

Tema 3. Data science cloud storage

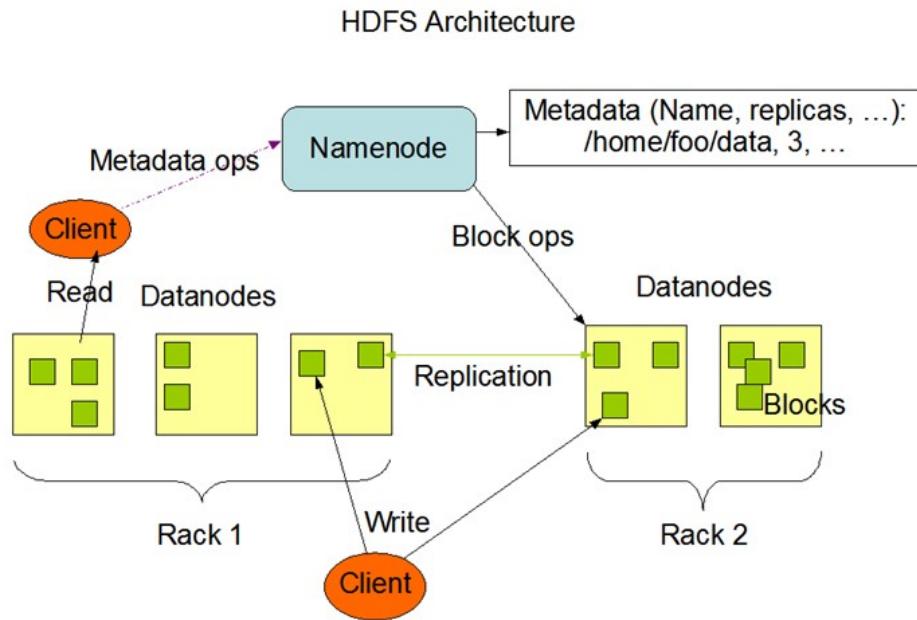


Figura 18. Arquitectura HDFS. Fuente: Mamani, 2023.

- ▶ **Replicación de datos.** HDFS utiliza la replicación de datos para garantizar la disponibilidad y la tolerancia a fallos. Por defecto, cada bloque de datos se replica en múltiples nodos DataNode (generalmente tres), lo que garantiza que los datos estén disponibles incluso en caso de fallo de un nodo.
- ▶ **API para acceso a datos.** HDFS proporciona API para acceder y manipular datos almacenados en el sistema de archivos, incluyendo una API Java, comandos de línea de Hadoop (Hadoop Shell) y API de acceso a través de Hadoop Streaming y Hadoop Pipes.

En resumen, **Apache HDFS** es un sistema de archivos distribuido diseñado para almacenar grandes conjuntos de datos de manera confiable y escalable en clústeres de servidores estándar. Ofrece distribución de datos, tolerancia a fallos, replicación de datos y API para el acceso a los datos, lo que lo convierte en una parte fundamental del ecosistema de Hadoop para el procesamiento distribuido de datos.

Tema 3. Data science cloud storage

Sistemas de archivos EMRFS

EMRFS (EMR File System) es una característica específica de Amazon Elastic MapReduce (EMR), que facilita el acceso a los datos almacenados en Amazon S3 desde clústeres de EMR. Se integra con Apache Hadoop y Apache Spark, lo que les permite a los usuarios leer y escribir datos en S3 como si estuvieran trabajando con un sistema de archivos tradicional, pero con mejoras específicas para el entorno de EMR.

Estas son algunas **características** importantes de EMRFS y se explica cómo se relacionan con HDFS en EMR:

- ▶ **Integración con S3.** EMRFS está diseñado específicamente para trabajar con Amazon S3 como un sistema de almacenamiento subyacente. Permite a los usuarios acceder y manipular datos en S3 de manera eficiente y confiable, aprovechando las características de S3 como escalabilidad, durabilidad y bajo costo.
- ▶ **Consistencia de metadatos.** EMRFS garantiza la consistencia de los metadatos entre el clúster de EMR y S3, lo que significa que los cambios realizados en los metadatos (como la creación, modificación o eliminación de archivos) son reflejados de manera consistente en ambos lados.
- ▶ **Optimizaciones de rendimiento.** EMRFS implementa diversas optimizaciones de rendimiento para mejorar la eficiencia de las operaciones de lectura y escritura en S3. Esto incluye el uso de cachés locales y remotos para metadatos y datos, así como técnicas de *prefetching* y *prefetching* de datos para acelerar el acceso a los datos.

Tema 3. Data science cloud storage

- ▶ **Consistencia fuerte.** A diferencia de HDFS, que utiliza una consistencia eventual para los metadatos, EMRFS ofrece consistencia fuerte para garantizar que los cambios en los metadatos sean visibles de manera inmediata en todo el clúster de EMR.
- ▶ **Soporte para características avanzadas de S3.** EMRFS es compatible con muchas características avanzadas de Amazon S3, como el cifrado de datos, control de acceso basado en políticas (IAM), versionado de objetos y Amazon S3 Cross-Region Replication (CRR).

En cuanto a **HDFS en EMR**, es importante tener en cuenta que los clústeres de EMR pueden utilizar tanto HDFS como EMRFS para el almacenamiento de datos. Sin embargo, EMRFS se considera la opción preferida para acceder a datos en S3 debido a sus ventajas en términos de escalabilidad, durabilidad y compatibilidad con las características avanzadas de S3. En resumen, EMRFS es la opción recomendada para acceder a datos en S3 desde clústeres de EMR, mientras que HDFS sigue siendo una opción válida para el almacenamiento de datos en el sistema de archivos local del clúster.



Figura 19. Flujo de trabajo ETL. Fuente: Chayel, 2014.

Tema 3. Data science cloud storage

3.4. Data warehouse en la nube

Data warehouse. Introducción

Un **data warehouse** (almacén de datos) es un sistema de almacenamiento y gestión de datos diseñado para posibilitar el análisis y la elaboración de informes sobre grandes volúmenes de datos, procedentes de múltiples fuentes. Su principal objetivo es centralizar y consolidar datos históricos para facilitar la toma de decisiones estratégicas en una organización.

Entre los principales **usos** de un *data warehouse*, tenemos los siguientes:

- ▶ **Análisis de datos.** Facilita el análisis en profundidad de los datos históricos de la empresa. Permite realizar análisis de tendencias y patrones a largo plazo.
- ▶ **Business intelligence (BI).** Provee una base para herramientas de inteligencia empresarial que ayudan a transformar datos en información útil para la toma de decisiones.
- ▶ **Soporta cuadros de mando, informes y visualización de datos.**
- ▶ **Mejora del rendimiento empresarial.** Ayuda a identificar oportunidades de negocio, mejorar procesos y aumentar la eficiencia. Permite la evaluación del rendimiento de distintas áreas de la empresa.
- ▶ **Integración de datos.** Centraliza datos provenientes de diversas fuentes y sistemas, facilitando su acceso y análisis conjunto. Mejora la coherencia y calidad de los datos.
- ▶ **Soporte a la toma de decisiones.** Proporciona una visión unificada y completa de la información empresarial. Facilita la toma de decisiones informadas basadas en datos precisos y actualizados.

