricta y pueden requerir un procesamiento adicional para extraer información útil.

- **Ingesta de datos no estructurados.** Implica la recopilación de datos que no siguen ningún formato predefinido o estructura organizada, como imágenes, vídeos, archivos de audio o texto libre. Estos datos suelen ser más difíciles de procesar y analizar debido a su falta de estructura, pero pueden contener información valiosa que puede ser aprovechada con técnicas de IA como el procesamiento de lenguaje natural o la visión por computadora.

- **Ingesta de datos de alta velocidad.** La ingesta de datos de alta velocidad implica la recopilación de datos que se generan a una velocidad muy rápida, como datos de sensores, registros de eventos o datos de transmisión en tiempo real. Estos datos deben ser procesados y analizados rápidamente para mantenerse al día con la tasa de generación de datos.

Cada tipo de ingesta de datos tiene sus propias características y desafíos, y la elección del enfoque adecuado dependerá de los requisitos específicos del proyecto y las necesidades de análisis de datos. Es común que los sistemas de ingesta de datos utilicen una combinación de estos enfoques para manejar diferentes tipos de datos y casos de uso

Desafíos en la ingesta de datos para IA

La ingesta de datos en entornos de *big data* enfrenta desafíos específicos debido al gran volumen, la variedad y la velocidad de los datos involucrados. A continuación, se presenta una lista de los **principales desafíos** de la ingesta de datos en *big data*, junto con una breve descripción de cada uno:

- **Escalabilidad.** Manejar grandes volúmenes de datos de manera eficiente y escalable es fundamental en *big data*. Los sistemas de ingesta deben poder procesar y almacenar enormes cantidades de datos de manera rápida y efectiva.

- **Velocidad de ingesta.** En entornos de *big data*, la velocidad de ingesta puede ser crítica, especialmente en aplicaciones de tiempo real. Los sistemas deben ser capaces de capturar y procesar datos rápidamente para garantizar que la información sea relevante y útil en el momento adecuado.

- **Variedad de fuentes y formatos de datos.** Los datos en entornos de *big data* pueden provenir de una amplia variedad de fuentes, incluidas las bases de datos, los archivos de registro, las redes sociales, los sensores IoT, entre otros. Integrar datos de diferentes formatos y estructuras puede ser complicado y requerir procesos de transformación y normalización.

- **Latencia.** La latencia en el procesamiento y la entrega de datos puede afectar la eficiencia y la utilidad de la información en aplicaciones de *big data*. Minimizar la latencia en la ingesta de datos es importante para garantizar que los datos estén disponibles para análisis y procesamiento en tiempo real.

- **Consistencia y confiabilidad.** Garantizar la consistencia y la confiabilidad de los datos en entornos de *big data* puede ser un desafío. Los sistemas de ingesta deben ser capaces de manejar errores, duplicados y pérdida de datos de manera efectiva para garantizar la integridad de los datos.

- **Seguridad y cumplimiento normativo.** La seguridad de los datos y el cumplimiento de regulaciones son preocupaciones importantes en la ingesta de datos en entornos de *big data*. Proteger los datos sensibles y cumplir con los requisitos normativos puede ser un desafío técnico y legal.

- **Complejidad de la arquitectura.** Diseñar, implementar y mantener una arquitectura de ingesta de datos robusta y escalable en entornos de *big data* puede ser complicado. La infraestructura y los sistemas de ingesta deben ser capaces de escalar y adaptarse a medida que crecen los volúmenes de datos y las necesidades del negocio.

- **Monitorización y gestión.** Monitorear y gestionar el proceso de ingesta de datos en entornos de *big data* es fundamental para garantizar la integridad y la disponibilidad de los datos. Implementar herramientas de monitoreo y alerta que permitan identificar problemas y tomar medidas correctivas es esencial para mantener la salud y el rendimiento del sistema.

Tema 2. Ingesta de datos para big data e inteligencia artificial en cloud

En resumen, la ingesta de datos en entornos de *big data* enfrenta una serie de desafíos, desde la escalabilidad y la velocidad hasta la variedad y la confiabilidad de los datos. Abordar estos desafíos de manera efectiva es crucial para garantizar el éxito de las iniciativas de *big data* y análisis de datos.

