Moisés Solano Espinoza-2021144322. Resumen de ***"Data Warehousing on AWS: AWS Whitepaper"***

Introducing Amazon Redshift

Amazon Redshift es una administrada, rápida y a escala de petabyte solución de warehousing que permite analizar grandes volúmenes de datos con el uso de herramientas BI actuales de forma simple y rentable. Ofrece velocidad de motores de almacenamiento columnar. Se puede ver como un almacen de datos en la nube. Tiene la capacidad de crecer hasta niveles de exabytes. Es uno de los servicios de AWS de más rápido crecimiento.

Modern analytics and data warehousing architecture

Los datos que llegan a los data warehouse pueden ser estructurados, semiestructurados o no estructurados. Los DW están optimizados para operaciones de escritura y lectura de grandes volúmenes de datos y las OLTP para grandes volúmenes de escrituras pequeñas. Los DW utilizan esquemas desnormalizados como STAR y Snowflake.

AWS anlaytics services: ayudan a convertir rápidamente datos en respuestas con servicios de análisis maduros e integrados. Ofrece un camino fácil para construir DW y DL, una infraestructura segura y el mejor rendimiento, estabilidad y menor costo. Se enfoca en proteger los datos en cuestión de días. Los datos están curados y catalogados y se elimina la información duplicada.

Analytics architecture: las analytics pipelines manejan grandes volúmenes de entradas de datos. Tienen etapas: recopilar datos, almacenarlos, procesarlos, analizarlos y visualizarlos.

Data collection: AWS proporciona soluciones para el almacenamiento de diferentes tipos de datos.

Transactional data: una base de datos no SQL se usa cuando los datos no están bien estructurados y una RDMBS cuando se requieren uniones complejas. Amazon ofrece soluciones de bases de datos como DynamoDB, Aurora y RDS.

Log data: captura confiable de registros que ayuda a solucionar problemas y análisis utilizando datos de registros.

Streaming data: las aplicaciones generan grandes cantidades de información que debe recopilarse y procesar. Se dice que es en tiempo real.

IOT data: los dispositivos mandan mensajes continuamente, esos datos deben ser capturados y manejados para obtener inteligencia de ellos. Con IoT los dispositivos se comunican fácilmente con la nube de AWS.

Data processing: se analiza la información en busca de inteligencia. La mejor forma de obtener esta inteligencia es con los data warehouse. Se puede hacer procesamiento por lotes o en tiempo real

Batch processing: se divide en ETL (extracción de datos para cargar en sistemas de almacenamiento de datos una vez han sido enriquecidos y transformados), ELT (en el que los datos primero se cargan en el sistema destino a diferencia del primero) y OLAP (que almacena esquemas multidimensionales de datos históricos agregados, utilizados para consultas, informes y análisis).

Real-Time Processing: la información que se obtiene del procesado en tiempo real ayuda a tener visibilidad de los datos, analizarlos y poder responder a situaciones. Requiere una capa de procesamiento altamente

concurrente y escalable. Amazon ofrece distintas herramientas para trabajar en Real-Time como AWS Lambda, KCL, Kinesis Data Firehouse, MSK, AWS Glue.

Data storage: los datos se pueden almacenar en un lake house (combina lo mejor de warehouse y data lake, permite consultar datos en WH, DL y bases de datos operativas para obtener informacion más rápido; además se pueden almacenar los datos en formatos de archivos abiertos), warehouse (se pueden ejecutar análisis rápidos en grandes volúmenes de datos y describir patrones) o data mart (es una forma simple de WH centrado en un área en específico. Son fáciles de diseñar, construir y administrar). Con Amazon Redshift se pueden crear las tres anteriores.

Analysis and visualization: una vez que ya se obtuvieron y procesaron los datos, se necesitan herramientas para análisis y visualización (BI). Un ejemplos es utilizar en MySQL Workbench Amazon Redshift utilizando ANSI SQL. Amazon ofrece servicios como QuickSight, Redshift, S3, Athena y RDS. También existen otros como Apache Zeppelin, Tableau y MicroStrategy.

Analytics pipeline with AWS services: AWS ofrece un amplio conjunto de servicios para implementar una plataforma de análisis integral.

Data warehouse technology options

Opciones disponibles para crear warehouse:

Row-oriented databases: suelen almacenar filas completas en un bloque fijo. Los índices secundarios brindan un alto rendimiento de operaciones de lectura. Son eficientes para el procesamiento transaccional (OLTP). Si se quiere utilizar como warehouse se deben construir vistas materializadas, tablas de resumen, particiones de datos y joins basados en índices. La desventaja es que para cada consulta se debe leer todas las columnas de todas las filas que satisfacen la consulta. Los WH y data marts que utilizan este tipo de DB están limitados a los recursos de la máquina y a posibles ralentizaciones.

Colum-oriented databases: organizan cada columna en su propio conjunto de bloques físicos. Son más eficientes en la (I/O) para consultas de lectura. Para el almacenamiento de datos las bases de datos por columnas son una mejor opción. También tienen una mejor compresión, porque las compresiones se pueden hacer con datos del mismo tipo, lo que resulta en menos uso del almacenamiento.