Tema 3. Data science cloud storage

Sobre las **características** fundamentales de un *data warehouse*, podemos enumerar las siguientes:

- ▶ **Orientación a sujetos.** Los datos están organizados por temas o áreas de interés, como ventas, finanzas, recursos humanos, etc., en lugar de por aplicaciones.
- ▶ **Integración.** Consolida datos de diferentes fuentes, garantizando consistencia y calidad. Estandariza formatos, unidades y estructuras de datos.
- ▶ **Variabilidad en el tiempo.** Almacena datos históricos, lo que permite análisis longitudinales y comparación a lo largo del tiempo. Los datos son etiquetados con un marco temporal, lo que facilita el seguimiento de cambios y tendencias.
- ▶ **No volatilidad.** Una vez que los datos se cargan en el *data warehouse*, no se modifican ni se eliminan. Permite mantener la integridad y exactitud de los datos históricos.
- ▶ **Optimización para consulta y análisis.** Está diseñado específicamente para ejecutar consultas y análisis rápidos y eficientes. Emplea técnicas como indexación, particionamiento y cubos OLAP (*online analytical processing*) para mejorar el rendimiento de las consultas.

Para entender la **arquitectura** de un *data warehouse*, se muestra el siguiente ejemplo:

- ▶ **Fuente de datos.** Sistemas transaccionales (ERP, CRM), archivos planos, bases de datos relacionales y datos externos.
- ▶ **ETL (extracción, transformación y carga).** Procesos para extraer datos de las fuentes, transformarlos según las necesidades de análisis y cargarlos en el *data warehouse*.
- ▶ **Área de staging.** Espacio temporal para limpiar y transformar los datos antes de cargarlos en el *data warehouse*.

Tema 3. Data science cloud storage

- ▶ **Almacén de datos central.** Base de datos principal donde se almacenan los datos integrados y consolidados.
- ▶ **Data marts.** Subconjuntos de datos específicos orientados a necesidades particulares de negocio o departamentos.
- ▶ **Herramientas de BI y análisis.** Aplicaciones y plataformas que permiten a los usuarios finales realizar consultas, generar informes y visualizar datos.

En resumen, un **data warehouse** es una herramienta clave para la gestión y el análisis de grandes volúmenes de datos, lo que proporciona una base sólida para la inteligencia empresarial y la toma de decisiones informada.

Data warehouse en la nube

En los últimos tiempos, existe una clara simbiosis entre un *data warehouse* y el *cloud* (la nube) mediante la integración de los almacenes de datos tradicionales con las tecnologías y servicios en la nube. Esta **combinación** ofrece una serie de ventajas significativas para las organizaciones en términos de flexibilidad, escalabilidad y costo-efectividad. A continuación, se detalla esta relación y sus beneficios.

- ▶ **Escalabilidad dinámica:**
 - **Data warehouses.** Tradicionalmente, los *data warehouses* se implementaban en infraestructuras físicas limitadas por la capacidad del *hardware* disponible.
 - **Cloud.** La nube permite escalar recursos de almacenamiento y procesamiento de manera dinámica según las necesidades, sin requerir inversiones iniciales en *hardware*.

Tema 3. Data science cloud storage

► Coste-efectividad:

- **Data warehouse.** La implementación y mantenimiento de un *data warehouse* on-premise (en las instalaciones de la empresa) puede ser costosa debido a la compra de *hardware*, licencias de *software* y personal especializado.
- **Cloud.** Los servicios de *data warehouse* en la nube, como Amazon Redshift, Google BigQuery y Azure Synapse Analytics, operan bajo un modelo de pago por uso, lo que reduce significativamente los costos iniciales y permite una mejor gestión del presupuesto al pagar solo por los recursos utilizados.

► Flexibilidad y agilidad:

- **Data warehouse.** Las actualizaciones y expansiones en un entorno *on-premise* pueden ser lentas y disruptivas.
- **Cloud.** Los servicios en la nube permiten implementar nuevas capacidades, actualizar *software* y ajustar configuraciones de manera rápida y sin interrupciones, lo que facilita una respuesta ágil a las necesidades cambiantes del negocio.

► Acceso global y colaboración:

- **Data warehouse.** El acceso a los datos puede estar limitado por la ubicación física del servidor y las restricciones de red.
- **Cloud.** Los *data warehouses* en la nube están accesibles desde cualquier lugar con conexión a internet, lo que facilita la colaboración global y el acceso remoto a los datos por parte de empleados, socios y clientes.

Tema 3. Data science cloud storage

► Seguridad y recuperación ante desastres:

- **Data warehouse.** Requiere una inversión significativa en medidas de seguridad y planes de recuperación ante desastres.
- **Cloud.** Los proveedores de servicios en la nube ofrecen avanzadas características de seguridad y recuperación ante desastres como parte de sus servicios, incluyendo cifrado de datos, redundancia y copias de seguridad automáticas.

Entre las **ventajas** de implementación de un *data warehouse* en la nube, podemos destacar las siguientes:

- **Implementación rápida.** Los *data warehouses* en la nube se pueden desplegar y configurar rápidamente sin la necesidad de instalar y configurar *hardware* físico.
- **Mantenimiento reducido.** Los proveedores de la nube gestionan el mantenimiento del *hardware* y las actualizaciones del *software*, lo que permite que las organizaciones se concentren en el análisis de datos y la toma de decisiones.
- **Integración con servicios de datos.** Los servicios de *data warehouse* en la nube suelen integrarse fácilmente con otros servicios en la nube, como herramientas de análisis de datos, aprendizaje automático y *big data*, creando un ecosistema completo para la gestión y análisis de datos.
- **Elasticidad.** La capacidad de ajustar automáticamente los recursos de procesamiento y almacenamiento en respuesta a las cargas de trabajo variables, garantizando un rendimiento óptimo sin necesidad de sobreaprovisionar recursos.
- **Innovación continua.** Los proveedores de la nube introducen regularmente nuevas características y mejoras, lo que les permite a las empresas beneficiarse de las últimas innovaciones tecnológicas sin esfuerzo adicional.

Tema 3. Data science cloud storage

Por tanto, la **integración de los data warehouses con la nube** ofrece una poderosa combinación de escalabilidad, flexibilidad, y eficiencia en costos. Las organizaciones pueden manejar grandes volúmenes de datos con mayor agilidad, garantizar la seguridad y la recuperación de sus datos, y aprovechar un ecosistema de servicios en la nube para potenciar su inteligencia empresarial y capacidad analítica. Esta simbiosis transforma el manejo de datos en una ventaja competitiva significativa en el entorno empresarial actual.