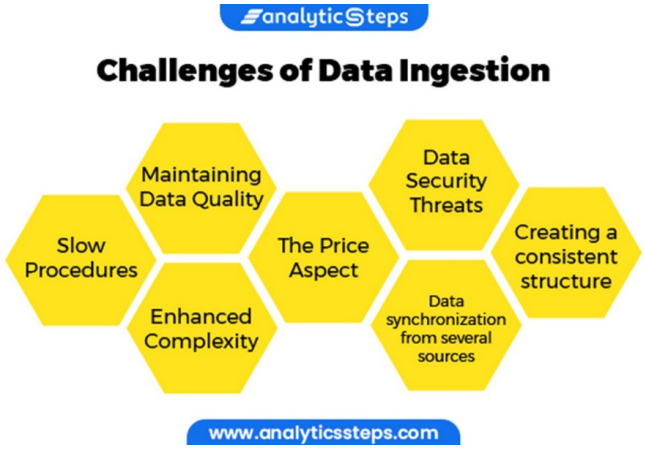


Figura 4. Desafíos de la ingesta de datos. Fuente: Vrinda Mathur, 2022.

Paradigmas de ingesta de datos

Los **paradigmas de ingesta de datos ETL** (*extract, transform, load*) y **ELT** (*extract, load, transform*) son enfoques utilizados en la preparación y procesamiento de datos antes de ser almacenados en un sistema de destino, como un almacén de datos o un lago de datos. A continuación, se desarrolla cada uno.

Tema 2. Ingesta de datos para big data e inteligencia artificial en cloud

ETL, que significa *extract, transform, load* (extraer, transformar, cargar), es un proceso fundamental en el campo del *big data* y la gestión de datos en general. A continuación, tienes una descripción detallada de cada una de las fases del proceso ETL:

- ***Extract* (extraer).** En la etapa de extracción, los datos se recopilan de múltiples fuentes, que pueden incluir bases de datos, archivos, sistemas en línea, API, entre otros. Los datos extraídos pueden ser crudos y no estructurados, y pueden estar en diferentes formatos y ubicaciones.

- ***Transform* (transformar).** En la etapa de transformación, los datos extraídos se procesan y se transforman en un formato adecuado para su almacenamiento y análisis. Esto puede implicar la limpieza de datos para eliminar valores atípicos y datos incompletos, la normalización de datos para asegurar la coherencia y la consistencia, y la agregación de datos para obtener resúmenes y estadísticas útiles. También puede implicar la conversión de datos a un esquema común y la combinación de datos de diferentes fuentes.

- ***Load* (cargar).** En la etapa de carga, los datos transformados se cargan en un sistema de destino, como un almacén de datos o un lago de datos, donde estarán disponibles para su análisis posterior. Los datos cargados suelen estar estructurados y listos para su uso en informes, análisis y aplicaciones de inteligencia empresarial.

El **proceso ELT** (*extract, load, transform*) en *big data* es una variante del proceso ETL en el que la transformación de datos se realiza después de la carga en el sistema de destino en lugar de antes. A continuación, se presenta una descripción de cada una de las fases del proceso ELT en el contexto del *big data***:**

- ***Extract* (extraer).** En el paradigma ELT, los datos se extraen de las fuentes de origen de la misma manera que en ETL, sin realizar transformaciones significativas en este punto.

- ***Load* (cargar).** Después de extraer los datos, se cargan directamente en el sistema de destino sin transformaciones significativas. Los datos se almacenan en su forma original, a menudo en un lago de datos o un almacén de datos sin procesar.

- ***Transform* (transformar).** En la etapa de transformación, los datos se transforman y se procesan en el sistema de destino según sea necesario. Esto puede implicar la ejecución de transformaciones complejas, consultas SQL o procesos de análisis en los datos cargados en el sistema de destino.

Diferencias entre ETL y ELT

- **ETL** se centra en transformar los datos antes de cargarlos en el sistema de destino, lo que puede ser beneficioso para limpiar y normalizar los datos antes de su almacenamiento.

- **ELT** carga los datos en el sistema de destino sin transformaciones significativas, lo que puede ser útil cuando se trabaja con grandes volúmenes de datos y se requiere una mayor escalabilidad.