Massively Parallel Processing (MPP) architectures: permite utiliar todos los recursos del clúster para procesar los datos. Permiten mejorar el rendimiento agregando nodos al clúster.

Amazon Redshift deep live

Amazon Redshifit es una tecnología MPP en columnas, con un almacenamiento de datos eficiente y rentable, I/O reducido, es una buena opción para los data warehouse, ofrece consultas rápidas, paralelismo y automatización. Se pueden crear data warehouses a escala de petabytes en minutos y consultas a escala de exabytes. También da la opción de escalar el cómputo y el almacenamiento por separado utilizando nodos RA3.

Integration with data lake: Redshift proporciona Spectrum que facilita la consulta y escritura de datos en en data lake en formatos de archivos abiertos como Parquet, ORC, JSON, Avro, CSV, ... Para exportar datos solo se ejecuta UNLOAD en el SQL y se especifica Parquet como formato de archivo. Se pueden guardar dato en tablas externas.

Performance: Redshift ofrece un rápido y flexible rendimiento, líder en la industria. Ofrece funciones como: hardware de alto rendimiento (diferentes tipos de nodos, anchod de banda y almacenamiento), AQUA (memoria caché distribuida y acelerada por hardware), almacenamiento eficiente y procesamiento de consultas de alto rendimiento (consultas desde giga hasta petabytes, almacenamiento en columnas, compresión de datos y zone maps), vistas materializadas, administración automática de la carga de trabajo para maximizar el rendimiento, autoaprendizaje que aumenta el rendimiento a medida que crece el uso, almacenamiento en caché de resultados (tiempos de respuesta en menos de un segundo).

Durability and availability: Redshift detecta automáticamente cuando un nodo falla y lo reemplaza en su clúster. El nodo de reemplazo está disponible de inmediato y se cargan primero los datos más accedidos. Siempre se trata de tener al menos tres copias de los datos (original, réplica y copia de seguridad). El clúster está en modo lectura hasta que se agrega un nodo de reemplazo. Se pueden configurar entornos sólidos de recuperación de desastres. Redshift hace copias de seguridad cada ocho horas o cinco GB.

Elasticity and scalability: se puede escalar el procesamiento y el almacenamiento de forma independiente y pagar solo por lo que se usa. Redshift proporciona dos formas de elasticidad: redimensionamiento elástico (se pueden agregar nodos al clúster cuando la carga de trabajo lo requiere y eliminarlos cuando el trabajo esté completo) y escalado de simultaneidad (se puedena admitir un número ilimitado de usuarios simultáneamente y consultas simultáneas, Redshift agrega automáticamente capacidad informática cuando lo necesita. Los usuarios siempre ven los datos más actuales). ***Amazon Redshift managed storage:*** permite escalar y pagar por el cómputo y el almacenamiento de forma independiente y así se adapta a las necesidades del usuario.

Operations

Redshift automatiza muchas tareas operativas como ClusterPerformance y optimización de costos con un buen manejo de los nodos.

Amazon Redshift Advisor: mejora el rendimiento y reduce los costos. Ofrece recomendaciones específicas personalizadas sobre cambios que se deben realizar. Advisor clasifica las recomendacione por orden de impacto.

Interfaces: Redshift tiene controladores personalizados de conectividad de bases de datos JAVA, y familias de clienes SQL. Proporciona un editor de consultas integrado en la consola web. Se pueden ejecutar consultas o diagnóstico de consultas. Tiene integraciones con herramientas BI como Kinesis Data Firehouse.

Security: se puede ejecutar dentro de una nube privada virtual. Se puede usar el modelo de redes de la VPC para definir reglas de firewall. Admite conexiones habilitadas para SSL y enrutamiento de VPC mejorado. En el clúster de Redshift cada nodo almacena información pero solo puede ser accedida desde el nodo líder. Admite cifrado AES-256 acelerado por hardware, las copias de seguridad también van a estar encriptadas. Se pueden usar servicios de contraseñas como AWS KMS. Los usuarios tienen un nivel de acceso que otorga diferentes privilegios. Redshift también proporciona autenticación.

Cost model: los cargos económicos se basan en el tamaño y la cantidad de nodos en el clúster. Si necesita potencia computacional extra se puede habilitar el escalado de simultaneidad. Cada 24 horas de ejecución acumula una hora de crédito para usar el escalado gratuitamente. No hay cargo extra por el almacenamiento de respaldo equivalente al almacenamiento del clúster. No hay cargo por transferencia de datos para la comunicación entre S3 y Amazon Redshift.

Ideal usage patterns

Redshift se utiliza para: ejecución de BI empresarial y generación de informes, analizar datos de ventas, almacenar datos históricos, analizar impresiones y clics, tendencias sociales y medir calidad. Utilizando Spectrum le permite a Redshuft utilizar datos semiestructurados lo que le da la opción de hacer análisis en eventos de gran volumen, descargar datos del historial a los que se accede con poca frecuencia y unir los datos externos con el data warehouse.

Anti-Patterns

Redshift no es ideal para los patrones: OLTP, datos no estructurados y datos BLOB (datos binarios de objetos grandes).