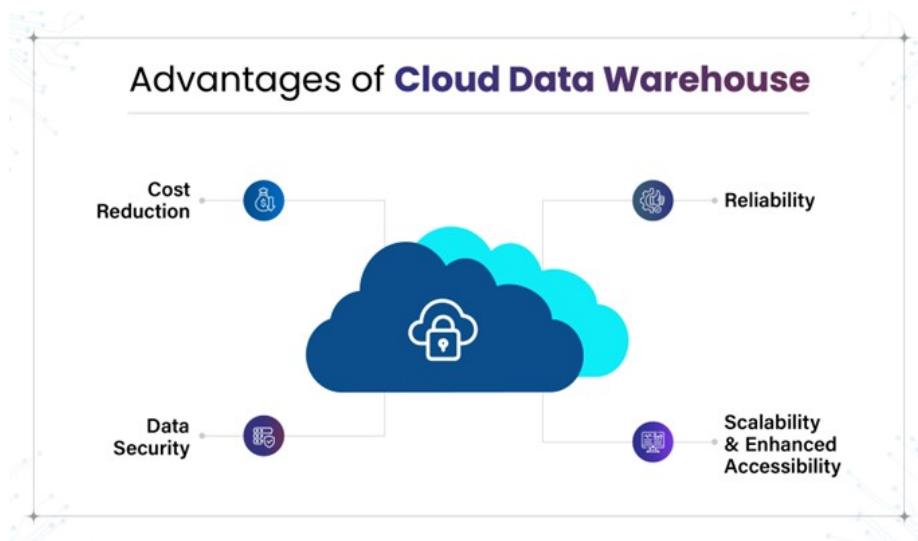


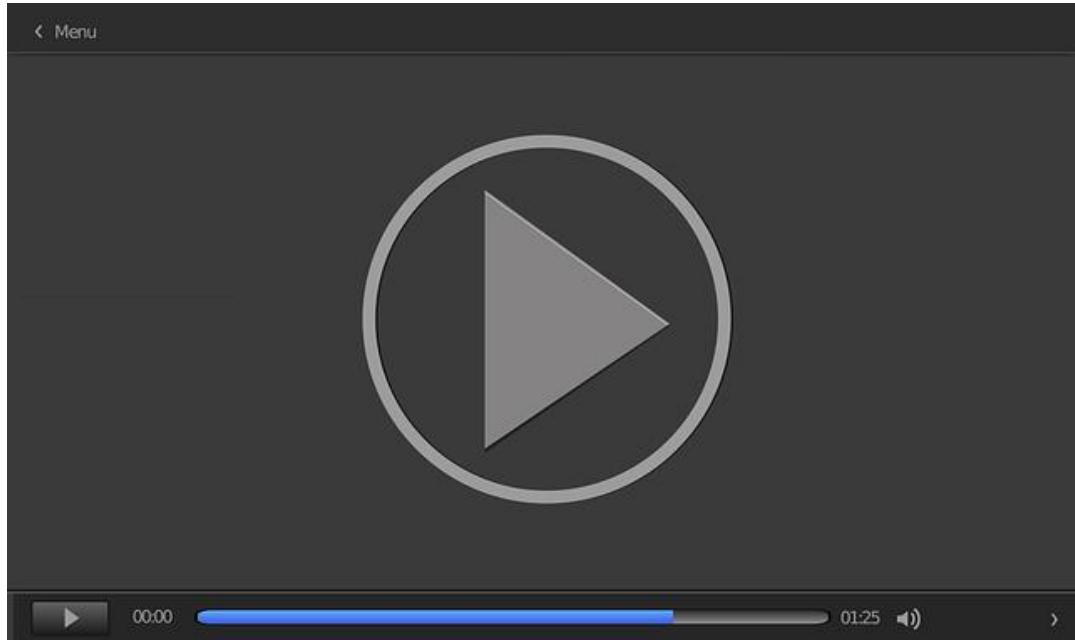
Figura 20. Ventajas del *cloud data warehouse*. Fuente: Cogent, 2024.

AWS Redshift

Amazon Redshift es un servicio de almacenamiento de datos (*data warehouse*) en la nube proporcionado por AWS. Está diseñado para manejar grandes cantidades de datos y realizar consultas analíticas rápidas y eficientes. A continuación, se describen sus características, funcionalidades, ventajas y casos de uso.

Tema 3. Data science cloud storage

En el vídeo *Introduction to Data Warehousing on AWS with Amazon Redshift | Amazon Web Services* (Amazon Web Services, 2021a), se realiza una introducción al servicio de AWS Redshift.



Introduction to Data Warehousing on AWS with Amazon Redshift | Amazon Web Services.

Accede al vídeo:

https://www.youtube.com/embed/lWwFJV_9PoE

Características de Amazon Redshift

Entre las principales características de Amazon Redshift, podemos enumerar:

- ▶ **Arquitectura distribuida.** Utiliza una arquitectura de clúster que permite distribuir las tareas de procesamiento de datos entre múltiples nodos, lo que mejora el rendimiento y la escalabilidad.
- ▶ **Compatibilidad con SQL.** Soporta SQL estándar, lo que facilita a los usuarios ejecutar consultas complejas y análisis sobre los datos almacenados.

Tema 3. Data science cloud storage

- ▶ **Integración con el ecosistema de AWS.** Se integra perfectamente con otros servicios de AWS, como S3, EMR, Kinesis y más, lo que permite una gestión y el análisis de datos más cohesiva.
- ▶ **Compresión de datos.** Redshift utiliza técnicas avanzadas de compresión de datos, lo que reduce el tamaño de almacenamiento necesario y mejora el rendimiento de las consultas.
- ▶ **Almacenamiento en columnas.** Emplea un almacenamiento en columnas, que optimiza las consultas de lectura intensiva, lo que permite un acceso rápido y eficiente a grandes conjuntos de datos.
- ▶ **Copias de seguridad automáticas.** Realiza copias de seguridad automáticas de los datos para garantizar la durabilidad y la recuperación en caso de fallos.
- ▶ **Data sharing.** Para compartir de datos entre clústeres de manera segura y eficiente.
- ▶ **Integración con Spark.**

La Figura 21 muestra las principales características y el flujo de funcionamiento de RedShift.

Tema 3. Data science cloud storage



Figura 21. Características y flujo de funcionamiento de RedShift. Fuente: Amazon Web Services, s. f.-e.