En resumen, ETL y ELT son dos paradigmas de ingesta de datos comúnmente utilizados en el procesamiento de datos antes de su almacenamiento en un sistema de destino. La **elección entre ETL y ELT** dependerá de los requisitos específicos del proyecto, incluida la naturaleza de los datos, el volumen de datos y los requisitos de transformación.

La Figura 5 ilustra las diferencias entre los dos paradigmas de acuerdo con el orden de acciones en el proceso de ingesta de datos.

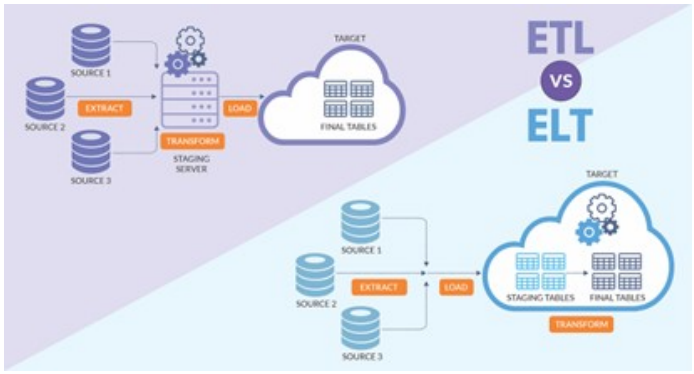


Figura 5. Diferentes entre ETL y ELT. Fuente: Halder, 2023.

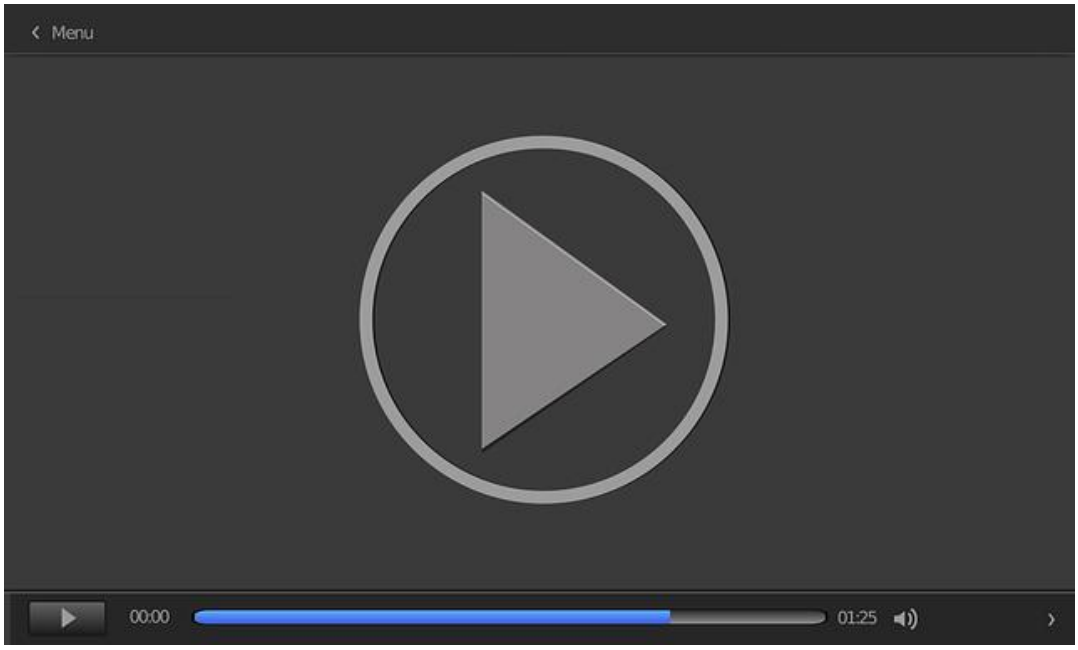
AWS Glue

Introducción a AWS Glue

En el mundo de la analítica de datos, la **preparación de datos** es una parte fundamental del proceso. AWS Glue es una herramienta desarrollada por AWS que facilita la preparación y el procesamiento de datos para su análisis en la nube. Con Glue, las empresas pueden automatizar tareas de extracción, transformación y carga (ETL) de datos, lo que les permite centrarse en el análisis y la generación de *insights* en lugar de en la gestión de infraestructura compleja.

