

Data Warehousing on AWS

Introducing Amazon Redshift.

Amazon Redshift es una solución de almacenamiento de datos en la nube totalmente administrada que ofrece Amazon Web Services (AWS). Está diseñada para proporcionar capacidades analíticas rápidas y escalables para grandes cantidades de datos. Redshift utiliza un almacén de datos en columnas y una arquitectura de procesamiento masivo en paralelo para gestionar consultas complejas y ofrecer resultados rápidamente. Con Redshift, los usuarios pueden crear y gestionar fácilmente almacenes de datos, cargar datos de diversas fuentes y realizar consultas complejas mediante SQL. También ofrece una serie de características para la seguridad de los datos, la optimización del rendimiento y la gestión de costos. Amazon Redshift es una opción popular para empresas y organizaciones que requieren una solución de almacenamiento de datos flexible y rentable que pueda escalar en función de sus necesidades.

Modern Analytics and data warehousing architecture.

AWS Analytics Services es un conjunto completo de herramientas y servicios basados en la nube que permiten a las organizaciones crear soluciones de análisis integrales. Pipeline está diseñado para enfrentar los desafíos de big data y análisis al proporcionar soluciones escalables y flexibles para logear, almacenar, procesar y visualizar datos. Incluye varios servicios como Amazon S3 para almacenamiento de datos, Amazon Redshift para almacenamiento de datos, AWS Glue para transformación de datos, Amazon EMR para procesamiento de datos y Amazon QuickSight para visualización de datos. El producto también se puede ampliar con herramientas y servicios de terceros, como Apache Spark y Jupyter Notebook, para crear soluciones personalizadas que satisfagan necesidades comerciales específicas.

Las organizaciones pueden aprovisionar recursos de forma rápida y sencilla cuando sea necesario para que puedan escalar hacia arriba o hacia abajo sus soluciones de análisis según sea necesario para cumplir con los requisitos cambiantes. Además, el modelo de precios de pago por uso de AWS permite a las organizaciones pagar solo por los recursos que utilizan, lo que lo convierte en una solución rentable para organizaciones pequeñas y grandes. Finalmente, AWS ofrece una amplia gama de características de seguridad, como cifrado, control de acceso y monitoreo, para garantizar la seguridad de los datos y el cumplimiento de las normas y regulaciones de la industria. En general, AWS Analytics Services Pipeline proporciona una solución poderosa y flexible para las organizaciones que desean crear soluciones de análisis modernas en la nube.

Data warehouse technology options.

Las bases de datos basadas en filas almacenan }datos en filas, lo que significa que todos los datos de un registro en particular se almacenan juntos. Esto facilita la recuperación de un registro completo a la vez, pero puede ser lento si realiza consultas que requieren agregación o cálculos en grandes conjuntos de datos. En cambio, las bases de datos basadas en columnas almacenan datos por columna, es decir, todos los valores de un determinado atributo se almacenan juntos. Esto permite consultas más rápidas al seleccionar columnas específicas o realizar uniones, pero puede ser más lento al recuperar registros completos.

Las bases de datos basadas en filas son adecuadas para sistemas transaccionales y situaciones en las que se deben recuperar registros completos con frecuencia. Por otro

lado, las bases de datos basadas en columnas son más adecuadas para sistemas analíticos y situaciones en las que el rendimiento rápido de las consultas y la agregación son fundamentales. Muchas soluciones modernas de almacenamiento de datos, incluido Amazon Redshift, utilizan el almacenamiento en filas y columnas para lograr lo mejor de ambos mundos.

Las arquitecturas de procesamiento paralelo masivo permiten el procesamiento de grandes conjuntos de datos en múltiples nodos, lo que permite un rendimiento de consulta más rápido y una mayor escalabilidad. Los datos se comparten entre varios nodos o clústeres, cada uno de los cuales puede procesar parte de los datos en paralelo. Esto permite tiempos de ejecución de consultas más rápidos y la capacidad de escalar el sistema según sea necesario agregando nuevos nodos o clústeres.

Amazon redshift deep dive.

Redshift una amplia gama de funciones que permiten a las organizaciones crear soluciones de servicios de datos flexibles y escalables, está diseñado para proporcionar escalabilidad y flexibilidad, lo que permite a las organizaciones ampliar o reducir fácilmente sus soluciones de almacenamiento de datos según sea necesario. Redshift permite a los usuarios aprovisionar y escalar recursos informáticos y de almacenamiento de forma independiente, lo que les permite optimizar su almacén de datos en cuanto a rendimiento y costo.

Se crean tres copias de todos los datos almacenados en múltiples zonas de disponibilidad (AZ) para la resiliencia. Además, Redshift admite la recuperación de un punto en el tiempo, lo que permite a los usuarios restaurar sus datos en un punto específico en el tiempo, lo que garantiza la disponibilidad de los datos y la continuidad del negocio. Su arquitectura de procesamiento paralelo masivo (MPP) ofrece un rendimiento de consulta rápido en grandes conjuntos de datos. Redshift también utiliza almacenamiento en columnas y técnicas de compresión avanzadas para optimizar el almacenamiento y la recuperación de datos.

Es importante recordar que Redshift admite la integración del lago de datos que permite a las organizaciones almacenar grandes cantidades de datos no estructurados o semiestructurados. Redshift Spectrum permite a los usuarios consultar datos almacenados en Amazon S3, lo que les permite acceder y analizar datos entre el lago de datos y el almacén de datos.

Operations.

Ese sistema de manejo cuenta con un asistente llamado Redshift Advisor que constantemente proporciona recomendaciones. Siempre optimizando operaciones, Amazon Redshift cuenta con gran automatización de tareas como lo son el auto analizar de los cluters para su debido funcionamiento y la optimización de costo ya incluida. En términos de seguridad, Amazon Redshift ofrece varias funciones para proteger los datos, incluido el aislamiento de la red, el cifrado de datos en tránsito y en reposo, y la capacidad de controlar el acceso mediante políticas de IAM y VPC. El modelo de costos de Amazon Redshift se basa en varios factores, como la cantidad de datos almacenados, la cantidad de nodos en el clúster y el nivel de uso. También hay disponibles diferentes modelos de precios, como bajo demanda, escala reservada y concurrente. Los usos ideales para Amazon Redshift incluyen escenarios en los que es necesario almacenar y analizar rápidamente grandes volúmenes de datos, como almacenamiento de datos, inteligencia comercial y análisis.