Funcionalidades de Amazon Redshift

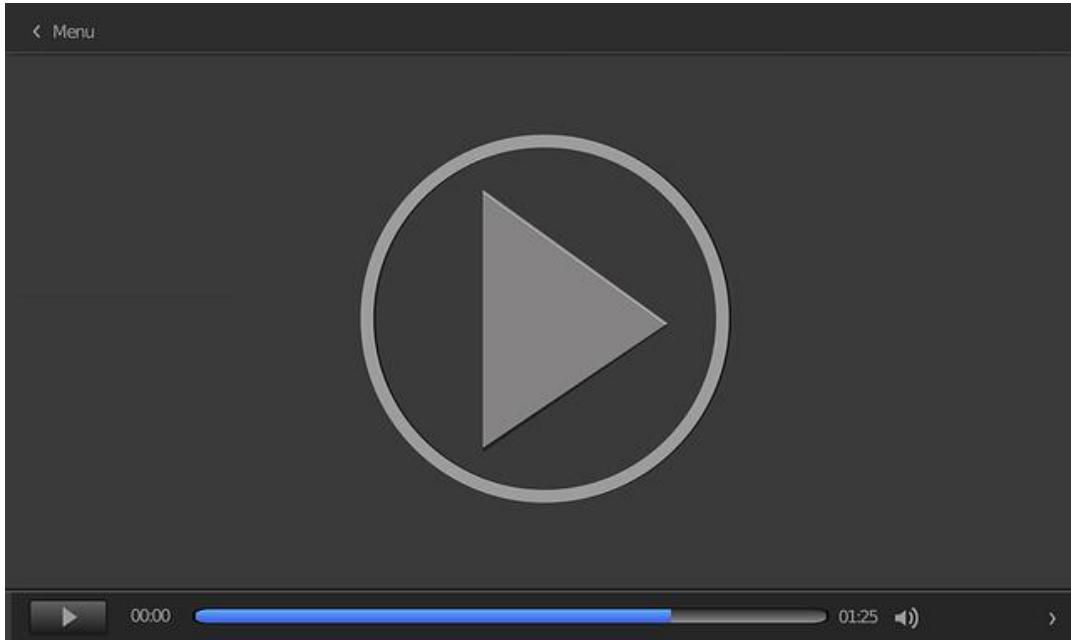
Entre las principales funcionalidades de Redshift, se pueden destacar las siguientes:

- ▶ **Redshift Spectrum.** Permite ejecutar consultas directamente sobre datos almacenados en Amazon S3 sin necesidad de cargar estos datos en Redshift, combinando el almacenamiento en columnas con la flexibilidad del almacenamiento en objetos.
- ▶ **Concurrencia y escalabilidad.** Redshift ofrece características para manejar múltiples consultas simultáneamente y ajustarse dinámicamente a las cargas de trabajo cambiantes.
- ▶ **Seguridad.** Soporta cifrado de datos en reposo y en tránsito, integración con AWS IAM para gestión de acceso y capacidades de auditoría y monitoreo a través de CloudTrail.
- ▶ **Optimización automática.** Redshift proporciona análisis y recomendaciones automáticas para optimizar las consultas y el rendimiento del clúster.

Tema 3. Data science cloud storage

- ▶ **Compatibilidad con *data lakes*.** Se integra con Amazon Lake Formation para construir *data lakes*, lo que permite que las organizaciones combinen *data warehouses* y *data lakes*.

En el vídeo *Getting Started with Amazon Redshift - AWS Online Tech Talks* (Amazon Web Services, 2020a), se realiza una introducción sobre cómo comenzar a trabajar en Amazon Redshift.



Getting Started with Amazon Redshift - AWS Online Tech Talks.

Accede al vídeo:

<https://www.youtube.com/embed/dfo4J5ZhlKI>

Tema 3. Data science cloud storage

Ventajas de Amazon Redshift

A continuación, se enumeran una serie de ventajas que aporta el uso de AWS Redshift:

- ▶ **Alto rendimiento.** La arquitectura de almacenamiento en columnas y la distribución de carga entre nodos garantizan un rendimiento elevado para consultas complejas y grandes volúmenes de datos.
- ▶ **Escalabilidad.** Redshift puede escalar fácilmente añadiendo o eliminando nodos según las necesidades de procesamiento y almacenamiento.
- ▶ **Costo-efectividad.** Ofrece un modelo de precios basado en el uso, lo que permite a las empresas pagar solo por los recursos que utilizan. Las opciones de pago por demanda y precios reservados a largo plazo permiten un mejor control de costos.
- ▶ **Fácil de usar.** La gestión y la configuración de Redshift son sencillas gracias a la integración con el ecosistema de AWS y las herramientas de gestión proporcionadas.
- ▶ **Integración y ecosistema.** La estrecha integración con otros servicios de AWS y herramientas de terceros facilita la creación de soluciones de análisis de datos completas y personalizadas.

Casos de uso de Amazon Redshift

Como casos de uso donde utilizar AWS Redshift, podemos describir los siguientes:

- ▶ **Análisis de datos empresariales.** Empresas pueden utilizar Redshift para analizar datos de ventas, *marketing*, finanzas y operaciones, lo que permite una toma de decisiones basada en datos.
- ▶ **BI y reporting.** Organizaciones utilizan Redshift para generar informes y cuadros de mando interactivos en tiempo real, integrándose con herramientas de BI como Tableau, Looker y Amazon QuickSight.

Tema 3. Data science cloud storage

- ▶ **Big data analytics.** Capaz de manejar grandes volúmenes de datos, Redshift es ideal para análisis de *big data*, lo que les permite a las empresas descubrir patrones y tendencias en sus datos.
- ▶ **Análisis de clickstream.** Empresas de comercio electrónico y medios digitales usan Redshift para analizar datos de *clickstream* y comprender mejor el comportamiento de los usuarios en sus plataformas.
- ▶ **Integración de data lakes.** Combinando Redshift con Amazon S3 y Redshift Spectrum, las organizaciones pueden analizar datos estructurados y no estructurados de manera conjunta, creando soluciones de análisis de datos más flexibles y poderosas.

En resumen, **Amazon Redshift** es una solución robusta y flexible para el almacenamiento y el análisis de grandes volúmenes de datos, y ofrece un alto rendimiento, escalabilidad y una integración fluida con otros servicios de AWS y herramientas de análisis.

Uso y funcionamiento de Redshift

A continuación, se describe el uso y funcionamiento de AWS Redshift. Comenzamos por la **configuración inicial** de AWS Redshift:

- ▶ **Crear un clúster de Redshift:**
 - Inicia sesión en la consola de administración de AWS.
 - Navega a Amazon Redshift y selecciona «Create Cluster».
 - Configura el clúster especificando el nombre, el tipo de nodos, el número de nodos y otras configuraciones de red y seguridad.

Tema 3. Data science cloud storage

► Configuración de redes y seguridad:

- Configura el VPC (*virtual private cloud*), las subredes y los grupos de seguridad para controlar el acceso al clúster.
- Asegúrate de configurar las reglas de entrada y salida para permitir el acceso desde las ubicaciones permitidas.

► Configurar usuarios y privilegios:

- Utiliza el entorno de SQL de Redshift para crear usuarios y roles, y asigna los privilegios necesarios para el acceso y la administración de los datos.

► Carga de datos:

- Los datos pueden cargarse en Redshift desde diversas fuentes, como Amazon S3, bases de datos relacionales, servicios de flujo de datos, como Kinesis, y a través de herramientas ETL.
- Utiliza el comando `COPY` para cargar datos desde Amazon S3, DynamoDB o archivos locales en las tablas de Redshift.

```
sql
COPY table_name
FROM 's3://bucket_name/file_path'
CREDENTIALS 'aws_access_key_id=...;aws_secret_access_key=...'
CSV;
```

En cuanto al funcionamiento general de Redshift, los principales **componentes** se describen a continuación:

- **Arquitectura.** Redshift usa una arquitectura MPP (procesamiento masivo en paralelo) que distribuye datos y consultas a través de múltiples nodos. Los datos se almacenan en un formato columnar, que es optimizado para las consultas analíticas. Un ejemplo de arquitectura columnar y MPP de Redshift se muestra en la Figura 22.