Tema 2. Ingesta de datos para big data e inteligencia artificial en cloud

El siguiente vídeo, *AWS Glue Overview | Amazon Web Services* (Amazon Web Services, 2022), muestra un *overview* del servicio AWS Glue.



AWS Glue Overview | Amazon Web Services.

Accede al vídeo:

https://www.youtube.com/embed/u14iVEc-C6E

Descripción de AWS Glue

AWS Glue es un servicio de ETL completamente administrado que ofrece una amplia gama de características y funcionalidades para ayudar a las organizaciones a preparar y procesar sus datos de manera eficiente y escalable. Está diseñado para trabajar con **grandes volúmenes de datos y flujos de trabajo complejos,** lo que lo hace ideal para aplicaciones en la nube que requieren análisis avanzados y procesamiento de datos en tiempo real.

Características principales de AWS Glue

Las **características principales** de AWS Glue son las siguientes:

- **Descubrimiento automático de esquemas.** Glue utiliza *crawlers* para explorar las fuentes de datos y descubrir automáticamente la estructura y el esquema de los datos. Esto simplifica el proceso de preparación de datos al eliminar la necesidad de definir manualmente los esquemas de datos.

- **Transformaciones ETL.** Glue proporciona una funcionalidad robusta para realizar transformaciones ETL en los datos, lo que permite limpiar, normalizar y enriquecer los datos antes de cargarlos en un almacén de datos o un lago de datos. Las transformaciones se pueden definir utilizando lenguajes de *script,* como Python o Scala, o utilizando una interfaz gráfica de usuario intuitiva.

- **Programación de trabajos ETL.** Los trabajos de Glue permiten programar y ejecutar flujos de trabajo de ETL de manera automatizada y periódica. Esto simplifica la gestión de flujos de trabajo complejos y permite manejar grandes volúmenes de datos de manera eficiente.

- **Integración con otros servicios de AWS.** Glue se integra estrechamente con otros servicios de AWS, como S3, Redshift y Athena, lo que permite cargar datos en almacenes de datos y realizar análisis avanzados. Esta integración facilita el acceso a una amplia gama de herramientas y servicios de AWS para el análisis de datos.

- **Completamente administrado.** Glue es un servicio completamente administrado, lo que significa que AWS se encarga de la infraestructura subyacente, incluida la escalabilidad, la disponibilidad y la seguridad. Esto les permite a los usuarios centrarse en el análisis de datos en lugar de en la gestión de infraestructura.

En resumen, AWS Glue es una herramienta poderosa y versátil que simplifica y automatiza el **proceso de preparación de datos** para análisis en la nube. Con su capacidad para descubrir automáticamente esquemas de datos, realizar

transformaciones ETL y programar flujos de trabajo de manera automatizada, Glue permite que las organizaciones aceleren el tiempo de obtención de información y obtener *insights* valiosos de sus datos.

Componentes de AWS Glue

Para el **descubrimiento e inferencia de esquemas,** se utiliza el **Glue Data Catalog,** que es el catálogo de datos de AWS Glue (AWS Glue Data Catalog). Este es un servicio fundamental que actúa como un repositorio centralizado y gestionado de metadatos sobre sus datos en AWS. A continuación, tienes una descripción detallada de los servicios y capacidades del catálogo de datos de AWS Glue:

- **Almacenamiento de metadatos.** El catálogo de datos de AWS Glue almacena metadatos sobre sus conjuntos de datos, incluidos esquemas, tablas, particiones y dependencias. Estos metadatos proporcionan información estructural y descriptiva sobre los datos, lo que facilita su descubrimiento, comprensión y uso.

- **Descubrimiento automático de esquemas.** El catálogo de datos de Glue utiliza *crawlers* para explorar las fuentes de datos y descubrir automáticamente la estructura y el esquema de los datos. Esto elimina la necesidad de definir manualmente los esquemas de datos y simplifica el proceso de preparación de datos.

  Los ***crawlers*** son herramientas que exploran las fuentes de datos para descubrir automáticamente la estructura y el esquema de los datos. Utilizan técnicas de inferencia para identificar metadatos relevantes, como tipos de datos, claves primarias y campos de partición. Los *crawlers* simplifican el proceso de preparación de datos al eliminar la necesidad de definir manualmente los esquemas de datos.

