Adriel Araya Vargas 2019312845 Instituto Tecnológico de Costa Rica Bases de Datos 2 IC-4302

Depósitos de datos en AWS.

Introduciendo Amazon Redshift.

Amazon Redshift es un sistema para depósitos de datos en la nube. Amazon Redshift presenta una solución rápida, completamente manejable, de alta capacidad y efectiva al costo para analizar grandes volúmenes de datos usando herramientas de inteligencia de negocios. Se ofrece este servicio para volúmenes de datos hasta petabytes. Para pasar a una escala de exabyte existe Amazon Simple Storage Service(Amazon S3) y funcionalidad Spectrum para Redshift. Amazon Redshift se lanzó como servicio de Amazon Web Services en 2013.

Analítica moderna y estructura de depósitos de datos.

Los datos usualmente fluyen a un depósito de datos de sistemas transaccionales y otras bases de datos relacionales. Los datos son procesados, transformados e y consumidos a una cadencia regular. Los usuarios lo accedan mediante herramientas de BI, clientes de SQL y entre otros. La diferencia entre un depósito de datos y una base de datos de procesamiento de transacciones (OLTP) es que el depósito esta optimizado para escritura por lotes de operaciones y lectura de altos volúmenes de información mientras que OLTP esta optimizado para escritura continua de operaciones y alto volumen de pequeña lectura de operaciones Para obtener los beneficios de un depósito de datos manejado como un almacenamiento de datos por separado de una base OLTP, se recomienda construir un pipeline de datos eficiente. Este pipeline extrae los datos de el sistema origen y lo convierte en un esquema para depósito de datos y lo carga en uno. A continuación, se discuten los bloques que construyen este pipeline y los servicios de AWS que se pueden usar.

Arquitectura de análisis.

Un pipeline de análisis suele contener las siguientes etapas: 1. Recolectar los datos 2. Guardar los datos 3. Procesar los datos. 4.Analizar y visualizar los datos. Recolección de datos: AWS provee soluciones para guardar todo tipos de datos. Primero, datos transaccionales de comercios o finanzas que se guardan en sistemas de base de datos relacional o NoSQL. Se ofrece Amazon DynamoBD como un servicio de base de datos NoSQL que puede ser usado como base OLTP para diferentes aplicaciones. También se ofrece implementaciones para bases de datos SQL tales como Amazon Aurora y Amazon RDS. Segundo, datos de registros generados por el sistema para problemas y analíticas. Amazon S3 es una solución para estos datos por su durabilidad. Tercero, datos en transmisión, Se ofrece Amazon Kinesis para su recolección y procesamiento. Por último, datos del internet de las cosas para dispositivos y sensores. Para el manejo de estos datos se ofrece AWS IoT.

Procesamiento de datos: Existen dos flujos de trabajo, procesamiento por lote y procesamiento en tiempo real. Los más comunes para cada uno respectivamente son, procesamiento de analíticas en línea(OLAP) y OLTP. Hay varios procesos que envuelven el procesamiento por lotes. Primero, extrae, transforma y carga(ETL) para el proceso de obtención de datos. Para este proceso se ofrece AWS Glue para controlar el servicio

ETL, o Amazon EMR para big data. Segundo, extrae, carga, transforma (ELT) como variante de ETL para cargar primero. Por último, OLAP para agregar datos históricos en esquemas multidimensionales. Para procesarla directamente de Amazon Kinesis Data Streams se usa AWS Lambda. Además se cuenta con Amazon Kinesis Client Library como alternativa, Amazon Kinesis Data Firehose para cargar datos a AWS, Amazon MSK y AWS Glue para ETL en stream de datos.

Almacenamiento de datos: La información puede ser guardada en una casa de lago, un depósito de datos o un mercado de datos. Una casa de lago es un patrón arquitectónico que incluye los mejores elementos del depósito y el mercado. Un depósito de datos puede correr largos volúmenes de datos y encontrar patrones. Un mercado de datos es una forma simplificada de depósito donde se concentra en un área o sujeto específico.

Análisis y Visualización: Para esto se usan herramientas como MySQL Workbench, yo otras soluciones como Tableu y MicroStrategy. Amazon Quicksight permite crear visualizaciones y realizar análisis. Entre otros servicios se incluye Amazon Athea, Apache Zeppelin y Spark SQL.

Opciones en tecnología para depósitos de datos.

Primero, bases de datos orientadas a filas. Estas son una fila completa en un bloque físico. Son mejores para OLTP. Estos se limitan por los recursos en un solo dispositivo. Segundo, bases de datos orientadas a columnas. Cada columna es su propio bloque físico. Es más eficiente para lectura. Son una mejor opción para depósito. Por último, procesamiento masivo paralelo. Permite usar todos los recursos en el cluster.

Enfoque profundo en Redshift.

Integración con lagos de datos: Amazon Redshift provee del servicio Redshift Spectrum para hacer más sencillo las consultas y la escritura para lagos de datos. Se pueden consultar archivos de formato abierto y exportar a un lago de datos directamente. Rendimiento: Amazon Redshift ofrece varias funciones para su rendimiento en esta sección se listan algunas. Hrdware de alto rendimiento, AQUA como sistema acelerador de consultas por hardware, almacenamiento eficiente y procesamiento de alto rendimiento, vistas materializadas, y automanejo de cargas de trabajo. Durabilidad y accesibilidad: Redshift detecta y reemplaza cualquier nodo fallido en el cluster de depósitos para su correcto funcionamiento. También se pueden hacer backups constantes. Elasticidad y escalabilidad: Se ofrecen dos formas de computarlo, cambiar el tamaño para añadir o eliminar nodos y escalado concurrente para soportar usuarios y solicitudes ilimitados mediante capacidad computacional adicional.

Operaciones.

Redshift automatiza muchas tareas operacionales como el rendimiento del clúster y la optimización de costos. Se se ofrece Amazon Redshift Advisor para recomendaciones y cambios. En interfaces Redshift tiene controladores JDBC y ODBC para descargar Connect Client. Además tiene un editor de consultas y muchas integraciones externas como las ya mencionadas. En seguridad se puede usar una nube privada con Amazon VPC. Se puede usar AWS Cloudtrail para auditar las llamadas al API de Redshift. También se ofrece AWS IAM para autentificación y acceso. En patrones de uso ideales se recomienda el uso para OLAP en empresas de inteligencia de negocios y reportes, datos de mercado, anuncios, tendencias, videojuegos y servicios médicos. Por otro lado no es ideal para usar con OLTP, datos no estructurados en esquemas y datos BLOB. Es el más eficiente.