Tema 3. Data science cloud storage

- ▶ **Consultas SQL.** Redshift soporta SQL estándar, lo que permite realizar operaciones de selección, inserción, actualización y eliminación. Las consultas se distribuyen entre los nodos para aprovechar el procesamiento paralelo.
- ▶ **Optimización.** Redshift incluye características de optimización automática, como el ajuste de consultas, almacenamiento en caché de resultados y recomendaciones de distribución de datos.
- ▶ **Query Editor de Redshift.** El Query Editor de Redshift es una herramienta integrada en la consola de AWS que les permite a los usuarios escribir y ejecutar consultas SQL directamente en el clúster de Redshift sin necesidad de instalar un *software* adicional.

Tema 3. Data science cloud storage

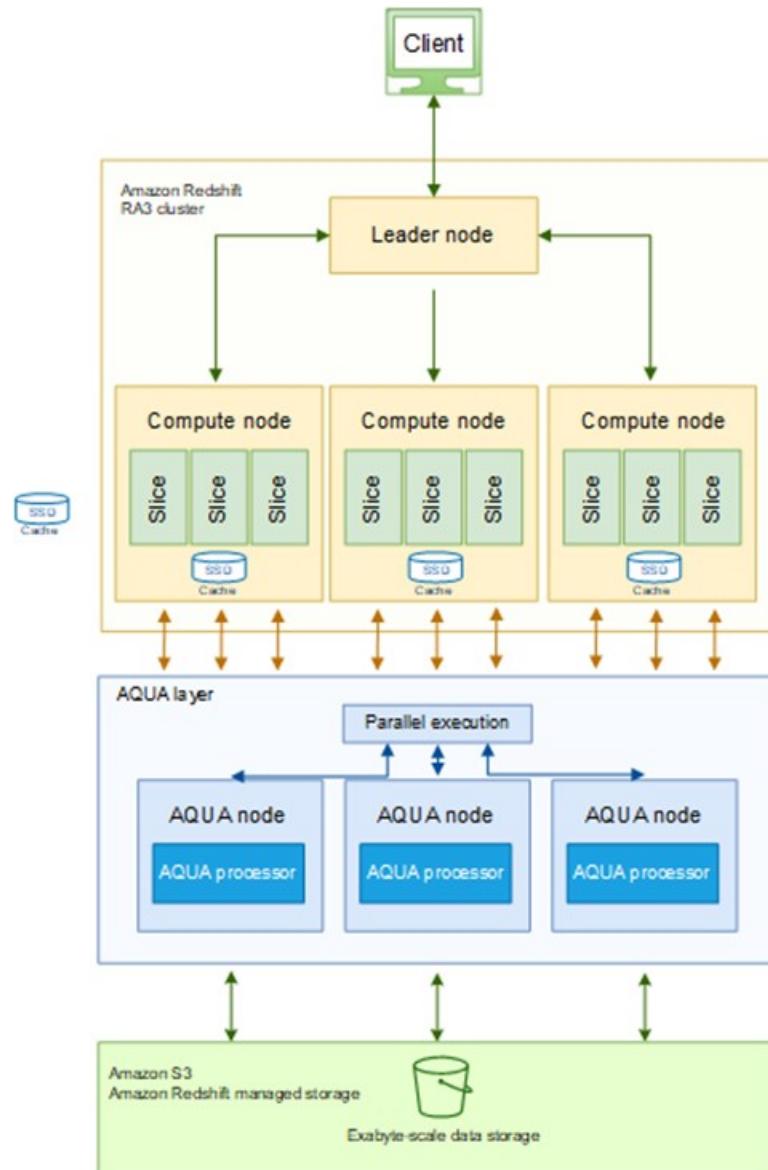


Figura 22. Arquitectura columnar y MPP de Redshift.

Tema 3. Data science cloud storage

A continuación, se muestra **cómo usar el Query Editor** y los pasos que hay que realizar:

► **Accede al Query Editor:**

- Inicia sesión en la consola de AWS y navega a Amazon Redshift.
- Selecciona el clúster de Redshift que deseas usar y haz clic en «Query Editor» en el panel de navegación.

► **Conexión al clúster:**

- Ingresa las credenciales necesarias para conectarte al clúster (nombre de usuario y contraseña).
- Selecciona la base de datos y el esquema con los que deseas trabajar.

► **Escribir y ejecutar consultas:**

- En el editor de consultas, escribe tus comandos SQL. Puedes realizar selecciones, inserciones, actualizaciones y eliminar datos, así como crear y modificar tablas.
- Haz clic en «Run» para ejecutar la consulta.

```
sql
SELECT * FROM users WHERE age > 30;
```

► **Visualización de resultados:**

- Los resultados de las consultas se mostrarán en una tabla debajo del editor.
- Puedes exportar los resultados en formatos CSV para su análisis posterior.

Tema 3. Data science cloud storage

► **Administración de consultas:**

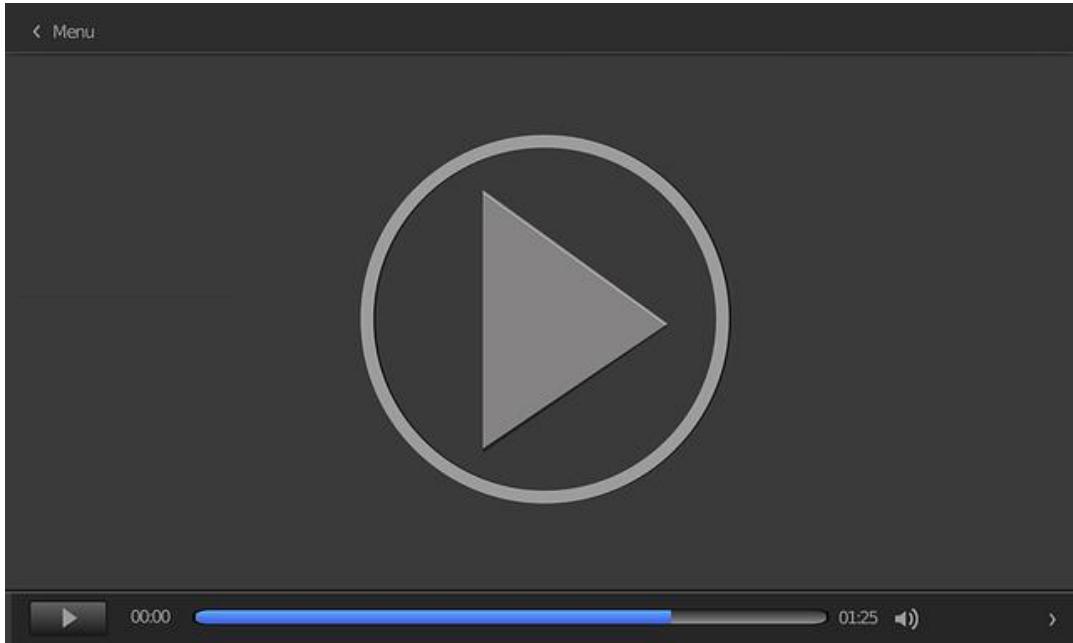
- El Query Editor permite guardar consultas frecuentes para reutilizarlas.
- Puedes ver el historial de consultas ejecutadas y reutilizar consultas anteriores.