En la Figura 6, se muestra cómo los rastreadores de AWS Glue interactúan con almacenes de datos y otros elementos para rellenar el catálogo de datos.

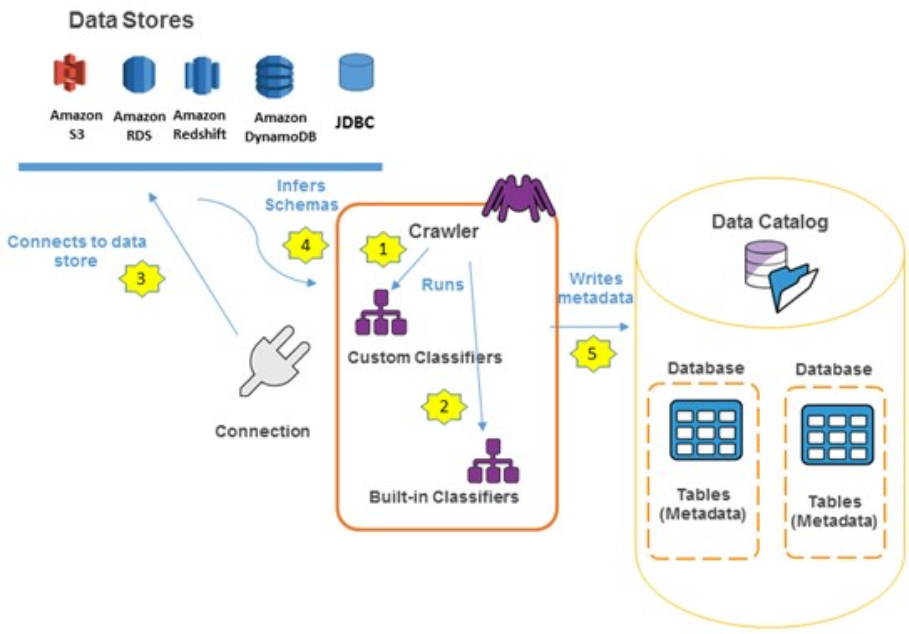


Figura 6. Catálogo. Fuente: Amazon Web Services, s. f-a.

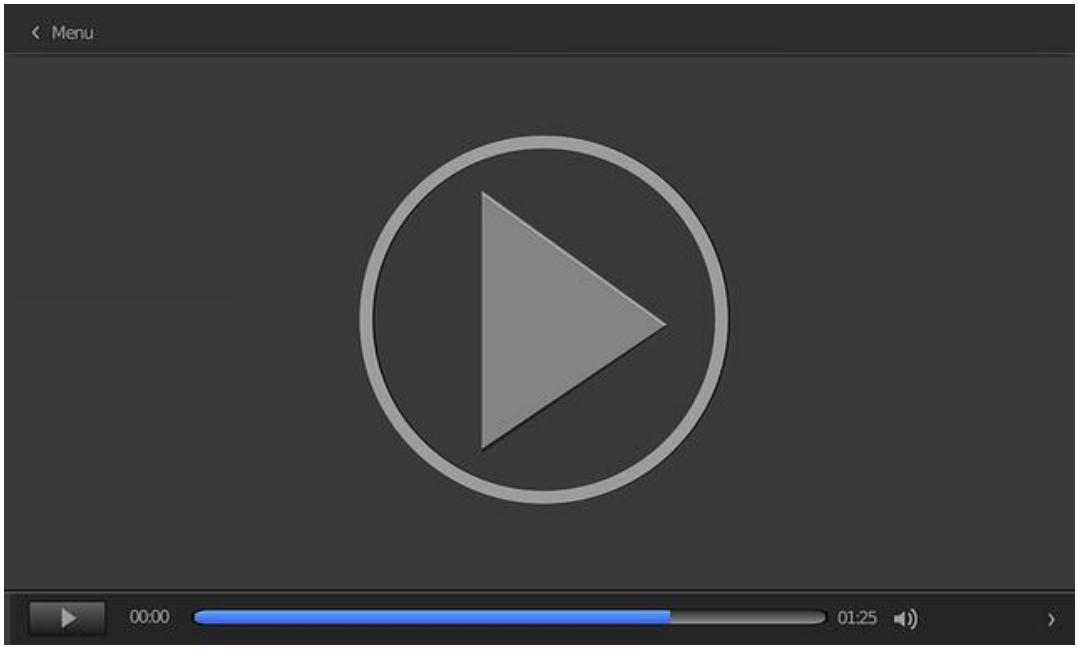
Este es el flujo de flujo de trabajo general de rellenado de AWS Glue Data Catalog por parte de un rastreador (Tabla 1).