Las **ventajas** del Query Editor son:

- **Facilidad de uso.** Proporciona una interfaz sencilla y accesible para escribir y ejecutar consultas SQL sin la necesidad de herramientas adicionales.
- **Acceso inmediato.** Permite el acceso inmediato a los datos del clúster de Redshift para realizar análisis *ad hoc*.
- **Sin instalación adicional.** Todo se maneja a través de la consola de AWS, lo que elimina la necesidad de instalar y de configurar clientes de SQL o herramientas ETL adicionales.
- **Integración completa.** El Query Editor está completamente integrado con el ecosistema de AWS, lo que facilita el acceso y la manipulación de datos en Redshift.

Tema 3. Data science cloud storage

A continuación, se muestra el vídeo *An Overview of Amazon Redshift Query Editor V2 / Amazon Web Services* (Amazon Web Services, 2021b), que explica el AWS RedShift Query Editor.



An Overview of Amazon Redshift Query Editor V2 | Amazon Web Services.

Accede al vídeo:

<https://www.youtube.com/embed/IwZNlroJUnc>

En resumen, **Amazon Redshift** es una potente solución de *data warehouse* en la nube que ofrece escalabilidad, rendimiento y facilidad de uso. Su integración con el ecosistema de AWS y herramientas como el Query Editor facilita la gestión, el análisis y la visualización de grandes volúmenes de datos, haciendo que las empresas puedan tomar decisiones informadas basadas en datos de manera eficiente.

Redshift Serverless

Amazon Redshift Serverless es una variante del servicio de *data warehouse* en la nube Amazon Redshift que elimina la necesidad de administrar infraestructura.

Tema 3. Data science cloud storage

Proporciona una experiencia totalmente gestionada, donde AWS gestiona automáticamente el aprovisionamiento, la configuración, la escalabilidad y el mantenimiento de los recursos de almacenamiento y procesamiento de datos.

Redshift Serverless permite a los usuarios ejecutar y escalar cargas de trabajo de análisis de datos sin preocuparse por la **administración de clústeres**. Los usuarios simplemente cargan sus datos y ejecutan consultas, y Redshift Serverless ajusta automáticamente los recursos necesarios para satisfacer las demandas de carga de trabajo.

Entre las principales **funcionalidades**, tenemos las siguientes:

- ▶ **Autoescalado.** Escala automáticamente los recursos de computación y almacenamiento en función de las necesidades de las cargas de trabajo, sin intervención manual.
- ▶ **Precios basados en el uso.** Cobra solo por el tiempo de consulta y los recursos utilizados, lo que puede resultar más económico para cargas de trabajo variables o intermitentes.
- ▶ **Simplicidad de configuración.** No requiere configuración de nodos, clústeres ni administración de infraestructura. Los usuarios pueden comenzar a cargar datos y ejecutar consultas rápidamente.
- ▶ **Compatibilidad completa con Redshift.** Total compatibilidad con las funciones y las características de Amazon Redshift, lo que incluye consultas SQL, integraciones con AWS y herramientas de BI.
- ▶ **Inicio rápido.** Proporciona un punto de partida rápido para ejecutar consultas sin necesidad de aprovisionar clústeres. Ideal para pruebas, desarrollo y análisis *ad hoc*.
- ▶ **Seguridad y cumplimiento.** Ofrece características avanzadas de seguridad, como cifrado de datos en reposo y en tránsito, integración con AWS IAM y cumplimiento con regulaciones y normativas de seguridad.

Tema 3. Data science cloud storage

Respecto a las **ventajas** que ofrece Redshift Serverless, tenemos las siguientes:

- ▶ **Administración simplificada.** Elimina la necesidad de gestionar la infraestructura del *data warehouse*, lo que permite que los usuarios se centren en el análisis y la obtención de *insights*.
- ▶ **Escalabilidad automática.** Ajusta automáticamente los recursos de computación y almacenamiento para satisfacer las demandas de las cargas de trabajo, lo que asegura un rendimiento óptimo.
- ▶ **Costes reducidos.** El modelo de pago por uso puede reducir significativamente los costos operativos, especialmente para cargas de trabajo intermitentes o variables.
- ▶ **Rapidez y flexibilidad.** Permite iniciar rápidamente con análisis de datos sin la necesidad de configuraciones complejas, lo que facilita la agilidad y la flexibilidad en el trabajo con datos.
- ▶ **Integración con el ecosistema AWS.** Se integra perfectamente con otros servicios de AWS, como S3, Glue, Kinesis y Lambda, lo que facilita la creación de *pipelines* de datos y soluciones de análisis complejas.

Por último, tenemos algunos ejemplos de **casos de uso** de Redshift Serverless:

- ▶ **Análisis ad hoc.** Ideal para análisis de datos no planificados donde la carga de trabajo es impredecible. Los analistas pueden ejecutar consultas *ad hoc* sin preocuparse por la capacidad de la infraestructura.
- ▶ **Entornos de prueba y desarrollo.** Permite a los desarrolladores y científicos de datos probar y desarrollar nuevas consultas y modelos analíticos sin necesidad de configurar y gestionar clústeres dedicados.