Flujo de trabajo	
1	Un rastreador ejecuta cualquier clasificador personalizado que elija para inferir el formato y el esquema de sus datos. Debe proporcionar el código para clasificadores personalizados, que se ejecutan en el orden especificado.
2	El primer clasificador personalizado en reconocer correctamente la estructura de sus datos se usa para crear un esquema. Los clasificadores personalizados que aparecen más abajo en la lista se omiten.
3	Si no coincide ningún clasificador con el esquema de sus datos, los clasificadores integrados intentarán reconocer el esquema de sus datos. Un ejemplo de un clasificador integrado es uno que reconoce JSON.
4	El rastreador se conecta al almacén de datos. Algunos almacenes de datos requieren propiedades de conexión para el acceso del rastreador.
5	El esquema inferido se crea para sus datos.
6	El rastreador escribe los metadatos en el catálogo de datos. Una definición de tabla contiene metadatos acerca de los datos de su almacén de datos. La tabla se escribe en una base de datos, que es un contenedor de tablas en el catálogo de datos. Entre los atributos de una tabla se incluye la clasificación, que es una etiqueta creada por el clasificador que determinó el esquema de tabla.

Tabla 1. Flujo de trabajo. Fuente: elaboración propia.

Tema 2. Ingesta de datos para big data e inteligencia artificial en cloud

El siguiente vídeo, *Descubre y cataloga tus datos con AWS Glue – Español* (AWS LATAM, 2021a), es un tutorial que explica cómo catalogar los datos con AWS Glue.



Descubre y cataloga tus datos con AWS Glue – Español.

Accede al vídeo:

https://www.youtube.com/embed/pbKnfjjVsx4

- **Integración con otros servicios de AWS.** El catálogo de datos de Glue se integra estrechamente con otros servicios de AWS, como S3, Redshift y Athena. Esto permite cargar datos en almacenes de datos y realizar análisis avanzados utilizando herramientas y servicios familiares de AWS.

- **Compatibilidad con particiones y vistas.** El catálogo de datos de Glue es compatible con particiones, lo que facilita la organización y el acceso a grandes volúmenes de datos. También admite la creación de vistas que pueden simplificar y mejorar el acceso a los datos al proporcionar una capa de abstracción sobre los datos subyacentes.

- **Gestión de metadatos.** El catálogo de datos de Glue proporciona herramientas para gestionar metadatos, incluida la edición, el borrado y la visualización de metadatos. También permite etiquetar y categorizar los datos para facilitar su organización y búsqueda.

- **Acceso controlado.** El catálogo de datos de Glue ofrece capacidades de acceso controlado que permiten controlar quién puede acceder y modificar los metadatos. Esto ayuda a garantizar la seguridad y la privacidad de los datos y cumple con los requisitos de cumplimiento normativo.

- **API y SDK.** Glue proporciona una API y SDK que permiten integrar el catálogo de datos con aplicaciones y herramientas personalizadas. Esto facilita la automatización de tareas y la integración con el ecosistema de herramientas existente.

Para el **motor de transformaciones ETL,** AWS Glue proporciona una funcionalidad robusta para realizar transformaciones ETL en los datos. Esto incluye limpiar, normalizar y enriquecer los datos antes de cargarlos en un almacén de datos o un lago de datos. Las transformaciones se pueden definir utilizando lenguajes de *script,* como Python o Scala, o utilizando una interfaz gráfica de usuario intuitiva. Glue permite que los usuarios definan transformaciones complejas y personalizadas para adaptarse a sus necesidades específicas.

Los trabajos de Glue permiten programar y ejecutar **flujos de trabajo de ETL** de manera automatizada y periódica. Esto simplifica la gestión de flujos de trabajo complejos y permite manejar grandes volúmenes de datos de manera eficiente. Los trabajos de Glue son completamente administrados y escalables, lo que garantiza un rendimiento confiable y consistente incluso con grandes cargas de trabajo.