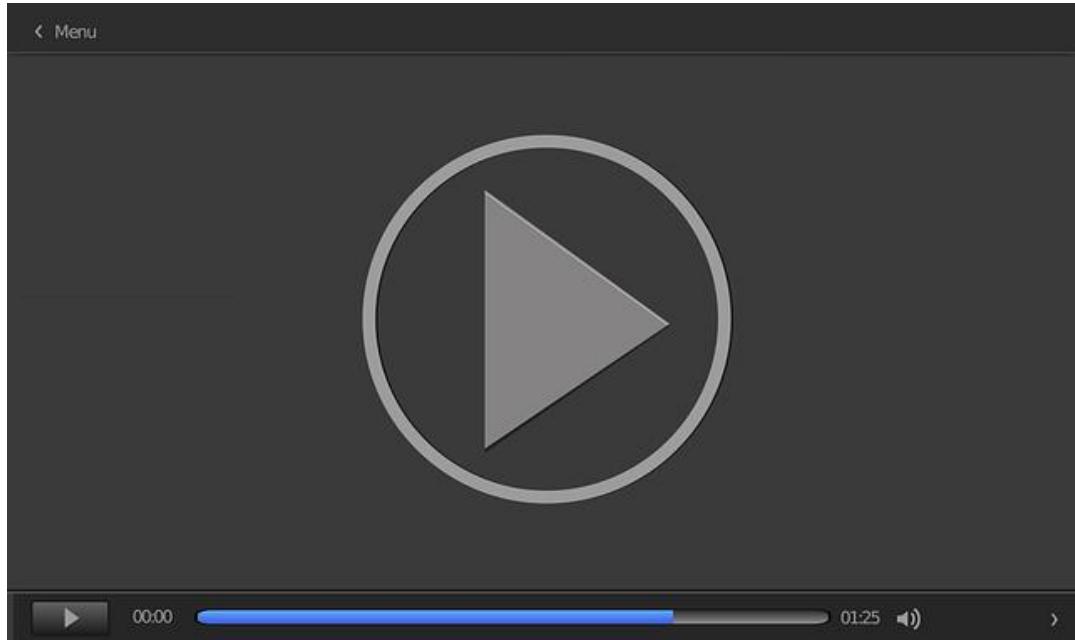
Tema 3. Data science cloud storage

- ▶ **Aplicaciones intermitentes.** Es adecuado para aplicaciones que requieren análisis de datos de manera intermitente, por lo que evita el costo de mantener clústeres en funcionamiento continuo.
- ▶ **Pequeñas y medianas empresas (pymes).** Las pymes que no tienen los recursos para gestionar infraestructura de *data warehousing* pueden beneficiarse de la simplicidad y el modelo de pago por uso de Redshift Serverless.
- ▶ **Análisis de eventos y datos de IoT.** Permite analizar grandes volúmenes de datos generados por eventos o dispositivos IoT de manera eficiente, sin necesidad de gestionar la infraestructura subyacente.

Amazon Redshift Serverless ofrece una solución de data warehousing poderosa y flexible que elimina la complejidad de la administración de infraestructura. Con su capacidad de autoescalamiento, precios basados en uso y fácil integración con el ecosistema de AWS, es una excelente opción para organizaciones que buscan agilizar sus procesos de análisis de datos y reducir costos operativos. Redshift Serverless proporciona las herramientas necesarias para que las empresas se centren en obtener *insights* valiosos de sus datos sin preocuparse por la gestión de recursos.

Tema 3. Data science cloud storage

A continuación, se muestra el vídeo *Amazon Redshift Serverless - End to End Use Case / Amazon Web Services* (Amazon Web Services, 2022a), que muestra el funcionamiento de AWS Redshift Serverless.



Amazon Redshift Serverless - End to End Use Case | Amazon Web Services.

Accede al vídeo:

<https://www.youtube.com/embed/gEAWSqjthXs>

Redshift Data Sharing

Amazon Redshift incluye una funcionalidad llamada **Redshift Data Sharing** que permite compartir datos de manera segura y eficiente entre distintos clústeres de Redshift. Esta característica es especialmente útil para organizaciones que necesitan colaborar entre diferentes equipos o departamentos, permitiendo un acceso rápido a los datos sin necesidad de mover o copiar grandes volúmenes de información.

Tema 3. Data science cloud storage

Como **características** del Redshift Data Sharing, se pueden enumerar las siguientes:

- ▶ **Acceso en tiempo real.** Permite compartir datos en tiempo real entre diferentes clústeres de Redshift sin necesidad de duplicar los datos. Los usuarios pueden acceder a los datos más actualizados en cualquier momento.
- ▶ **Seguridad y control de acceso.** Los permisos de acceso se pueden gestionar de manera granular. Los propietarios de los datos pueden controlar quién puede acceder a qué datos, por lo que de este modo se garantiza la seguridad y la privacidad.
- ▶ **Eficiencia en el uso de recursos.** Al compartir datos sin duplicarlos, se reduce la necesidad de almacenamiento adicional y se optimiza el uso de los recursos de almacenamiento y procesamiento.
- ▶ **Escalabilidad.** Facilita la colaboración entre múltiples equipos o departamentos, cada uno con su propio clúster de Redshift, sin preocuparse por los límites físicos de un solo clúster.
- ▶ **Simplicidad.** El proceso de compartir datos es sencillo y no requiere configuraciones complicadas ni la implementación de ETL adicionales.

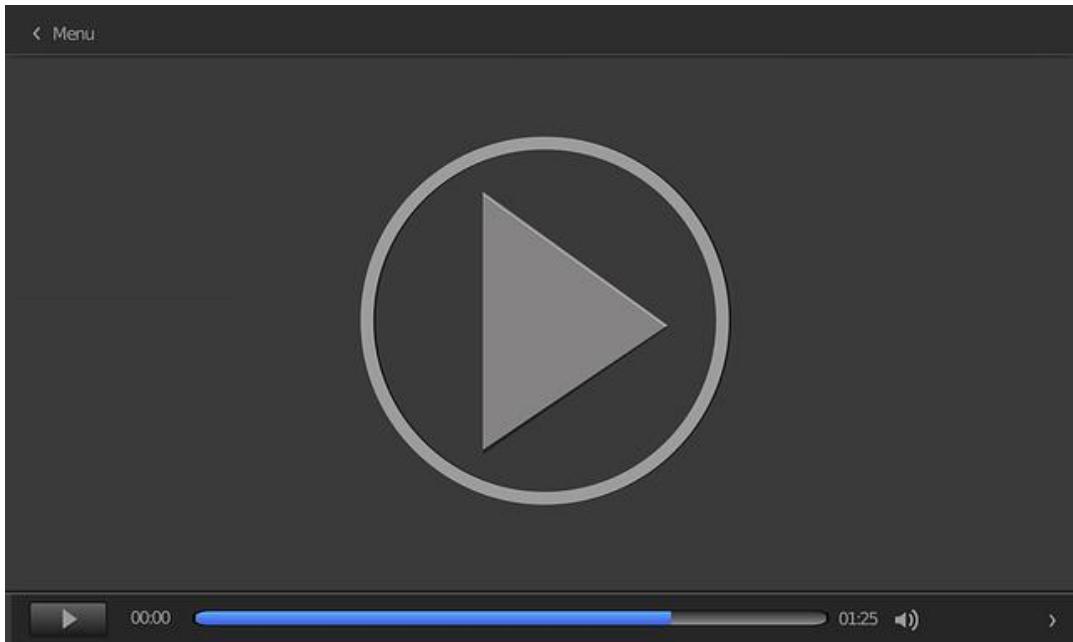
Las principales **funcionalidades** de Redshift Sharing son:

- ▶ **Compartición entre clústeres.** Permite a un clúster de Redshift (clúster productor) compartir datos con uno o más clústeres (clústeres consumidores) de manera directa.
- ▶ **Compatibilidad con SQL.** Los usuarios pueden ejecutar consultas SQL en los datos compartidos de la misma manera que lo harían con los datos locales en sus clústeres.

Tema 3. Data science cloud storage

- ▶ **Soporte para diferentes regiones.** Redshift Data Sharing soporta la compartición de datos entre clústeres en diferentes regiones de AWS, lo que facilita la colaboración global.
- ▶ **Metadatos y estadísticas.** Los clústeres consumidores tienen acceso a los metadatos y las estadísticas de los datos compartidos, lo que mejora el rendimiento de las consultas.

A continuación, se muestra el vídeo *Amazon Redshift Data Sharing Workflow* (Amazon Web Services, 2020b), que muestra un ejemplo del funcionamiento de AWS Redshift Data Sharing.



Amazon Redshift Data Sharing Workflow.

Accede al vídeo:

<https://www.youtube.com/embed/EXioFirlnA>

Tema 3. Data science cloud storage

Las **ventajas** que ofrece Redshift Data Sharing son las siguientes:

- ▶ **Colaboración mejorada.** Facilita la colaboración entre diferentes equipos y departamentos al permitir el acceso compartido a datos críticos sin necesidad de copiar grandes volúmenes de información.
- ▶ **Aceleración del análisis.** Los datos están disponibles para análisis en tiempo real, lo que acelera los procesos de toma de decisiones basadas en datos.
- ▶ **Reducción de costos.** Al evitar la duplicación de datos, se reducen los costos de almacenamiento y se optimiza el uso de recursos.
- ▶ **Flexibilidad operacional.** Los equipos pueden trabajar de manera independiente en sus propios clústeres mientras comparten datos según sea necesario, lo que mejora la flexibilidad y la eficiencia operativa.

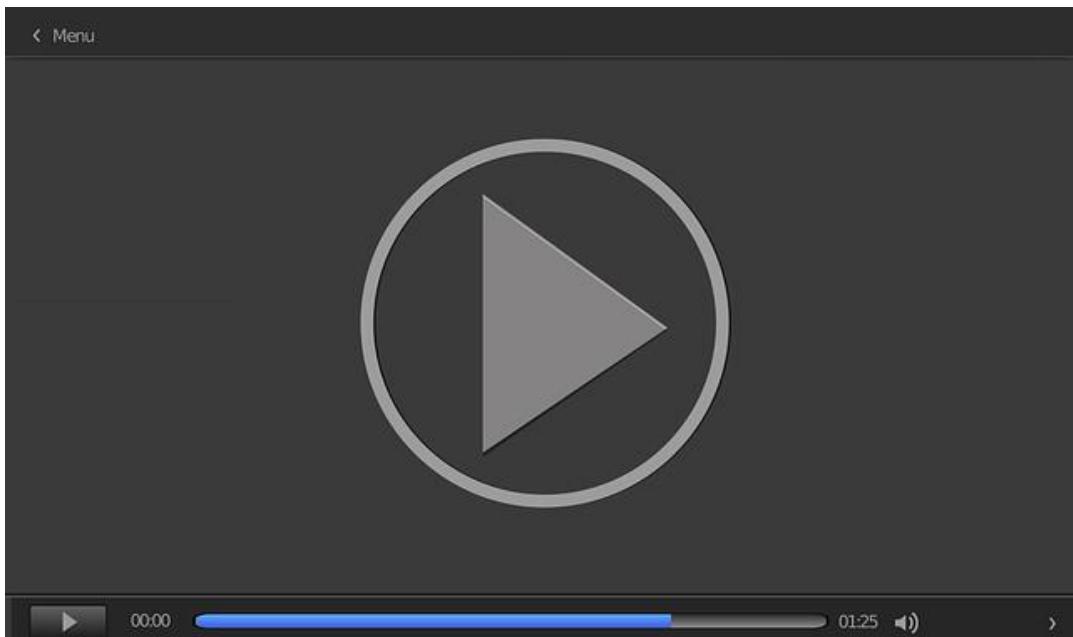
Los principales **casos de uso** de Redshift Data Sharing son:

- ▶ **Análisis conjunto entre equipos.** Equipos de *marketing*, ventas y finanzas pueden compartir datos entre sí para obtener una visión holística del rendimiento de la empresa y tomar decisiones informadas.
- ▶ **Colaboración entre departamentos.** Diferentes departamentos dentro de una empresa pueden acceder a los mismos datos para realizar análisis específicos sin necesidad de copiar los datos a cada clúster.
- ▶ **Integración con aplicaciones externas.** Empresas que trabajan con socios o clientes externos pueden compartir datos de manera segura para permitir el acceso a información relevante sin comprometer la seguridad.
- ▶ **Multirregional data analysis.** Organizaciones globales pueden compartir datos entre clústeres ubicados en diferentes regiones geográficas para realizar análisis distribuidos y coordinados.

Tema 3. Data science cloud storage

En resumen, **Redshift Data Sharing** es una poderosa funcionalidad que mejora la colaboración y eficiencia en el análisis de datos al permitir compartir datos de manera segura y en tiempo real entre distintos clústeres de Amazon Redshift. Esta característica ofrece ventajas significativas en términos de costos, rendimiento y flexibilidad operativa, haciendo que las organizaciones sean más ágiles y colaborativas.

En el vídeo *Amazon Redshift Data Sharing Use Cases* (Amazon Web Services, 2020c), se muestran casos de uso de aplicación de Redshift Data Sharing.



Amazon Redshift Data Sharing Use Cases.

Accede al vídeo:

<https://www.youtube.com/embed/sloTB8B5nn4>

Tema 3. Data science cloud storage

Redshift Machine Learning

Amazon Redshift ML es una funcionalidad de Amazon Redshift que permite a los usuarios crear, entrenar y desplegar modelos de ML directamente dentro de su *data warehouse*. Esta integración simplifica el proceso de incorporar capacidades de ML en las consultas de datos, ya que elimina la necesidad de mover datos a servicios externos para el entrenamiento y la inferencia de modelos.

El **concepto** detrás de Redshift ML es llevar el poder del ML directamente al *data warehouse* mientras aprovecha el entorno de SQL familiar para los analistas de datos y científicos de datos. Utiliza Amazon SageMaker, el servicio de ML de AWS, para el entrenamiento de modelos, donde los resultados se integran de vuelta en Redshift, lo que permite que las predicciones se realicen directamente dentro de las consultas SQL.

Entre las principales **funcionalidades** de Redshift ML, se puede destacar lo siguiente:

- ▶ **Creación de modelos desde SQL.** Permite a los usuarios crear modelos de ML utilizando comandos SQL, lo que facilita la creación de modelos sin necesidad de conocimientos profundos en ML.
- ▶ **Entrenamiento automático.** Redshift ML utiliza Amazon SageMaker para entrenar los modelos automáticamente. Los usuarios pueden especificar los datos y las características a utilizar, y SageMaker se encarga del resto.
- ▶ **Inferencia directa en SQL.** Una vez que los modelos están entrenados, las predicciones se pueden realizar directamente en las consultas SQL, lo que facilita la integración de *insights* de ML en los análisis de datos.
- ▶ **Gestión de modelos.** Ofrece herramientas para gestionar el ciclo de vida de los modelos de ML, incluyendo actualizaciones y reentrenamiento directamente desde Redshift.